Computer Science and Information Technology

`

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

5th International Conference on Computer Science and Information Technology (COMIT 2021), October 29 ~ 30, 2021, Vienna, Austria
International Conference on Cryptography and Blockchain (CRBL 2021)
International Conference on Big Data, IoT and Machine Learning (BIOM 2021)
8th International Conference on Wireless and Mobile Network (WiMNET 2021)
10th International Conference on Signal & Image Processing (SIP 2021)
7th International Conference on Artificial Intelligence and Soft Computing (AISO 2021)
International Conference on NLP & Data Mining (NLDM 2021)

Published By

AIRCC Publishing Corporation

`

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-51-0
DOI: 10.5121/csit.2021.111701- 10.5121/csit.2021.111721

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

`

Preface

The International Conference on 5th International Conference on Computer Science and Information Technology (COMIT 2021), October 29 ~ 30, 2021, Vienna, Austria, International Conference on Cryptography and Blockchain (CRBL 2021), International Conference on Big Data, IoT and Machine Learning (BIOM 2021), 8th International Conference on Wireless and Mobile Network (WiMNET 2021), 10th International Conference on Signal & Image Processing (SIP 2021), 7th International Conference on Artificial Intelligence and Soft Computing (AISO 2021), International Conference on NLP & Data Mining (NLDM 2021) was collocated with 5th International Conference on Computer Science and Information Technology (COMIT 2021). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The COMIT 2021, CRBL 2021, BIOM 2021, WiMNeT 2021, SIP 2021, AISO 2021 and NLDM 2021 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, COMIT 2021, CRBL 2021, BIOM 2021, WiMNeT 2021, SIP 2021, AISO 2021 and NLDM 2021 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the COMIT 2021, CRBL 2021, BIOM 2021, WiMNeT 2021, SIP 2021, AISO 2021 and NLDM 2021.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

`

General Chair	Organization
David C. Wyld,	Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds)	Wireilla Net Solutions, Australia

Program Committee Members

Abdel-Badeeh M. Salem,	Ain Shams University, Egypt
Abdelhadi Assir,	Hassan 1st University, Morocco
Abdelhalim Kessal,	University of Bordj Bou Arreridj, Algeria
Abderrahmane EZ-Zahout,	Mohammed V University, Morocco
Abdullah,	Adigrat University, Africa
Abdulraqeb Alhammadi,	Multimedia University, Malaysia
Ablah AlAmri,	King Abdulaziz University, Saudi Arabia
AbtoyAnouar,	AbdelmalekEssaadi University, Morocco
AdrianOlaru,	University Politehnica of Bucharest, Romania
Ágnes Vathy-Fogarassy,	University of Pannonia, Hungary
Ahmed Alsabbagh,	University of Babylon, Iraq
Ahmed Kadhim Hussein,	Babylon University, Iraq
Ahmed Yaseen Mjhool,	University of Kufa, Iraq
Ahmet CIFCI,	Burdur Mehmet Akif Ersoy University, Turkey
Aishwarya Asesh,	Adobe, USA
Ajay Anil Gurjar,	Sipna College of Engineering and Technology, India
Akashdeep Bhardwaj,	University of petroleum and energy studies, India
Akhilesh K. Sharma,	Manipal University Jaipur, India
Alex Mathew,	Bethany College, West Virginia
Alexander Gelbukh,	Instituto Politécnico Nacional, Mexico
Ali A. Amer,	Taiz University, Yemen
Alireza Valipour Baboli,	University Technical and Vocational Babol, Iran
Alireza Valipour Baboli,	University Technical and Vocational, Iran
Allel Hadjali,	Professeur des Universités en Informatique, France
Amal Zouhri,	Sidi Mohammed Ben Abdellah University, Morocco
Aman Jatain,	Amity University, India
Amando P. Singun Jr,	University of Technology and Applied Sciences, Oman
Amanpreet Kaur,	Chandigarh University, India
Anand Nayyar,	Duy Tan University, Viet Nam
Anuj Singal,	GJU S&T, India
Aridj Mohamed,	Hassiba Benbouali University, Algeria
Arjav A. Bavarva,	RK University, India
Ashraf Elnagar,	University of Sharjah, United Arab Emirates
Assia Djenouhat,	University Badji Mokhtar Annaba, Algeria
Atik Kulakli,	American University of the Middle East, Kuwait
Atul Garg,	Chitkara University, India
Azah Kamilah Muda,	UTeM, Malaysia
Benyamin Ahmadnia,	University of California, USA
Beshair alsiddiq,	Prince Sultan University, Saudi Arabia
Boukari Nassim,	Skikda Universiy, Algeria
Brahim Lejdel,	University of EL-Oued, Algeria
Caitong Yue,	Zhengzhou University, China

`

Carlos Guardado da Silva,	University of Lisbon, Poland
Carlos Westphall,	Federal University of Santa Catarina, Brazil
Chang-Yong Lee,	Kongju National University, South Korea
Chemesse ennehar Bencheriet,	University of Guelma, Algeria
Cheng Siong Chin,	Newcastle University, Singapore
Christian Mancas,	DATASIS ProSoft srl, Bucharest, Romania
Christos J. Bouras,	University of Patras, CEID, Greece
Claude Tadonki,	MINES ParisTech, France
Cong-Conog Xing,	Nicholls State University, USA
Cristina Freitas,	University of Coimbra, Poland
Dadmehr Rahbari,	Tallinn University of Technology, Estonia
Dário Ferreira,	Univerity of Beira Interior, Portugal
Dário Ferreira,	Universidade da Beira Interior, Portugal
Dariusz Jacek Jakobczak,	Koszalin University of Technology, Poland
Dibya Mukhopadhyay,	University of Alabama, USA
Dinesh Reddy Vemula,	SRM University, Amaravati, India
Dokuz,	Eylül University, Turkey
Domenico Rotondi,	FINCONS SpA, Italy
Ekbal Rashid,	RTC Institute of Technology, India
El Murabet Amina,	Abdelmalek Essaadi University, Morocco
Elżbieta Macioszek,	Silesian University of Technology, Poland
Essam Sourour,	Alexandria University, Egypt
Ez-zahout Abderrahmane,	Mohammed V University, Morocco
F. Abbasi,	Islamic Azad University, Amol, Iran
Faiza Tabbana,	Military Academy, Tunisia
Felix J. Garcia Clemente,	University of Murcia, Spain
Fernando Zacarias Flores,	Universidad Autonoma de Puebla, Mexico
Francesco Zirilli,	Sapienza Universita Roma , Italy
Froilan D. Mobo,	Philippine Merchant Marine Academy, Philippines
Fu Jen,	Catholic University, Taiwan
G. Rajkumar,	N.M.S.S.Vellaichamy Nadar College, India
GálZoltán,	University of Debrecen, Hungary
Gang Wang,	University of Connecticut, United States
Ghasem Mirjalily,	Yazd University, Iran
Giuliani Donatella,	University of Bologna, Italy
Gniewko Niedbała,	Poznań University of Life Sciences, Poland
Govindraj Chittapur,	Basaveshwar Engineering College, India
Grigorios N. Beligiannis,	University of Patras ,Greece
Grzegorz Sierpiński,	Silesian University of Technology, Poland
Hamed Taherdoost,	West University, Canada
HamedTaherdoost,	University West Canada, Canada
Hamid Ali Abed AL-Asadi,	Iraq University College, Iraq
Hao-En Chueh,	Chung Yuan Christian University, Taiwan
Harisha A, Sahyadri,	College of Engineering & Management, India
Hedayat Omidvar,	National Iranian Gas Company, Iran
Hemashree,	Hindusthan College of Arts and Science, India
Hiba Zuhair,	Al-Nahrain University,Iraq
Hiromi Ban,	Sanjo City University, Japan
Hossein Bavarsad,	Mechanical Engineer and Project Manager, Iran
Husam Suleiman,	Applied Science Private University, Jordan
Ilham Huseyinov,	Istanbul Aydin Universitesi, Istanbul, Turkey
Isa Maleki,	Islamic Azad University, Iran

`

Israa Shaker Tawfic,	Ministry of Science and Technology, Iraq
Iyad Alazzam,	Yarmouk University, Jordon
Iyd Eqqab Maree,	Salahddine University, Iraq
Janet Walters,	University of Technology, Jamaica
Jasmin Cosic,	DB AG, Germany
Javad Khamisabadi,	Islamic Azad University, Tehran, Iran
Jawad K. Ali,	University of Technology, Iraq
Jesuk Ko,	Universidad Mayor de San Andres (UMSA), Bolivia
Jianyi Lin,	Khalifa University, UAE
João Calado,	Instituto Superior de Engenharia de Lisboa, Portugal
Joey S. Aviles,	Panpacific University , Philippines
Jorge Bernardino,	Polytechnic of Coimbra, Portugal
José Alfredo F. Costa,	Federal University, Brazil
José Luis Abellán Miguel,	UniversidadCatólica De Murcia, Spain
K.Sujatha,	Dadi Institute of Engineering & Technology, India
Kamel Jemai,	University of Gabes, Tunisia
Karim Mansour,	Université Salah Boubenisder, Algeria
Kazuyuki Matsumoto,	Tokushima University, Japan
Ke-Lin Du,	Concordia University, Canada
Kerem Elibal,	BCS Metal Co., Turkey
Kirti Patel,	Chemic Engineers & Constructors, USA
Kirtikumar Patel,	Chemic Engineers, USA
Klenilmar L. Dias,	Federal Institute of Amapa -IFAP, Brazil
Laith Abualigah,	Amman Arab University, Malaysia
Lal Pratap Verma,	Moradabad Institute of Technology, India
Larisa Ofelia Filip,	Petroșani University, Romania
Ljubomir Lazic,	Union University Belgrade, Serbia
Loc Nguyen,	Independent scholar, Vietnam
Luís Corujo,	University of Lisbon, Poland
Luisa Maria Arvide Cambra,	University of Almeria, Spain
M V Ramana Murthy,	Osmania University, India
M. Zakaria Kurdi,	University of Lynchburg, VA, USA
M.A. Jabbar,	Vardhaman College of Engineering, India
M.K.Marichelvam,	Mepco Schlenk Engineering College, Tamilnadu, India
MA.Jabbar,	Vardhaman College of Engineering, India
Maad M. Mijwil,	hdad College of Economic Sciences University, Iraq
Mahendra Bhatu Gawali,	Savitribai Phule Pune University, Pune
Mahesh Swami,	Guru Gobind Singh Indraprastha University, India
Mahsa Mohaghegh,	Auckland University of Technology, New Zealand
Malka N. Halgamuge,	The University of Melbourne, Australia
Mamoun Alazab,	Charles Darwin University, Australia
Marcin Paprzycki,	Polish Academy of Sciences, Poland
Mario Versaci,	DICEAM - University Mediterranea, Italy
Maumita Bhattacharya,	Charles Sturt University, Australia
Md. Sadique Shaikh,	Aimsr, Maharashtra, India
Mehdi Gheisari,	Iau, Iran
Meisam Abdollahi,	University of Tehran, Iran
Menaouer Brahami,	National Polytechnic School of Oran, Algeria
Michail Kalogiannakis,	University of Crete, Greece
Mihai Horia Zaharia,	Gheorghe Asachi Technical University, Romania
Ming An Chung,	National Taipei University Of Technology, Taiwan
Mohamed Arezki Mellal,	M'Hamed Bougara University, Algeria

`

`

`

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

`

5th International Conference on Computer Science and Information Technology (COMIT 2021)

International Conference on Cryptography and Blockchain (CRBL 2021)

`

International Conference on Big Data, IoT and Machine Learning (BIOM 2021)

8th International Conference on Wireless and Mobile Network (WiMNET 2021)

10th International Conference on Signal & Image Processing (SIP 2021)

7th International Conference on Artificial Intelligence and Soft Computing (AISO 2021)

`

International Conference on NLP & Data Mining (NLDM 2021)

CONFIDENTIALITY AND INTEGRITY MECHANISMS FOR MICROSERVICES COMMUNICATION

Lenin Leines-Vite, Juan Carlos Pérez-Arriaga and Xavier Limón

School of Statistics and Informatics,
Universidad Veracruzana, Xalapa, Ver, Mexico

ABSTRACT

The microservices architecture tries to deal with the challenges posed by distributed systems, such as scalability, availability, and system deployment; by means of highly cohesive, heterogeneous, and independent microservices. However, this architecture also brings new security challenges related to communication, system design, development, and operation. The literature contains spread information regarding security related solutions for microservices-based systems, but this spread makes difficult for practitioners to adopt novel security related solutions. In this study, we aim to present a catalogue of security solutions based on algorithms, protocols, standards, or implementations; supporting principles or characteristics of information security, also considering the three possible states of data, according to the McCumber Cube. Our research follows a Systematic Literature Review, synthesizing the results with a meta-aggregation process. We identified a total of 30 primary studies, yielding 71 security solutions for the communication of microservices.

KEYWORDS

Microservices, Software architecture, Secure communication, Information security.

1. INTRODUCTION

The development of applications based on microservices has been gaining more and more momentum in enterprise IT [1], due to its low coupling or dependency on each other, flexibility, and scalability gains. This type of software architecture solves the development and scalability problems that were present in monolithic or service-oriented systems (SOA). The microservice architecture bring desirable characteristics: service isolation, functional independence, only responsibility, independent implementation, and light communication [2]. However, the microservice architecture is relatively new, so also new challenges arise in the development of applications based on this type of architecture [3]. These challenges or "pains" as [4] defines them, appear in the design, development, and operation stage of the application.

When integrating an application in a microservice architecture, there are problems related to the communication of its systems, entities, or processes, highlighting confidentiality and integrity issues. Failure to address these issues could compromise the architecture's internal infrastructure, as issues related to confidentiality entail vulnerabilities such as spoofing, illegal access, and replay attacks; and regarding integrity, there are problems such as data interception, manipulation, and leakage [3]. Identifying these problems in the communication of microservices-based systems is important when designing security policies for the development and deployment of the software; it is crucial not to compromise assets and high-value information

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 01-16, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111701

that are generally exposed on endpoints, and which tend to proliferate. This type of architecture tends to be more susceptible, since it requires opening more ports, exposing more APIs, and distributing their access control, thus exposing a more extensive attack surface [4].

This study is a collection of technologies, mechanism and solutions related to confidentiality, authentication, authenticity, integrity, and authorization for the communication of microservices-based systems. In the same way, this study contributes to the understanding of problems in the communication of microservices, highlighting how developers can address them during the development and deployment of the software according with security solutions identified in the Systematic Literature Review (SLR).

This paper is divided into the following sections. Section 2 presents the background and related work, highlighting the common opinions of other authors on the challenges faced by the microservice architecture, and the need to motivate the reinforcement of security in the systems. We address the Systematic Literature Review method in section 3, which considers the review protocol, the results obtained, and the synthesis of the findings. Section 4 presents the discussion of our findings. In Section 5 we conclude the study, setting lines of research and future work.

2. BACKGROUND AND RELATED WORK

During the communication of microservices there are several security problems, as can be seen in Figure 1; so, it is necessary to strengthen the systems internally, since, in its network infrastructure, the services and data contained in the devices connected to the network, are usually very important business and personal assets [5].

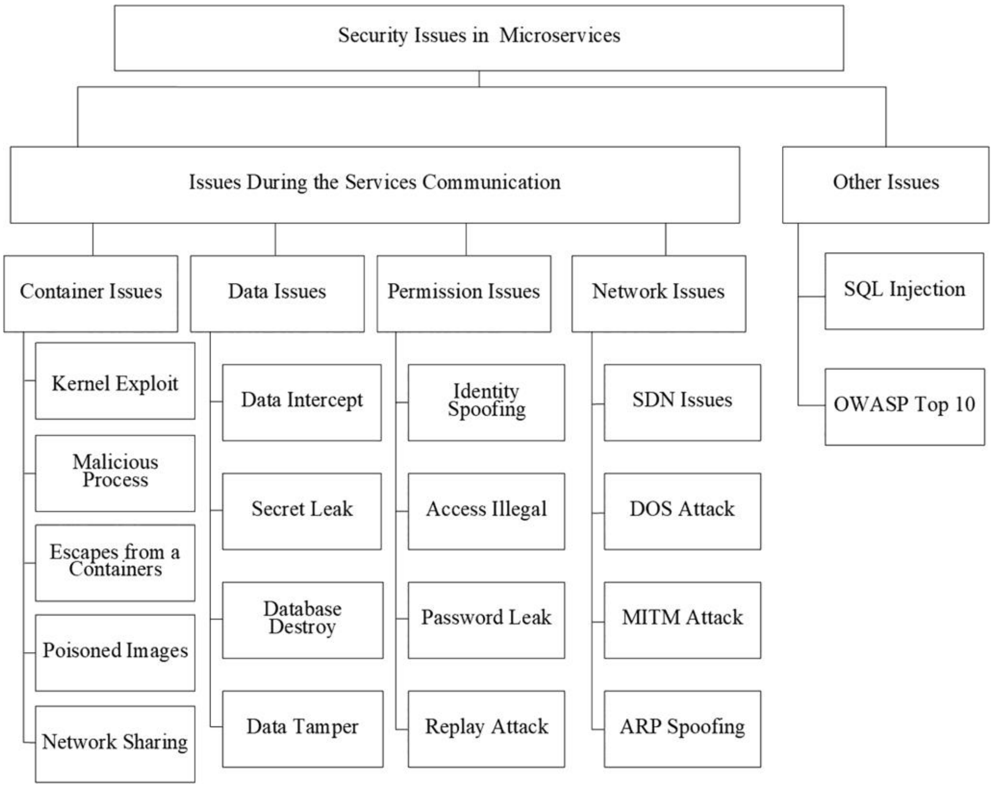


Figure 1. Taxonomy of security problems in microservices.

Note. Adapted from "A survey on security issues in services communication of Microservices-enabled fog applications" by [3], Concurrency and Computation: Practice and Experience, 31. All rights reserved [2018] by John Wiley and Sons. Adapted with permission of the author.

Microservices require more complex communication due to their fine granularity, so [3] highlights that there is not only the risk that the data can be intercepted, but also that malicious entities can infer commercial operations from the shared information. Therefore, [3] mentions that the microservices architecture must verify the authenticity of each service, in addition to verifying the legitimacy of the shared messages, and the valid authorization of origin service. The authors mention that in the study carried out by [6], it is proposed to assign only necessary rights to a subject who requests access to a resource, and that it is valid for the shortest possible time.

The microservices' architecture has had a wide acceptance since it emerged in the industry, therefore, its research has been increasing since then. The literature presents a large number of solutions related to security for the communication of microservices-based systems, but the fact that there isn't a research and compilation on the security methods that can address these problems, makes difficult for practitioners to adopt novel security related solutions.

We identified three studies that address security for communication in microservices as the main concern:

- [3] performs an analysis on the security vulnerabilities related to the communication of microservices-based systems, considering security problems in four aspects: containers, data, permissions, and network. In addition, the authors mention that, with respect of confidentiality, integrity and availability of the data, consideration should be given to offering minimum data security capabilities, including an encryption scheme, strict access, and storage controls safe. The authors' article addresses the security problems of microservices communication, mentioning resource isolation solutions, container protection, data security, permission security, and network security issues; compared to our article, the solutions exposed in the proposed catalogue present different solution approaches for the communication of microservices, but considering the principles and characteristics of information security, the state of the data according to the McCumber Cube, and extending the solutions to be used.
- [7] mentions that the problems related to microservices security are multifaceted, so the authors present a security taxonomy in this type of architecture, breaking it down into six categories: hardware, virtualization, cloud, communication, service, and orchestration. His study places microservices and their security in the broader context of SOA and distributed systems, however, the study only considers the security of microservices communication from components deployed in containers, without considering communication with other hardware components, as well as processes and entities. In contrast to our work, we include solutions focused on the protection of the communication of microservices, processes and entities; without considering another layer of security for the architecture, such as hardware protection or virtualization as is considered in the authors' paper.
- [8] aims to provide a useful guide for developers on recognized threats in microservices and how these can be detected, mitigated, or prevented. The study addresses a systematic mapping to discover the main security threats in microservices, introducing an ontology that can serve as a guide for developers to learn about the threats to detect and the security mechanisms to use. Proposing a general security ontology in the microservice architecture leaves research gaps about the parts of the software that developers must protect. As a result of this, the authors mention the need for studies that intervene in the

security of microservices, related to the protection of communication and its individual defence.

Our study's main objective is to discover solutions that reinforce security for microservices communication and endpoint protection; Therefore, it contributes to the need for studies in the literature that cover these research gaps, as mentioned by the authors.

3. SYSTEMATIC LITERATURE REVIEW

To guide the study, we followed a Systematic Literature Reviews (SLR) based on the guidelines for performing Systematic Literature Reviews in Software Engineering proposed by [9], this includes planning the review, conducting the review, and data synthesis.

[9] mentions that the guidelines for performing SLR in software engineering are used to conduct rigorous reviews of the current empirical evidence in the software engineering community. It is relevant for our work to include an SLR, since the SLR allows us to use a well-defined methodology to identify, analyse and interpret all the available evidence related to a specific research question. That in the case of this study is appropriate for the investigation, documentation and classification of mechanisms, to help reinforce the confidentiality, authentication, authenticity, integrity, and authorization in the communication of microservices-based systems, in order to provide a catalogue that allows to recognize solutions to guide a secure integration of applications based on this type of architecture; since, as has been mentioned, there are security gaps in the communication of microservices that could allow the exploitation of vulnerabilities and information leakage within the infrastructure, which could be used as a guide for the total control of the system [4].

The exhibition of these solutions is intended to serve as a line of research and application in development projects that lack security frameworks in software, and that both the academic and professional fields can consult as a catalogue of references to guide an environment. secure in the interaction of microservices during their deployment.

3.1. Planning the review

The objective of the SLR is to analyse solutions related to the principles and characteristics of information security [10], in conjunction with the 2nd dimension of the McCumber Cube [11], concerning the communication of systems, entities and processes built in a microservices-based systems, with the purpose to catalogue and expose the security methods discovered given their similarities.

Initially [9] mentions carrying out a planning of the review, raising research questions as a separator of doubts; the proposed SLR considers three research questions about communication security for microservices-based systems, with the objective of identifying solutions that work to mitigate problems related to confidentiality and integrity, as well as discovering communication protocols used within this context, revealing security mechanisms.

- **Q1.** What security mechanisms are related to the confidentiality of communication in the microservices-based systems?
- **Q2.** What communication mechanisms for communication integrity between microservices are reported in the literature?
- **Q3.** What communication protocols are used in the context of the microservice architecture?

3.2. Conducting the review

Table 1. Keywords and related concepts used in the search query.

Keyword	Related concepts
Microservices	Microservice, micro-service
Mechanism	Mechanism, algorithm, protocol, standard, framework
Communication	Communication, interaction, connection
Authentication	Authentication
Authorization	Authorization, consent, permit, permission
Confidentiality	Confidentiality, secret
Integrity	Integrity, wholeness
Information	Information, data
Request	Petition, request

- **Q1.** (“Microservice” OR “Micro-service”) AND (“Mechanism” OR “algorithm” OR “standard” OR “framework”) AND (“confidentiality” OR “authorization” OR “Authentication”) AND (“information” OR “data”) AND (“communication” OR “interaction” OR “connection”)
- **Q2.** (“Microservice” OR “Micro-service”) AND (“Communication” OR “interaction” OR “connection”) AND (“mechanism” OR “algorithm” OR “technology” OR “standard” OR “framework”) AND (“integrity” OR “wholeness”)
- **Q3.** (“Microservice” OR “Micro-service”) AND (“protocol”) AND (“communication” OR “interaction” OR “connection”)

The review conduction considers carrying out a search strategy establishing keywords, as shown in Table 1, to later formulate search strings for determining information sources and databases for the collection of articles. The search process for the SLR was an automated search of conference proceedings and articles from information sources and databases. The sources of information consulted were Emerald, Science Direct, Springer Link, Editorial Wiley, ProQuest, ACM Digital Library, IEEEXplore Digital Library, and Google Scholar.

As [9] suggests, we established a systematic search strategy with inclusion and exclusion criteria to identify the most relevant studies in the literature. Table 2 summarizes the criteria applied in the SLR, presented as selection filters.

3.2.1. Study Selection

We carried out the collection of articles with the search strings in the aforementioned information sources and databases, we replicated each search string in each source, and we applied the filtering stages of Table 2. During the first collection of articles, we identified 17 primary studies. To enrich the collection of articles, we carried out a second search, conjugating the search strings to cover the maximum number of articles, managing to increase the collection to 30 studies in total between the first and second collection, these studies are in Table 3.

Table 2. Filters for the selection of studies.

Filters	Criteria
Without filters	Exclusion criteria are not applied.
	Publication date <5 years old from 2020.
1st Filter	English language.
	Publication: congresses, conferences, journals.

2**nd** Filt**er**	Title: At least one keyword answers the research question.
3**rd** Filt**er**	Context: the keywords in the abstract or in the conclusion respond to of the research questions directly or indirectly.
4t**h** Filt**er**	Quick reading to confirm the relationship of the study with the question of investigation.

Table 3. Selected studies

No.	Studies
1	Defense-in-depth and Role Authentication for Microservice Systems. [12]
2	A Cluster of CP-ABE Microservices for VANET. [13]
3	A Web Service Security Governance Approach Based on Dedicated Micro-services. [14]
4	A survey on security issues in services communication of Microservices-enabled fog applications. [3]
5	Capabilities for Cross-Layer Micro-Service Security. [15]
6	Mechanisms for Mutual Attested Microservice Communication. [16]
7	Towards Automated Inter-Service Authorization for Microservice Applications. [17]
8	eZTrust: Network-Independent Zero-Trust Perimeterization for Microservices. [18]
9	An optimized control access mechanism based on micro-service architecture. [2]
10	Authentication and authorization orchestrator for microservice-based software architectures. [19]
11	Towards Multi-party Policy-based Access Control in Federations of Cloud and Edge Microservices. [20]
12	Graph-based IoT microservice security. [21]
13	Overcoming Security Challenges in Microservice Architectures. [7]
14	Identity and Access Control for micro-services based 5G NFV platforms. [22]
15	DNS/DANE Collision-Based Distributed and Dynamic Authentication for Microservices in IoT. [23]
16	Applying Spring Security Framework and OAuth2 To Protect Microservice Architecture API. [24]
17	Design of a micro-service based Data Pool for device integration to speed up digitalization. [25]
18	Hybrid Blockchain-Enabled Secure Microservices Fabric for Decentralized Multi-Domain Avionics Systems. [26]
19	Microservice Security Agent Based On API Gateway in Edge Computing. [27]
20	Secure end-to-end processing of smart metering data. [28]
21	Secure Cloud Processing for Smart Meters Using Intel SGX. [29]
22	BlendSM-DDM: BLockchain-ENabled Secure Microservices for Decentralized Data Marketplaces. [30]
23	Performance Analysis of RESTful API and RabbitMQ for Microservice Web Application. [31]
24	A platform-independent communication framework for the simplified development of shop-floor applications as microservice components. [32]
25	Component-Based Refinement and Verification of Information-Flow Security Policies for Cyber-Physical Microservice Architectures. [33]
26	Design and implementation of a decentralized message bus for microservices. [34]
27	Interface Quality Patterns: Communicating and Improving the Quality of Microservices APIs. [35]
28	Securing IoT microservices with certificates. [36]
29	Implementing a Microservices System with Blockchain Smart Contracts. [37]
30	Building Critical Applications Using Microservices. [38]

3.2.2. Data Extraction

For data extraction, we used meta-aggregation as a qualitative synthesis method. This method makes it possible to synthesize the results by aggregating findings into categories [39], easing the grouping of findings according to attributes or properties in common.

To carry out the meta-aggregation process, the JBI (Joanna Briggs Institute) evidence synthesis manual was partially considered, which involves the extraction of findings, which are then grouped into categories. We combined the categories to construct synthesis statements, taking the findings as conclusions reached by researchers, often presented as topics [40]. The process for meta-aggregation is divided into three stages as mentioned by [41]:

- Extract the findings.
  The first step in this process involves the reviewers extracting all the findings from each of the included articles and defining one illustration per finding. A finding is defined as a topic, category, or metaphor reported by the authors of original articles.
- Categorize the findings.
  The second step in the meta-aggregation involves an evaluation of the similarity in the meaning of the findings, which cross the different original articles.
- Synthesize categories.
  The reviewer should review the full list of categories developed and identify sufficient similarity in meaning to generate a complete set of synthesized findings.

As first step in meta-aggregation, we classified each article in a table, exposing the identified findings and associating the following properties per finding:

- Security model as a reference for security concepts (CIA as the default security model).
- Related layer of the OSI model.
- STRIDE threat model classification [42].
- Security concept based on the principles and characteristics of information security [10]. With relation of the STRIDE model [42], we considered the following security properties mapped with corresponding threats:

    - Confidentiality = Information disclosure
    - Authentication = Spoofing
    - Authorization = Elevation of privilege
    - Authenticity = Spoofing and Repudiation
    - Integrity = Tampering).
- Security technique (s) associated with the discovered algorithm, protocol, standard, or implementation (possible finding ID).
- Name of the discovered algorithm, protocol, standard, or implementation (finding ID).

Subsequently, we categorize each finding according to their related security property and their appearance frequency in the literature, as shown in Table 4. Finally, we catalogued each finding given its nature (protocol, algorithm, standard or implementation), in addition to categorizing the finding according to the 2nd dimension of the McCumber Cube [11], concerning the state in which data security is carried out (transmission, storage, process). This catalogue is shown in Table 5.

Table 4. Grouping of findings in relation to the concept of security.

Security solution	**Safety concept**	**Frequency of mention in the literature**	
		Methods proposed in the articles	**Methods mentioned in articles**
Finding ID (name of discovered algorithm, protocol, standard, or implementation).	• Confidentiality. • Authentication. • Authenticity. • Authorization. • Integrity.	Total mentions of solutions as proposed by the authors.	Total mentions of solutions addressed by the authors and techniques mostly used by developers or autonomous DevOps teams.

3.2.3. Results

The results of the frequency of the findings according to the abstraction of their nature can be seen in Figure 2, while the result of the frequency of the findings in relation to the principles and characteristics of information security can be seen in Figure 3.

The analysis of these 30 studies served as the basis to understand the security problems faced by microservices communication, and how the solutions presented can address these protection needs for the system infrastructure. Of the 71 findings identified in the literature, the authors mention them with different frequency, highlighting, for example, algorithms such as SGX, which is used to provide a Trusted Execution Environment, since it protects memory regions, data, and encryption keys [43]. Protocols such as TLS, which was mentioned in 8 studies as a security solution for the protection of the communication channel, in addition to finding 7 implementations based on TLS. And standards such as the Jason Web Token (JWT), to control the authorization of services and resources within an application.

In the same way, very interesting security solutions were mentioned, for example [15] mentions "capabilities" as a way of indicating actions that can be carried out through relationships between two subjects, as a security measure for controlling access to cache memory or resources. In the context of security for microservices communication, microservices exchange information when they create a relationship to determine what actions can be carried out according to the identity of the microservice with which the connection was formed. Secondly, the implementation of Smart Contracts is also a novel security related solution, these are programs that live and run on a blockchain backbone; written for the purpose of enforcing agreements between two parties in a decentralized and unreliable environment without the control of a central authority. [37] aims to demonstrate that it is possible to fully implement a system based on microservices with Smart Contracts, taking advantage of the principle that each change and operation is permanently and transparently recorded in the blockchain ledger.

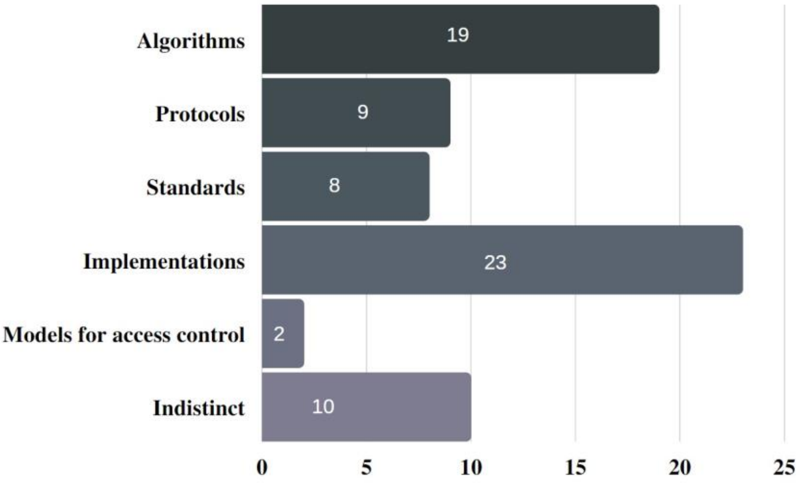


Figure 2. Frequency by category of the findings according to their nature

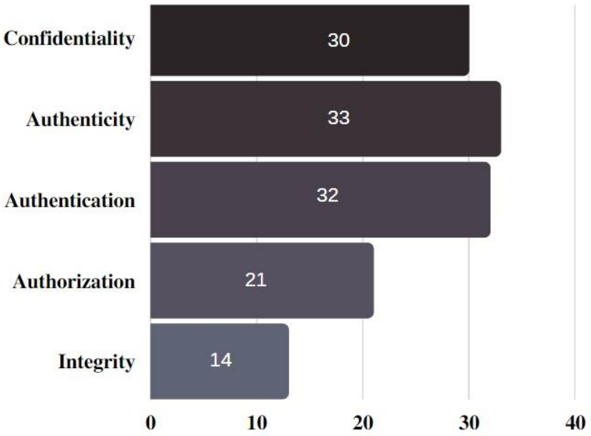


Figure 3. Frequency by category of the findings in relation to the principles and characteristics of information security

3.3. Data Synthesis

The diagram in Figure 4 shows the 71 findings identified and their frequency in the literature. We grouped the results according to their security property; we also considered the existing relationships between the security concepts presented by the findings. The synthesis of the findings is outlined in two colour tones, blue and red, as there are two classes. The first class are the methods proposed by the authors, identified by blue tones; while the second class are methods mentioned by the authors, identified by red tones. The grouping for the findings related to the concept of confidentiality involved for the most part heterogeneity with other concepts of security, for this reason, the findings located in the center of the diagram in Figure 4 are groupings related to the concept of confidentiality, while the findings that are on the edge of the diagram, are groupings that are directly associated with their concept. Also, the relationship with other security concepts can exists, we make this distinction using labels in the diagram.

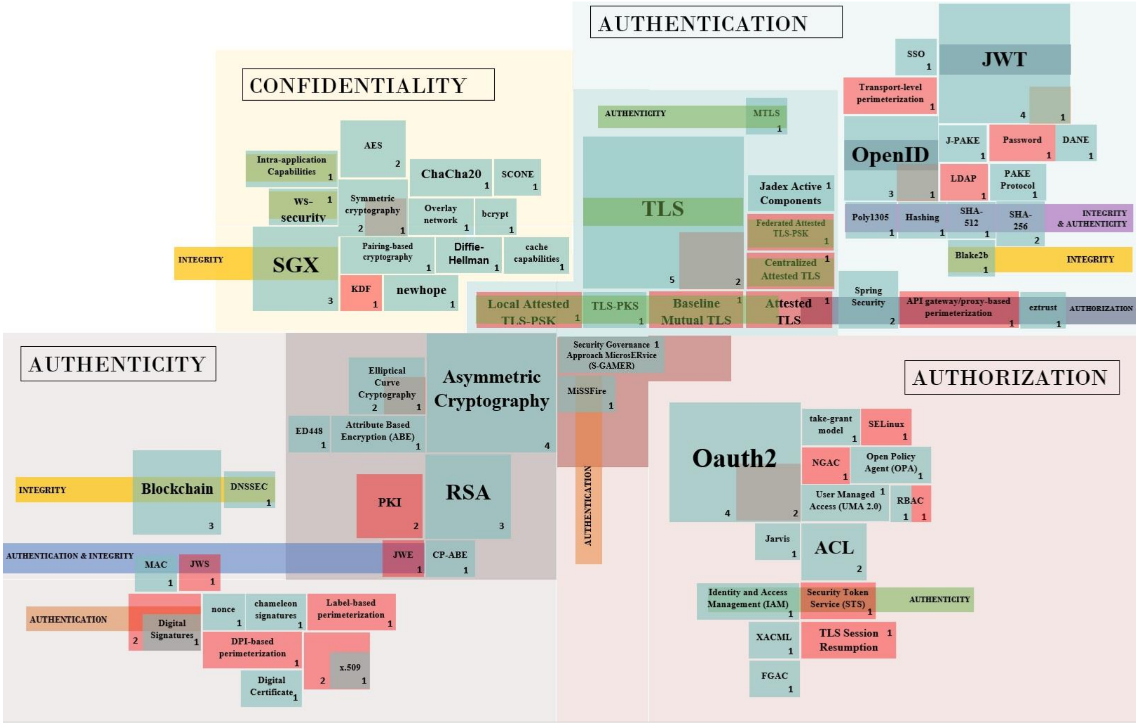


Figure 4. Diagram of findings identified in the literature through meta-aggregation

3.4. Threats to the validity of the study

This subsection poses threats to the validity of the proposed SLR, since regarding the criteria of the Guidelines for performing Systematic Literature Reviews in Software Engineering proposed by [9], the criterion for evaluating the quality of the papers to be included in the SLR, it was not considered. We decided to exclude this criterion, since the sources consulted from where the primary articles were extracted are reliable sources with a track record in computer science, or in various academic areas, so the quality of the studies is high. However, during the discussion of the findings, the lack of discussion of quality metrics of each study included in the SLR was noted, assisted by criteria aimed at questioning their veracity, performance, efficiency, among other critical properties. Therefore, it is proposed to address this issue as future work, and it is discussed in the final part of the study.

4. Discussion

In a network infrastructure, there are three categories of network components: Devices, Media, and Services [14]. In a microservices architecture, these components can be abstracted since this type of architecture allows interactions within the same system or cluster of microservices. Therefore, the use of security techniques is different from those used in a normal network infrastructure, as the medium is not only communication channels, but it also involves interconnections between processes of microservices, which increases the number of inter-process communications, the number of context switches, and the number of I/O operations [44].

Given the diversity of approaches that we found of security solutions for microservices communication in the literature, we present the following classification:

- Security methods aimed at reinforcing security in some layers of the OSI model, for data security, as well as providing Defence in Depth or considering a Zero Trust building approach, as recognized in the literature. Possibly considering a security model, such as the McCumber Cube [11], the STRIDE threat taxonomy [42], or the principles and characteristics of information security [10].
- Solutions aimed at a specific microservices communication problem. For example, in IoT, VANET networks, NFV (Network Function Virtualization), among others.
- Solutions based on existing security methods.
- Solutions proposed by the authors.
- Solutions oriented to a computing paradigm. For example, Cloud Computing or Edge Computing.
- Solutions focused on data security.
- Solutions focused on the security of the communication of entities or software components.

After having synthesized the findings through meta-aggregation, it was possible to answer the research questions. Table 5 presents a catalogue divided by categories according to the nature of the finding. These solutions represented as findings are susceptible to the implementation context, so their approach is important. To clarify the differences between these solutions, we made a distinction according to the state of the data, or if the security approach is associated with security for the interaction between systems and entities. In relation to **Q1** we found that all the solutions associate with the principle of confidentiality, or with the information security characteristics (authenticity, authentication, or authorization). Concerning **Q2**, we identified 14 solutions related to information integrity.

NA (Not Applicable). It is because the method is not involved with information security but with the communication of the entities, services or processes that communicate in the microservices architecture.

Table 5. Catalogue of security solutions for microservices communication.

	Security method	**Safety concept**					**Security applied according to the state of the data**		
		█	▓	█	░	▒	░	▒	░
Algorithms	Intra-application capabilities	X	°	°	°	X	NA(X)		
	Cache capabilities	X	°	°	°	°	°	°	X
	New Hope	X	°	°	°	°	NA		
	ChaCha20	X	°	°	°	°	X	X	°
	Bcrypt	X	°	°	°	°	X	X	°
	Hashing	°	X	X	°	X	X	X	°
	SHA-256	°	X	X	°	X	X	X	°
	SHA-512	°	X	X	°	X	X	X	°
	RSA	X	°	X	°	°	X	X	°
	Poly1305	°	X	X	°	X	X	X	°
	Blake2b	°	X	°	°	X	X	X	°
	MAC	°	X	X	°	X	X	°	°
	Chameleon Signatures	°	°	X	°	°	X	°	°
	CP-ABE	X	°	X	°	°	X	X	°
	Attribute-based encryption	X	°	X	°	°	X	X	°

	(ABE)								
	ED448	X	°	X	°	°	X	°	°
	AES	X	°	°	°	°	X	X	°
	KDF	X	°	°	°	°	NA		
	SGX	X	°	°	°	X	°	°	X
Protocols	TLS	X	X	X	°	°	X	°	°
	TLS-PKS	X	X	X	°	°	X	°	°
	Mutual authentication TLS (MTLS)	X	X	X	°	°	X	°	°
	Diffie-Hellman	X	°	°	°	°	NA(X)		
	WS-Security	X	°	°	°	X	X	°	°
	LDAP	°	X	°	°	°	NA		
	PAKE Protocol	°	X	°	°	°	X	°	°
	J-PAKE	°	X	°	°	°	X	°	°
	User Managed Access (UMA 2.0)	°	°	°	X	°	NA		
Standards	JWT	°	X	°	X	°	NA(X)		
	JWE	X	X	X	°	X	X	°	°
	JWS	°	X	X	°	X	X	°	°
	Oauth 2.0	°	°	°	X	°	NA		
	OpenID	°	X	°	X	°	NA		
	Security token service (STS)	°	°	X	X	°	NA		
	DANE	°	X	°	°	°	NA		
	x.509	°	°	X	°	°	NA		
Implementations	Security Governance Approach MicrosERvice	X	°	°	X	°	NA		
	Attested TLS	X	X	X	X	°	X	°	°
	Centralized Attested TLS	X	X	X	°	°	X	°	°
	Centralized Attested TLS-PSK	X	X	X	°	°	X	°	°
	TLS Session Resumption	°	°	°	X	°	X	°	°
	Local Attested TLS-PSK	X	X	X	°	°	X	°	°
	Federated Attested TLS-PSK	X	X	X	°	°	X	°	°
	Baseline Mutual TLS	X	X	X	°	°	X	°	°
	Jadex Active Components	X	X	°	°	°	NA		
	API gateway/proxy-based perimeterization	°	X	°	X	°	NA		
	eztrust	°	X	°	X	°	NA		
	Open Policy Agent (OPA)	°	°	°	X	°	NA		
	Spring Security	°	X	°	X	°	NA		
	SCONE	X	°	°	°	°	°	X	°
	Transport-level perimeterization	°	X	X	°	°	NA		
	Label-based perimeterization	°	°	X	°	°	NA		
	DPI-based perimeterization	°	°	X	°	°	NA		
	Jarvis	°	°	°	X	°	NA		
	blockchain	°	°	X	°	X	°	X	°
	DNSSEC	°	°	X	°	X	NA(X)		
	PKI	X	X	X	°	°	NA		
	MiSSFire	X	X	°	X	°	X	°	°

	SELinux	°	°	°	X	°	NA (X)		
Models for access control	Take-grant model	°	°	°	X	°	NA		
	RBAC	°	°	°	X	°	NA		
Indistinct solutions	Password	°	X	°	°	°	NA		
	ACL	°	°	°	X	°	NA		
	Identity and Access Management	°	°	X	X	°	NA		
	Overlapping network	X	°	°	°	°	X	°	°
	XACML	°	°	°	X	°	NA		
	NGAC	°	°	°	X	°	NA		
	SSO	°	X	°	°	°	NA		
	Digital certificate	°	°	X	°	°	NA		
	Digital signature	°	°	X	°	X	X	°	°
	nonce	°	°	X	°	°	X	°	°

■ **Confidentiality** ■ **Authentication** ■ **Authenticity** ■ **Authorization** ■ **Integrity**

■ **Transmission** ■ **Storage** ■ **Process**

Finally, in response to **Q3**, we identified four protocols to establish secure communication, and ten protocols only oriented to the communication of systems. The findings include cryptographic protocols such as Diffie-Hellman and TLS; protocols for secure communication, such as gRPC and WS-Security; communication and message protocols such as RESTful, RabbitMQ, AMQP, ZeroMQ, Google Protobuf serializer, OPC Unified Architecture (OPC UA), Extendable Machine Connector (XSC), Pasty Protocol; and, Pastry or Scribe, as protocols for the discovery of services. In addition, we found six articles that discuss protocols and API construction patterns for the communication of systems, entities, and processes in a microservices architecture: [27], [31], [32], [33], [34] and [35].

5. CONCLUSIONS

With the help of the Guidelines for performing Systematic Literature Reviews in Software Engineering proposed by [9], and the meta-aggregation process [39], we achieved the objective of this study. We identified solutions related to the principles and characteristics of information security [10]. We grouped the solutions discovered based on its properties. We identified protocols and API construction patterns for the communication of systems, entities, and processes in a microservices architecture, both insecure and secure.

At the conclusion of the SLR, we identified lines of research for a reliable and secure deployment of an application based on a microservices architecture. It is worth noting that in order to carry out a safe development and deployment of microservices based-systems, the processes should preferably adhere to a security model, as well as aim to provide global data protection. The lines of research identified can be briefly noted as:

- Considering the McCumber Cube [11], extending the principles of information security with the attributes of authentication, authenticity, and authorization, as a framework for building software around a microservices architecture, and thus contributing to security evaluation and auditing, as an aid to develop security policies, and determine education, training, and awareness requirements [45].

- Collecting solutions oriented to the origin of data by source, such as databases. It is critical to approach them as a security complement for the integration of a reliable deployment for the communication of systems and other elements in a microservices architecture.

Some selected studies of the SLR mention complementary solutions for the communication of microservices. These studies relate to availability, monitoring, and optimization of resources, due to the close relationship with security for the communication of microservices. Monitoring complements some principles or information security [10], in the communication of microservices and interaction with other components and entities of the application, as monitoring aids to verify that the security attributes are fulfilling their function within the application and at the same time, it allows keeping under observation the behaviour of the parts of the application; likewise, the optimization of resources in the deployment of applications based on a microservices architecture is critical, since depending on the security implementation in the architecture, the authors mostly agree on the problem of resources, due to the security layers that developers must implement to respect the principles and characteristics of security in the communication of entities and software components. Finally, it is important to stress that availability, is crucial to deploy a complete amalgam of security in the communication of microservices.

For future work, as mentioned in section 3.3, we do not considered all the criteria from the Guidelines for performing Systematic Literature Reviews in Software Engineering proposed by [9], as we considered that the criterion for evaluating the quality of the papers to be not necessary for this study. However, a discussion including quality metrics on the studies is desirable to better assess their veracity, performance, efficiency, among other critical properties.

REFERENCES

[1] R. Macdonal, (2015) “Microservices. Pros & Cons of Using Microservices On A Project”. MadeTech. Available in: https://www.madetech.com/blog/microservices-pros-and-cons

[2] G. Fu, J. Sun, and J. Zhao, (2018) “An optimized control access mechanism based on micro-service architecture”. 2018 2nd IEEE Conference in Energy Inter-net and Energy System Integration (EI2). pp. 1-5.

[3] D. Yu, Y. Jin, Y. Zhang, and X. Zheng, (2018) “A survey on security issues in services communication of Microservices-enabled fog applications”. Concurrency and Computation: Practice and Experience. 31(22), 1–19 (2019).

[4] J. Soldani, D. A. Tamburri, and W. V. Den Heuvel, (2018) “The pains and gains of microservices: A Systematic grey literature review”. Journal of Systems and Software, vol. 146, pp. 215-232.

[5] Cisco Networking Academy, (2012) “CCNA v5.0 Routing and Switching”. Introduction to Networks. Networking explore, pp. 15-60.

[6] M. Gegick, and S. Barnum, (2005) “Least Privilege”. CISA. Available in: https://www.us-cert.gov/bsi/articles/knowledge/principles/least-privilege

[7] T. Yarygina, and A. H. Bagge, (2018) “Overcoming Security Challenges in Microservice Architectures”. IEEE Symposium on Service-Oriented System Engineering (SOSE). pp. 11-20.

[8] A. Hannousse, and S. Yahiouche, (2020) “Securing Microservices and Microservice Architectures: A Systematic Mapping Study” [abs/2003.07262].

[9] B. A. Kitcheham and S. M. Charters, (2007) “Guidelines for performing Systematic Literature Reviews in Software Engineering”. Keele University and Durham University Joint Report, EBSE 2007-001, pp. 1-44.

[10] L. Bass, P. Clements, and R. Kazman, (2012) “What Is Software Architecture?”. Software Architecture in Practice, Third Edition, pp. 1-19.

[11] J. McCumber, (2004) “Assessing and Managing Security Risk in IT Systems: A Structured Methodology” (1st. ed.). Auerbach Publications. pp. 131-153

[12] K. Jander, L. Braubach, and A. Pokahr, (2018) “Defense-in-depth and Role Authentication for Microservice Systems”. Procedia Computer Science. 130, pp. 456-463.

[13] M. B. Taha, C. Talhi, and H. Ould-Slimanec, (2019) “A Cluster of CP-ABE Microservices for VANET”. Procedia Computer Science, 155, pp. 441-448.

[14] S. Abidi, M. Essafi, C. G. Guegan, M. Fakhri, H. Witti, and H. H. Ben Ghezala, (2019) “A Web Service Security Governance Approach Based on Dedicated Micro-services”. Procedia Computer Science, 159, pp. 372-386.

[15] R. Sprabery, (2018) “Capabilities for cross-layer micro-service security”. ProQuest Dissertations & Theses Global.

[16] K. Walsh, and J. Manferdelli, (2017) “Mechanisms for Mutual Attested Microservice Communication”. Companion Proceedings of the10th International Conference on Utility and Cloud Computing (UCC '17 Companion). Association for Computing Machinery, pp. 59–64.

[17] X. Li, Y. Chen, and Z. Lin, (2019) “Towards Automated Inter-Service Authorization for Microservice Applications”. Proceedings of the ACM SIGCOMM 2019 Conference Posters and Demos (SIGCOMM Posters and Demos '19). Association for Computing Machinery, pp. 3–5.

[18] Z. Zaheer, H. Chang, S. Mukherjee, and J. V. der Merwe, (2019). EZTrust: Network-Independent Zero-Trust Perimeterization for Microservices. Proceedings of the 2019 ACM Symposium on SDN Research (SOSR '19). Association for Computing Machinery, pp. 49–61.

[19] A. Bánáti, E. Kail, K. Karóczkai, and M. Kozlovszky, (2018) “Authentication and authorization orchestrator for microservice-based software architectures”. 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1180-1184.

[20] D. Preuveneers, and W. Joosen, (2019) “Towards Multi-party Policy-based Access Control in Federations of Cloud and Edge Microservices”. IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). Stockholm, Sweden, pp. 29-38.

[21] M. Pahl, F. Aubet, and S. Liebald, (2018) “Graph-based IoT microservice security”. NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, pp. 1-3.

[22] D. Guija, and M. S. Siddiqui, (2018) “Identity and Access Control for micro-services based 5G NFV platforms”. Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018). Association for Computing Machinery, Article 46, pp. 1–10.

[23] D. Díaz-Sánchez, A. Marín-Lopez, F. Almenárez Mendoza, and P. Arias Cabarcos, (2019) “DNS/DANE Collision-Based Distributed and Dynamic Authentication for Microservices in IoT”. Sensors 19. no. 15: 3292.

[24] Q. Nguyen, and O. Baker, (2019) “Applying Spring Security Framework and OAuth2 To Protect Microservice Architecture API”. Southern Institute of Technology, Invercargill, New Zealand. pp. 257-264.

[25] J. C. García Ortiz, D. Todolí-Ferrandis, J. Vera-Pérez, S. Santonja-Climent, and V. Sempere-Payá, (2019) “Design of a micro-service based Data Pool for device integration to speed up digitalization”. 27th Telecommunications Forum (TELFOR), pp. 1-4.

[26] R. Xu, Y. Chen, E. Blasch, A. Aved, G. Chen, and D. Shen, (2020) “Hybrid Blockchain-Enabled Secure Microservices Fabric for Decentralized Multi-Domain Avionics Systems”. Proc. SPIE 11422, Sensors and Systems for Space Applications XIII, 114220J.

[27] R. Xu, W. Jin, and D. Kim, (2019) “Microservice Security Agent Based On API Gateway in Edge Computing”. Sensors (Basel, Switzerland), 19(22), 4905.

[28] A. Brito, C. Fetzer, S. Köpsell, M. Pasin, K. Fonseca, M. Rosa, L. Gomes, R. Riella, C. Prado, L. F. da Costa Carmo, D. Lucani, M. Sipos, L. Nagy, and M. Fehér (2019) “Secure end-to-end processing of smart metering data”. Journal of Cloud Computing, 8. Article 19.

[29] M. V. Araújo, C. B. do Prado, L. F. Rust C. Carmo, A. E. Rincón, and C. M. Farias, (2018) “Secure Cloud Processing for Smart Meters Using Intel SGX”. 2019, National Institute of Metrology, Quality and Technology (INMETRO).

[30] R. Xu, G. Sankar Ramachandran, Y. Chen, and B. Krishnamachari, (2019) “BlendSM-DDM: BLockchain-ENabled Secure Microservices for Decentralized Data Marketplaces”. IEEE International Smart Cities Conference (ISC2), pp. 14-17.

[31] X. J. Hong, H. S. Yang, and Y. H. Kim, (2018) “Performance Analysis of RESTful API and RabbitMQ for Microservice Web Application”. International Conference on Information and Communication Technology Convergence (ICTC), pp. 257-259.

[32] M. M. Strljic, T. Korb, T. Tasci, E. Tinsel, D. Pawlowicz, O. Riedel, and A. Lechler, (2018) "A platform-independent communication framework for the simplified development of shop-floor applications as microservice components". IEEE International Conference on Advanced Manufacturing (ICAM), pp. 250-253.

[33] C. Gerking, and Schubert (2019) "Component-Based Refinement and Verification of Information-Flow Security Policies for Cyber-Physical Microservice Architectures". IEEE International Conference on Software Architecture (ICSA), pp. 61-70.

[34] P. Kookarinrat, and Y. Temtanapat, (2016) "Design and implementation of a decentralized message bus for microservices". 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1-6.

[35] M. Stocker, O. Zimmermann, U. Zdun, D. Lübke, and C. Pautasso, (2018) "Interface Quality Patterns: Communicating and Improving the Quality of Microservices APIs". Proceedings of the 23rd European Conference on Pattern Languages of Programs (EuroPLoP '18). Association for Computing Machinery, New York, NY, USA, Article 10, pp. 1–16.

[36] M. Pahl, and L. Donini, (2018) "Securing IoT microservices with certificates". NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, pp. 1-5.

[37] R. Tonelli, M. I. Lunesu, A. Pinna, D. Taibi, and M. Marchesi, (2019) "Implementing a Microservices System with Blockchain Smart Contracts". IEEE International Workshop on Blockchain Oriented Software Engineering (IWBOSE), pp. 22-31.

[38] C. Fetzer, (2016) "Building Critical Applications Using Microservices". IEEE Security and Privacy 14, 6, pp. 86–89.

[39] E. Aromataris, and Z. Munn, (2020) "JBI Manual for Evidence Synthesis". JBI. Available in: https://synthesismanual.jbi.global. https://doi.org/10.46658/JBIMES-20-01

[40] S. F. Johnson, and R. L. Woodgate, (2017) "Qualitative research in teen experiences living with food-induced anaphylaxis: A meta-aggregation". Journal of Advanced Nursing, Volume 73, pp. 2534-2546.

[41] K. Hannes, and A. Pearson, (2011) "Obstacles to the Implementation of Evidence-Based Practice in Belgium: a worked example of meta-aggregation". K. Hannes y C. Lockwood (Eds.), Synthesizing Qualitative Research, pp. 21-39. John Wiley & Sons.

[42] M. Howard, and S. Lipner, (2006) "The Security Development Lifecycle Process". The Security Development Lifecycle, SDL: A Process for Developing Demonstrably More Secure Software, pp. 114-116. Microsoft press.

[43] S. Arnautov, B. Trach, F. Gregor, T. Knauth, A. Martin, C. Priebe, J. Lind, D. Muthukumaran, D. O'Keeffe, M. L. Stillwell, D. Goltzsche, D. Eyers, R. Kapitza, P. Pietzuch, and C. Fetzer, (2016). SCONE: secure Linux containers with Intel SGX. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16). USENIX Association, USA, pp. 689–703.

[44] B. Christudas, (2019) "Microservice Performance. Practical Microservices Architectural Patterns". Apress, Berkeley, CA. pp. 279-314

[45] J. P. Myers, and S. Riela, (2008) "Taming the diversity of information assurance & security". Journal of Computing Sciences in Colleges, 23, 4, pp. 173–179.

CARENVISION: A DATA-DRIVEN MACHINE LEARNING FRAMEWORK FOR AUTOMATED CAR VALUE PREDICTION

TianGe (Terence) Chen[1], Angel Chang[1], Evan Gunnell[2], Yu Sun[2]

[1]Rancho Cucamonga High School, Rancho Cucamonga, CA, 91701
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

When people want to buy or sell a personal car, they struggle to know when the timing is best in order to buy their favorite vehicle for the best price or sell for the most profit. We have come up with a program that can predict each car's future values based on experts' opinions and reviews. Our program extracts reviews which undergo sentiment analysis to become our data in the form of positive and negative sentiment. The data is then collected and used to train the Machine Learning model, which will in turn predict the car's retail price.

KEYWORDS

Machine Learning, Polynomial Regression, Artificial Neural Network.

1. INTRODUCTION

The struggle of buying a new car can also bring the conflict of knowing when to sell it for maximum profit. Thousands of car owners ranging from multiple age groups have this problem. According to I. Wagner, in 2019, around 91.3 million motor vehicles were sold globally [1].Nonetheless, the key component in finding the perfect time is the popularity and the voice of previous car owners over a series of years. This thought occupied us for the majority of our project. How are we going to extract millions of data values to help car owners around the globe? Through the use of top-of-the-line technology, we were able to bring the benefits of CarEnvision for everyone's usage. Without CarEnvision, thousands of car owners are hit with the question of: is this the right time? CarEnvision allows car owners to answer this simple question. However, the car industry changes every day and on, making this assumption quite difficult. So, the solution? CarEnvision. It is easy to use through the use of machine learning and automation [9, 10, 11]. With a short simple survey on the car, a single press of a button can tell you if it is the right time. It gathers thousands of information from the vast interweb to determine the price of the car next year. With this information, it can then help you decide if it's the right time to sell the car. The difference between profit and no profit can determine the future of the car industry.

Throughout the car market industry, many car predicting software have been presented to the public to help ease the struggle of selling your own car. These software are engineered to use the pricing of car dealerships without regard to the profit, policies, or managers that determine the price points. Some dealerships may have had different amounts of customer interaction and therefore have fewer or more sales. This can greatly impact the companies' choice of the price of cars. With changing price tags depending on the dealership, the value of the car may not be stable

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 17-25, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111702

and not be reliable. In addition, constant bargaining for cheaper price points may also affect the value of the car. According to Edmunds, “the more car deals the car salesperson makes, the more money that salesperson takes in.” The pricing of the car uses different factors such as the manufacturer’s price. With hopes of gaining maximum profit, dealerships may input policies or change the price tag. A second practical problem is that the car value predictor may not use the opinions of the buyers. When selling a car, knowing that the population has a positive view on this model can increase the value of the car. Nonetheless, the most important aspect of selling a car is the opinion of the buyer. As a result, our program, CarEnvision, allows for users to easily extract information that is not artificially produced, but straight from the opinions of hundreds of car owners.

Though CarEnvision is unique, it is not the only predictor in the field of artificial intelligence. For example, Market Insider is a website that collects data from the stock market, and creates a two-dimensional graph depicting the growth and decay of car values over time. It also provides live changes of every company’s stock values whether they are dropping or climbing. But our program, CarEnvision, assembles big data together from a public car website, analyzes the sentiments to see how positive or negative the opinions are, and uses Machine Learning algorithms to predict a future car value. This prediction is not a feature of Market Insider as they only provide the historical activities of the car company’s values. Our goal is for the program to predict an accurate retail price for the car in the upcoming year using enough collected information. That is why we used jdpower.com as our main source to gather the chosen car’s past five years of public opinions and reviews. This data is able to then train our Machine Learning model to be as accurate as possible. With more and more data provided, the predicted retail price will fluctuate accordingly which is why our program provides trusted information.

In order to keep our data up-to-date and reasonable, we relied on jdpower.com. This website includes universal information to almost every old and new car on the market. Whenever new cars are hinting to hit the market, new data will begin to appear on the website for users to see. Using the first part of our algorithm code, the chosen car’s opinions are downloaded directly off of the website itself. For example, each opinion is evaluated, producing a level of positivity and negativity. The sentiments are used to train the Machine Learning Model. For every new prediction, the program pulls data live off of the website. These are examples of different car models we personally chose which input the sentiments into our model. We used four different car models with four different sets of sentiments which make our prediction as accurate as possible. Furthermore, the output of our model would be the car’s value which is the quotient of the current price (used) divided by the release price.

The added paper’s structure is organized in different sections that include the process and details about CarEnvision. Section 2 is about the challenges we faced during our development and procedure. Section 3 provides the solution we came up with in order to resolve the conflicts mentioned in section 2. Section 4 includes in-depth explanation on how our program was created and the reason behind the goal, along with a presentation that connects to our work in section 5. Lastly, section 6 wraps up the essay and prefigures the future step of our program.

2. CHALLENGES

When starting a new coding project, you are always faced with different barriers and challenges along the way. These may include using reliable data and picking the right model. We also wanted to make this project useful for the real world. Some other complications range from choosing reliable data points from the vast internet and making it easy to use for inexperienced users. Below is an overview of the different barriers we had to overcome to execute our project.

One issue we had at the beginning was searching for a reliable website source. When looking at the economy of the car market, every digit must be precise and can make a great change in how cars are exchanged. Some sources are biased while others have inaccurate information. Since every website may vary and may not be 100% accurate, we chose to use the most universal website. This reliable source comes straight from the consumer car market and uses actual car exchange prices. We agreed that extracting information from the actual market exchange is the most organic form of information we could put our hands on.

During our first stage of programming, we had trouble gathering big data for our program model. This brought us more issues as we continued further into development. Originally, we made the program to collect only five years of data from one car. But as we began testing our model, it resulted as an inaccurate and unreasonable value. So in order for the Machine Learning model to provide a more legitimate value, we must input more data into it. Ten years historical data from one car was the final decision for the model. This meant we had to revise our code's format as well as extend the lines of our input data.

Creating automation for the entire program was the biggest conflict we ran into during the process. In order for the user to be able to answer a few questions about their chosen car, we had to write a set code that somehow takes the information provided by the user and communicate with the website we chose. But we could not determine how to split apart the website's URL in the program code in order for the given information by the user to be sent. Plus, the names of the car and models must be very specific without any spelling errors since it connects with the website's URL. Developing the automation part of our program consumed the most time throughout our coding process.

3. SOLUTION

A. An Overview of the Solution

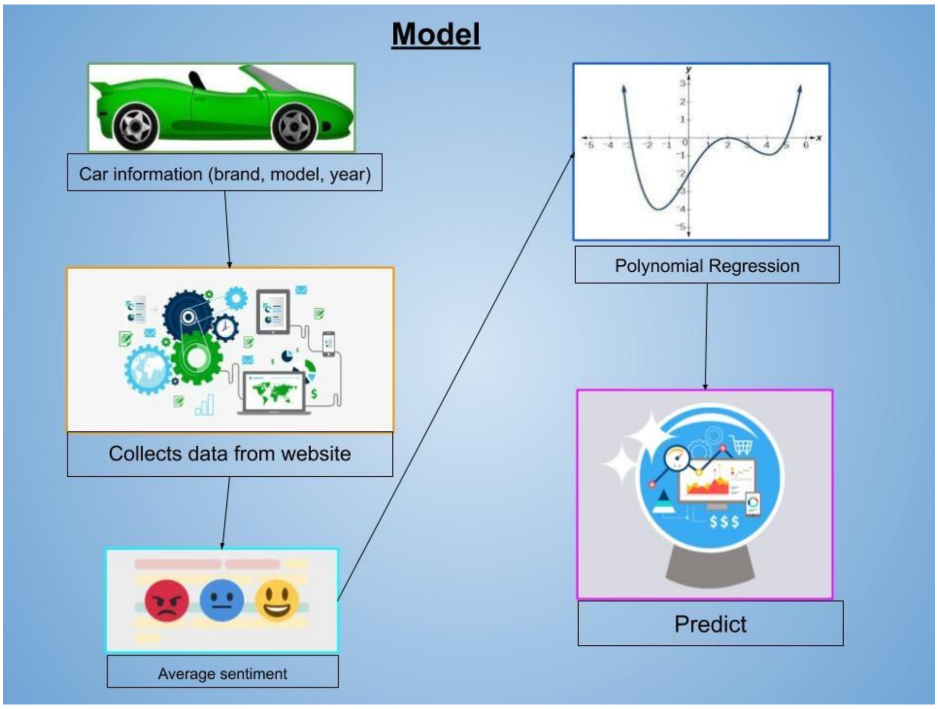
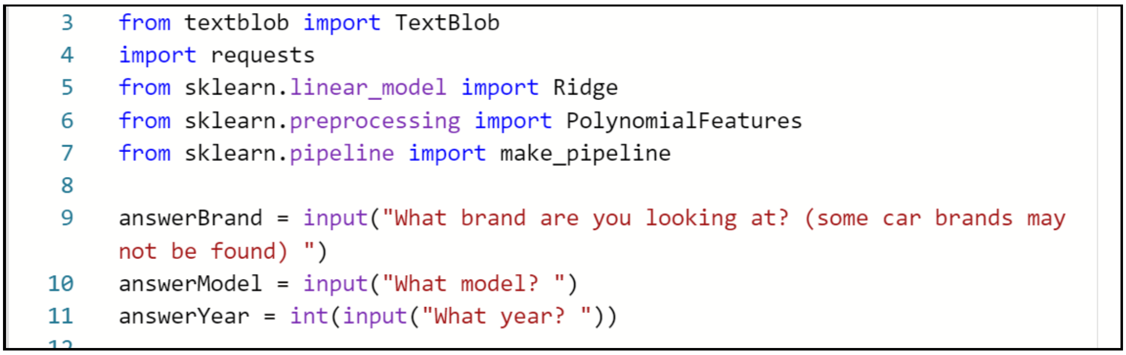

Model
Car information (brand, model, year)
Collects data from website
Average sentiment
Polynomial Regression
Predict
$$$

Figure 1. Overview of the solution

CarEnvision is able to pull the pricing of a broad variety of cars. These brands can then be specified through the model and the year it came out. Through the use of jdpower.com, we were able to pull information straight from the consumers themselves. In a sense, jdpower.com is a consumer operated system where people from the car market can sell their cars and know how much they’re car is going for at the moment. After answering a short survey, CarEnvision starts running. Using TextBlob, the code was able to extract only information that was needed. We extracted thousands of reviews of different car models and implemented sentiment analysis to determine how positive the comment was. The automated system loops this code until all reviews are calculated. The grand result would be all the review’s results average. This loop is then repeated for the last 5 years of the same model. Then, using Polynomial Regression we were able to predict the next year. Since the model was trained to fit cars from a variety of basic and luxurious brands, this model is able to fit most cars in the market. (Illustration shown in Figure 1).

B. Automation and Machine Learning Data

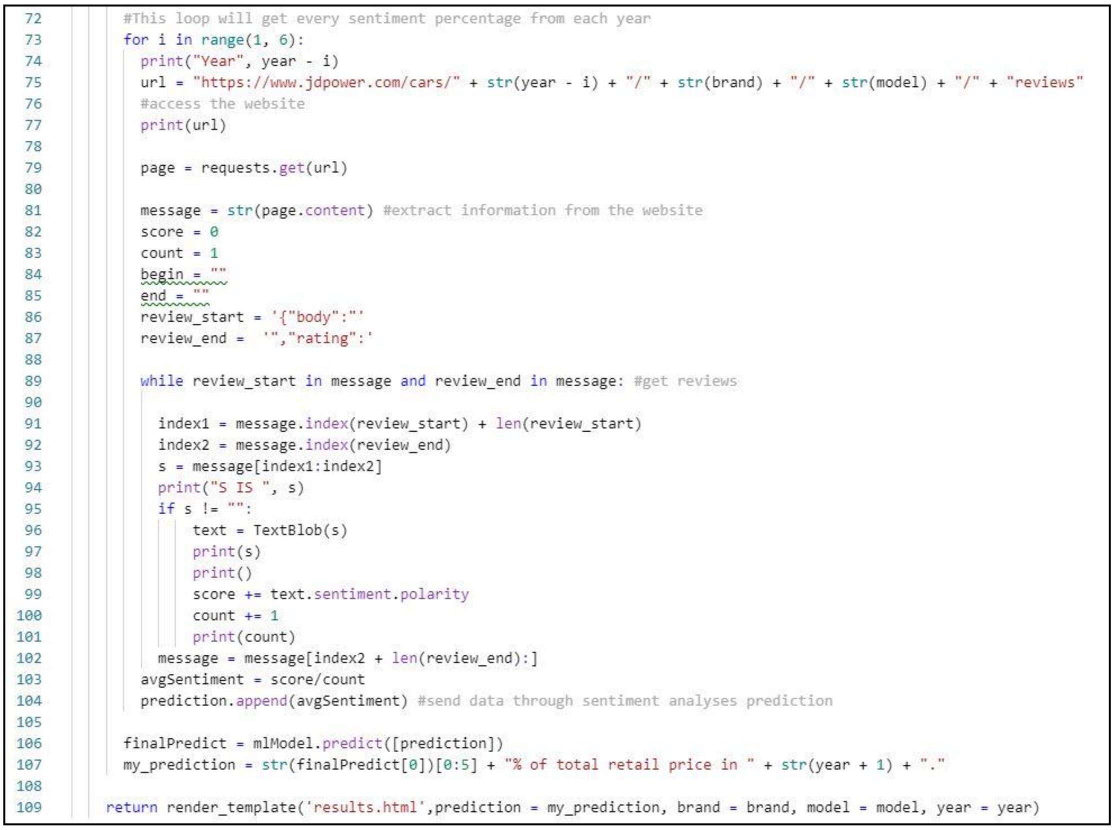
The following segments of code shows the entire program running as if a user is using it. Each component of the process is presented with a set of captured code in order to provide visual representation for the audience. The implemented code presented in Figure 2 shows the importing of textblob and different models from Machine Learning. A few lines passed, the input functions are there to collect the user’s data as in the car’s brand, model, and year. Each function contains different questions being asked to the user.



```
from textblob import TextBlob
import requests
from sklearn.linear_model import Ridge
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline

answerBrand = input("What brand are you looking at? (some car brands may not be found) ")
answerModel = input("What model? ")
answerYear = int(input("What year? "))
```

Figure 2. the importing of textblob and different models from Machine Learning

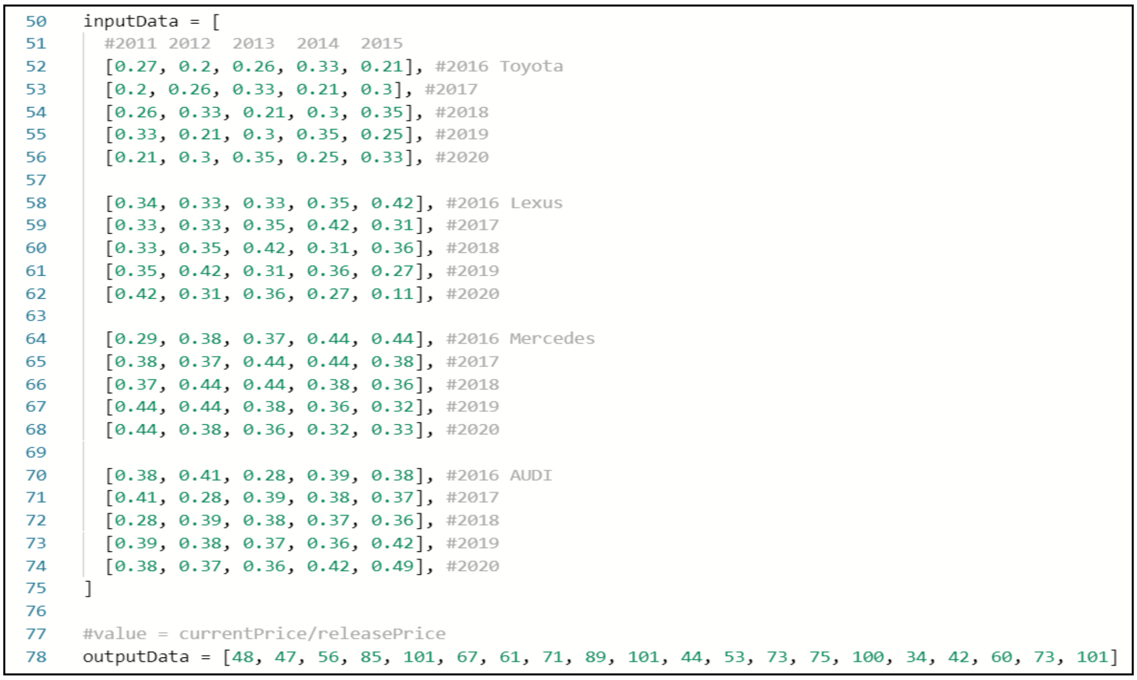
This set of code in Figure 3 is the main algorithm that operates the automation. After gathering the information from the user, each function from the previous set of code is carried to line 19 where it splits apart the URL of jdpower.com. This will take the program straight to the website and start searching for sentiments there. The loop runs for five times because it needs to gather five years of history from the car off of the website. Every year’s sentiment from the car is shown to the user as an average. For example, year one represents one year before the user’s chosen year and so on.



```
    #This loop will get every sentiment percentage from each year
    for i in range(1, 6):
      print("Year", year - i)
      url = "https://www.jdpower.com/cars/" + str(year - i) + "/" + str(brand) + "/" + str(model) + "/" + "reviews"
      #access the website
      print(url)

      page = requests.get(url)

      message = str(page.content) #extract information from the website
      score = 0
      count = 1
      begin = ""
      end = ""
      review_start = '{"body":"'
      review_end =  '","rating":'

      while review_start in message and review_end in message: #get reviews

        index1 = message.index(review_start) + len(review_start)
        index2 = message.index(review_end)
        s = message[index1:index2]
        print("S IS ", s)
        if s != "":
            text = TextBlob(s)
            print(s)
            print()
            score += text.sentiment.polarity
            count += 1
            print(count)
        message = message[index2 + len(review_end):]
      avgSentiment = score/count
      prediction.append(avgSentiment) #send data through sentiment analyses prediction

    finalPredict = mlModel.predict([prediction])
    my_prediction = str(finalPredict[0])[0:5] + "% of total retail price in " + str(year + 1) + "."

  return render_template('results.html',prediction = my_prediction, brand = brand, model = model, year = year)
```

Figure 3. the main algorithm that operates the automation

Machine Learning models operate only if there is enough data to train it. We chose five different example car models with five sets of different historical sentiments as shown in Figure 4. In order to train the polynomial regression model, the sentiments are used as the input data. The output data would be the current value of the vehicle compared to its first release-- found by dividing the used current price by the release price. Each car's data was also extracted from the same website in order to maximize the accuracy of the prediction. This was the only set of data we used since we only used one Machine Learning model which was polynomial regression. We chose this model because it is capable of graphing various curves on an x and y coordinates system. Since sentiments constantly fluctuate, the graph of a polynomial would represent the set of data the best.



```
inputData = [
  #2011 2012  2013  2014  2015
  [0.27, 0.2, 0.26, 0.33, 0.21], #2016 Toyota
  [0.2, 0.26, 0.33, 0.21, 0.3], #2017
  [0.26, 0.33, 0.21, 0.3, 0.35], #2018
  [0.33, 0.21, 0.3, 0.35, 0.25], #2019
  [0.21, 0.3, 0.35, 0.25, 0.33], #2020

  [0.34, 0.33, 0.33, 0.35, 0.42], #2016 Lexus
  [0.33, 0.33, 0.35, 0.42, 0.31], #2017
  [0.33, 0.35, 0.42, 0.31, 0.36], #2018
  [0.35, 0.42, 0.31, 0.36, 0.27], #2019
  [0.42, 0.31, 0.36, 0.27, 0.11], #2020

  [0.29, 0.38, 0.37, 0.44, 0.44], #2016 Mercedes
  [0.38, 0.37, 0.44, 0.44, 0.38], #2017
  [0.37, 0.44, 0.44, 0.38, 0.36], #2018
  [0.44, 0.44, 0.38, 0.36, 0.32], #2019
  [0.44, 0.38, 0.36, 0.32, 0.33], #2020

  [0.38, 0.41, 0.28, 0.39, 0.38], #2016 AUDI
  [0.41, 0.28, 0.39, 0.38, 0.37], #2017
  [0.28, 0.39, 0.38, 0.37, 0.36], #2018
  [0.39, 0.38, 0.37, 0.36, 0.42], #2019
  [0.38, 0.37, 0.36, 0.42, 0.49], #2020
]

#value = currentPrice/releasePrice
outputData = [48, 47, 56, 85, 101, 67, 61, 71, 89, 101, 44, 53, 73, 75, 100, 34, 42, 60, 73, 101]
```

Figure 4. five sets of different historical sentiments

4. Experiment

For our experiment to test the accuracy of our model, we implemented three different models of Machine Learning to be trained by our data. Machine Learning provides several regressions for predicting different types of data. That is why we chose to compare Polynomial Regression, Linear Regression, and SVM to see which one can produce the most precise and reasonable prediction.

A. Comparison of Different Machine Learning Algorithms

In Figure 5 below, it depicts a comparison of the three models we tested. We used a 2020 Toyota Prius and a 2020 Chevrolet Camaro for the experiment. When using SVM, the prediction value is 101 for both cars which is not reasonable since it is way too high. Linear regression is not precise as well because one car got 74 percent and the other got 59 percent, and the graph is a straight line. Finally, Polynomial Regression shows the best prediction value because the result is precise and reasonable where the car value does not increase nor decrease intensely.

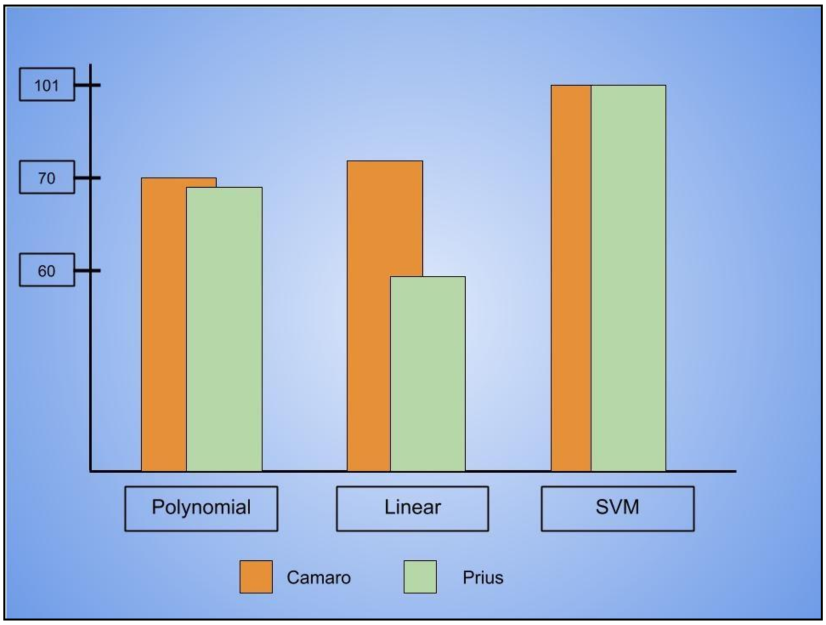


Figure 5. a comparison of the three models

5. Related Work

There are also related works and programs that perform similar functions compared to our program. For example, “Car Price Prediction using Machine Learning Techniques” by Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić is a program that predicts car values using data from autopijaca.ba written in PHP programming language [6]. They used a total of three models: Artificial Neural Network, Support Vector Machine and Random Forest. In contrast, CarEnvision uses only Polynomial Regression from Machine Learning and we collected data from a very universal and professional website-- jdpower.com. Cars that they predicted were also only from both Bosnia and Herzegovina, whereas our program focuses on cars within the United States. CarEnvision’s unique feature that is different from the compared example is the training data which are sentiment reviews on the cars. The experts’ opinions are what supports our prediction which is a major contrast from other projects.

Another related program that uses the same techniques as CarEnvision is “Flood Prediction Using Machine Learning Models” by Amir Mosavi [7]. His program gathers data from “rainfall and water level, measured either by ground rain gauges, or relatively new remote-sensing technologies such as satellites, multisensor systems, and/or radars” (62). Several algorithms they used were multiple linear regressions, quantile regression, and Bayesian forecasting models (34). Both of our programs used Machine Learning models but we used Polynomial regression and they included several different regression techniques. We both had different data as well as predictions; Mosavi predicted floods and we predicted car values. CarEnvision provides a much faster way to gather data since our automation just pulls the data off of a website, but Mosavi’s flood prediction had to get a piece of data daily.

Snow avalanche hazard prediction using machine learning methods by Bahram Moslem Borjia, Amir Mosavibc, Farzaneh Sajedi, Hosseinia Vijay, and P.Singhd Shahaboddin Shamshirband [8] is the last program that is similar to CarEnvision. Their avalanche prediction is based on data from “avalanche occurrence locations, meteorological factors, and terrain characteristics.” They implemented both Support Vector Machine (SVM) [12, 13, 14, 15] and Multivariate

Discriminant Analysis (MDA) as their models to predict avalanches. The only similarity our program has with theirs is we both used models from Machine Learning Algorithms. Other than that, the models fitted for the prediction was different as well as the topic of the programs. The main difference between the two programs is that our sentimental data is used to train both of our models but in different ways, while the avalanche-related data comes from three different areas.

6. CONCLUSION AND FUTURE WORK

In today's world, we are rapidly developing new technology to appease the workload of tens of millions of people around the globe, especially transportation. With modeling technology, CarEnvison makes predicting car prices seem easy. With our experiment, we were able to decide which model from Machine Learning accurately represents the data we want to input. Since the economy of the car industry changes by the minute, we needed Polynomial Regression to accommodate the different dips and ridges that may be created. Since this model represents both high and low data points, the accuracy and precision of our model greatly increased.

As with any invention, there will always be improvement as we develop. We find new ways to make the invention even better and effective. Like other programs, it may not correctly represent all brands or models. CarEnvision is limited in the fact that it cannot pull information from cars that are either not sold in the car industry or too new. In addition, the exchange of cars can include other factors and not only the opinions of the buyers, but also the appearance of the car, features, safety, or speed. As developers of CarEnvison, we solely rely on the opinions of car producers and consumers.

Due to current limitations, CarEnvison needs new ways to improve and increase its precision to help car consumers make the best decision to keep up in this rapidly increasing world of technology. Our future plans include the increase of data and inclusion of more cars and models to make CarEnvison available to more of the public.

REFERENCES

[1] Rietmann, Nele, Beatrice Hügler, and Theo Lieven. "Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO2 emissions." *Journal of Cleaner Production* 261 (2020): 121038.

[2] Ostertagová, Eva. "Modelling using polynomial regression." Procedia Engineering 48 (2012): 500-506.

[3] Peixoto, Julio L. "A property of well-formulated polynomial regression models." The American Statistician 44, no. 1 (1990): 26-30.

[4] Bradley, Ralph A., and Sushil S. Srivastava. "Correlation in polynomial regression." The American Statistician 33, no. 1 (1979): 11-14.

[5] Heiberger, Richard M., and Erich Neuwirth. "Polynomial regression." In R Through Excel, pp. 269-284. Springer, New York, NY, 2009.

[6] Gegic, Enis, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, no. 1 (2019): 113.

[7] Mosavi, Amir, Pinar Ozturk, and Kwok-wing Chau. "Flood prediction using machine learning models: Literature review." Water 10, no. 11 (2018): 1536.

[8] Choubin, Bahram, Moslem Borji, Amir Mosavi, Farzaneh Sajedi-Hosseini, Vijay P. Singh, and Shahaboddin Shamshirband. "Snow avalanche hazard prediction using machine learning methods." Journal of Hydrology 577 (2019): 123929.

[9] Mair, Carolyn, Gada Kadoda, Martin Lefley, Keith Phalp, Chris Schofield, Martin Shepperd, and Steve Webster. "An investigation of machine learning based prediction systems." Journal of systems and software 53, no. 1 (2000): 23-29.

[10] Mackenzie, Adrian. "The production of prediction: What does machine learning want?." European Journal of Cultural Studies 18, no. 4-5 (2015): 429-445.

[11] Weiss, Sholom M., and Nitin Indurkhya. "Rule-based machine learning methods for functional prediction." Journal of Artificial Intelligence Research 3 (1995): 383-403.

[12] Noble, William S. "What is a support vector machine?." Nature biotechnology 24, no. 12 (2006): 1565-1567.

[13] Suthaharan, Shan. "Support vector machine." In Machine learning models and algorithms for big data classification, pp. 207-235. Springer, Boston, MA, 2016.

[14] Joachims, Thorsten. "Svmlight: Support vector machine." SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund 19, no. 4 (1999).

[15] Pisner, Derek A., and David M. Schnyer. "Support vector machine." In Machine Learning, pp. 101-121. Academic Press, 2020.

AUTHORS

My name is **Terence Chen**, one of the founders of CarEnvision. Before high school, I had zero coding experience and did not even know of this field. I started coding in Coding Minds Academy starting freshman year of high school. That was the time when the field of computer science sparked my interest. From then on, I was able to create ideas beyond my imagination through coding. In the future, I see myself going deeper into this field and formulating bigger programs as well as cooperating with many intelligent people.

My name is **Angel Chang**, one of the founders of CarEnvision. Despite having coding experience for only a year, I have been part of the Coding Minds Academy program and it has been a great accomplishment. When I first got my feet wet in the coding world, it has ever since amused me and became a hobby. Besides coding, however, I have devoted my time to playing video games, swimming, and playing water polo. I am currently in my sophomore year of high school in Rancho Cucamonga and aiming to become a future doctor.

AN INTELLIGENT SYSTEM TO IMPROVE VOCABULARY AND READING COMPREHENSION USING EYE TRACKING AND ARTIFICIAL INTELLIGENCE

Harrisson Li[1], Evan Gunnell[2], Yu Sun[3]

[1]Friends Select School, Philadelphia, PA, 19103
[2]3218 Napoli Way, Philadelphia, PA, 19145
[3]California State Polytechnic University, Pomona, CA 91768

ABSTRACT

When reading, many people frequently come across words they struggle with, and so they approach an online dictionary to help them define the word and better comprehend it. However, this conventional method of defining unknown vocabulary seems to be inefficient and ineffective, particularly for individuals who easily get distracted. Therefore, we asked ourselves: "how we could develop an application such that it will simultaneously aim to help define difficult words and improve users' vocabulary while also minimizing distraction?". In response to that question, this paper will go in depth about an application we created, utilizing an eye-tracking device, to assist users in defining words, and enhance their vocabulary skills. Moreover, it includes supplemental materials such as an image feature, "search" button, and generation report to better support users' vocabulary.

KEYWORDS

Eye-tracking, Artificial Intelligence, Vocabulary, Reading Comprehension.

1. INTRODUCTION

The background of my topic arose from the fact that my reading comprehension skills are so poor due to the struggle in vocabulary. When reading an article or book, I constantly must go back and forth between the book and the online dictionary to help me better understand the words in the context. I knew this issue didn't only pertain to me; therefore, I created an app to help individuals, particularly students, improve their vocabulary skills whether it be for educational purposes or just in their daily life. Currently, similar dictionary tools are popularizing and frequently used amongst high school students because humanities teachers often assign students challenging comprehension texts to read.

Some existing tools for this topic include the most basic: going on to the internet to look up the definition of the word. Another tool is this chrome extension created by GoodWordGuide.com called "Instant Dictionary". This extension allows users to double click on a word using their cursors—on any website—they are struggling with, and a mini dictionary will pop up on the same exact screen. There is an app called "Quick Dictionary" which is like Instant Dictionary by the fact that the definition also appears on the same screen. However, the app has an additional

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 27-36, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111703

feature of allowing users to study words they previously searched which is designed to reinforce the user's understanding of the word.

Using online dictionaries is currently the most popular or classic method people use to approach unknown words they come across. The method is generally time-consuming if one wants to find a good definition for their word by checking several websites. Additionally, it's very distracting to readers' focus when they are reading an online article or book. The "Instant Dictionary" extension just doesn't have a very good dictionary, especially for very complicated words, which is really bugging because you would have to look the word through online dictionaries. The "Quick Dictionary" app is only found in the google play store, meaning it is limited to android users, particularly those using the phone. Similarly, the "Quick Dictionary" is also equipped with a less reliable dictionary.

Our application incorporates an eye-tracking device, or simply using the cursor, that provides a side-by-side dictionary with some sort of book or article the user is reading. For that to even happen, users must use either the eye-tracking device or cursor to pause on the word they are struggling with for a specific amount of time to signal that they are having trouble on that particular word. Features of this app help provide definitions from a good-quality, popular dictionary, alongside it carries images (only if the word has one) for an even better understanding of the word. Furthermore, if the dictionary already provided doesn't actually give users a "not so good" definition of the word, there is a "search" button that allows for users to look up the word by providing other dictionary website URLs. This product compared to the other mentioned is simple and efficient. First, users don't need to look up any definitions on the internet because a satisfying definition is most likely already provided; however, if that isn't the case, then it's just simply finding other URLs through the search button. Therefore, the app not only brings forth great definitions but does it in a very efficient way. Users won't have to be concerned about easily being distracted since the dictionary is side by side with whatever the user is reading. All in all, this app is a major progression, especially with the inclusion of an eye-tracking device, making user-experiences much finer compared to that of similar tools/methods!

Successfulness => Yes, the app is not only able to track and define what words users are struggling with, but also provide supplementary materials that can enhance their understanding of those words. Some of the supplementary materials that are implemented consist of an image feature that is complementary to the definition, a "search" button, and a generation report. The additional image feature is provided to reinforce users' understanding of the word's definition. (A quick disclaimer: the image displayed may not always be an accurate representation of the word's definition.) Similar usage to the image feature is the "search" button. This magical button designates users to a dictionary webpage, Merriam Webster, if they are sincerely struggling to understand the word's definition after having read the initial definition and seen the image. Lastly, the generation report is a report on the words users struggled on— meaning a definition and image appeared when a word was either stared at/hovered on for two seconds— throughout the duration users use the application. The report appears after the application is closed by users, and aims to help them efficiently study for the words they struggled with so the words don't become a barrier the next time they come across it.

The application is very successful as it meets the goal, we initially set out to accomplish: using an eye-tracking device to assist in defining challenging words for users. Furthermore, it incorporates additional features that look to increase user experience while improving and enhancing users' vocabulary in reading.

2. Format Guide

The rest of the research paper is organized as follows: Section 3 will go in depth about the challenges that we faced while creating the program/app on different coding platforms; Section 4 focuses on the details of how we overcame and resolved the corresponding challenges mentioned in Section 3; Section 5 presents the specifics about the experiment we completed, following with the related work in Section 6. Finally, Section 7 gives the conclusion remarks, and states how I will continue to improve upon the app I created in the future.

3. Challenges

Challenge 1. Figuring out which words are problem words. We don't know when users are struggling with a word because many people aren't outspoken when in such a situation. Furthermore, the goal of our program is to help users define words they struggle with, so it is vital that our program knows that information. When users usually encounter challenging words, they are given access to either an online or physical dictionary alongside the reading. To make the solution more efficient, we designed the application so that when users hover over a word they are struggling with for exactly two seconds, it will process that information and define the word.

Challenge 2. Text and Image Recognition (Natural Language Processing) often takes time to run, hindering the goal of the program. Very frequently, we see many programs that take an extensive time to load so that they may properly function when users utilize the application. This extended wait time shifts the user's attention away where they will find a similar functioning application with a shorter load time. In our application we used Amazon Web Services' Textract which, though relatively fast, still requires time to preload the textract. This time leads to a situation that is not an optimal user experience. A very widespread solution when dealing with this kind of problem is designing a splash screen that would capture the user's attention, psychologically making them feel like they aren't having to wait for the program to run. Similarly, our solution was to implement multithreading as well as a splash screen to overlap and conveniently pace the time needed to run the textract in the background.

Challenge 3. Creating the UI and application to be as quick and smooth as possible, so as to encourage the user to utilize it. Something that is supposed to assist users in their reading, needs to be quick and responsive so that it doesn't become more of a hindrance. You can move on to it specifically being a challenge. We want this program to help user's learn words faster and more efficiently. If the overall program/ UI is too slow, the program ends up costing them more time. Making the program aesthetically simple/clean as well as very responsive encourages users to use it. We designed the program to have a very simple design with very obvious ways of utilizing it. Simply looking at something or tracking your mouse over it, does the lookup for you. We also used the Python Tkinter library since it is lightweight and usable on most computers.

4. Methodology/Solution

A. Overview of the Solution

An overview of the system is depicted in the figure below. Users simply run the e- Dictionary application on their desktop while reading. Depending on their comfortability and accessibility, they can choose to either hover over words with their mouse cursor (for those who do not have the eye-tracking software) or an eye-tracking device. Regardless, it will display the word's definition and image to the users. Finally, after users finish doing their reading, a generation report, which gathers the words the user struggled with for the duration of using the application, will appear to further assist them with their vocabulary learning.

B. Step/Components in this System

In order to provide a more efficient setting for users' readings and the potential definitions they seek, we designed an application in Python Tkinter that utilizes the Tobii eye tracking hardware. The application provides a Graphical User Interface (GUI), which was used to give the application a general landscape from viewing the definitions to navigating pages akin to an eReader. After completing the GUI, we implemented Amazon Textract to request the word image recognition and labelling from their site to save to our application. As previously mentioned, there are two modes for word identification: one that tracks the user's mouse location to simulate eye tracking and one that receives an eye tracking position. Depending on our decision heuristic (if they hover too long) we use a dictionary library, web page request library, and imaging library, to pull the appropriate definition, image, and relevant searches. Our multithreaded application will poll (check) for any changes to the user's eye position and update it accordingly in the app. Furthermore, we also preload the app during the splash screen. In order to include the Tobii eye tracking hardware for positioning, our eye Tobii Tracking software (in C# and Unity) will constantly save the user's current eye tracking position. When everything successfully operates, all information is displayed back to the user.

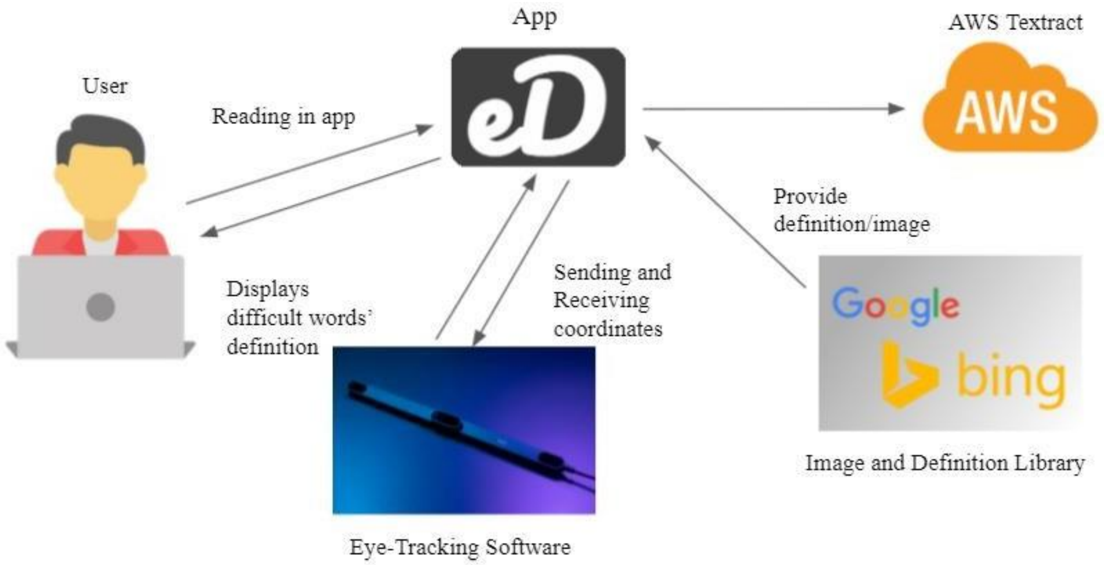
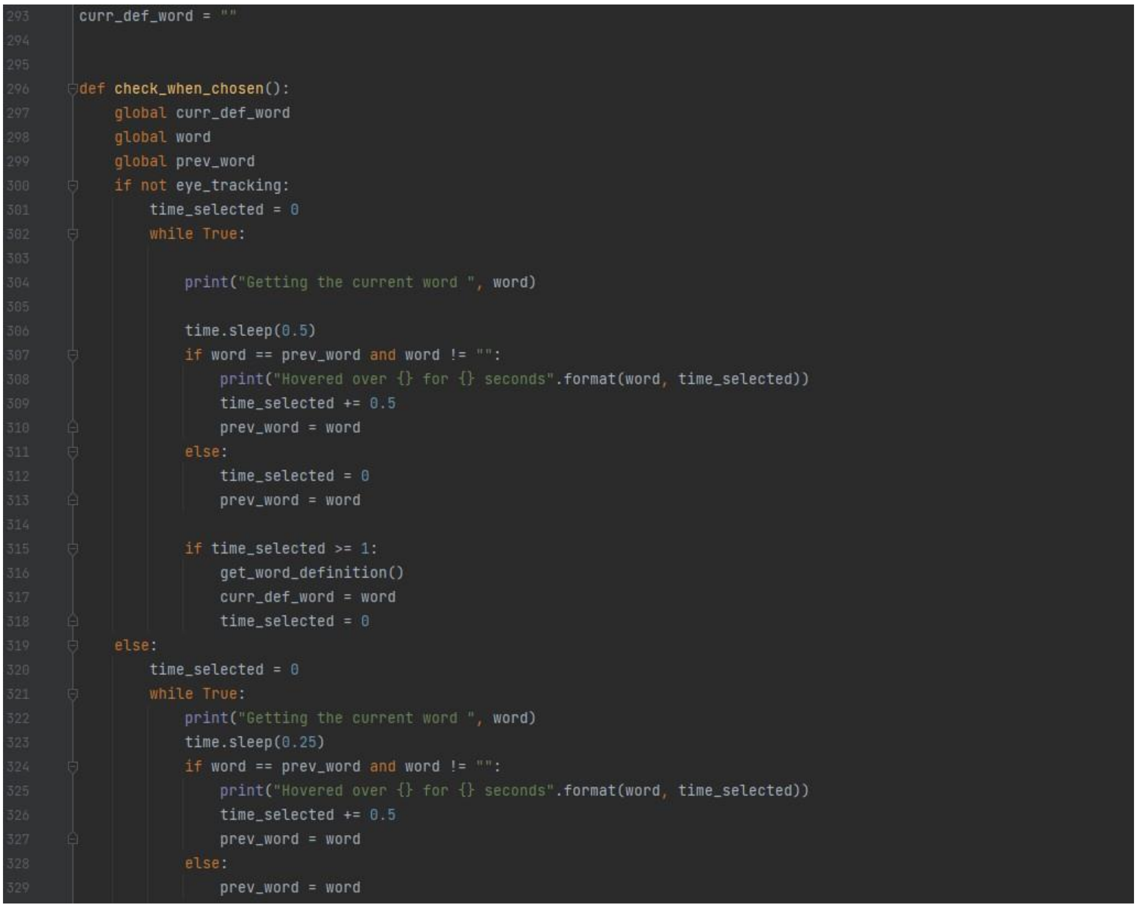

App
eD
AWS Textract
AWS
User
Reading in app
Provide definition/image
Displays difficult words' definition
Sending and Receiving coordinates
Google
bing
Image and Definition Library
Eye-Tracking Software

Figure 1. Overview of the Solution

C. Each Component

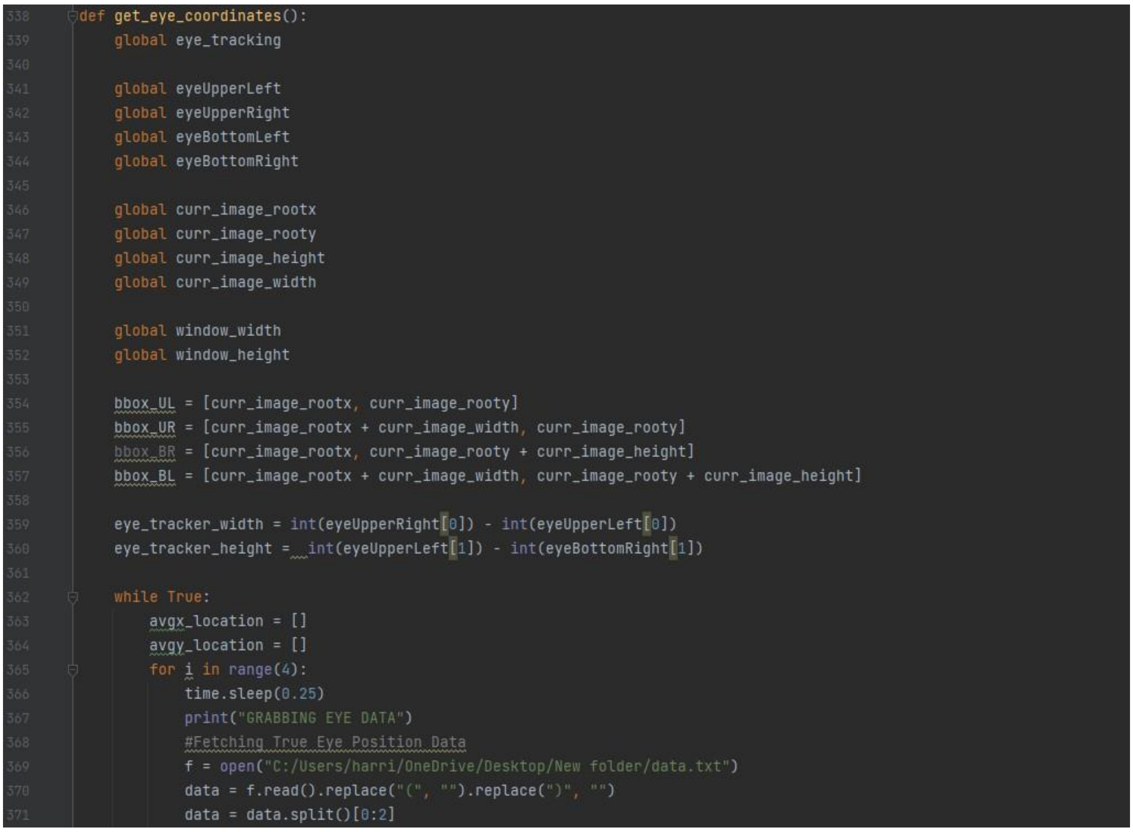
Python Tkinter is a GUI development library built into Python that allows for simple grid-based app development. The GUI development allows users to navigate pages akin to an eReader using the forward/backward buttons. Users can also view the definitions and images for a specific vocabulary through the inclusion of a dictionary and image library. Furthermore, it increases user experience by giving users the option to click a "search" button, bringing them to a designated dictionary webpage, Merriam Webster, if they have further confusion with a word. After designing the application's overall structure in Python Tkinter, we implemented AWS Textract. AWS Textract or Amazon Textract is an Amazon Web Service that takes in pdfs or other pages and extracts the text into bounding boxes for each word. It utilizes Optical Character Recognition

(OCR) to scan through the documents and provide a quick result of the words to the user. Therefore, using either the mouse or eye-tracker location to identify which words users are struggling on, we check their x, y coordinates (starting from the top left) to see which image from the textract library they are focused on. If users use the eye-tracking software, our Tobii Eye-Tracking software (in C# and Unity) will constantly save the user's current eye tracking position. The Tobii Eye tracker uses the Tobii API from the Tobii Unity SDK. It provides highly accurate monitoring of the user's eye positioning. The eye tracker itself grabs the position by monitoring eye and face position along with IR light to help it more accurately see. The API provides the "gaze data" which is the x, y coordinate of the user's gaze relative to the screen. We then save it to a file at a constant rate. To improve the practicability of our application, we inserted multithreading to the code, which will poll (check) for any changes to the user's eye position and update it accordingly in the app. Additionally, it consists of a splash screen that creates an animation while the application is running in the background to increase the overall speed, resulting in better user experience. For the thread reading the eye tracking software, we are using a library called Watchdog that is able to constantly check for file modifications and updates in a specific sub folder.



```
curr_def_word = ""


def check_when_chosen():
    global curr_def_word
    global word
    global prev_word
    if not eye_tracking:
        time_selected = 0
        while True:

            print("Getting the current word ", word)

            time.sleep(0.5)
            if word == prev_word and word != "":
                print("Hovered over {} for {} seconds".format(word, time_selected))
                time_selected += 0.5
                prev_word = word
            else:
                time_selected = 0
                prev_word = word

            if time_selected >= 1:
                get_word_definition()
                curr_def_word = word
                time_selected = 0
    else:
        time_selected = 0
        while True:
            print("Getting the current word ", word)
            time.sleep(0.25)
            if word == prev_word and word != "":
                print("Hovered over {} for {} seconds".format(word, time_selected))
                time_selected += 0.5
                prev_word = word
            else:
                prev_word = word
```

Figure 2. e-Dictionary coding



```python
def get_eye_coordinates():
    global eye_tracking

    global eyeUpperLeft
    global eyeUpperRight
    global eyeBottomLeft
    global eyeBottomRight

    global curr_image_rootx
    global curr_image_rooty
    global curr_image_height
    global curr_image_width

    global window_width
    global window_height

    bbox_UL = [curr_image_rootx, curr_image_rooty]
    bbox_UR = [curr_image_rootx + curr_image_width, curr_image_rooty]
    bbox_BR = [curr_image_rootx, curr_image_rooty + curr_image_height]
    bbox_BL = [curr_image_rootx + curr_image_width, curr_image_rooty + curr_image_height]

    eye_tracker_width = int(eyeUpperRight[0]) - int(eyeUpperLeft[0])
    eye_tracker_height =  int(eyeUpperLeft[1]) - int(eyeBottomRight[1])

    while True:
        avgx_location = []
        avgy_location = []
        for i in range(4):
            time.sleep(0.25)
            print("GRABBING EYE DATA")
            #Fetching True Eye Position Data
            f = open("C:/Users/harri/OneDrive/Desktop/New folder/data.txt")
            data = f.read().replace("(", "").replace(")", "")
            data = data.split()[0:2]
```

Figure 3. e-Dictionary coding

5. EXPERIMENT

Author names are to be written in 13 pt. Times New Roman format, centred and followed by a 12pt. paragraph spacing. If necessary, use superscripts to link individual authors with institutions as shown above. Author affiliations are to be written in 12 pt. Times New Roman, centred, with email addresses, in 10 pt. Courier New, on the line following. The last email address will have an 18 pt. (paragraph) spacing following.

A. Does e-Dictionary solve the problem of assisting users?

Our application showcases extremely quick response times for bringing up the appropriate data that users are expecting. The responsiveness of the application dramatically reduces the plentiful distraction towards users, particularly younger kids, when reading. Furthermore, it can determine what the user is struggling with through the decision heuristic: if they sit on a word for 2 seconds. In our own testing, this application had a positive impact on vocabulary-comprehensive reading. Additionally, it carries out useful data from the generation report to reinforce vocabulary learning for users, hoping it will make reading easier for users as they are exposed to more vocabulary (and their definitions) in the long-term.

B. Are we able to integrate eye tracking to an application like this? Does it work?

Yes, we were able to integrate eye-tracking into our application through using the Tobii API from the Tobii Unity SDK and Watchdog library. Unfortunately, due to the inconsistency of the Tobii eye-tracking software, the device does not work well with the eDictionary application. When experimenting with the eye-tracking software several times, we noticed that the coordinates displayed from the eye-tracking software of the word were incompatible with the coordinates displayed when utilizing the mouse cursor. This posed a challenging problem for us, so we decided to have users, for now, use the traditional mouse cursor as they wouldn't frequently encounter problems with the eye-tracking software.

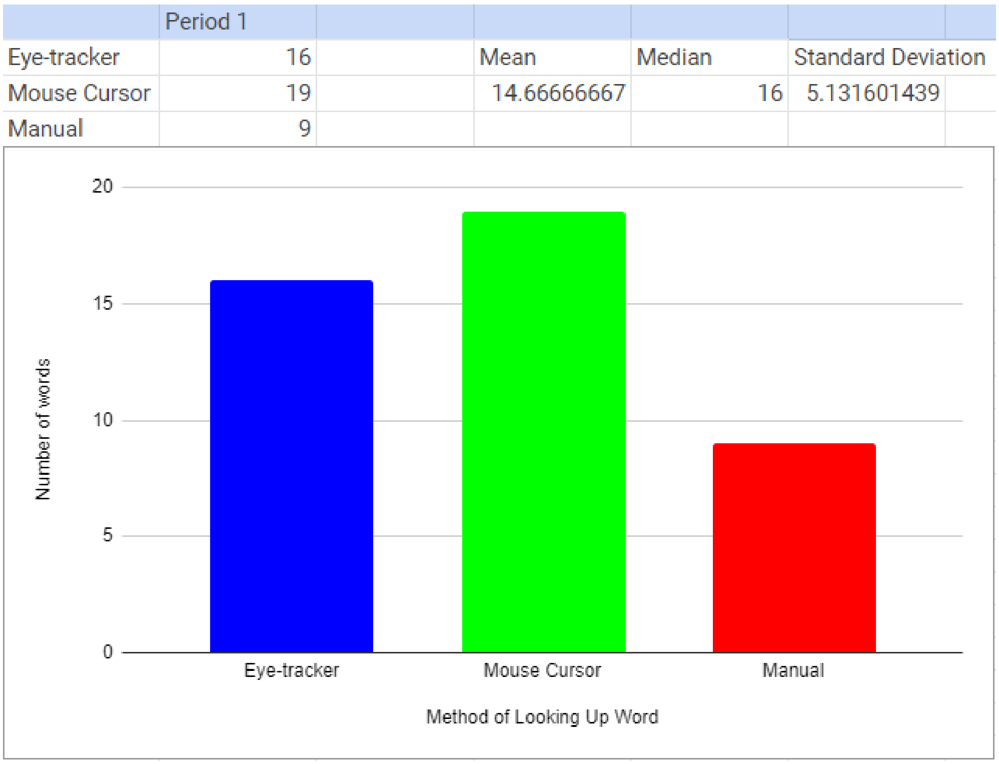
C. Did the final program meet the requirements we set out at the beginning of the paper? Did we manage to accomplish anything?

The implementation of an automatic eye-tracking device to the application did indeed meet the needs of the problem we were looking to solve. It is able to analyze specific words the user wants to comprehend; using that information, the remainder of our application will utilize Amazon Textract to provide users with both a visual and text-based definition to the challenging vocabulary, instantly. Overall, our application is able to successfully provide not only assistance in a difficult reading language, but doing so while being fully automated. Through our own testing, this application serves as a wonderful digital assistance to those looking to learn complicated words in new fields or even in a new language, English. The ease of not needing to do anything beyond reading the text while still maintaining active searching of complex words truly helps users with their readings.

D. Summary

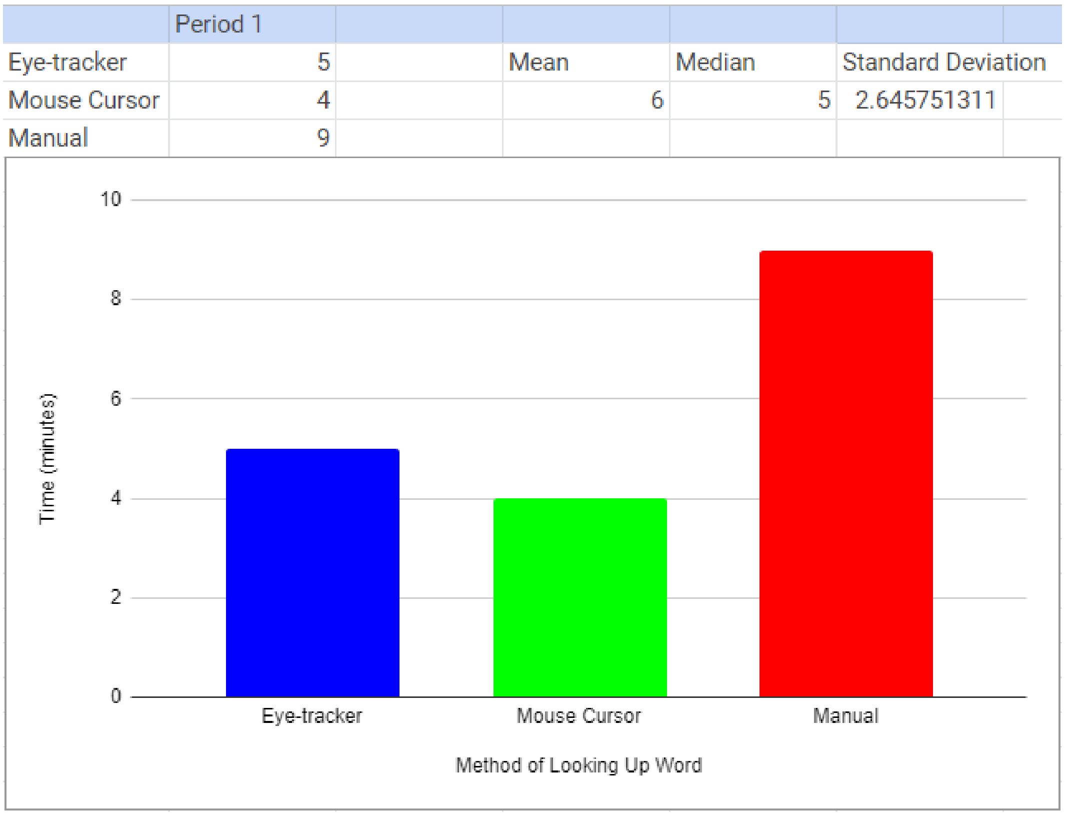
The responsiveness and high user experience of e-Dictionary effectively undermines user's quirk of getting frequently distracted while looking up words on a dictionary application. Furthermore, with the inclusion of an eye-tracking software, it further reinforces the efficiency of users having to look up very demanding words and also improves their vocabulary. All in all, the attributes of this application as mentioned above, do satisfy the criteria of the problem we initially set out to solve.

E. Data



	Period 1					
Eye-tracker	16		Mean	Median	Standard Deviation	
Mouse Cursor	19		14.66666667	16	5.131601439	
Manual	9					

Figure 4. Number of words looked up in 5 minutes.



	Period 1					
Eye-tracker	5		Mean	Median	Standard Deviation	
Mouse Cursor	4		6	5	2.645751311	
Manual	9					

Figure 5. Time taken to get through one SAT Reading passage with challenging vocabulary.

6. RELATED WORKS

Related Work 1. Katsuyu Fujii and Jun Rekimoto created an application called SubMe, a smart subtitle system with a machine learning algorithm for estimating users' English levels. The goal of this application was to estimate English levels through using an eye tracker and gathering eye gaze movement data. Unlike my work which was designed to help improve user’s English skill, specifically in concentration of their vocabulary, Subme was used to estimate English levels through using a smart subtitle system and eye tracking data. One strength of this application is being able to gather data, which hypothetically is more convenient when trying to improve a problem than that of an application that primarily aims to help improve a user's English.

Related Work 2. Elizabeth Krupinski and Josh Borah researched how the use of an eye tracking technology correlates with the improvements of speed and reading of radiological readings. They concluded that there was a positive correlation between the two; different eye tracking systems had the ability to track, record, and analyze eye movement which not only improved image reading but also could easily detect flaws during the radiological readings. My work was geared towards creating an application, while Elizabeth and Josh researched how eye tracking technology helped improve radiological readings. However, both our work demonstrated the benefits of an eye tracking technology’s functionality in a variety of fields.

Related Work 3. A group of 6 people developed an application called Private Reader; it implements an eye tracker that shows only the portion of text that the user is focused on while reading. The goal was to maintain privacy from nosy people in public work spaces while the users were reading. The application created by the group, I would say, is very close in difficulty to the application I created. Furthermore, their application in some ways is definitely better than mine. For example, one strength their work acquired was having 6 people collaborate together on the project, which did not allow for more ideas but also higher efficiency when creating the application. Additionally, before creating the application they did a user study that would give users a better user experience, which is beneficial when promoting and selling the products to others.

7. CONCLUSION AND FUTURE WORK

In this project, we had the idea of assisting people who struggled with vocabulary in their daily lives. One of the main problems with learning vocabulary is that of actually looking up and understanding the definition of a word. The deeper one delves into more complex readings at a greater level of education, the more frequent one will pause to look up the definition of a word. Therefore, to ameliorate this problem, we designed an application that not only will make looking up the word’s definition faster for users, but also automate the process for them. We do this in two different ways: 1. having the definition pop up for the user on the side when they hover over a word with the eye-tracking device 2. having the user click the “search” button, bringing them to a dictionary webpage, Merriam Webster, if they have further confusion for the word. Through frequent experimenting with the application, results show that it is very successful. Although the project is not perfect, it carries many features that make it unique from other dictionary webpages or applications, such as implementing an eye-tracking device, providing a combination of images and definitions, and equipping a “search” button that offers an alternative for looking up definitions.

One big limitation of the project is the accuracy of the image API, which is supposed to increase efficiency of vocabulary learning for users. However, it often provides images that are not compatible with the word users are struggling with, resulting in further confusion for them.

Another limitation is the practicability of the eye-tracking application; whether it’s the splash screen or tracking of a word, a good portion of the application remains to be somewhat choppy when users are using it. Finally, even though the optimization of the application is definitely not the greatest (as there could be more features added to it to make it seem more multifaceted), it is certainly usable for users.

In the future, I hope to find a better image API to implement into the application that will result in better user experience. Furthermore, I hope to increase the responsiveness and speed of the application by executing future touches to the programming. Last but not least, expanding the use cases to improve user experience, such as adding features like translations of words, analyzing data for users’ vocabulary, a thesaurus accompanied by the dictionary, all will require further programming and time dedicated to the application.

REFERENCES

[1] Engdahl, Sylvia. “AWS Machine Learning Blog.” Amazon, Greenhaven Press/Gale, 30 May 2019, aws.amazon.com/blogs/machine-learning/automatically-extract-text-and-structured-data- from-documents-with-amazon-textract/.

[2] Developers, Tobii. “Unity Sdk.” Tobii Developer Zone, 14 July 2021, developer.tobii.com/pc-gaming/unity-sdk/.

[3] Amos, David. “Python Gui Programming with Tkinter.” Real Python, Real Python, 3 Apr. 2021, realpython.com/python-gui-tkinter/.

[4] Khandelwal, Renu. “Monitoring Your File System Using Watchdog.” Medium, Analytics Vidhya, 11 Jan. 2021, medium.com/analytics-vidhya/monitoring-your-file-system-using- watchdog-64f7ad3279f.

[5] Katsuya Fujii and Jun Rekimoto. 2019. SubMe: An Interactive Subtitle System with English Skill Estimation Using Eye Tracking. In Proceedings of the 10th Augmented Human International Conference 2019 (AH2019). Association for Computing Machinery, New York, NY, USA, Article 23, 1–9. DOI:https://doi.org/10.1145/3311823.3311865

[6] Krupinski, E., & Borah, J. (2006). Eye tracking helps improve accuracy in radiology. Biophotonics International, 13(6), 44-49. https://arizona.pure.elsevier.com/en/publications/eye- tracking-helps-improve-accuracy-in-radiology [3] Amos, David. “Python Gui Programming with Tkinter.” Real Python, Real Python, 3 Apr. 2021, realpython.com/python-gui-tkinter/.

[7] Kirill Ragozin, Yun Suen Pai, Olivier Augereau, Koichi Kise, Jochen Kerdels, and Kai Kunze. 2019. Private Reader: Using Eye Tracking to Improve Reading Privacy in Public Spaces. In Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19). Association for Computing Machinery, New York, NY, USA, Article 18, 1–6. DOI:https://doi.org/10.1145/3338286.3340129

[8] Mehvish, Mehvish. “7 Best Dictionary Extensions for Chrome.” Guiding Tech, 22 June 2020, www.guidingtech.com/best-dictionary-extensions-chrome/.

[9] “Quick Dictionary – Apps on Google Play.” Google, Google, 18 July 2020, play.google.com/store/apps/details?id=com.yaki.wordsplash&hl=en_US&gl=US.

AUTHORS

Hi! I'm **Harrisson Li**, the inventor and creator of e-Dictionary. I'm a high school senior that resides in Philadelphia; I’m passionate about basketball and math. When being asked about problems that I face in my daily life, I immediately thought of my poor vocabulary skills and lazy personality. From there, I thought of creating something that would make using dictionaries more efficient for people. With my enthusiasm in technology and programming, I decided to create e-Dictionary.

CLOUD COMPUTING STRATEGY AND IMPACT IN BANKING/FINANCIAL SERVICES

Prudhvi Parne

Information Technology, Bank of Hope,
1655 E Redondo Brach Blvd, Gardena, CA, USA

ABSTRACT

With recent advances in technology, internet has drastically changed the computing world from the concept of parallel computing to distributed computing to grid computing and now to cloud computing. The evolution of cloud computing over the past few years is potentially one of the major advances in the history of computing. Unfortunately, many banks are still hesitant to adopt cloud technology. New technologies such as cloud and AI will have the biggest impacts on the banking industry. For banks and credit unions wanting to achieve greater business agility, cloud technology enables organizations to respond instantly to changing market conditions, leveraging data and applied analytics to achieve customer experience and operational productivity benefits. As a result, cloud computing comes in to provide a solution to such challenges making banking a reliable and trustworthy service. This paper aims at cloud computing strategy, impact in banking and financial institutions and discusses the significant reliance of cloud computing.

KEYWORDS

Cloud Computing, Technology, Finance, Security.

1. INTRODUCTION

Banks have always relied on legacy systems from historic times, and in most cases, they are usually reluctant to embrace changes to their technology infrastructure. Nevertheless, with increasing concern for customer data, many banking institutions are now considering shifting to cloud computing. Although the shift is quite intimidating, cloud computing is promising banking institutions to award them immense benefits, which will assist them in overcoming the problems they have been facing with the legacy systems in the recent past. Fintech's recent survey indicated that 46% of bankers perceive these trials as the most significant obstruction to the development of commercial banks [1]. Therefore, as the banking institutions embark on the journey, the paper will reflect on how the banks will be impacted by the implementation of cloud computing into their daily operations. Cloud computing consists of a set of resources and services offered through the Internet. Hence, cloud computing is also called Internet computing [18].

2. CLOUD COMPUTING

Cloud computing is defined as a model for enabling banks to have convenient and on-demand network access to a communal pool of configurable computing resources. Such include servers, networks, applications, storage, and services that can be swiftly released and provisioned with little administration effort or the level of interaction with the service provider. Cloud computing promotes accessibility through various attributes, including on-demand services where the client

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 37-45, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111704

can enjoy unilaterally providing computing abilities [4]. These encompass network storage and server time depending on the organization's requirements without necessarily calling for human interaction with the service provider. Another attribute is broad network access, which is strengthened over the network and can be easily accessed via standard mechanisms that encourage various thick or thin client podiums, including laptops, mobile phones, and PDAs. Resource pooling is the situation where the service provider allows the banking institution to pool their resources such that they can be used to serve numerous clients through the multi-tenant model that is empowered with different virtual and physical resources. These resources can be assigned or reassigned dynamically depending on the consumer demands.

Looking at Cloud Computing Software as a Service (SaaS), we must acknowledge that, this is the package that is used all over the world. In fact, this package allows the provider to license associate use to customers either as a demand or as a service, and this is done through subscription in what is seemingly termed as pay-and-go model. Sometimes this is also increased at no charge in cases where there could be the likelihood of revenue induction from diversified streams apart from the user [6]. For instance, from the promotion or user list sales SaaS which might be growing in the market as already experienced. This anecdote is critical because it is heralding the fact that this software has become a commonplace and so it is critical that various customers and technology users get to know what SaaS connotes and its applications. SaaS offers solutions for organizations and companies which are providing access to various flow of labor, CRM, CMS, analytic and even the third-party services and all these are pegged upon a pay-per-use model [8]. It is also a platform which is professionally administered and protected 24/7 and 365 days a year throughout a remote datacenter. This provides state-of-the-art of Cloud computing technologies [13].

3. Cloud Computing Models

Cloud computing promises to improve banking institutions by providing them with a variety of software to select from. The entire cloud architecture [16] is aimed at providing the users with high bandwidth, allowing users to have uninterrupted access to data and applications, on-demand agile network with possibility to move quickly and efficiently between servers or even between clouds and most importantly network security.

The first option is using the software as a service that helps develop banking institutions. Through this software, the micro banking institutions can easily access the internet-hosted software services using the browser instead of relying on traditional applications that have been stored in their server or computer. Depending on software as a service, the application host is responsible for maintaining and controlling the application comprising software settings and updates. This software package has been revered and it has become very popular and fashionable recently for various reasons. For example, it is engaging to the clients because of its shifts in cumbrance and value both the hardware and software package preparation and how it is maintained from the perspective of the client to the seller. SaaS is conjointly offering numerous benefits to the seller. In fact, WHO has recently developed and maintained this application on their platform and have permitted the customer use option of their applications.

The second option is infrastructure as a service which means that the banking institution can buy or rent the disk space and computers from an internet service provider. This will enable them to access information over the internet or private network. The provider maintains the physical computer hardware such as memory, CPU processing, network connectivity, and data storage. Examples of this software include Windows Azure. This is regarded as a fashion for delivering Cloud Computing through an infrastructure-like servers, storage, network associate degrees operative systems and as associate degree-on demand service [7]. It prevents the clients for shopping for the servers, datacenters, and software or network instrumentation, rather they can only buy various resources like sometimes an extremely outsourced service on demand.

The last option is the platform as a service that will enable the banking institution to rent operating systems, hardware, network capacity, and storage [17] provided by infrastructure as a service and the corresponding software application and servers environments. Platform as a service offers banks top-notch control over their technical aspects of the computing setup and customize to suit their needs. However, this must be done or applied in the package development environments. PaaS is similarly printed to be a computing platform that allows the establishment of web applications faster and easily and while not the quality of buying and maintenance of the package and infrastructure at a lower end [6]. It is also regarded as analogous to SaaS only that, instead of being package delivered over the web, it is a platform for the creation of the package, delivered over the web.

4. Different Types of Cloud Deployments

In contrast to the models discussed above, which define how services are offered via the cloud, these different cloud deployment types [11] have to do with where the cloud servers are and who manages them.

The most common cloud deployments are:

Private cloud: A private cloud is a server, data center, or distributed network wholly dedicated to one organization.

Public cloud: A public cloud [12] is a service run by an external vendor that may include servers in one or multiple data centers. Unlike a private cloud, public clouds are shared by multiple organizations. Using virtual machines, individual servers may be shared by different companies, a situation that is called "multitenancy" because multiple tenants are renting server space within the same server.

Hybrid cloud: hybrid cloud deployments combine public and private clouds and may even include on-premises legacy servers. An organization may use their private cloud for some services and their public cloud for others, or they may use the public cloud as backup for their private cloud.

Multi-cloud: multi-cloud is a type of cloud deployment that involves using multiple public clouds. In other words, an organization with a multi-cloud deployment rents virtual servers and services from several external vendors to continue the analogy used above, this is like leasing several adjacent plots of land from different landlords. Multi-cloud deployments can also be hybrid cloud, and vice versa.

5. Factors Influencing the Adoption of Cloud Computing in Banks

From a different perspective, the shift to adopt cloud computing technology in the banking institution is influenced by various factors. The first factor is a classification of information sensitivity in the banking system. Banks have the freedom to select the kind of system suitable for them depending on the data the institution will be processing as permitted by the government regulations and data security [19] necessities. As a result, the banking organization must ensure that the cloud computing they will decide on meets all the qualifications [3]. For instance, the

United States carefully evaluates their outsourcing procedures and ensure that the banks apply them strictly. The second factor is the competitive advantages and differentiation capacities availed by Cloud Computing. For this, the bank will decide on the choice of computing for the organization depending on the level of competition existing in the industry. As a result, the organization will consider cloud computing an effective alternative to turn to outdo the competitors successfully. Similarly, if there is a need for the bank to differentiate its products, in-house development is more preferred by the institution owing to its swiftness in marketing its expectations.

6. Positive Impacts of Cloud Computing on Banks

Cloud computing has continued to gain popularity as it acts as a catalyst through which banks and financial institutions can rely on transforming the features of their monetary services and tailoring them according to the customers' needs. As banks continue to adopt cloud technology, it becomes essential to propel the institutions towards a positive future. One of Cloud technology's paramount importance to banks and financial institutions is that it creates no need to develop an up-front investment in infrastructure such as software licenses and does not attract the risk of unused licenses. Similarly, cloud technology does not demand investment in hardware components and other associated maintenance services. Therefore, capital expenditure that would have been spent in meeting these qualifications is turned into operational cost, thus enabling the institutions to achieve their goals. The consumers of cloud technology are only needed to use the number of IT resources they initially need and will only be required for the volume of technology they have used.

The second impact of cloud computing on banks and financial institutions includes that cloud computing facilitates fast and easy scaling of the computing resources required to sustain the organization's cloud operations. Cloud technology also plays a crucial role in diminishing ongoing operating, upgrades, and maintenance costs through the utility model. This payback is often described as an instant outcome of the technology. Scaling [15] the technology up and down network capacity, hardware, and cost based on demand can be expensive for large-scale financial institutions because they deal in a wide range of data. Therefore, cloud computing enables banks to easily add or eliminate resources at a fine grain along with a lead time of few minutes instead of several weeks as they wait for the matching of resources to workload closer. Peaks mainly influence this demand for cloud technology resources. The process of waiting can end up attracting a lot of complications for the organization in determining the definite number of servers that an institution requires to execute its tasks without disruption or experiencing data breaches. In most situations, the response regarding the organization's needs is usually based on the cost-benefit analysis. However, with the concept of cloud computing, organizations should forget all about the investigation because the technology offers them a flexible solution for the consistent change in IT resources demand.

Another impact of cloud computing is that it has increased in availability compared to other in-house solutions by intensifying the accessibility of Virtual Machine, which strengthens the ability of the organization to create a customized environment above the physical infrastructures [2]. Cloud technology enables organizations to access a wide variety of numerous applications and attributes. One of them is Software as a service, a virtualization technique fully exploited when employing the cloud computing model. This is an indication that software applications can be easily accessed via the web interface. The significance is that this application portfolio is getting more dynamic concerning the transformations of the banks as a way of meeting consumer behaviour change. Through cloud computing, applications can be quickly deleted or added to the firm's portfolio within a short period. Nevertheless, cloud computing attracts minimal maintenance costs because it acts as the primary catalyst for innovation. Cloud computing is becoming more affordable and more universal, thus creating the opportunity for innovation to continue to grow to new heights.

Besides, cloud computing comes with high data security, thus eliminating the thought of losing valuable data to hackers, which can be drastic to the bank and the customers. In the long run, data breaches tend to be costly, racking up in millions and dollars, thus increasing the opportunity that the institution that has been rigged will be expelled from the market as customers will lose faith in them [5]. while data breach threats are on the rise owing to numerous advancements in technology providing hackers with immense techniques to crack organizations' passwords, cloud computing is paving a long-lasting solution to these problems to hinder hackers from tampering with important information. Cloud computing ensures that the bank has access to an up-to-date customer-centered platform strengthened by complicated password combinations to protect banking data.

Cloud computing has strengthened collaborations where partnership in a cloud environment gives a firm the capability to share and communicate frequently and efficiently outside the old techniques. Cloud computing can provide all workers, sub-contractors and third parties access to all the files, primarily if the banks work in various locations. There is a possibility of choosing a cloud computing model that seems too easy for the banks to share their archives with their advisers. On the contrary, cloud technology is providing banks with improved efficiency in their operations. The financial services provided by these organizations are aided by cloud technology to streamline all the processes using enhanced efficiency. Both sellers and buyers are connected in the payment procedures on a communal application. This is very necessary because it improves the speed of transactions and tracking of data becomes simple. Besides this, cloud computing helps in the continuity of the business since it can assist the financial services firms and banks with fault tolerance, data protection, and recovery from disasters for financial companies. Data Mining is used for extracting potentially useful information from raw data. The implementation of data mining [20] techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.

Moreover, cloud computing incurs lower prices compared with traditional solutions. Nevertheless, agility and transformation are on the list. This involves the financial administrations experiencing shorter development cycles for the novel products via the supple cloud-based operating replicas. The associated technology ropes the quicker and more efficient replies to the requirements of the latest banking customers [5]. It helps the firm shift non-critical facilities, including software patches, maintenance, and other computing matters. This aids the financial companies to concentrate more on business growth and expansion. Also, the flexibility of the work practices allows the employees to be flexible, especially in their work practices. For instance, increase the capacity to access data from homesteads, on public holidays, or even through the commute to and from the job. If there is a need to access the data while off-site, it is possible to connect to the virtual office faster and straightforwardly. Finally, is the access to automatic updates. The necessities can be involved in the service fee. The system will often be updated with modern technology depending on the cloud computing service provided. The current date versions of software can be included and the advancements to servers and computer dispensation power.

7. NEGATIVE IMPACTS OF CLOUD COMPUTING

Despite the various aids of cloud computing knowledge, most financial institutions are still not able to adopt it. Some of the challenges that are barriers to the banks from implementing it are data and security privacy. Sensitive information is contained in the bank data and keeping it safe from a cyber-breach is a must for all banking sectors. There is no exception regardless of the technology; security has to be tight and remains an issue. The security breaches occurrences are inventible but avoidable [2]. Also, regulatory and compliance are considered where all banks are authorized to comply with strict standards. A lot of the banking regulators need the client's financial information situated in the same nation. Specific compliance guidelines need treasured data not to be mixed with the other data on the database or shared servers. The Data Cloud [14] allows organizations to unify and connect to a single copy of all of their data with ease. The result is an ecosystem of thousands of businesses and organizations connecting to not only their own data, but also connecting to each other by effortlessly sharing and consuming shared data and data services. Lastly, there is no complete control of severe firms' submissions, and data is an essential concern for financial organizations. If a third-party handles cloud service providers, they might lessen the ability to be supple and elegant. Therefore, not having control of an enormous volume of data dissuades organizations from moving to the cloud.

8. CASE STUDY

Deutsch Bank is a financial institution offering investment and commercial banking, transaction banking, retail banking along with wealth administration services and products to organizations, organizational financiers and government, private individuals and large and small businesses. Deutsch Bank is among the Germany’s well performing banking institutions with a robust brand position across Europe as well as robust presence in Asia Pacific and American regions. Google on the other hand is a platform that offers organizations with leading infrastructure, industrial solutions and also enhances platform capabilities. The platform’s objective is to deliver top-notch cloud solutions that leverages Google’s superior technology to assist corporations to function more effectively and adjust to the transforming wants of their consumers as a way of providing them with a firm foundation for their future financing decisions. customers across 150 countries have been turning to Google cloud as their reliable partner to address the dire business challenges facing them.

Google Cloud and Deutsch bank entered a multi-year contract in July and with the ink already dried up now, these organizations have proceeded to outline a number of ways that the bank will be using the digitalization of their processes. The 150-year-old banking institution is looking forward to utilizing the Google cloud for several number of their upcoming products. These comprises of innovative loaning products to offer upkeep to ‘pay per use’ systems as a way of providing alternative to procuring assets outright and developing a freshly innovated interface for consumption by their clients. The company’s Autobahn podium which is tasked with providing access to Deutsch Bank research, analytics and commentary will also be getting a makeover to establish more individualized experiences.

On a more typical theme, providing robust security is a slogan for Deutsch Bank as it will provide them with high level functionality and administer their encryption keys and select the data analytics region that can be used in deploying the applications. With the improvement, resilience and flexibility is expected to improve significantly without compromising the organization’s goal on providing privacy and security to secure customer information as well as the Deutsch Bank information assets. This collaboration will be a significant win for both Google cloud and Deutsch Bank releasing proposal in February. The utilization of shift to the cloud to the press materials proposes a significant overhaul. Regarding to the new statement released in October of 2019, an internal memo highlighted a total of $ 12 billion investments for the technology with an assurance to strengthen cloud computing strategy. For Google, Deutsch Bank was a perfect match into one of the three chief consumer verticals of healthcare, finance and retail.

According to Deutsche Bank's chief technology, innovation and data officer, Leukert, the new chapter to the bank as it would open numerous opportunities to the institutions in the long run. The officer continued that with Google Cloud by their side, they will be able to have a strategic partner that will play a significant role in accelerating their technological transformation enabling them to utilize a safe and flexible environment for them to swiftly deliver new services and products to their users [9]. The officer continued that this transformation to join the cloud computing is a blueprint to bring together the respective strengths inside technology and banking for the benefits of the Deutsche clients.

Deutsch Bank and Google Cloud are developing a plan that will precisely co-innovate with numerous encouraging establishments to Fintech by providing accessibility to a wide implementation of the bank's cloud-native offerings. The bank's client will benefit significantly from the innovation as it will play a great role in reshaping how the bank will be designing and delivering their financial products and services. With rapid application development, along with the effective usage of forward-thinking artificial intelligent and wide-ranging data analysis gears, the organization will be in a profitable position to respond to the increasing flexibility and accurately solve pressing problems, customer needs and trends.

Additionally, Robin Enslin the president of Google cloud pointed that the cloud computing will also focus on advancing the company's Mobile self-service selections, artificial intelligence-oriented endorsements and other helpful novelties, the company expects that the customer's banking experience will transform greatly [10]. The president continues that the partnership will play a great role in reviving numerous innovations and further develop the bank to emerge as the best early expertise adopter. Deutsch Bank is a pioneer in the service industry and the Google Cloud computing pointed out that they could not be more ecstatic to partner with such an important market leader.

These partners are also considering various areas to explore to make the best out of their partnerships. Such include providing new lending commodities to provide ample support to the pay-per-use systems as a brilliant alternative to secure the right asset outright. The partners are also developing an intuitive and unified interface for the retail consumers across their member countries to more easily access the various products offered by Postbank and Deutsche bank products. Lastly, Google Cloud and Deutsch banks are exploring numerous strategies to aid in enhancing the Autobahn banking platform to become an award-winning electronic service that will be providing organizational and company clients with the power to establish more individualized experiences and recommendations. In attempt to expand their customer outreach, Deutsch bank is making strategies to list all Google Cloud merchandises on the Google Cloud marketplace to propel a wider implementation of the institution's new cloud native solutions and services. Significantly, the shift to the cloud computing will provide the Deutsch Bank with an upper hand of exploiting an up-to-date information and fully administered atmosphere for their applications. With the new partnership underway, Deutsch Bank will be able to select the type of datacenter region that will work well with their applications while they are being deployed to accommodate for data location roles or preferences. The bank's applications will be in a position to encrypt data both in transit and at rest. Google Cloud will also offer the Deutsche bank with the functionality it needs to allow the bank to effectively manage their chief encryption keys.

9. Conclusions

As a concluding remark, it is clear that cloud computing has completely transformed the face of banking. Cloud computing have brought more affordable and reliable means through which banks can rely on to ensure that their customers information is well protected. Other positive impacts include its high scalability levels and vast capacities to support collaboration between different shareholders in the organization by allowing free exchange of information among them. Despite the positivity, cloud computing has also brought along some among them including that the costs are still extremely high for micro enterprises to adopt it.

REFERENCES

[1] Tiwari, S., Bharadwaj, S., and Joshi, S. (2021). A Study of Impact of Cloud Computing and Artificial Intelligence on Banking Services, Profitability and Operational Benefits. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(6), 1617-1627.

[2] Kaya, F., Van Den Berg, M., Wieringa, R., and Makkes, M. (2020, June). The Banking Industry Underestimates Costs of Cloud Migrations. In 2020 IEEE 22nd Conference on Business Informatics (CBI) (Vol. 1, pp. 300-309). IEEE.

[3] Rieger, P., Gewald, H., and Schumacher, B. (2013). Cloud-computing in banking influential factors, benefits and risks from a decision maker's perspective.

[4] Parry, R., and Bisson, R. (2020). Legal approaches to management of the risk of cloud computing insolvencies. Journal of Corporate Law Studies, 20(2), 421-451.

[5] Kshetri, N. (2010, January). Cloud computing in developing economies: drivers, effects, and policy measures. In Proceedings of PTC (pp. 1-22).

[6] Marcel Decker, "Security of the Internet", The Froehlich/kent reference work of Telecommunications vol, 15.

[7] www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/vk5/ report.html.

[8] Rerns Grobauer Tobias Walloschek, Elmar Stocker,Co published by IEEE Computer and Reliability Societies" March-2011, Pg 53.

[9] Ali, A. (2021). Case Study-Deutsche Bank.

[10] Villar, A. S., & Khan, N. (2021). Robotic process automation in banking industry: a case study on Deutsche Bank. Journal of Banking and Financial Technology, 1-16.

[11] Ali, A. (2021). Case Study-Deutsche Bank.

[10] Villar, A. S., & Khan, N. (2021). Robotic process automation in banking industry: a case study on Deutsche Bank. Journal of Banking and Financial Technology, 1-16.

[11] S Sokolov, O Idiriz, M Vukadinoff, S Vlaev (2020). Scaling and automation in cloud deployments of enterprise applications, Journal of Engineering Science and Technology Review jestr.org.

[12] B Hayes (2008). Cloud computing, https://dl.acm.org/doi/fullHtml/10.1145/1364782.1364786 dl.acm.org.

[13] L Wang, G Von Laszewski, A Younge, X He, M Kunze, J Tao & C Fu (2010). Cloud computing: a perspective study generation computing, 2010. Springer.

[14] N Antonopoulos, L Gillam (2017). Cloud computing. Springer, 191-233.

[15] M Armbrust, A Fox, R Griffith, AD Joseph, R Katz, A Konwinski, G Lee, D Patterson, A Rabkin, I Stoica, M Zaharia (2010). A view of cloud computing, dl.acm.org. Volume 53, Number 4 (2010), Pages 50-58.

[16] Y Jadeja, K Modi (2012). Cloud computing-concepts, architecture and challenges. International Conference on Computing, Electronics and Electrical Technologies (ICCEET). IEEE.

[17] DC Marinescu - 2017. Cloud computing: theory and practice, second edition. 195-230.

[18] MNO Sadiku, SM Musa, OD Momoh (2014). Cloud computing: opportunities and challenges. IEEE potentials, Volume: 33, Issue: 1, Jan.-Feb. 2014. IEEE.

[19] JF Ransome (2017). Cloud Computing: Implementation, Management, and Security. taylorfrancis.com. 153-182.

[20] M Rambabu, S Gupta, RS Singh (2021). Data Mining in Cloud Computing: Survey. Innovations in Computational Intelligence and Computer Vision. Springer, pp 48-56.

AUTHORS

Prudhvi Parne received the Master's (MS) degree in Computer Science from University of Louisiana, Lafayette, LA, USA. His expertise spans in the areas of Cloud Architecture, Software Development, Finance, Banking, Hybrid clouds, Product Management, and Product leadership.

AN ADAPTIVE AND INTERACTIVE 3D SIMULATION PLATFORM FOR PHYSICS EDUCATION USING MACHINE LEARNING AND GAME ENGINE

Weicheng Wang[1] and Yu Sun[2]

[1]Arnold O. Beckman High School, 3588 Bryan Ave, Irvine, CA 92602
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

When undergraduate students just got into the physics field, it might be difficult for them to understand, think, and imagine what is happening in certain phenomenons [6]. For example, when two objects have different masses and velocity collide into each other, how are they going to act? Are they going to stop, bounce away from each other, or stick together? This simulation helps the students who do not feel comfortable imagining these scenarios.

Currently we only have the gravitation lab, the trajectory lab, and the collision lab. The gravitation lab is a planet orbiting a sun, where the users can input different masses for the sun and planet, and the radius (in AU), the program will then calculate the gravitational force and orbital period while the planet starts orbiting its sun at a certain speed [7]. The trajectory lab is an object doing projectile motion, where the user input variables like initial velocity, angle, height, and acceleration, the program will present current position and velocity on the screen as the object doing projectile motion [8]. The collision lab is where the user input the masses and velocities for the two objects, and after the user decide the collision is going to elastic or not, set the lab and press start, the program will calculate the total momentum and kinetic energy and have it on the right side of the screen while the objects starts colliding [9].

KEYWORDS

Physics, Simulation, Problem Solving, Animation.

1. INTRODUCTION

Recently I have been studying physics, learning different phenomenons, and solving different kinds of questions. However, I do sometimes find out that some phenomenons, like a planet orbiting its sun or when two different objects colliding into each other either in elastic or inelastic conditions, are very hard to imagine or even think of. Therefore, we developed a program to demonstrate how an object, or objects, looks in different situations.

For example, we first applied our application to a situation where you set up the angle, initial velocity, acceleration, and the height of an object you want to start off. Then, we conducted a qualitative evaluation of the approach; the results came out the same as how we calculated it, and it fully demonstrates what will happen after a period of time.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 47-62, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111705

Some of the techniques and systems like “myPhysicsLab” by Erik Neumann that have been proposed to demonstrate how objects work, which also allows the user to change the conditions of labs, to test out and see how objects act in different conditions [15]. However, the proposal assumes people who intended to use this to know variable names that are not commonly known or used in physics like damping [2]. It is also hard for people who just got into physics to use, because normally we learn kinematics, which is associated with projectile motions, and this proposal is mainly focused on wave topics, which won’t be taught until nearly the end of the course.

Other physics simulations, such as the Cliff Diver in exploration series by the CK-12 Foundation, are using simple terms and have very nice graphics [14]. However, it has barely any variables that can be changed [3]. They are more like a game instead of a lab, which is what I wanted.

I used unity as a tool and C# as a coding language to make this, and it’s mainly because Unity is very effective while rendering 2D and 3D scenes, and the quality offered is also relatively good compared to other apps [10]. The reason I used C# is because it is somewhat like java, but faster. The program I created is more of a tool that helps people to understand physics better, to demonstrate and help people with physics problems.

There are three labs I have created so far, the gravitation lab, the trajectory lab, and the collision lab. The gravitation lab is a planet orbiting its sun, the user can change the masses of the sun and the planet, and the radius between them, and after they press “set” and “play”, the planet starts orbiting. It also shows some outputs like the current gravitational force and the current orbital period. The trajectory lab demonstrates the projectile motions, the user can set the conditions like angle, the height, the initial velocity and acceleration of the object. After clicking “set”, it will predict the path the object will go through, with outputs like current position and current velocity. You can also change the time interval, so that it goes by that time interval every time you click “step”. The collision lab is going to be a lab with two objects inside, you can set the mass and velocity of them, and on the right side it will show the total kinetic energy and momentum. You can also set the condition to elastic or perfectly inelastic. Compared to some other methods, this tool is surely easier to use for those beginners who just got into physics, and also more like a lab instead of a game.

In this paper, we follow the same line of research by solving different physics problems. Our goal is to make sure it’s correct and shows exactly what is going on. Our method is inspired by physicslab, and there are some good features, such as clear and easier to understand. Therefore, we believe that it helps people to get into physics much easier.

To prove that the labs are evaluating the results correctly, I first randomly generated twenty lists of numbers, calculated them by hand, and inputted them into the labs. The outputs were very accurate, so were the demonstrations. However, because I used float instead of double, when it comes to some numbers like the square root of two or ⅓, there might be an almost irrelevant difference between. Second, I asked some random students who are either taking, or took physics to fill up the survey. The result came up that most of them think it would be really helpful, especially the gravitation and trajectory lab, because they think it could have saved so much time if they had it.

In two application scenarios, we demonstrate how the above combination of techniques increases ... First, we found some physics problems that had to do with the labs and made sure they all gave the right output we wanted. Second, we let some people test it out and make sure there are no errors or bugs that are unexpected and need to be fixed.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Self-learning coding

The first challenge I faced was the fact that I am very new to coding, and have only used python and a little java. Therefore I asked my teacher to see if he had any recommendations. He told me that he doesn't recommend me to use python, and provided me with a few other programming languages. After looking over them, I decided to use c#, on unity. Then I started learning how to use c# by watching YouTube videos and googling. At first, it was super hard because it was my first time learning a programming language on my own, but then I got better and better. I also realized that c# is very similar to python and java, especially java: for example, the semicolon you have to put at the end of a line, use parentheses and functions etcetera.

2.2. Equations to methods

The second challenge was the fact I also had to self teach some physics. The reason behind it was because some of the things (like the period equation in the gravitation lab) we didn’t really learn in physics A, and some equations needed to be changed so I can put them into code, and provide a correct output. I solved this problem by countless tries and by reviewing and understanding more about physics. However, there are still some issues like the fact that the collision lab only gives the total momentum and kinetic energy, that was mostly because I didn’t have enough time to re- code for the momentum and kinetic energy for each one, I will probably fix it in the future [11].

2.3. Modeling

The third challenge is when I try to combine the code with unity. Unity is harder to use than it looks, it has a lot of features, which is difficult for new users like me to use smoothly. I first tried to explore myself, but then after I turned my files into a disaster and had to delete and re-download unity, I decided to watch the tutorial that teaches you how to use unity “properly”. After a list of trial and error, for example forgetting where I put the object and ended up lots of planets crashing into each other, or a block that just won’t stop going down, also there’s that time I accidentally put in too much stuff so the unity keep on crushing on me, I finally combined the script and models together.

3. Solution

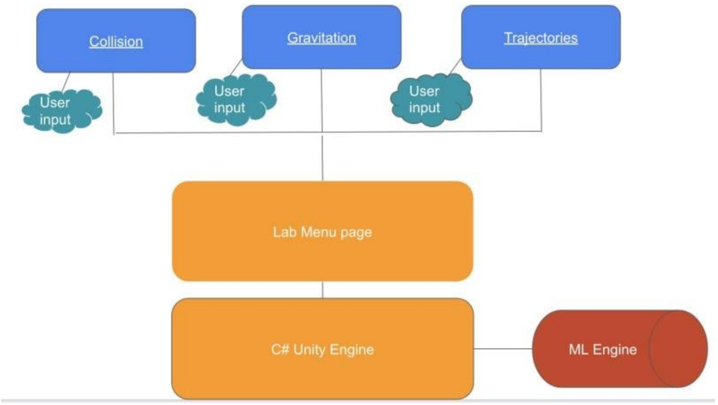
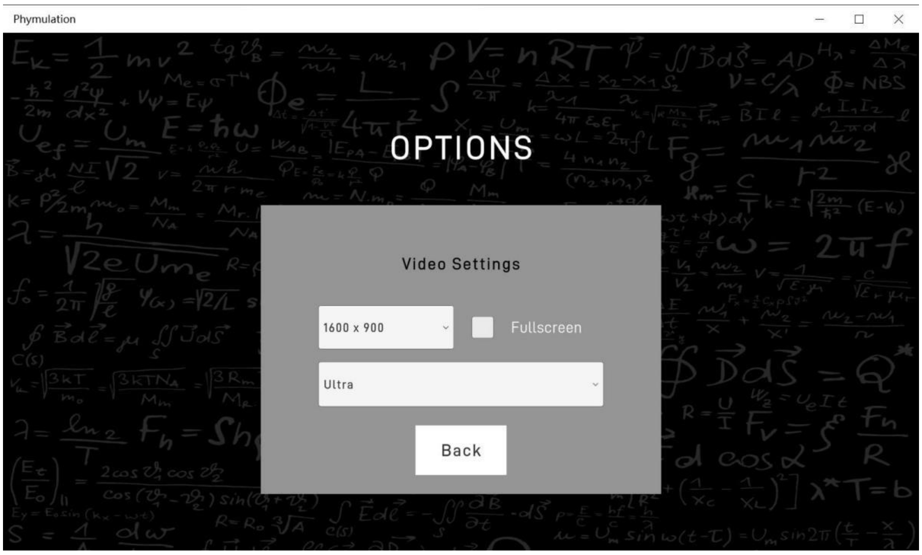


Figure 1. The overview of the project

Phymulation is a program that helps students who find it hard to imagine certain situations in physics to get a better understanding. It's written in C# and run by ML Engine in Unity. When the user first loads in, they will see the menu page, and then they can choose which lab they want to use. For the collision lab, you input the masses and velocities of the two objects, and set it. It will provide the total momentum and total kinetic energy on the right while the objects collide with each other. Then, for the gravitation lab, the user input the masses of the sun and the planet, and then input the radius(in AU), the program will calculate the gravitational force and orbital period, and show it on the screen. The trajectory lab is when the user inputs the initial velocity, angle, height, and acceleration, the program will output the current position and velocity on the screen as the object doing projectile motion.



1600 x 900

□ Fullscreen

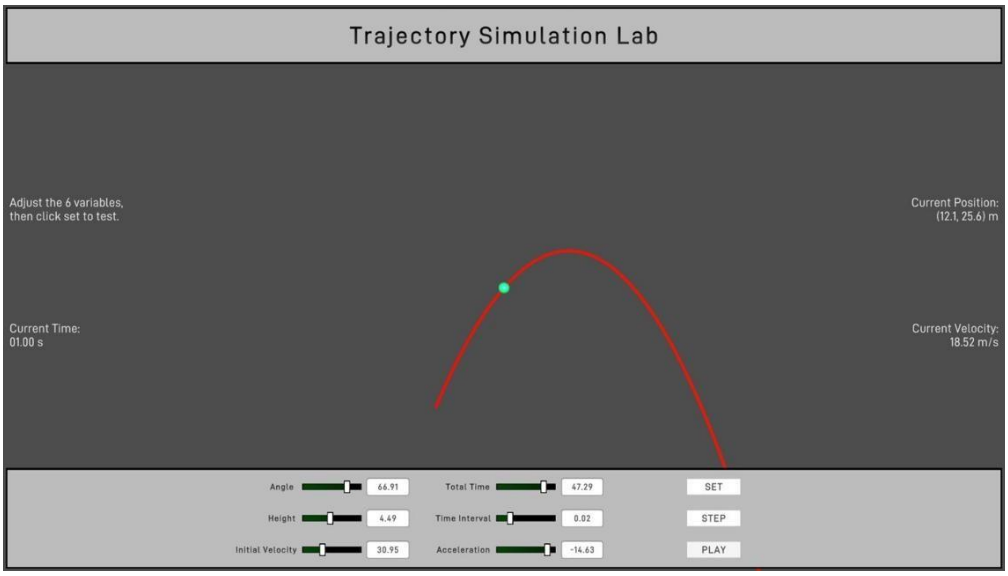
Ultra

Figure 2. The screenshot of settings.

```
  public void Start()
  {
    resolutions                    =
    Screen.resolutions;
    resolutionDropdown.ClearOptio
    ns();
    List<string>   options   =   new
    List<string>(); for(int i = 0; i <
    resolutions.Length; i++)
    {
      options.Add(resolutions[i].width + " x " + resolutions[i].height);
      if(resolutions[i].width          ==          Screen.currentResolution.width          &&
resolutions[i].height ==Screen.currentResolution.height)
      {
        resIndex = i;
      }
    }

    resolutionDropdown.AddOptions(options);

    if (PlayerPrefs.HasKey("resValue"))
    {
      resolutionDropdown.value                         =
      PlayerPrefs.GetInt("resValue");
      resolutionDropdown.RefreshShownValue();
      resValue                      =
      PlayerPrefs.GetInt("resValue");
      resWidth                      =
      PlayerPrefs.GetInt("resWidth");
      resHeight                     =
      PlayerPrefs.GetInt("resHeight");
    }
    else
    {
      resolutionDropdown.value = resIndex;
      resolutionDropdown.RefreshShownVal
      ue();
    }


    if (PlayerPrefs.HasKey("isFullscreen"))
    {
      isFullscreen                       =
      PlayerPrefs.GetInt("isFullscreen");    if
      (PlayerPrefs.GetInt("isFullscreen")  ==
      1)
        fullscreenToggle.i
      sOn = true;else
```

```
            fullscreenToggle.isOn = false;
        }

        if (PlayerPrefs.HasKey("qualityLevel"))
        {
          qualityDropdown.value = PlayerPrefs.GetInt("qualityLevel");
          qualityDropdown.RefreshShownValue(
          );                qualityLevel                =
          PlayerPrefs.GetInt("qualityLevel");
        }


        resolutionDropdown.onValueChanged.AddListener((value) =>
        {
          resWidth                       =
          resolutions[value].width;
          resHeight                      =
          resolutions[value].height;
          Screen.SetResolution(resWidth,               resHeight,
          Screen.fullScreen);resValue = value;
        });

        qualityDropdown.onValueChanged.AddListener((value) =>
        {
          QualitySettings.SetQualityLevel(value)
          ;qualityLevel = value;
        });

        fullscreenToggle.onValueChanged.AddListener((value) =>
        {
          Screen.fullScre
          en  =  value;  if
          (value)
          {
            isFullscreen = 1;
          }
          else
          {
            isFullscreen = 0;
         }
         }
)
;

}
```

This part of the code is to set the quality, resolution, and full-screen if the user wished to. I created a new list “options” to save the preset length and width resolutions, and added them into “drop- down”, so the user can choose from different resolutions like 1920x1440. Then I created a full- screen toggle so the user can set the simulation into full-screen. I also made a quality drop-down to let the user choose the quality level they would prefer.



Adjust the 6 variables, then click set to test.

Current Position: (12.1, 25.6) m

Current Time: 01.00 s

Current Velocity: 18.52 m/s

Angle 66.91

Total Time 47.29

Height 4.49

Time Interval 0.02

Initial Velocity 30.95

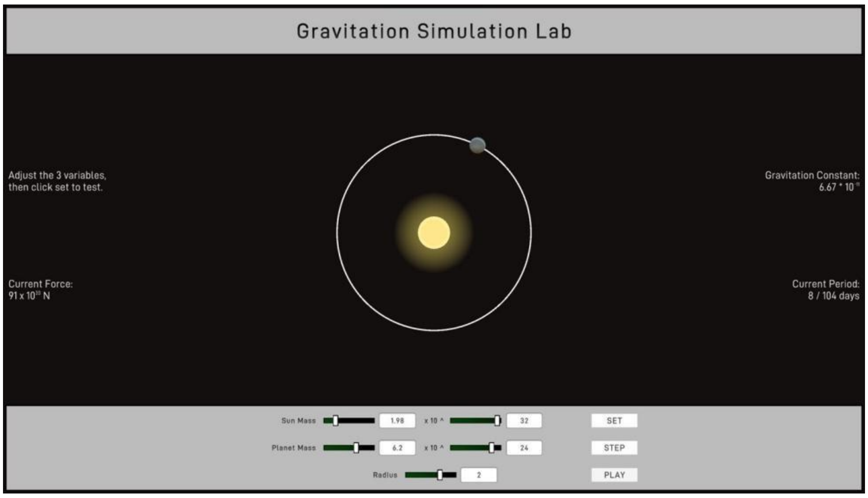
Acceleration -14.63

Figure 3. Screenshot of trajectory Simulation Lab

```
   public void SetVariables()
   {
     if(!float.TryParse(angleField.text, out float result) || !float.TryParse(heightField.text, out float
result2) || !float.TryParse(velInitField.text, out float result3)
|| !float.TryParse(accelerationField.text, out float result4) || !float.TryParse(timeIntField.text, out
float result5) || !float.TryParse(totalTimeField.text, out float result6))
     {
       return;
     }

     angle = float.Parse(angleField.text);
     height = float.Parse(heightField.text);
     velInit = float.Parse(velInitField.text);
     acceleration = float.Parse(accelerationField.text);
     timeInterval = float.Parse(timeIntField.text);
     totalTime = float.Parse(totalTimeField.text);

     PlayerPrefs.SetFloat("angle", angle);
     PlayerPrefs.SetFloat("height", height);
     PlayerPrefs.SetFloat("velInit", velInit);
     PlayerPrefs.SetFloat("acceleration", acceleration);
     PlayerPrefs.SetFloat("timeInterval", timeInterval);
     PlayerPrefs.SetFloat("totalTime", totalTime);

     segments = (int)(totalTime / timeInterval);
     points = new Vector3[segments];
     arcPoints = new Vector3[segments];
```

```
        float x = 0;
        float y = 0;

        velXInit = velInit * Mathf.Cos(Mathf.Deg2Rad * angle);
        velYInit = velInit * Mathf.Sin(Mathf.Deg2Rad * angle);

        for (int i = 0; i < segments; i++)
        {
            x = velXInit * timeInterval * i;
            y = (velYInit * timeInterval * i) + height + (0.5f * acceleration * Mathf.Pow(timeInterval
* i, 2));
            points[i] = new Vector3(x, y, 0);
            arcPoints[i] = new Vector3(x*0.1f, y*0.1f, 0);
        }

        currentPoint = 0;
        sphere.transform.position = arcPoints[currentPoint];

        lr.positionCount = segments;
        lr.SetPositions(arcPoints);
        SetText();
    }
```

In “setVariables”, I first check if the user input can be turned into a float, if it can, then return it. Then I set the angle, height, initial velocity, acceleration, total time, and time interval from a string (where the user imputed) into floats, and also set up the lab. Next step was to calculate the initial velocity in x and y directions to calculate the current position. The last thing (in shown codes, not the actual code) I did was to count and set the positions, and also printing it out.



Gravitation Simulation Lab

Adjust the 3 variables,
then click set to test.

Gravitation Constant:
6.67 * 10 [illegible]

Current Force:
91 x 10 [illegible] N

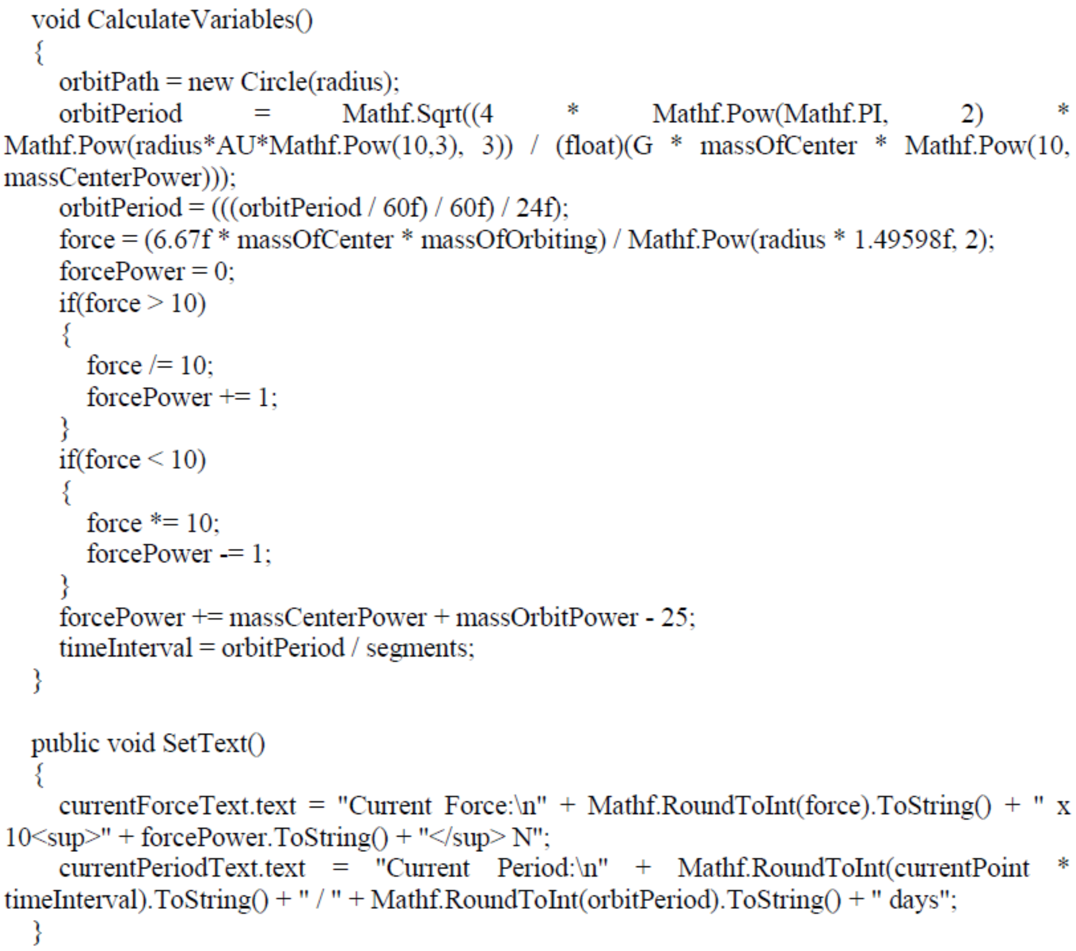
Current Period:
8 / 104 days

Sun Mass 1.98 x 10 ^ 32

Planet Mass 6.2 x 10 ^ 24

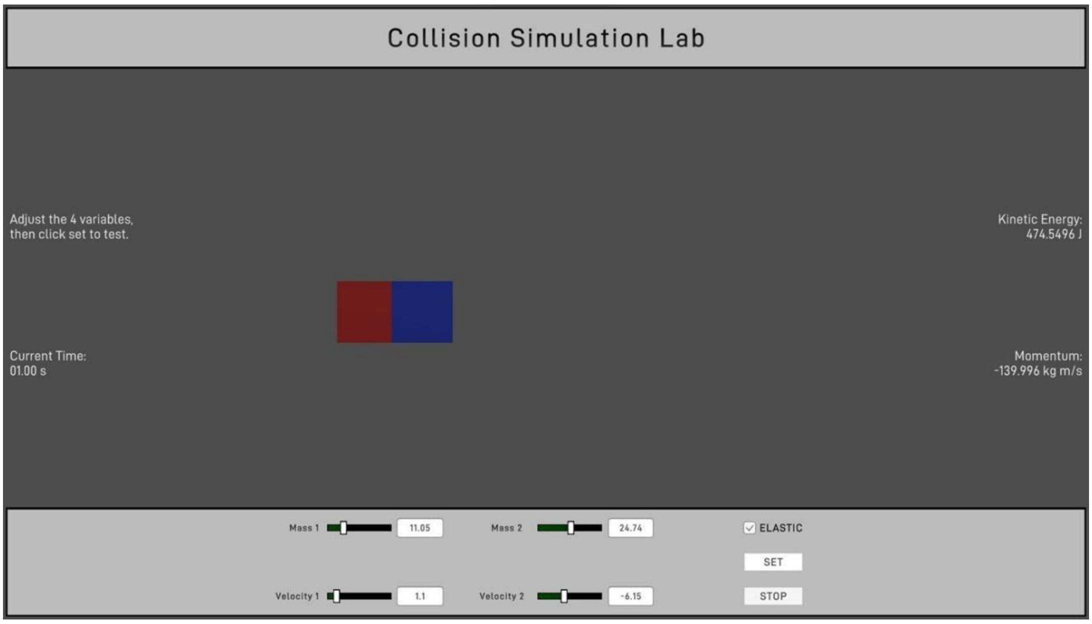
Radius 2

Figure 4. Screenshot of gravitation Simulation Lab



```
    void CalculateVariables()
    {
        orbitPath = new Circle(radius);
        orbitPeriod = Mathf.Sqrt((4 * Mathf.Pow(Mathf.PI, 2) *
Mathf.Pow(radius*AU*Mathf.Pow(10,3), 3)) / (float)(G * massOfCenter * Mathf.Pow(10,
massCenterPower)));
        orbitPeriod = (((orbitPeriod / 60f) / 60f) / 24f);
        force = (6.67f * massOfCenter * massOfOrbiting) / Mathf.Pow(radius * 1.49598f, 2);
        forcePower = 0;
        if(force > 10)
        {
            force /= 10;
            forcePower += 1;
        }
        if(force < 10)
        {
            force *= 10;
            forcePower -= 1;
        }
        forcePower += massCenterPower + massOrbitPower - 25;
        timeInterval = orbitPeriod / segments;
    }

    public void SetText()
    {
        currentForceText.text = "Current Force:\n" + Mathf.RoundToInt(force).ToString() + " x
10<sup>" + forcePower.ToString() + "</sup> N";
        currentPeriodText.text = "Current Period:\n" + Mathf.RoundToInt(currentPoint *
timeInterval).ToString() + " / " + Mathf.RoundToInt(orbitPeriod).ToString() + " days";
    }
```

The orbitPath is to create an orbit based on the radius(in AU) provided. Using that we can calculate the orbit period by using the equation(T=42r3GM), and then convert it into days. Then I used another equation (Fg=GGm1m2r2) to calculate the gravitational force, and then I used a list of concepts, like keep dividing it by 10 until it’s less than 10 and add a power to 10 every time, to convert it into scientific notation form. At last I used “Set Text” to set the text that would be printed out.


Collision Simulation Lab

Adjust the 4 variables,
then click set to test.

Kinetic Energy:
474.5496 J

Current Time:
01.00 s

Momentum:
-139.996 kg m/s

Mass 1 11.05

Mass 2 24.74

☑ ELASTIC

Velocity 1 1.1

Velocity 2 -6.15

Figure 5. Screenshot of collision Simulation Lab.

```
 public void InitSetVariables()
{
    If(!PlayerPrefs.HasKey("mass1")||!PlayerPrefs.HasKey("mass2")
|| !PlayerPrefs.HasKey("vel1") || !PlayerPrefs.HasKey("vel2") || !PlayerPrefs.HasKey("elastic"))
    {
        return;
    }

    cube1.transform.position      =      new
    Vector2(-4, 0); cube2.transform.position
    = new Vector2(4, 0); rb1.mass = mass1;
    rb2.mass = mass2;

    if (isElastic)
    {
        material.bounciness = 1f;
    }
    else
    {
        material.bounciness = 0f;
    }

    CalculateVariables(
    );SetText();
}
void CalculateVariables()
{
    kineticEnergy = (0.5f * mass1 * Mathf.Pow(vel1,2)) + (0.5f * mass2 * Mathf.Pow(vel2,
    2));momentum = (mass1*vel1) + (mass2 * vel2);
}
```

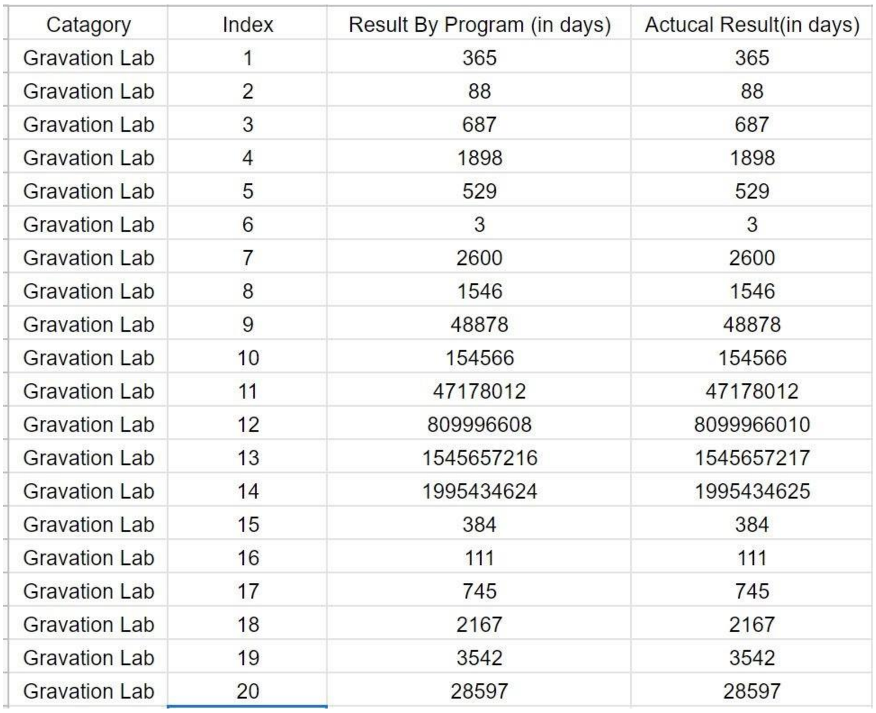
```
  public void SetVariables()
  {
    if (!float.TryParse(mass1Field.text, out float result) || !float.TryParse(mass2Field.text,
out float result2) || !float.TryParse(vel1Field.text, out float result3) ||
!float.TryParse(vel2Field.text, out float result4))
    {
      return;
    }

    mass1                         =
    float.Parse(mass1Field.text);
    mass2                         =
    float.Parse(mass2Field.text);
    vel1                          =
    float.Parse(vel1Field.text);
    vel2                          =
    float.Parse(vel2Field.text);
    isElastic                     =
    elasticToggle.isOn;

    PlayerPrefs.SetFloat("mass1",          mass1);
    PlayerPrefs.SetFloat("mass2",          mass2);
    PlayerPrefs.SetFloat("vel1",            vel1);
    PlayerPrefs.SetFloat("vel2",            vel2);
    PlayerPrefs.SetString("elastic",
    isElastic.ToString());

    if (isElastic)
    {
      material.bounciness = 1f;
    }
    else
    {
      material.bounciness = 0f;
    }


    cube1.transform.position     =     new
    Vector2(-4, 0); cube2.transform.position
    = new Vector2(4, 0); rb1.mass = mass1;
    rb2.mass = mass2;

                        CalculateVariables();SetText();
```

First of all, I used initSetVariables to check if the user set any variables like mass and velocity, and check if the user set the system into elastic. Secondly, I made the objects(cubes) transform into positions (4,0) and (-4,0). If the user clicks the set button, the simulation will set the objects' masses exactly like how the user inputted it, and set the system into elastic if the user set it to elastic. The third thing was to calculate the kinetic energy and momentum, then display them. After all that I let the variables, which were set by the user, to show on the screen, and wait for the player to either change them, or start the lab.

4. EXPERIMENT

4.1. Experiment 1

To test if each lab is accurate, I found 20 questions for each lab to test their accuracy. For example, I used the actual value from some solar systems to test out the gravitation lab, they worked out perfectly. However, sometimes(especially when including numbers like square root of 2 etc) the result comes out to be approximate. I think it is because the float only saves values up to 6 to 7 desmos, therefore when facing some numbers that have more than 7 desmos or continues going, it won't be as accurate. The same problem exists in the Trajectory lab. However, unlike the other two labs, the collision lab actually worked more accurately. I think the reason might be that there are not many divisions, especially during the calculation of total momentum(the equation was mass times velocity for individual one, and the total momentum was the sum of those), which only includes mutualisation.



Catagory	Index	Result By Program (in days)	Actucal Result(in days)
Gravation Lab	1	365	365
Gravation Lab	2	88	88
Gravation Lab	3	687	687
Gravation Lab	4	1898	1898
Gravation Lab	5	529	529
Gravation Lab	6	3	3
Gravation Lab	7	2600	2600
Gravation Lab	8	1546	1546
Gravation Lab	9	48878	48878
Gravation Lab	10	154566	154566
Gravation Lab	11	47178012	47178012
Gravation Lab	12	809996608	8099966010
Gravation Lab	13	1545657216	1545657217
Gravation Lab	14	1995434624	1995434625
Gravation Lab	15	384	384
Gravation Lab	16	111	111
Gravation Lab	17	745	745
Gravation Lab	18	2167	2167
Gravation Lab	19	3542	3542
Gravation Lab	20	28597	28597

Table 1. The output, total period, of the gravitation lab

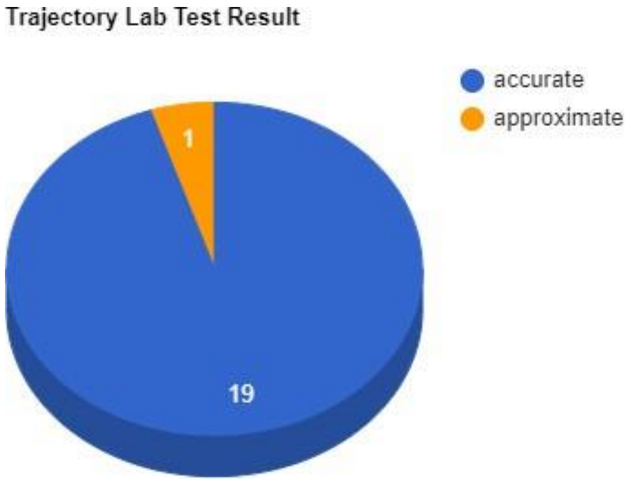


Figure 6. The test result of Trajectory Lab

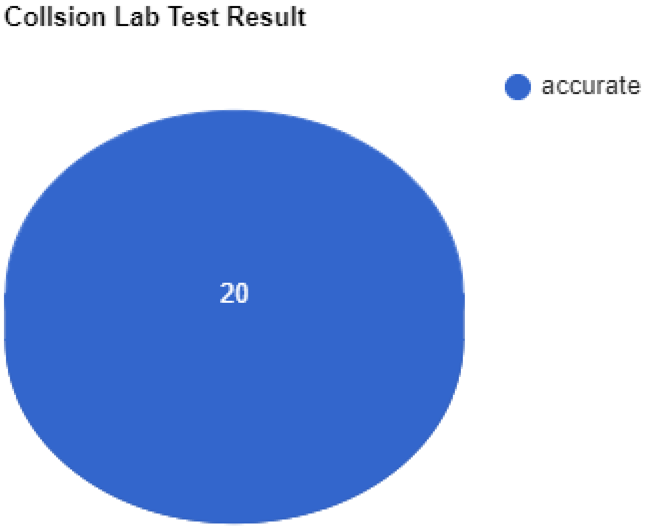


Figure 7. The test result of Collision Lab

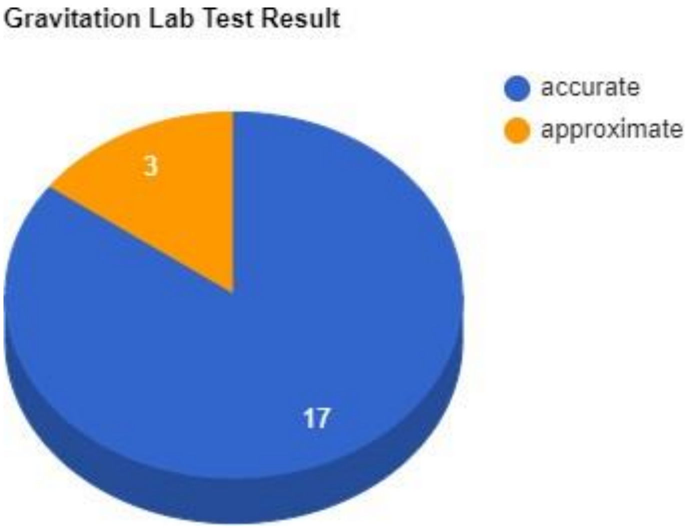


Figure 8. The test result of Gravitation Lab

4.2. Experiment 2

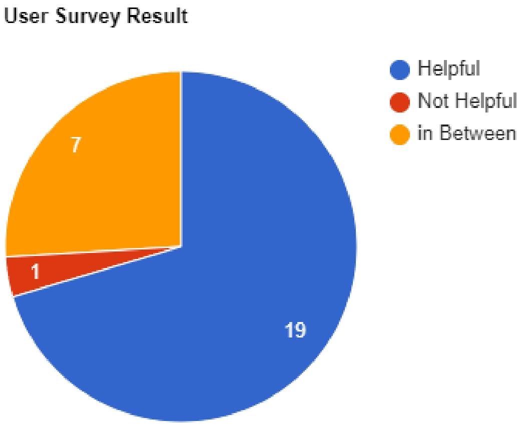


Figure 9. The result of user survey

I presented the program to twenty-seven people that had physics last year, and the result came up above as a pie chart. Most people are saying that it's helpful, especially the gravitation and trajectory lab because they think it could have saved so much time if they had it. The most common reason for those who were saying "in-between" was because it can only calculate certain values in certain conditions. For example, if you had the gravitational force and the radius and wanted to find the mass of the sun, it doesn't really help. The only one that said "not helpful" gave out three reasons: first, the radius of the gravitation lab was in au, and we didn't learn it during school year; secondly, the collision lab on gives the total kinetic energy and total momentum, which is actually very easy to calculate; third reason was that the trajectory lab, you can only do the problems that provide the angle, height, initial velocity and acceleration.

For the first experiment, I calculated and compared the answer between the output where the program gave me and the actual values. Although there were differences between the answers, it only exists if the value is too large, and the differences were not very significant. The most common reason why people think that the program is helpful is because it can only calculate certain outputs with certain inputs. However, the main goal of this program is to let the people understand more about physics, and to help students to imagine certain situations better instead of a calculator that helps you to solve problems. As long as it works similar to what actually happens, it will reach my expectations.

5. Related Work

Chytracek et al. presented a special language to describe the geometries of detectors related with the physics measurement, in order to analyze the solids and materials [1]. Our work focuses on creating a lab and demonstrating how objects act in different conditions using a game engine. Our work does not rely on a modeling language, but uses a 3D game engine for visual simulation.

K. Binder et al. created a simulator for those already familiar with the theory of critical phenomena, but it would be difficult, and mostly impossible for the undergraduate physics students to use [4]. Our work is much simpler, and is good for students who just got into physics to use, to help them get more into physics.

José M. Carcione et al. develop a numerical algorithm for a wave propagation in linear nonisothermal poroelastic media simulation [5]. Our work demonstrates projectile motion, how planets orbits, and what would happen when two objects with different mass and velocity collide into each other in three different labs.

6. CONCLUSIONS

This is a program for better understanding of physics for students who are undergraduates and are new to the physics field. We used three different labs: Gravitation Lab, Trajectory Lab, and Collision Lab to demonstrate how planets orbit, projectile motions, and how objects collide. To test if the program is efficient, I randomly generalized 20 different lists of inputs, and made sure the output is accurate, and the objects are behaving how they should behave [12]. Then I asked twenty-seven students that had physics last year to finish a survey on whether they think the app is helpful, most of them think that it is helpful, some others said it needs improvement. However, they all admit I have reached my goal, which is to demonstrate how objects behave under certain conditions.

One of the limitations was the accuracy, even though it doesn’t make a big difference in between [13]. To solve this, I think I can change the float into double, because the precision of float is only six or seven decimal digits, while double variables have a precision of about 15 digits. The other one was that in the gravitation lab, the planet rotates too slow. I am thinking of adding a “speed up” button so the users can speed it up. For the collision lab, I will add in the kinetic energy and momentum of each object so it could be more helpful. I can also add in other labs like “rotation lab” and “oscillation lab” for people that find it hard to imagine these things. The fact that it only can calculate certain outputs by putting in certain inputs might also be solved, although this wasn’t really a part of my goal, but since people wanted this feature, I can probably make it real.

REFERENCES

[1] Chytracek, Radovan, Jeremy McCormick, Witold Pokorski, and Giovanni Santin. "Geometry description markup language for physics simulation and analysis applications." IEEE Transactions on Nuclear Science 53, no. 5 (2006): 2892-2896.

[2] myPhysicsLab https://www.myphysicslab.com/

[3] Cliff. Diver, https://interactives.ck12.org/simulations/physics/cliff-diver/app/index.html?screen=sandbox&lang=en&referrer=ck12Launcher&backUrl=https://interactives.ck12.org/ simulations/physics.html

[4] Binder, Kurt, et al. "Monte Carlo simulation in statistical physics." Computers in Physics 7.2 (1993): 156-157.

[5] Selvadurai, Antony PS, ed. Mechanics of poroelastic media. Vol. 35. Springer Science & Business Media, 1996.

[6] Heisenberg, Werner. Physics and beyond. London: Allen & Unwin, 1971.

[7] Boynton, Paul E., et al. "Gravitation physics at BGPL." New Astronomy Reviews 51.3-4 (2007): 334-340.

[8] Zheng, Yu. "Trajectory data mining: an overview." ACM Transactions on Intelligent Systems and Technology (TIST) 6.3 (2015): 1-41.

[9] Goldberger, Marvin L., and Kenneth M. Watson. Collision theory. Courier Corporation, 2004.

[10] Harman, Philip V., et al. "Rapid 2D-to-3D conversion." Stereoscopic displays and virtual reality systems IX. Vol. 4660. International Society for Optics and Photonics, 2002.

[11] Jegadeesh, Narasimhan, and Sheridan Titman. "Momentum." Annu. Rev. Financ. Econ. 3.1 (2011): 493- 509.

[12] Bushnell, David S. "Input, process, output: A model for evaluating training." Training & Development Journal 44.3 (1990): 41-44.

[13] Diebold, Francis X., and Robert S. Mariano. "Comparing predictive accuracy." Journal of Business & economic statistics 20.1 (2002): 134-144.

[14] Christensen, Neil, et al. "A comprehensive approach to new physics simulations." The European Physical Journal C 71.2 (2011): 1-57.

[15] Jeary, A. P. "Damping in structures." Journal of wind engineering and industrial aerodynamics 72 (1997): 345- 355.

POLITICAL CORRECTNESS: THE EFFECTS OF GAMING IN THE SOCIETY AND THE SOCIAL DIMENSION

Zhengye Shi

Obridge Academy, NY 11801, USA

ABSTRACT

In this study my approach is the dispute and the structure of political correctness in terms of sociological questions, as follows. 1. Why this apparently centered in creative output on achieving social change through the gaming industry? 2. How are we to comprehend the association among the chaos of inequality in the gaming industry and putting character disfigurement (gender, race, ethnicity, sexual orientation)? 3. How do we connect globalization - political correctness to video games? The study conclude with a discussion and tactics for contesting critiques.

KEYWORDS

Culture, Discourse, Political Correctness, Video Games

1. INTRODUCTION

Politically correctness has been the focal point of controversy among international gamers and game enthusiasts. Some people think a healthy piece of political correctness is expected in a topography that has a long, ugly history of inequality. Some are more cautionary, knowing that worrying about being politically correct is ruining the artisticliberation of developers and can be acognizant as the turn-away of forms of expression or action that debar, marginalize, or insult certain racial, cultural, or other groups as defined by Oxford, and they are a big part of our multicultural community and therefore their encompassing in games can be preceded as an act of political correctness. However, in our numérique society, this term is often disdained. Nonetheless, in this study I will use it as a neutral term because I believe that there is indeed asurge of these groups, and it is often used in discourse about this topic. Ironically, the term was brought into the modern day by Toni Bambara to call out those who hid behind the wall of being publicly "courteous or polite" to continue holding their bigoted comportment and avoid the conference of social transformation. As we know it today, it has been contorted to simply mean "coherent to public customary and mind what you say," which is exactly the adverse of what Bambara ought. In the cyber or gaming community, political correctness predominantly has two sides: you're for or against it. It's easy to think of the anti-politically correct crowd as one that is anti-progress, stuck in the 1800s with how they think people should or should not behave. For the liberal, they are the Conservatives "ruining" the country's potential and stuck in a past where it was OK to be discriminatory. In the gaming world, the anti-politically correct crowd is half of the problem.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 63-70, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111706

2. THEORY

We might see the altercation around ‘political correctness’ as a political and anthropological contentious in which both labelled ‘political correctness’ and those who labeled them ‘politically correct’ are engaged in a centered representation, values, and identities - in short, ‘cultural chaos’. An instantaneous caveat is that the homogeneity of ‘’politically correct people’ is no more than a constructed homogeneity through the category but shall leave until later. The objective on both side is cultural diversity. As trigger for boarder social change.

2.1 Games are Fantasy so being Realistically Vivid is Unnecessary.

When it comes to games set in true and real-life settings, this debate pops up that the video game world is fantasy. If one complains about how someone is represented, whether that be body image or stereotypes, people are quick to say that it's "just a game" and no one should take them seriously. Developers have a constructive take on the antiquity and can swivel as they please to make a remarkable game. However, the problem with this steam of consciousness is that it throws out any being responsible for actualy issues that phlebotomize into the gaming world and fortifyprehistorc and antiquiated ideas.

It's a psychological truth that perceived the same ideas in our amusement without condemning it does, in fact, affect us. This study is a serene model of how race is illustrated and or portrayed in games and is only one of the plentitudes of studies that come out yearly on the subject. All games don't need to be pragmatic, but they don't need the same fantasy either.

2.2. Some Games are Real, and that Incorporate Politically Incorrect Activity.

If someone grumble about the insufficiency of women military soldiers or ethnic minorities in the game called Call of Duty, netizens are quick to say that it's preordained to be realistic, as there are far less women than men in real combat. That's true. But it's also a fact that the game Call of Duty is a hyper-frenzied game. If it were meant to be representational in any way, there would be a lot less shooting, a lot more sitting around standby for commands, and no way to heal your shot wounds by just solely ducking behind a counter. This argument falls on the perception that your quintessential video game is pragmatic in any way. People can't pick and choose what they deem to be a necessary "pragmatism" factor without appearing hypocritical.

2.3. If You Don't Like a Game, Don't Buy It.

This argument is the most understandable. If you don't support an organization's ideals for any reason, you're in your rights to boycott said organization. Nonetheless, a thing about entertainment and amusement is that it is constructed and built by people who are flawed and have opinions. If someone were to only buy, watch, or playthings that perfectly lined up with their worldview, the entertainment industry would have shriveled up decades ago and we'd all be outside.

The world is not a perfect place, but it is one full of the means to better it. With unbolted communication and the web, it is easy to vociferate your concerns and find others to sustain you. In the end, this argument falls flat because it's one against constructive criticism. Without constructive criticism, games can't get better. People will buy what they enjoy, but they can also be aware of its shortcomings.

2.4. Make Your Own Personal Computer Games.

Possibly the most political answer of them all is this one. People are already making their own diverse games that they feel represent them better. The problem is that the gaming industry is like any other American industry—discriminatory, biased, and under the thumb of mostly white men. It's difficult for an indie company to break into the "AAA" side, but statistically, it's even harder for those who are in the minority.

A more appropriate response would be to tell people to support the already existing alternatives or tell companies to take diverse initiatives if they haven't already. Ultimately, this response is one of ignorance. There are already personal computer games, but it is up to the gamer, whether pro- or anti-politically correct, to seek them out in a world that typically does not support them and relies on the consumer to vote with dollars.

2.5. Political Correctness Ruins Creative Freedom

This is the most ironic response of them all, considering creative freedom is already ruined. The creative process is stifled by publishers forcing developers to appeal to their wanted demographic. Typically, this is a white man. Being politically correct in this case would encourage creative freedom. Having to create the same, heroic, white, and muscular man with a scruffy beard, "hard" personality, and who probably lost a loved one is an old trope that is mostly prevalent not because developers like creating the same person over and over, but for safe marketing purposes. There's nothing wrong with creating a character you identify with, but when that character becomes indistinguishable from the rest, it's hard to believe that it's anything but intentional. The white, scruffy, heroic man can work - but not every game has a Joel from The Last of Us. In conclusion, this response is an insult not just to the creatives, but to all white men who don't think of themselves as a singular audience who need pandering to.

On the flip side, some can find the pro-politically correct crowd as one that is ridiculous, touting freedom of expression while also demonizing those who don't agree with them. For the conservative, they are the Liberals "ruining" the country with their need to feel fawned over and don't care about the realities of things like money or politics. In the gaming world, the pro-politically correct crowd are the other half of the problem. Arguments include games can have social consequences, games are fantasy and should not be subject to real-life statistics, if the game industry wants to be taken seriously it should conduct itself in a more responsible manner and appealing to more demographics increases profitability.

3. DISCUSSION

PART 1

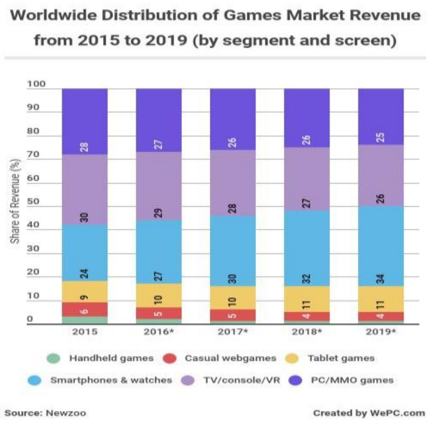
3.1. Personal Computer Games are in Decline

Arguments clouded those games are fantasy so being realistic is unnecessary, some games are realistically vivid and that includes politically incorrect things and political correctness ruins creative liberation. The modern gaming industry is a behemoth that generated an estimated $135 billion in revenue last year (according to TechJury), with a gaming community of more than 2.5 billion people from all over the world. It wasn’t always mainstream, though — the industry started as a niche market that targeted students, nerds, and individuals willing to put up with atrocious load times and horrendous graphics because they were enchanted by the medium.

As computers are concerned a Commodore 64 that featured games with 16-color graphics, 320 x 200 resolution, and chiptune sound effects. Followed by the Amiga 2000, upgraded this experience many times over, thanks to its 4,096 colors, 640 x 256 resolution, and far better audio quality. Then, in the 1990s, Intel INTC, -1.26% Pentium processor, and the rest is, as they say, history. Those older computer games had only one goal: to entertain and excite.

Early computer games were a special art form that could awe you with presentation, jog your brain with puzzles, immerse you in their world with engaging stories or simply help you hone your reflexes as you blasted dozens of baddies off the screen. Fast-forward to the present, and this concept of personal compiuter gaming is crumbling. Many avid Steam user with over 400 games under its belt, find it increasingly hard to find new titles that could provide the same level of fun as the older ones. And that's not because of a lack of visual or audio fidelity — modern games look and sound realistic and breathtaking. Nor is it because of the amount of content — unlike older games, modern titles provide hundreds of hours of in-game activities to keep you invested. One reason is the gaming industry has become far more focused on generating large profits than creating good products.

When you stop caring about quality and focus on money-making mechanics, games lose their most important purpose and become shiny, but ultimately shallow and often half-baked, gambling simulators. If you look at the chart below, you can see that personal computer gaming isn't even the biggest revenue maker anymore — in 2019, it constitutes a projected 25% of the industry's total and is constantly overshadowed by mobile games and continuously moving forward – present. Console games are also going through a market-share decline. I don't see this declining trend for personal computer gaming slowing any time soon unless something changes drastically.



Source: Newzoo

Created by WePC.com

Figure 1.0

PART 2

3.2. Gender, Ethnicity, Sexual Orientation,

The diversification of the world has not always been shown in games. It took a little bit longer for this medium to have characters that mirror our globalized society. But why is the globalization and the diversification of the world only now showing in video games?

3.2.1. Gender

You might wonder why diversity issues are controversial in gaming since diversity has already been somewhat apparent in other means of entertainment, like television and film. One reason for the issue of female representation in games could be that gaming is still mostly seen as a masculine hobby, even though about 40% of gamers are female (Entertainment Software Association, 2015).

A common justification for this stereotype is that, although women might play games, they should not be considered "true" or "hard-core" gamers because they play more casually and less skilfully compared to their male counterparts. (Paaßen, Morgenroth & Stratemeyer, 2016) Games would therefore be more directed to 'typical' male players, presumably a young, heterosexual audience. This is shown in the small number of female video game characters and the hyper sexualization of those characters that do exist. It is because of ideas like these a lot of male main characters in games are presented as strong, and females as sexy (and straight), since that would appeal to the audience (Paaßen et al., 2016).

According to several older studies, stereotypical white male characters reigned in mainstream gaming. However, nowadays, "results show that female characters appeared as often in leading parts as male characters did. They were portrayed with a sexualized emphasis on female features" (Jansz & Martis, 2007). This study seems to indicate that female characters' inclusion in recent games is far larger than it was in earlier games. Jansz and Martis (2007) call the development of strong female characters the 'Lara Phenomenon'. Lara Croft (figure 6) is a name that you might recognize from the movie Tomb Raider, which is based on a video game (first realeased in 1996) that has become a big nostalgic name in the community. Although she was a strong and independent character, she nonetheless had a sexualized appearance that has become a brand for the game. Initially, it seems that this was done to boost sales, since most females were depicted heavily objectified like that in games. However, when we compare depictions like these to the new and diverse characters that populate video games nowadays, we see how stereotypical these depictions truly are. I will elaborate on that later.

3.2.2. Ethnicity

"Research have shown that most leading roles in games are of white race, the heroes exclusively so.'' (Jansz & Martis, 2007). Jansz and Martis (2007) partially back this up by arguing that the medium has a different way of how players identify with characters. In games you can change and influence the world you are playing in and therefore the character creation in games is a lot different than with other media. A study on the representations of gender, age and ethnicity in 150 games of the same year showed that other races than white were largely underrepresented. Their results showed that 80% of the leading characters in games were white. Other ethnicities filled the remaining 20% of the games (Williams, Martins, Consalvo & Ivory, 2009). A very low number indeed.

Non-white characters are also often stereotypical and for example portrayed as aggressive, as thugs or as athletes. These depictions are dangerous according to some researches, because they normalize these stereotypes. This can provoke negative social judgements that are not necessarily true (Burgess, Dill, Stermer, Burgess & Brown, 2011). The stories told about minorities in games come "(…) with underrepresentation and overreliance on stereotypes'' (Burges et al, 2011). It is primarily a white male world and that does not seem right since ''we are moving toward a more global and therefore more racially diverse society'' (Burges et al, 2011). Even though many of these games were brilliant pieces, like Grand Theft Auto: San Andreas.

3.2.3. Sexual Orientation

Unfortunately, there is not a lot of research on sexual orientation in video games. However, successful games like Gone Home (2013), Life is Strange (2015), The Last Of Us (2013) and Dragon Age: Inquisition (2014) show that having a leading character who isn't heterosexual is a rising phenomenon. If we take a look at a list containing video games that feature LGBT characters, we see that those characters were also present in games in the past. However, when I looked more closely, I found that these instances almost always were comprised of subtle references or of characters being made fun of because of their sexual orientation.

In the past, this underrepresentation was because of censoring laws that were present for several mediums (Hays code). Later, homosexuality was carefully introduced in the media in ways that were not out there. While in other media this has already changed, video game writers for a long time still allowed ‘’space for an audience member to overlook or deny the homosexuality of a particular character if that’s the way they would prefer to see things’’ (Gravning, 2014). In Gravning’s article we see examples of these careful references in games, one of them being Lara Croft. The writer of the game states how she would love for Croft to be (openly) gay, but played in to expectations of gamers (and their parents) instead. This subtleness might be connected to the masculine culture I already spoke of. LGBT rights are still a sensitive topic and although this has become more widely accepted in Western countries, it is still a taboo for a lot of gamers, like we saw in the controversy surrounding The Last Of Us. But still, slowly but certainly we are seeing an increase of LGBT representation in games.

PART 3

3.3. Connecting Globalization to Video Games

Globalization is not a new phenomenon; it has always been present in our history. But what we see now is that it has accelerated a lot. It includes more people; it goes faster, and it happens more often. Part of why this is the case can be explained by the new digital infrastructures of our Western society (Wang, Spotti, Juffermans, Cornips, Kroon & Blommeart, 2013). Through the internet with its Web 2.0 that sparked the culture of connectivity and the addition of smartphones, it has become fairly easy to connect to people all over the world in different ways. Because of that, we are more easily confronted with other norms, opinions, and ideas from different parts of the globe (van Dijck, 2013). Society has become information-driven, with new forms of global flows and networks that exist online and offline. The internet is now the main infrastructure of globalization; it is the thing that globalizes us into who we are today (Castells, 2010). New global 'identities', that are more in the open because of the digitalization of our society, are slowly also showing up in games, in the form of new diverse characters.

But why with new diverse characters? Our identities have changed with globalization. "People define their ‘identity’ (singular) in relation to a multitude of different niches" (Blommaert & Varis, 2015). These different niches, such as urban culture, hipster culture, LGBT culture, gaming culture etc. have become more easily accessible because of new technologies. This influences the way we organize our lives and how we think about them. There is a wide range of new cultural phenomena due to globalization, and these new phenomena also seem to be more present in video games. Different ethnicities, genders and sexual orientations appear more in mainstream games, and we can likely connect these politically correct depictions to our new digitalized global life that is influencing our way of being. Our video games now match the super-diverse society we live in more, due to globalization.

4. CONCLUSION

gender stereotypes can be seen when looking at the physical features of two characters whose roles are very much alike. Croft (figure 6) seems to be a hyper-sexualized character. In other words, she is presented in a 'sexy' way (big breasts, tiny waist, revealing clothing etc.). As a result, powerful women like her are seen as sex objects. If we compare this to Nadine (figure 5), from the Uncharted series, that is (co-)written by the same man who wrote The Last Of Us, we see a non-sexualized woman with a strong physique that fits her character. Moreover, it is worth mentioning that she has a South-African background that reflects in her appearance. She is not a stereotypical sexy white female at all, but rather a globalized super-diverse one.

This is just one example of a new type of character in a Western video game that is influenced by globalization. When it's looked at a list of the most anticipated games, still to be released, or some Triple A games released in the past two years or so, we see a lot of games that have diverse characters, may this be main characters, or side characters. This means that the phenomenon is not just about some games that have happened to play over the past couple of years. Games like The Last of Us: Part ll, Death Stranding, Ghosts of Tsushima, Detroit: Become Human, Assassins's Creed Origins, Red Dead Redemption 2, Dishonored 2 and Uncharted: The Lost Legacy all have diverse characters in leading roles.

These type of games show that the stereotypical white male main character is no longer as dominant a character in the gaming community as it used to be. They also show that we are slowly turning towards a more accurate representation of our globalized society (and past) in the characters of video games. The gaming community has long been a niche culture within our society. It had its own culture, which used to be a masculine one that did not leave much room for diverse characters. Gaming is now more global and popular than ever, and this might be why games are also becoming more diverse, a development that has provoked conversations about political correctness, since it includes groups of people who in the past were a neglected subject in gaming. Whether this development is 'ruining' the games, story, or gameplay wise, is another discussion. But it is so that our globalized and diverse society is now indeed influencing the stories of video games.

REFERENCES

[1] Blommaert, J. & Varis, P. (2015). Enoughness, accent and light communities: Essays on contemporary identities. Tilburg Papers in Culture Studies 139.

[2] Burgess, M., Dill, K., Sterner, P., Burgess, S., & Brown, B. (2011). Playing With Prejudice: The Prevalence and Consequences of Racial Stereotypes in Video Games. *Media Psychology* Volume 14, Issue 3.

[3] Castells, M. (2010). *The rise of the network society.* Malden: Wiley Blackwell.

[4] Entertainment Software Association (2015). Essential facts about the computer and video game industry.

[5] Gravning, J. (2014), *How Video Games Are Slowly, Quietly Introducing LGBT Heroes*, The Atlantic.

[6] Jansz, J. & Martis, R. (2007). The Lara Phenomenon: Powerful Female Characters in Video Games. *Sex Roles* Volume 56, Issue 3–4, pp 141–148.

[7] Lalonde, R., Doan, L. & Patterson L. (2000). Political Correctness Beliefs, Threatened Identities, and Social Attitudes. *Group Processes & Intergroup Relations*: Volume 3, Issue 3, pp 317–336.

[8] Oxford Dictionary, political correctness.

[9] Paaßen, B., Morgenroth, T. & Stratemeyer, M. (2016). What is a True Gamer? The Male Gamer Stereotype and the Marginalization of Women in Video Game Culture. *Sex Roles* Volume 76, Issue 7–8, pp 421–435.

[10] Tamburo, P. (2017), *The Last of Us 2 and Why "Personal Politics" Belong in Video Games.* Crave Entertainment.

[11] Van Dijck, J. (2013). The culture of connectivity: A critical history of social media. New York, NY: Oxford University press.

[12] Vertovec, S. (2006). The emergence of super-diversity in Britain. Centre of Migration, Policy and Society, Paper 25.

[13] Wang, X., Spotti, M., Juffermans, K., Cornips, L., Kroon, S. & Blommaert, J. (2013), Globalization in the margins, Tilburg Papers in Culture Studies 73.

[14] Williams, D., Martins, N., Consalvo, M. & Ivory, J. (2009). The virtual census: representations of gender, race and age in video games. *New Media & Society* Volume 11, Issue 5, pp 815–834 [DOI: 10.1177/1461444809105354]

ADVANCED DEEP LEARNING MODEL

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China

ABSTRACT

Deep learning is a type of machine learning that trains a computer to perform human-like tasks, such as recognizing speech, identifying images or making predictions. Instead of organizing data to run through predefined equations, deep learning sets up basic parameters about the data and trains the computer to learn on its own by recognizing patterns using many layers of processing. This paper aims to illustrate some of the different deep learning algorithms and methods which can be applied to artificial intelligence analysis, as well as the opportunities provided by the application in various decision making domains.

KEYWORDS

Artificial Intelligence, Machine Learning, Deep Learning.

1. INTRODUCTION

Deep learning is one of the foundations of artificial intelligence (AI), and the current interest in deep learning is due in part to the buzz surrounding AI. Deep learning techniques have improved the ability to classify, recognize, detect and describe – in one word, understand [1]. For example, deep learning is used to classify images, recognize speech, detect objects and describe content.

Several developments are now advancing deep learning:

- Algorithmic improvements have boosted the performance of deep learning methods.
- New machine learning approaches have improved accuracy of models.
- New classes of neural networks have been developed that fit well for applications like text translation and image classification.
- We have a lot more data available to build neural networks with many deep layers, including streaming data from the Internet of Things, textual data from social media, physicians notes and investigative transcripts.
- Computational advances of distributed cloud computing and graphics processing units have put incredible computing power at our disposal. This level of computing power is necessary to train deep algorithms.

At the same time, human-to-machine interfaces have evolved greatly as well. The mouse and the keyboard are being replaced with gesture, swipe, touch and natural language, ushering in a renewed interest in AI and deep learning [2]. This paper will look at some of the different deep learning algorithms and methods which can be applied to AI analysis, as well as the opportunities provided by the application in various decision making domains.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 71-76, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111707

2. HOW DEEP LEARNING WORKS

Deep learning changes how you think about representing the problems that you are solving with analytics. It moves from telling the computer how to solve a problem to training the computer to solve the problem itself.

A traditional approach to analytics is to use the data at hand to engineer features to derive new variables, then select an analytic model and finally estimate the parameters (or the unknowns) of that model. These techniques can yield predictive systems that do not generalize well because completeness and correctness depend on quality of the model and its features [3]. For examples, if you develop a fraud model with feature engineering, you start with a set of variables, and you most likely derive a model from those variables using data transformations. You may end up with 30,000 variables that your mode depends on, then you have to shape the model, figure out which variables are meaningful, which ones are not, and so on. Adding mode data requires you to do it all over again.

The new approach with deep learning is to replace the formulation and specification of the model with hierarchical characterizations (or layers) that learn to recognize latent features of the data from the regularities in the layers. The paradigm shift with deep learning is a move from feature engineering to feature representation. The promise of deep learning is that it can lead to predictive systems that generalize well, adapt well, continuously improve as new data arrives, and are more dynamic than predictive systems built on hard business rules. You no longer fit a model. Instead, you train the task.

Deep learning is making a big impact across industries. In life sciences, deep learning can be used for advanced image analysis, research, drug discovery, prediction of health problems and disease symptoms, and the acceleration of insights from genomic sequencing. In transportation, it can help autonomous vehicles adapt to changing conditions [4[. It is also used to protect critical infrastructure and speed response.

More deep learning methods use neural networks architectures, which is why deep learning models are often referred to as deep neural networks. The term “deep” usually refers to the number of hidden layers in the neural network. Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150. Deep learning models are trained by using large sets of labelled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.

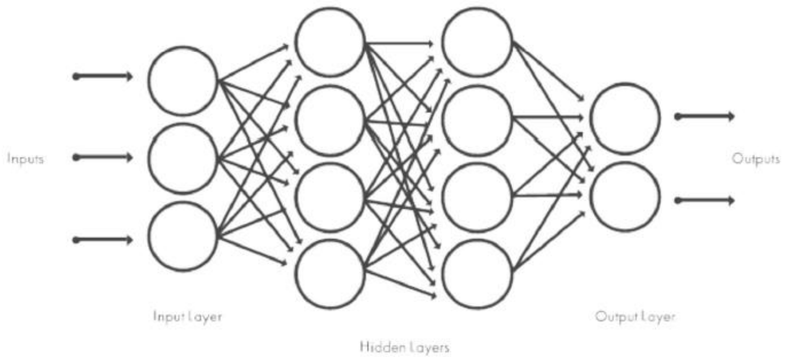


Figure 1: Neural networks, which are organized in layers consisting of a set of inter connected nodes. Networks can have tens or hundreds of hidden layers.

3. HOW DEEP LEARNING BEING USED

To the outside eye, deep learning may appear to be in a research phase as computer science researchers and data scientists continue to test its capabilities. However, deep learning has many practical applications that businesses are using today, and many more that will be used as research continues [5]. Popular uses today include:

Speech Recognition

Both the business and academic worlds have embraced deep learning for speech recognition. Xbox, Skype, Google Now and Apple's Siri®, to name a few, are already employing deep learning technologies in their systems to recognize human speech and voice patterns.

Natural Language Processing

Neural networks, a central component of deep learning, have been used to process and analyse written text for many years. A specialization of text mining, this technique can be used to discover patterns in customer complaints, physician notes or news reports, to name a few.

Image Recognition

One practical application of image recognition is automatic image captioning and scene description. This could be crucial in law enforcement investigations for identifying criminal activity in thousands of photos submitted by bystanders in a crowded area where a crime has occurred. Self-driving cars will also benefit from image recognition through the use of 360degree camera technology.

Recommendation Systems

Amazon and Netflix have popularized the notion of a recommendation system with a good chance of knowing what you might be interested in next, based on past behaviour. Deep learning can be used to enhance recommendations in complex environments such as music interests or clothing preferences across multiple platforms.

Recent advances in deep learning have improved to the point where deep learning outperforms humans in some tasks like classifying objects in images [6]. While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

1. Deep learning requires large amounts of labelled data. For example, driverless car development requires millions of images and thousands of hours of video.
2. Deep learning requires substantial computing power. High-performance GPUs have a parallel architecture that is efficient for deep learning. When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

When choosing between machine learning and deep learning, consider whether you have a high-performance GPU and lots of labelled data. If you don't have either of those things, it may make more sense to use machine learning instead of deep learning. Deep learning is generally more complex, so you'll need at least a few thousand images to get reliable results. Having a high-performance GPU means the model will take less time to analyse all those images [7].

4. DEEP LEARNING OPPORTUNITIES AND APPLICATIONS

A lot of computational power is needed to solve deep learning problems because of the iterative nature of deep learning algorithms, their complexity as the number of layers increase, and the large volumes of data needed to train the networks.

The dynamic nature of deep learning methods – their ability to continuously improve and adapt to changes in the underlying information pattern – presents a great opportunity to introduce more dynamic behaviour into analytics [8]. Greater personalization of customer analytics is one possibility. Another great opportunity is to improve accuracy and performance in applications where neural networks have been used for a long time. Through better algorithms and more computing power, we can add greater depth.

While the current market focus of deep learning techniques is in applications of cognitive computing, there is also great potential in more traditional analytics applications, for example, time series analysis. Another opportunity is to simply be more efficient and streamlined in existing analytical operations. Recently, some study showed that with deep neural networks in speech-to-text transcription problems [9]. Compared to the standard techniques, the word-errorrate decreased by more than 10 percent when deep neural networks were applied. They also eliminated about 10 steps of data preprocessing, feature engineering and modelling. The impressive performance gains and the time savings when compared to feature engineering signify a paradigm shift.

Here are some examples of deep learning applications are used in different industries:

Automated Driving: Automotive researchers are using deep learning to automatically detect objects such as stop signs and traffic lights. In addition, deep learning is used to detect pedestrians, which helps decrease accidents.
Aerospace and Defence**:** Deep learning is used to identify objects from satellites that locate areas of interest, and identify safe or unsafe zones for troops.

Medical Research**:** Cancer researchers are using deep learning to automatically detect cancer cells. Teams at UCLA built an advanced microscope that yields a high-dimensional data set used to train a deep learning application to accurately identify cancer cells [10].

Industrial Automation**:** Deep learning is helping to improve worker safety around heavy machinery by automatically detecting when people or objects are within an unsafe distance of machines.

Electronics**:** Deep learning is being used in automated hearing and speech translation. For example, home assistance devices that respond to your voice and know your preferences are powered by deep learning applications.

5. HOW TO CREATE AND TRAIN DEEP LEARNING MODELS

The three most common ways people use deep learning to perform object classification are:

Training from Scratch

To train a deep network from scratch, you gather a very large labelled data set and design a network architecture that will learn the features and model. This is good for new applications, or applications that will have a large number of output categories. This is a less common approach because with the large amount of data and rate of learning, these networks typically take days or weeks to train [11].

Transfer Learning

Most deep learning applications use the transfer learning approach, a process that involves finetuning a pre-trained model. User can start with an existing network, such as AlexNet or GoogLeNet, and feed in new data containing previously unknown classes [12]. After making some tweaks to the network, user can now perform a new task, such as categorizing only dogs or cats instead of 10,000 different objects. This also has the advantage of needing much less data (processing thousands of images, rather than millions), so computation time drops to minutes or hours.

Feature Extraction

A slightly less common, more specialized approach to deep learning is to use the network as a feature extractor. Since all the layers are tasked with learning certain features from images, user can pull these features out of the network at any time during the training process [13]. These features can then be used as input to a machine learning model such as support vector machines (SVM).

6. CONCLUSIONS

So this study was concerned by understanding the interrelation between deep learning and AI, what frameworks and systems that worked, and how deep learning can impact the learning process whether by introducing new innovations that foster advanced deep learning process and escalating power consumption, security issues and replacing human in workplaces [14]. The advanced deep learning algorithms with various applications show promising results in artificial intelligence development and further evaluation and research using smarter deep learning models are in progress.

REFERENCES

[1] Shi, Z., (2019). Cognitive Machine Learning. *International Journal of Intelligence Science*, 9, pp. 111-121.

[2] Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B., (2019). Human-Level Concept Learning through Probabilistic Program Induction. Science, 350, pp. 1332-1338.

[3] Silver, D., Huang, A., Maddison, C.J., et al., (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, 529, pp. 484-489.

[4] Fukushima, K., (1980). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36, pp. 193-202.

[5] Lecun, Y., Bottou, L., Orr, G.B., et al., (1998). Efficient Backprop. Neural Networks Tricks of the Trade, 1524, pp. 9-50.

[6] McClelland, J.L., et al., (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. Psychological Review, 102, pp. 419-457.

[7] Kumaran, D., Hassabis, D. and McClelland, J.L., (2016). What Learning Systems Do Intelligent Agents Need? Complementary Learning Systems Theory Updated. Trends in Cognitive Sciences, 20, pp. 512-534.

[8] Wang, R., (2019). Research on Image Generation and Style Transfer Algorithm Based on Deep Learning. *Open Journal of Applied Sciences*, 9, pp. 661-672.

[9] Krizhevsky, A., Sutskever, I., Hinton, G.E., et al., (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems, 141, pp. 1097-1105.

[10] Long, J., Shelhamer, E., Darrell, T., et al., (2015). Fully Convolutional Networks for Semantic Segmentation. Computer Vision and Pattern Recognition, Boston, pp. 3431-3440.

[11] Noh, H., Hong, S., Han, B., et al., (2015). Learning Deconvolution Network for Semantic Segmentation. International Conference on Computer Vision, Santiago, 7-13, pp. 1520-1528.

[12] Cheng, Z., Yang, Q., Sheng, B., et al., (2015). Deep Colorization. International Conference on Computer Vision, Santiago, pp. 415-423.

[13] Mahendran, A. and Vedaldi, A., (2015). Understanding Deep Image Representations by Inverting Them. Computer Vision and Pattern Recognition, Boston, pp. 5188-5196.

[14] Gatys, L.A., Ecker, A.S. and Bethge, M., (2015). Texture Synthesis Using Convolutional Neural Networks. In: Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, Quebec, pp. 262-270.

AUTHOR

Prof. Yew Kee Wong (Eric) is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, HewlettPackard (HP) and Unisys as an AI consultant.
His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.

GLOBAL RESEARCH DECENTRALIZED AUTONOMOUS ORGANIZATION (GR-DAO): A DAO OF GLOBAL RESEARCHERS

Kelly L. Page[1] and Adel Elmessiry[2]

[1]LWYL Studio, Chicago, IL, USA
[2]AlphaFin, Nashville, TN, USA

ABSTRACT

The latest trend in Blockchain formation is to utilize decentralized autonomous organizations (DAO) in many verticals. To date, little attention has been given to address the global research domain due to the difficulty in creating a comprehensive framework that can marry the cutting edge of academic grade scientific research with a decentralized governance body of researchers. A global research decentralized autonomous organization (GR-DAO) would have a profound impact on the research community academically, commercially, and the public good.

In this paper, we propose the GR-DAO as a global community of researchers committed to collectively creating knowledge and sharing it with the world. Scientific research is the means for knowledge creation and learning.

The GR-DAO provides the guidance, community and technological solutions for the evolution of a global research infrastructure and environment. Through its design, the GR-DAO embraces, enhances and extends the model of research, research on decentralization and DAO as a model for decentralised and autonomous organizing. This design, in turn, improves most of the uses for and applications of research for the greater good of society.

The paper examines the core motivation, purpose and design of the GR-DAO, its strategy to embrace, enhance and extend the research ecosystem, and the GR-DAO design uses across the DAO ecosystem.

KEYWORDS

Scientific Research, Researcher, Research, Knowledge, Learning, Cocreated Knowledge, Applied Research, Decentralized Autonomous, Organization, DAO, Research Model, Research Activity, Blockchain, Emerging Technology, Incentive Design, Reputation Staking, Distributed Ledger Technology, Decentralized Infrastructure.

1. INTRODUCTION

While much scientific research is publicly funded or in the public interest, it is not available to the public. This is a global problem fueled by the increasing centralization of scientific research and knowledge in our academies and foundations; and the private commodification of research activities and outputs.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 77-90, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111708

Now imagine a different world. A world where cutting edge scientific research is publicly owned and available to everyone. Where researchers are paid fairly, equitably, and transparently for their contributions to science, their participation in the global research community, and the impact and attribution of their research is effectively and efficiently tracked and remunerated over time, and openly in a distributed digital ledger.

Let's take this world even further. You are an early career researcher working on a cutting edge research idea. You live and work in Kandahar, Afghanistan. You identify as female with three children and English is not your first language. Everyday you too go to work to do scientific research, contributing to a global research community for the public good. You too are a cutting edge scientific researcher and your contribution is as equally valued and remunerated in the global research community as that of your peer and collaborator, who lives in Boston, and is tenured at a world leading research institution.

The creation of these research worlds to address the shortcomings of the current organizational design of scientific research, especially its centralized and private commodification, is socially, financially and technically possible. The latest trend in Blockchain formation is to utilize decentralized autonomous organizations (DAO) in many verticals. To date, little attention has been given to address the global research domain and its design due to the difficulty in creating a comprehensive framework that can marry the cutting edge of academic grade scientific research with a decentralized governance body of researchers.

To address this, we propose the creation of a Global Research Decentralized Autonomous Organization (GR-DAO). A GR-DAO would have a profound impact on the research community both academically, commercially and for the public good. Decentralized Autonomous Organizations (DAO) represent the next evolution in global organizational governance. They are on the rise, and it is an exciting time for research scholars globally and organizational and technology scholars in particular, to address this emerging phenomenon with new theory and solid empirical research for a global research community.

The paper examines the motivation, purpose and design of the GR-DAO, its strategy to embrace, enhance and extend the research ecosystem, and the GR-DAO design uses across the DAO ecosystem.

2. MOTIVATION

We are experiencing the rise in the private commodification and centralization of scientific and/or academic research. These two phenomena have contributed to a number of challenges for the global research community. The most pressing challenge being that while much scientific research is publicly funded or in the public interest, it is not available to the public. The creation of a Global Research Decentralized Autonomous Organization (GR-DAO), would have a profound impact on the research community both academically, commercially and for the public good.

In this section we discuss the two driving motivations for the establishment of a GR-DAO including the rise in: 1) Research Commodification and 2) Research Centralization.

2.1. Research commodification

Research is any creative systematic activity undertaken in order to increase the stock of knowledge, including knowledge of humankind, culture and society, and the use of this knowledge to devise new applications [1]. It is an activity motivated for 1) the public interest, that is for the betterment of society and the advancement of knowledge for all humankind [2][3] and/or 2) commercial interest and application, that is advancing self or an entity for interest, compensation or commercial gain [4][5].

To serve both motives, two complementary forms of research activity have emerged in our modern culture [4]: *Scientific* which is also called basic or academic (or of the academy) and *Applied* or professional research designed for practical purposes and commercial advancement. The two differ in terms of methodology, methods and the interests of the sponsoring parties.

Scientific research is more often motivated for academic or general interests of the public, is performed by researchers in research institutions applying systematic and constructed scientific methods and/or protocols to obtain, analyze, and interpret data, where the intention is to identify facts and/or opinions that will assist in solving the problem or dealing with the situation [6][7]. It is also academic or *of the academy*, in this it is a core activity conducted or supported by members of public and private, largely non profit universities, and funded by public tax money, foundations or donors. Applied research in contract is designed to answer specific questions aimed at solving practical or use problems, be they technical, social or for commercial purposes. Conducted by private entities, both nonprofit and for profit, it is more often funded privately and conducted for proprietary purposes[6][7].

The purpose of these forms of research may serve to advance knowledge, the devising of new applications and for the betterment of society. Yet, while we expect the commodification of applied research as a ‘private good’, since the 1980’s, we are increasingly seeing the rise of the commodification of publicly funded scientific research, and shift from a public good to a private one that can be traded [8].

A complex phenomenon, commodification is identified with commercialization, that is, the pursuit of profit by academic or media institutions through selling the expertise of researchers and the results of their inquiries [9]. The commodification of scientific research is part of a comprehensive and long-term social development [9]. This development is often described as the economization, or economic instrumentalization, of human activities and institutions, or even entire social subsystems [10][11]. In the higher education sector, we further see this activity through the lens of the marketization of higher education - the growing influence of market forces on higher education, resulting in what is defined by Fairclough (1993) as the marketization of academic discourse [12][13].

Commodification implies the expropriation of goods from the particular communities that produced them by reducing the intrinsic, community value of these goods to their pecuniary exchange value on an independent market [14]. Increasingly we are seeing the rise in higher education and its activities, moving from a once held belief of it as a public good of benefit to the individual and to the public, to it being a private commodity for sale.

As a private commodity for sale it benefits the interest of the few and can take many forms such as research that is commercially funded, strategic research alliances or partnerships with private firms, or interest groups, paid or sponsored positions such as research or endowed Chairs, as well as the commercial activities of many of the major scientometric databases and media companies with a vested interest. The acquisition and exploitation of intellectual property and patents on the results of scientific research is another form [9].

This rise in the private commodification of scientific research as a private good for sale contributes to a number of challenges for a global research community committed to research as a public good, in the public's interest and the public domain.

2.2. Research Centralization

The organization of research and research cultures is becoming increasingly centralized, hierarchical and top-down as well as privatised. This is in contrast to over twenty or thirty years ago, when many research and higher education institutions were relatively decentralised, flatter and with more autonomy at the department, school or faculty level and in research activity. [15][16][17].

The centralization of scientific research includes: the central organization of institutional research activity, research evaluation and funding, publishing, as well as a centralist view of the global research community.

- *Institutional Research Activity:* The organization of institutional research activity is becoming increasingly centralised such as in the creation of administrative units centrally to oversee research policy and activities; consolidation and marketization of research training programs for the education of early career researchers; central decision making as to what and whom to fund and support, and research activity evaluation.
- *Research Evaluation/Funding*: This is coupled with the formalization of bureaucratic research bodies for the conduct of research assessment exercises (e.g., REF in the UK) in order to receive public funding, centralising the decision as to what research is of value to the public good (money) and the indicators upon which this research is evaluated.
- *Research Publishing:* We've experienced the rise in private centralization of essential research activities such as the peer-review, publishing and the dissemination process, today called the 'academic publishing industry'. Once overseen by researchers and institutions themselves, it is today a robust business sector that economically benefits greatly from researchers and the public. Worldwide it has sales amounting to more than USD 19 billion. This positions it between the music industry and the film industry in revenue [25]. Yet, it is a closed system with high barriers to entry and limited public access. It is dominated by five large publishing houses: Elsevier, Black & Wiley, Taylor & Francis, Springer Nature and SAGE, which control more than 50 % of the market between them. In its current business model, public funds fund all stages of research production, the research faculty who manage the peer-review process for free and esteem, and then pays through an institution again to have member access to the research articles and results archived behind a paywall.
- *Research Community*: There is an inherent centralism in the view of participation in research in the global research community. The community is dominated by and centralised from a western, developed and euro-centric world-view in knowledge creation, value and its dissemination. For example, often science is not not encouraged in developing countries because it is expensive, yet it is of inherent value and impact to their development and progress, as much if not more as developed ones. Further, researchers in these countries do not have the same level or type of support for their contribution.

This rise in the centralization of scientific research is in contrast to what is stressed in much management and organization research as to what fosters cultures of innovation and sharing-- all critical ingredients for scientific research [18][19][20].

2.3. Emerging Challenges

The rising centralisation of scientific research, coupled with its private commodification, is contributing to the challenges for a global research community to do research for the public and in the global domain (Table 1). These include (but are not limited to): Unsustainable business model, funding is inflexible and inefficient, rising inequity, fixed boundaries and social norms, Inaccessibility, Poor Transparency.

Table 1. Current Challenges of Our Global Research Community

Challenge	**Description**
Unsustainable Business Model	● The labor of research business model is not economically workable or sustainable for researchers or institutions participating in the knowledge creation economy. ● Researchers can not sustain the work product or load expected of them, esp. with a rise in 'work for hire' doctrines resulting in limited to no long term fiscal benefit for the creative work contributed.
Funding is Inflexible and Inefficient	● How, who and when is research funded, who makes these decisions and the evaluation criteria upon which they are made are not transparent. ● Science funding is a mess with academic researchers having to rely on outside grants in order to pay salaries and buy their equipment. ● This results in many leading researchers spending some 40-50% of their time writing research grants and responding to grant administration and evaluation processes.
Rising Inequity	● Researchers are not paid equitably or enough for their labor, skills and contribution to the knowledge and research economy. ● The rising inequality continues to disadvantaged people of color, women and those of less economic means are sidelined or not given access to research training, research and denied due attribution for their research labor and work.
Fixed Boundaries and Social Norms	● The culture is fixed with boundaries and high barriers to entry. It is of benefit to only those who have access to a "certain" education or the institutions which are deemed "worthy" or "elite" in research terms to participate. ● The current structure de-incentivises interdisciplinary research across boundaries, and/or researchers playing at the edges or outside the community norms.
Inaccessibility	● Scientific research work is not in the public domain. It is archived behind a paywall and only accessible to those who can afford to pay membership fees or have access through institution affiliation.
	● Those research works which are publicly accessible, are often not valued or ranked as worthy for the promotional / tenure system universities put in place to evaluate the research deliverables of a researcher.
Poor Transparency	● The agreements and contracts between researchers and institutions, and private interests are not transparent. It is difficult to know who is funding research, who owns it and who has the rights of access to it. ● There is also limited transparency in how much the industry or sector is actually worth or of value and the impact or access to that value in the public domain.

In short, while much scientific research is publicly funded or in the public interest, it is not available to the public for advancing the knowledge of humankind.

3. THE PROPOSED SOLUTION

To address the shortcomings of the current organizational design of scientific research, especially its centralized and private commodification, we propose the creation of a Global Research Decentralized Autonomous Organization (GR-DAO).

A DAO is a non-hierarchical organizations that perform and record routine tasks on a peer-to-peer, cryptographically secure, public network, and rely on the voluntary contributions of their internal stakeholders to operate, manage, and evolve the organization through a democratic consultation and voting process [26][27]. DAOs coordinate routine tasks through cryptographic routines (as opposed to human routines). Blockchain-based organizing and the resulting DAOs have the ability to replace centralized intermediaries in other applications requiring complex coordination [29]. Blockchain has the clear potential to substantially upgrade the processes and organization traditionally underpinning academic science [35].

3.1. Blockchain Applications to Scientific Publishing

The goal of blockchain applications for scientific research has been to establish origins of research outputs, and tracking how the assets change through the publishing lifecycle. A summary of applications are summarised in Table 2.

Table 2. Blockchain-based Scientific Publishing Applications

Application	Description
ARTIFACTS	● Records immutable chains of scholarly artifacts (e.g., figures, images, etc.) to establish attribution and proof-of-existence of early scientific work. ● Focuses on research asset creation, tracking and sharing of publishing processes.
Manubot	● Manuscript version control needed during the editing and publishing process.
Orvium	● Focuses on integrating blockchain technology into the publication life cycle while also encouraging open-science and research dissemination aims. ● Ability to create Decentralized Autonomous Journals (DAJs) with their own governance rules and licensing and subscription models.
Pluto	● Manage transfers of value through the research lifecycle by using smart contracts and tokens on the blockchain and smart contract environment Ethereum. ● Allows users to submit and store different types of scientific information/data (with Digital Object Identifiers) and retain copyright control.
Sciencematters and Eureka	● Open-access web-based OA publishing platform that focuses on single observation studies (and also encourages negative and replication studies); works in tandem with a journal submission and token reward system powered by blockchain.
Scienceroot	● Utilizing its own token [called “Science Token” (ST)], a digital wallet, and smart contracts operating on a proof-of-stake consensus. ● Decentralized collaboration platform, marketplace, and repository. ● Relies on tokenization to drive the research process.

Most blockchain applications like those shared in Table 2 focus on blockchain applications for the peer-review, publishing and citation activities of scientific research. This is not comprehensive of the knowledge creation process. The GR-DAO proposal is a community wherein blockchain application and token economy supports and actuates the entire knowledge creation, workflow process and validates the contribution of researchers throughout the knowledge generation experience.

3.2. DAO Design Benefits

While DAO as a type of organization is relatively new, researchers in the fields management and organizational design have long documented the numerous social and cultural benefits of more flatter organisational structures, decentralisation of decision making and local autonomy over more structured and centralised ones. This is especially apparent for the fostering of cultures of innovation [18], a culture which is critical for the development of research, its evaluation and dissemination.

A global research community would benefit from being a decentralised organization as it would provide for more effective communication horizontally as well as vertically [19], encourage creativity among and between members [20], generate imaginative solutions to problems [21], increase levels of research motivation and satisfaction [22], as well as increase member responsiveness to changes in the external environment [23]. A detailed meta-analysis of the determinants of organisational innovation further confirm the significant negative influence of centralisation and of formalisation on organisational innovation [24], providing support for a more decentralized organization design.

In the following sections of this paper we outline the social design, legal and economic design of the GR-DAO, as well as its strategy to embrace, enhance and extend the global research ecosystem, as well as research across the DAO ecosystem.

4. GR-DAO Social Design

4.1. About

The GR-DAO is a global community of researchers committed to collectively creating knowledge and sharing it with the world. Scientific research is the means for knowledge creation and learning.

4.2. Mission

The mission of the GR-DAO is to embrace, enhance and extend scientific research in the public interest, as a public good and for advancing the knowledge of humankind publicly.

The GR-DAO will provide the guidance, community and technological solutions for the evolution of a global research infrastructure and community. Through its design, the GR-DAO embraces, enhances and extends the model of research, dedicates resources for scientific research on DAO's as a model for decentralised and autonomous organizing in research and technology communities. Our aim is to improve most of the uses for and applications of scientific research for the greater good of society.

4.3. Community Members

The GR-DAO is a global community of researchers committed to collectively creating knowledge and sharing it with the world. Scientific research is the means for knowledge creation and learning. The community includes (but is not limited to): scientists, researchers, philosophers, engineers, developers, and those with an active participatory interest in the pursuit and sharing of knowledge.

4.4. Foundational Beliefs and Values

At the heart we are the 'Founding Mothers', the 'Creators of Intelligence' and the 'Makers of Intelligent Life.' The GR-DAO community shares the following core beliefs as foundational elements to articulate their values:

- ***We are a collective*** - means we are a group of individuals who collectively as a community create and share knowledge with the world.
- ***We nurture the pursuit of knowledge* -** meaning we support research activities, education and the ways people learn, share and advance understanding and knowledge of the world.
- ***We hold space for research as a public good*** - available, accessible, transparent and verifiable.
- ***We champion equity*** - meaning equitable community participation and access to knowledge, as well as the equitable distribution of power and decision-making
- ***We honor individual freedom of choice* -** meaning an individual's power to choose, be it of research participation or ownership of one's personal information, individual security or privacy.
- ***We aim for efficiency*** - meaning the ratio of the useful work performed by a human, machine, or process to the total energy expended.
- ***We collaborate for the common good* -** meaning improving power distribution by raising the power of the most individuals without harming the least powerful.

The GR-DAO recognizes that these community values and core beliefs provide cohesion and longevity in the GR-DAO and in decentralized systems at large.

4.5. Community Activities

The GR-DAO will support research activities such as: research funding, research review and evaluation, research education, as well as the dissemination of scientific research via open access and creative commons licenses in the public domain. At start-up, the GR-DAO will focus on research pertaining to decentralization, DAO, blockchain applications and cryptocurrency, and expand its focus as the community, it's member design and its activities grow.

4.6. Community Agreements

Community agreements are a set of basic ground rules or policies that are asked of participants in the community to follow. A living constitution if you will, that is regularly revised and updated with the goal is to create an open and inclusive space so that every individual has the ability to participate, flourish and be heard, as well as to know what is expected of them and how they will exchange value as a member of the community.

To execute the agreement the GR-DAO will use smart contracts to execute the agreement logic in response to events, executing the performance of various tasks, processes or transactions that have been programmed into them to respond to a given set of conditions as set-down in the community agreement.

4.7. Community Development

The GR-DAO will be committed to the health and well-being of its community and members with ongoing development, education and learning activities. From the on-boarding of new members, to learning about advancements in technology, infrastructure and policies that could have an impact on the GR-DAO as a global research community committed to embracing, enhancing and extending research for the public good.

5. GR-DAO LEGAL CONSIDERATION

One important aspect of creating a successful DAO is to consider the legal structure and wrapper for how it will interact with the external entities. Current research organizations depend on centralized entities for the grant solicitation, allocation and management. This structure is largely inefficient and requires a huge amount of human intervention while creating a single point of failure [30].

In a DAO framework, the DAO member governance is conducted through the decentralized consensus mechanism which is more efficient and much more resilient.

5.1. Legal Structure

To be able to operate in both the real world and the virtual space in loosely coupled a structure in both must exist. This is accomplished by creating a legal rapper in the non-DAO world that would function as the representative of the DAO in a jurisdiction allowing it to conduct operations within their jurisdiction and in collaboration with other jurisdictions while maintaining its separate virtual existence. For example an association could be created under Swiss law as a nonprofit with a mandate to be engaged by the DAO so that it carries out three major functions namely: engage in contractual agreements, receive funds and issue appropriate receipts, and provide employment structures for those where it is required.

5.2. Governance

In most DAOs Governance is a function allowing the DAO to vote and reach decisions based on the collective will of the DAO. It is an important mechanism to make sure that the function of the DAO is going into the right direction as seen by the collective. An important aspect of the governance is to be able to weigh the difference in votes eternity and figuring out a way to make each vote has some sort of accountability and consequences to the voter.

5.3. Contribution-Based Internal Administration

The question that we'll have to answer is how are we going to make those decisions within the DAO itself. One such approach is to use a contribution based system. The contribution system in research there will be correlated with the amount of knowledge each individual individual actually contributes to the DAO. This is a way for each researcher to acquire more contribution tokens as they contribute more to the actual system itself and to the mission of the DAO. When taking a vote on a certain proposal for research or a decision each individual can stick an amount of this gained contribution to emphasize their support for this decision.

When the decision is voted on the member risks that stick to contribution on the outcome of the collective. Thus it allows the collective to steer towards a common vision for the research being funded, reviewed, conducted or shared with the end of the DAO.

6. GR-DAO GAME THEORY AND RESILIENCE

Game theory can be defined as the study of mathematical models of strategic interaction between rational decision-makers [31]. Our governance structure should continue a game theory-like incentive mechanism that allows the decision makers to have incentives for voting in the right direction as set by the collective and to be penalized for misleading the collective.

6.1. Game-Theoretical Design

One of the interesting game theory works is established in the cooperation through social influence done by Molinero et. al. (2015) and (2021) [31][32]. The main contribution is to show the relationship between the influence spread phenomenon coming from social network analysis and the binary decision-making in the voting system. We utilize the same basis but replacing the gains with our version of knowledge based reputation.

6.2. Equal Distribution of Knowledge

Since the main objective of the DAO is to promote knowledge, it is natural to utilize knowledge as the basis of how the system works. To create an equatibal system we would need to assure the equal distribution of knowledge within the system. That means that access to knowledge is available for everyone. It is the capability of each individual to achieve additional knowledge while contributing to the overall research activities of the system that gains them additional points and power.

6.3. Attack Resistance

In any system with value attacks or text or nothing but assured. The system must create a mechanism. It isn't to make the coolest of the attack, it is the attack much higher than the gains acquired by a successful attack. The GR-DAO is no different in that it tries to mitigate those attacks on the overall collective.

6.3.1. Sock Puppet Attacks

The 1st mechanism here is to assure that the only way for individuals to gain weight in the system, the system, meaning they are able to affect the decisions of the collective is by contributing to the actual output of the collective. The output is basically research and adding to the knowledge of the collective. The more people contribute the more power they will have to affect the decisions of the collective And since that power is not financially driven it cannot be bought but it can be only acquired.

Finally the requirement of creating a substantial body of work will create a deterrent for the players that want to attack the actual collective because in the event of losing they will lose all the hard work that they have put into the season.

This is akin to the proof of work however in our case we call it proof of research.

6.3.2. DOS Attacks

Denial of service attacks are another way of trying to sabotage a system that is working. The most effective way of defending against a denial of attack is by creating a cost for each request of service. This cost is set so that an entity cannot create an infinite number of requests and as tying up the resources of the collective while effectively shutting down the services.

7. GR-DAO ECONOMIC DESIGN

Unlike the traditional DAO models our model will use non-fungible tokens (NFT's) as the base for onboarding members into the collective. A non-fungible token (NFT) is a representation of a unique digital asset, essentially a digital certificate of authenticity, that cannot be equally swapped or traded for another NFT of the same type. They are stored on a blockchain or a distributed ledger and are used to represent ownership of unique items [34].

Each member of the GR-DAO will receive an NFT for their membership with a wallet. The wallet associated with the NFT, and the NFT will work as a reputation holder for the knowledge reputation tokens held for this member. This dual token system allows the user to vote on different proposals in the DAO [33].

The NFT based DAO membership model is depicted in Figure 1.

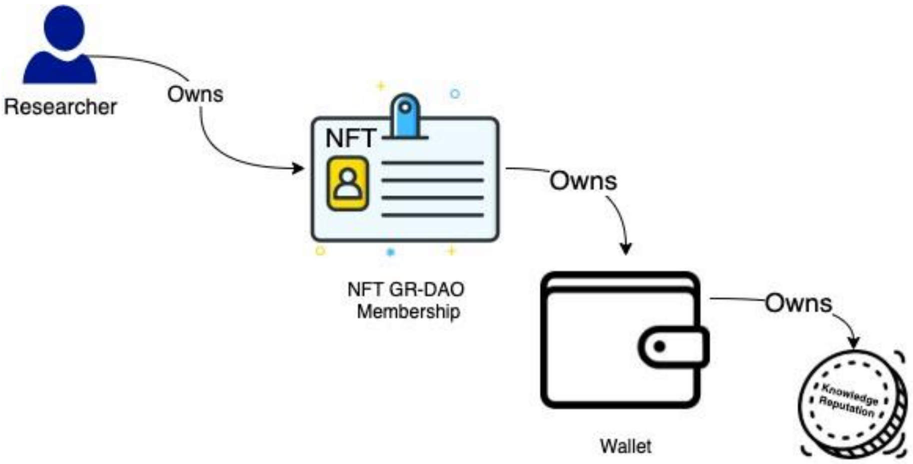


Figure 1. NFT Based DAO Membership Model

7.1. NFTs Model

Each of the NFTs hold the knowledge reputation tokens on behalf of the member who owns the NFT. Those NFTs are non transferables which lock in the reputation with this member. The only way for a remember to gain those knowledge reputation tokens is by contributing through the research and activities required to advance the mission of the DAO.

This is a quintessential part of how the research DAO will be able to function without being dominated by one single entity. Naturally the individuals or group of researchers who contribute the most work we'll be able to get higher gains out of the digital assets (e.g., research grants, papers, reviews etc), that will pass through the collective.

This is the actual desired outcome because it sets up a competitive landscape for the researchers in the DAO so that they can produce more research and gain more decision power in the DAO.

7.2.Knowledge Contribution Token

Knowledge Contribution tokens (KCT) are non transferable tokens. They can be only acquired by producing research work. New IP and data can be time-stamped and indisputably filed as NFTs on the blockchain as proof-of-knowledge for firmly claiming author-, and ownership, possibly backed by blockchain-based (self-sovereign) identity management (SSI). This means that the only benefit of gaining KCT is to participate in the DAO governance. They function like governance tokens except that they can not be traded. They represent a concrete proof of research contribution, thus we consider them proof of research.

Blockchain-enabled token economy may efficiently and transparently incentivize and coordinate an integrative and community-inclusive participatory approach to fuel crowdsourcing of collective intelligence [35]. The gaming aspect of staking the KCTs during the voting process ensures that the member is voting their conventions to the best of the entire DAO. They stand to lose those hard earned KCTs if they are voting in the wrong direction.

7.3. Weighted NFTs

An important consideration here is that we would need to set up some weights allowing different individuals to start from the same point. Weighted NFT's is an approach to assure that individuals in areas that do not have a lot of research facilities going through we'll be able to gain a voice in how the DAO is operated.

For example, members belonging to the same institution will have a reduced weight when voting and members of different institutions will have a slightly elevated weight.

8. CONCLUSIONS

Research is by far the most valuable aspect of human civilizations. Throughout human history, those civilizations capable of creating, using and passing on knowledge ended up with a lasting impact on the entire planet. Yet only if it is publicly available and accessible. While much scientific research is publicly funded or in the public interest, it is increasingly not a public good, available to the public, nor advancing the knowledge of humankind publicly. This is a global problem fueled by the increasing centralization of scientific research and knowledge in our academies and foundations; and the private commodification of research activities and outputs.

Decentralized Autonomous Organizations (DAO) represent the next evolution in global organizational governance. They are on the rise, and it is an exciting time for research scholars globally and organizational and technology scholars in particular, to address this emerging phenomenon with new theory and solid empirical research for a global research community.
We have presented a conceptual framework for adopting the best of both worlds of a global research community and of DAOs in creating the Global Research DAO (GR-DAO). We hope to be able to adopt this model and realize it in our next efforts.

ACKNOWLEDGEMENTS

The authors would like to thank the DEVxDAO and Learning Economy Foundation (LEF) for ongoing discussions about DAO solutions for the global research community.

REFERENCES

[1] OECD, (2007) *OECD Glossary of Statistical Terms – Research and development UNESCO Definition*, Accessed from: stats.oecd.org. Archived from the original on 19 February 2007, Retrieved 24 September 2021.

[2] Kitcher, P, (2001) *Science, truth, and democracy*, New York: Oxford.

[3] Habermas, J, (1971) *Knowledge and Human Interest*, Beacon Press, Boston.

[4] Sintonen, M, (1990) "Basic and Applied Sciences-Can the Distinction (Still) Be Drawn?" *Science Studies* 3:2 (1990), 23–31.

[5] Krimsky, S, (2003). *Science in the private interest*. Lanham, MD: Rowman and Littlefield.

[6] Mody, C. C. M, (2006) "Corporations, universities, and instrumental communities: Commercializing probe microscopy, 1981–1996," Technology and Culture, Vol. 47, No. 1, pp 56–80.

[7] Montgomery, K., and A. L, Oliver. (2009) "Shifts in guidelines for ethical scientific conduct: How public and private organizations create and change norms of research integrity" *Social Studies of Science,* Vol. 39, No. 1, pp 137–55.

[8] Tilak, J.B.G, (2008) "Higher education: a public good or a commodity for trade?" *Prospects,* Vol. 38, pp 449–466.

[9] Radder, Hans, (2010) "Chapter 1: The Commodification of Academic Research." In Hans Radder (ed.), *The Commodification of Academic Research: Analyses Assessments, Alternatives,* University of Pittsburgh Press.

[10] Calıskan, Koray and Callon, Michel, (2009) "Economization, part 1: shifting attention from the economy towards processes of economization" *Economy and Society,* Vol. 38, No. 3, pp 369-398.

[11] Calıskan, Koray and Callon, Michel, (2010) "Economization, part 2: A research programme for the study of markets" *Economy and Society,* Vol. 39, No. 1, pp. 1-32.

[12] Fairclough, Norman, (1993) "Critical Discourse Analysis and the Marketization of Public Discourse: The Universities" *Discourse and Society,* Vol. 4, No. 2, pp 133-168.

[13] Etzkowitz, H, (1998) "The norms of entrepreneurial science: Cognitive effects of the new university-industry linkages" *Research Policy,* Vol. 27, No. 8, pp 823–33.

[14] Kleinman, Daniel Lee (2010) "The Commercialization of Academic Culture and the Future of the University" In Hans Radder (ed.), *The Commodification of Academic Research: Analyses Assessments, Alternatives,* University of Pittsburgh Press.

[15] Alderman, G, (2009) "Higher education in the United Kingdom since 1945", *Times Higher Education,*
30 July, available fromhttps://www.timeshighereducation.com/books/higher-education-in-theunited-kingdom-since-1945/407560.paper?storycode=407560 [accessed January 2015]

[16] Dearlove, J, (1997) "The academic labour process: from collegiality and professionalism to managerialism and proletarianisation?" *Higher Education Review,* Vol. 30, pp 56–75.

[17] AGB, (1996) "Renewing the Academic Presidency: Stronger Leadership for Tougher Times", *Association of Governing Boards of Universities and Colleges,* Washington DC.

[18] Ben R. Martin, (2016) What's happening to our universities?, *Prometheus,* 34:1, 7-24.

[19] Burns, T. and Stalker, M, (1961) *The Management of Innovation*, Tavistock Publications, London.

[20] Khandwalla, P, (1977) *The Design of Organizations*, Harcourt Brace Jovanovich, New York, NY.

[21] Deal, T. and Kennedy, A, (1982) *Corporate Culture*, Addison-Wesley, Reading MA.

[22] Dewar, R. and Werbel, J, (1979) "Universalistic and contingency predictions of employee satisfaction and conflict", *Administrative Science Quarterly*, Vol. 24, pp 426–48.

[23] Schminke, M., Ambrose, M. & Cropanzano, R, (2000) "The effect of organizational structure on perceptions of procedural fairness", *Journal of Applied Psychology*, Vol. 85, pp 294–304.

[24] Damanpour, F, (1991) "Organizational Innovation: A Meta-Analysis of Effects of Determinants and Moderators" *The Academy of Management Journal,* Vol. 34, No. 3 pp 555-590.

[25] Buranyi S, (2017). Is the staggeringly profitable business of scientific publishing bad for science? *The Guardian* Date: 27.6.2017. Accessed 25.9.2021.

[26] van Valkenburgh P, Dietz J, De Filippi P, Shadab H, Xethalis G, Bollier D (2015), "Distributed collaborative organisations: distributed networks and regulatory frameworks" *Harvard Working Paper,* Accessed 01 Aug 2016.

[27] Dietz J, Xethalis G, De Filippi P, Hazard J (2016), "Model distributed collaborative organizations" *Stanford Working Group,* Accessed 01 Aug 2016.

[28] Nakamoto S, (2008) *Bitcoin: a peer-to-peer electronic cash system.* New York.

[29] Hsieh, YY., Vergne, JP., Anderson, P. et al., (2018) "Bitcoin and the rise of decentralized autonomous organizations" *Journal of Organizational Design,* Vol. 7, No. 14.

[30] McGregor-Lowndes, I. (2019). "The rise of the DAO disrupting 400 years of corporate structure" *The Proctor,* Vol. 39, No. 1, pp 32–33.

[31] Molinero, X., & Riquelme, F. (2021) "Influence decision models: From cooperative game theory to social network analysis" *Computer Science Review*, Vol. 39.

[32] Molinero, X., Riquelme, F., & Serna, M, (2015) "Cooperation through social influence." *European Journal of Operational Research*, Vol. *242,* No. 3.

[33] ElMessiry, M., ElMessiry, A., & ElMessiry, M. (2019). "Dual token blockchain economy framework" *In the International Conference on Blockchain,* Springer, Cham. pp. 157-170.

[34] Popescu, A. (2021). "Non-Fungible Tokens (NFT) - Innovation beyond the craze." *5th International Conference on Innovation in Business, Economics & Marketing research (IBEM-2021) Proceedings of Engineering & Technology – PET* - Vol 66. pp. 26-30.

[35] Ducrée, J. (2020). "Research – A blockchain of knowledge?" *Blockchain: Research and Applications*, Vol 1. (1-2).

AUTHORS

Kelly, L. Page, Ph.D.

A social design ethnographer, social and digital innovator, and learning entrepreneur committed to developing truly social cultures, people, and organizations with emerging and social technology. Kelly has a Ph.D. in the Psychology of Web (Hypermedia) Knowledge from UNSW and an obsession with learning innovation and digital social storytelling. She has over 18 years of experience working at the intersection of social innovation, social design, and learning of mediated social experiences for Startups, Universities, Schools, and School Districts, to Fortune 500 companies. She believes that innovative and entrepreneurial thinking is at the heart of creating truly social cultures, organizations, and leaders.

Her work has been published in leading peer-reviewed academic education, technology, and business journals, such as *Journal of Business Research, Studies in Higher Education, Computers in Human Behavior,*
International Journal of Interactive Marketing, International Journal of Human-Computer Studies, Psychology& Marketing, Behavior & Information Technology .... and been featured in The New York Times, Fast Company, Wall Street Journal. Nominated for an Edison Innovation Award, her work has received awards from IDMA and a BIMA for Best in British Digital.

Adel Elmessiry, Ph.D.

Tech entrepreneur, published expert on AI and Blockchain with 20+ years Healthcare, Mentor, Advisors & Speaker. Adel is a serial entrepreneur with three successful technology companies taken from inception to acquisition. He has a proven executive experience with a solid track record that includes over 10 years at HealthStream and 7 years at InVivoLink/ HealthTrust. Academically, he is holding a Ph.D. in Computer Science at NCSU Natural Language Processing. He serves as the president chief technology officer for AlphaFin, a Draper Goren Holm portfolio company. Together we are on a mission to build the next financial technology ecosystem that will empower the global economy.

FAST IMPLEMENTATION OF ELLIPTIC CURVE CRYPTOGRAPHIC ALGORITHM ON GF(3^M) BASED ON FPGA

Tan Yongliang, He Lesheng, Jin Haonan and Kong Qingyang

Information Institute, Yunnan University, Kunming, China

ABSTRACT

As quantum computing and the theory of bilinear pairings continue being studied in depth, elliptic curves on GF(3^m) are becoming of an increasing interest because they provide a higher security. What's more, because hardware encryption is more efficient and secure than software encryption in today's IoT security environment, this article implements a scalar multiplication algorithm for the elliptic curve on GF(3^m) on the FPGA device platform. The arithmetic in finite fields is quickly implemented by bit-oriented operations, and then the computation speed of point doubling and point addition is improved by a modified Jacobia projection coordinate system. The final experimental results demonstrate that the structure consumes a total of 7518 slices, which is capable of computing approximately 3000 scalar multiplications per second at 124 Mhz. It has relative advantages in terms of performance and resource consumption, which can be applied to specific confidential communication scenarios as an IP core.

KEYWORDS

GF(3^m), Elliptic Curve Cryptography, Scalar Multiplication, FPGA, IoT Security

1. INTRODUCTION

Ellipse Curve Cryptography (ECC) has advantages including a relatively short key size, a high security and the applicability to resource-constrained embedded products, which is widely used as an Advanced Encryption Standard (AES) in symmetric encryption algorithms[1] and has been a hot research topic in the application of cryptography in recent years. The current version 1.3 of the transport layer protocol highlights the growing importance of elliptic curve cryptographic algorithms[2], and with an increasing demand for security of data privacy in the IoT of modern network society, the key length of ECC required is getting longer and longer, which makes the implementation of traditional software more and more difficult while less and less efficient to achieve. Therefore, in today's IoT environment, software encryption is only applicable to general network security[3], and hardware encryption has undoubted advantages over software encryption in some areas that are extremely sensitive to security. According to the current status, the hardware implementation of the scalar multiplication algorithm on the binary and prime fields has been relatively well studied, while relatively little work has been done on GF(3^m). While Galbraith[4] experimentally pointed out that for Weil or Tate pairing-based cryptosystems, the security of GF(3^m) is higher in terms of bandwidth efficiency and security. Shen Shao[5] and other authors also proved that elliptic curves on GF(3^m) also had some properties similar to the fast computation of those on GF(2^m) in terms of computational efficiency. Other studies on GF(3^m) have only improved to reduce the computational complexity at the algorithm level[6,7]. Therefore, due to the lack of research on the hardware implementation of the scalar multiplication algorithm for elliptic curves on GF(3^m), the scalar multiplication algorithm for elliptic curves on GF(3^m) is

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 91-100, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111709

designed and implemented in this paper on the FPGA device platform. The arithmetic is quickly implemented by bit-oriented operation like binary fields, and a modified Jacobian projection coordinate system is used to increase the speed of point addition and point doubling. Meanwhile, point addition and point doubling operation are designed as a whole to improve the resource reuse rate. The finally implemented scalar multiplication structure has certain advantages over traditional fields in terms of resource consumption and computing speed, and is suitable for use as a secure cryptographic algorithm carrier in various communication scenarios. Next, this paper introduces the relevant theoretical knowledge in the second part, and the third part focuses on the arithmetic design and implementation over a finite field of characteristic 3. The fourth part focuses on the general design and implementation of the scalar product structure. And in the fifth part we find some related work to do the performance comparison analysis. At the end of this paper, we have made a summary of our work. It also gives directions for further improvement.

2. Related Knowledge

The security of elliptic curve cryptosystems is based on the computational Diffie-Hellman problem in order-n subgroups (ECDLP security) and that in finite fields (MOV security)[8]. The Weierstrass equation[9] for an elliptic curve is an elliptic curve E(K) defined over a field K with the following expression:

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_5 \qquad (1)$$

When K = GF(3^m), we call E(K) an elliptic curve defined on GF(3^m). An abelian Group can be formed based on the set of all solutions on an elliptic curve E(K) plus one infinity point, and the algorithm in the exchange group formed should satisfy the basic theory of groups. The following describes the group operators on elliptic curves defined on field K and char (K) = 3. The expression of E(K) is given in equation (2):

$$E(K) = \{(x, y) \in K \times K \mid y^2 = x^3 + ax + b\} \bigcup \{o\} \qquad (2)$$

Set $P = (x_1, y_1)$, $Q = (x_2, y_2)$ as two points on the elliptic curve, and according to the definition of group theory, there are $-O = O$, $-P = (x_1, -y_1)$ and $P + O = O + P = P$. Therefore, the expression of the third point obtained by adding two points on the elliptic curve on GF(3^m) has the following operating rules.

1) Point Addition 2) Point Doubling

$$\begin{cases} x_3 = \lambda^2 - x_1 - x_2 \\ y_3 = y_1 + y_2 - \lambda^3 \end{cases}, \quad \lambda = \frac{y_2 - y_1}{x_2 - x_1} \qquad \begin{cases} x_3 = \lambda^2 + x_1 \\ y_3 = -\lambda^3 - y_1 \end{cases}, \quad \lambda = -\frac{a}{y_1}$$

3. Design and implementation algorithms on GF(3^M)

3.1. Addition and Subtraction

In binary fields, hardware implementation of addition and subtraction operation among polynomial elements is simply a matter of dissimilarity by corresponding bits. So, in order to improve the performance of the algorithm in the finite fields of characteristic there, the coefficients of each polynomial element in GF(3) are stored as two values q2 and q1, of which q2 stores the higher of the polynomial coefficients, while q1 stores the lower of them. q1 and q2 increase in size as the number of polynomials increases. By saving the high and low bits of each coefficient as two separate values, it is possible to perform arithmetic operation that is directly bit-oriented as arithmetic operation in the binary fields, and this basic logic is easy for FPGAs to implement, improving the efficiency of operation. For example, the polynomial element $2x^6+x^5+x^3+2x^2+2x+1$ is represented in Figure 1.

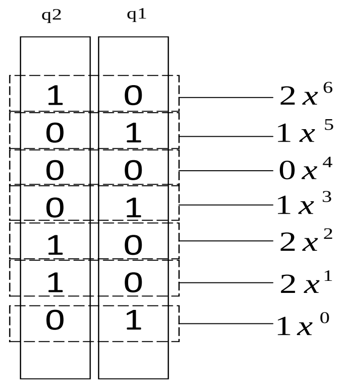


Figure 1. Polynomial element representation

The arithmetic operation of adding two polynomial elements with their coefficients is done in field GF(3), and the operation of adding two polynomials expressed through the above new method can be done using the basic operating logic. For example, the specific hardware flow implementation of adding polynomial A(x) to polynomial B(x) to obtain C(x) is shown in Figure 2.

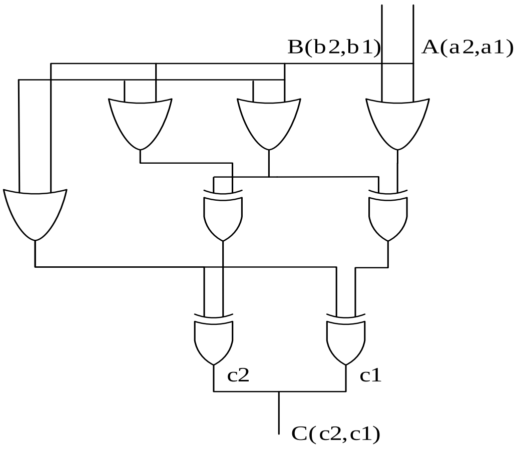


Figure 2. Hardware structure of addition

3.2. Multiplication

Multiplication in finite field is the most critical module in the entire design. Its operation is based on the principle of multiplying two polynomials and then taking the modulus P(x). P(x) is an irreducible polynomial defined in a finite field. A fully parallel modulo multiplier was initially considered to be implemented in this design, but the area and power consumption would be far more than expected, which would not be suitable for most practical cryptographic applications. So, an all-bit serial design was used to implement the multiplication operation. Repeatedly shifted the multiplicative polynomial down to one place while shifting the multiplied polynomial up to the other to perform the multiplication operation. Then, a corresponding bit of the multiplied polynomial was added or subtracted from the output value in each iteration depending on whether the lowest significant bit in multiplicative polynomial q1 or q2 was set as 1. This is shown in Figure 3.

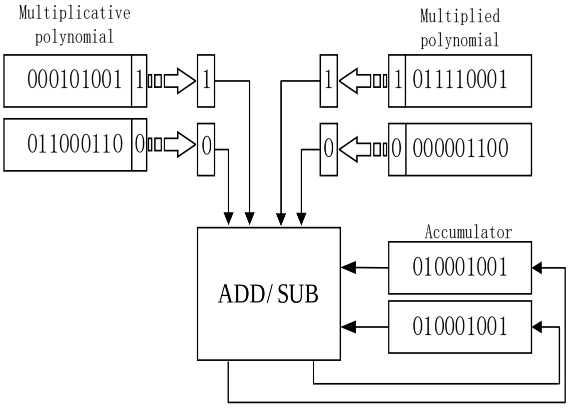


Figure 3. Multiplication of polynomial illustration

The advantage of this all-bit serial approach is that it does not require large intermediate storage, and a large amount of shift operation is more suitable for FPGAs to implement. Using a basic iterative structure and simple logic cells, neither a direct multiplier nor addition circuits are required, saving hardware resources. However, its disadvantage is that the calculation speed is relatively slow.

3.3. Inversion

Inverse is the most time and resource consuming operation in finite fields. Since the extended Euclidean algorithm effectively avoids the division operation, the use of this algorithm for inversion can effectively improve the calculation speed. The pseudocode for computing the inversion in hardware using the extended Euclidean algorithm is shown in Figure 4.

```
Algorithm 1
Input： A(X)
Output： $A(X)^{-1}$
S:=P(x);  R:=A(x);  U:=1;  V:=0;  d:=0;
q= S[`MOST]/R[`MOST]
FOR i:=1 to 2m DO
IF  R[`MOST]=0
 THEN R:=x*R;  U:=(x*u)modP(x)；  d:=d+1；
ELSE IF  d[0]=0
THEN R:=x*(S-q*R);  S:=R;
U:=x*(V-qU)modP(x) V:=U;  d:=d+1;
ELSE IF  d[0]!=0
THEN S:= x*(S-q *R); V:=V-q*U;
U:=(U/x)modP(x);  d:=d-1;
END;
( $A(X)^{-1}$=U/R[`MOST])
```

Figure 4. Algorithm for inversion in GF(3^m)

4. Scalar Multiplication Structure Design and Implementation

The most central operation on elliptic curves is the calculation of C*P, which can be expressed as $c \cdot P = P + P + \cdots + P(c\ times)$. So, the scalar product is actually a multiplication of the same point on the elliptic curve, which can be implemented through the algorithm shown in Figure 5.

```
Algorithm 2
Input： C,P
Output： B
B←O,   A←P;
WHILE  C>0  DO{
IF c is odd THEN B←B+A
   A←A+A
C ← floor(C/2)
}
RETURN B
```

Figure 5. Scalar multiplication algorithm

Because the scalar multiplication algorithm calls point operation in every loop, in order to avoid the time-consuming inverse operation, the point doubling and point addition operation can be done using a projection coordinate system to increase the speed. At the beginning of the calculation, it is necessary to convert point (x, y) in the affine coordinate system to point (X, Y, Z) on the Jacobian projection coordinates, and the conversion process is as follows.

$$X = x,\ \ Y = y,\ \ Z = 1$$

Thus, at the end of the scalar multiplication operation, we only need to convert the projection coordinates to affine coordinates, which requires only one inverse operation, as follows.

$$x = X / Z^2, y = Y / Z^3$$

In this paper, a modified Jacobia projection coordinate system[10] is used for point addition and point doubling calculation, in which a quadratic representation of (x, y) transformed into (X,Y,Z,aZ^4) is used. Because the modified Jacobia projection coordinate system can further improve the overall arithmetic performance. Figure 6 and 7 show the specific computation process and data flow of the designed point addition and point doubling operation as well as the data dependencies among each operation. The quadrilateral represents cubic operation, the ellipse represents multiplication operation, and the rectangle represents addition and subtraction operation. We can see that in this way there is no inverse operation in each loop of the scalar multiplication, and only one inverse operation is needed at the end of the calculation to transfer the points back to the affine coordinate system. On the other hand benefits from the fact that the cubic operation on GF(3^m) is much faster than the multiplication operation, largely reducing the computational complexity compared to the paper[11] where the cubic result is computed by two multiplication steps, thus making the overall scalar multiplication algorithm run much faster.

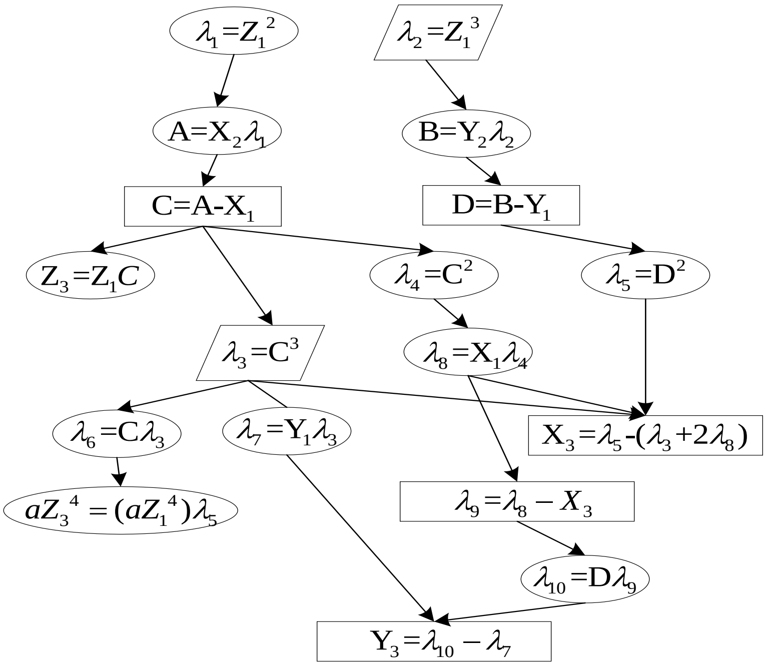

$\lambda_1 = Z_1^2$
$\lambda_2 = Z_1^3$
$A = X_2\lambda_1$
$B = Y_2\lambda_2$
$C = A - X_1$
$D = B - Y_1$
$Z_3 = Z_1 C$
$\lambda_4 = C^2$
$\lambda_5 = D^2$
$\lambda_3 = C^3$
$\lambda_8 = X_1\lambda_4$
$\lambda_6 = C\lambda_3$
$\lambda_7 = Y_1\lambda_3$
$X_3 = \lambda_5 - (\lambda_3 + 2\lambda_8)$
$\lambda_9 = \lambda_8 - X_3$
$aZ_3^4 = (aZ_1^4)\lambda_5$
$\lambda_{10} = D\lambda_9$
$Y_3 = \lambda_{10} - \lambda_7$

Figure 6. Point addition steps and data flow

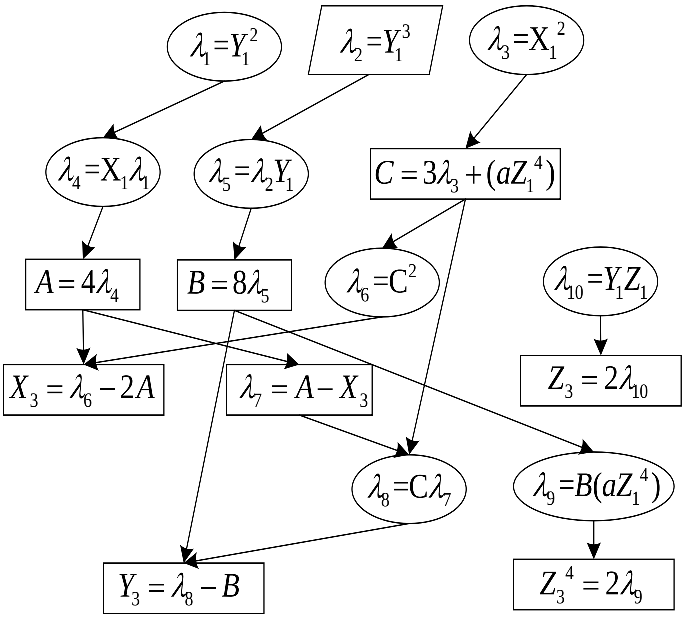


Figure 7. Point doubling steps and data flow

The general structure of the scalar multiplication module is given in Figure 8. The master controller is designed to control the final scalar multiplication operation by judging whether the input signal is valid or not. All intermediate variables are stored in the register heap, and because they will be read frequently, dual-port registers are used for the maximum reuse of hardware resources. The data after operation is stored in the data unit.

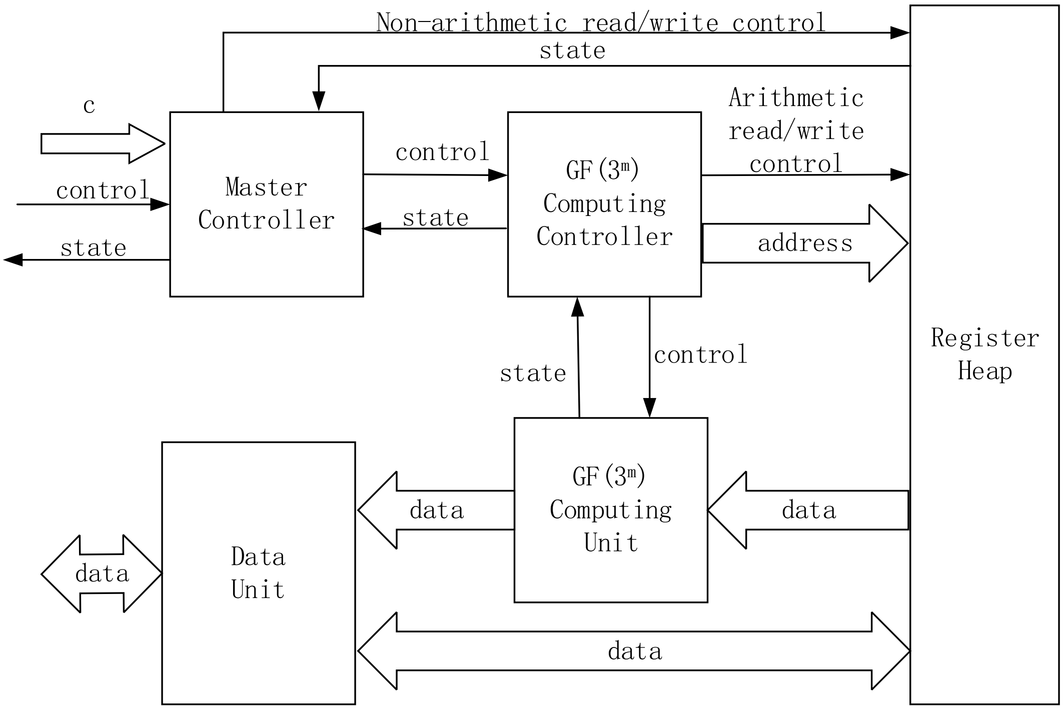
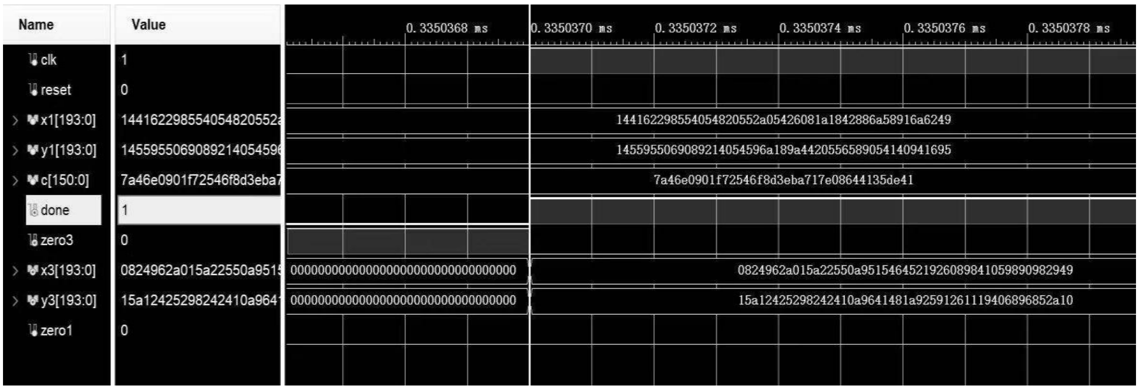


Figure 8. General framework of scalar multiplication

5. Experimental Results and Performance Analysis

5.1. Experimental Results

The choice of the hyperelliptic curve $y^2=x^3-x+1$ is defined on GF(3^m). Compared with the conventional elliptic curve cryptosystem, the Hyperelliptic Curve Cryptosystem (HCC)[12] has a higher security in the context of the general trend of quantum computer development and quantum attacks. Simulate and verify the elliptic curve with base field on GF(3^{97}) on the target device. The results of the inputs and outputs associated with the scalar multiplication module at a system clock frequency of 124Mhz are shown in Figure 9, where c is selected as the parameter of the order of this curve. To ensure the accuracy of the experimental results, we implemented the arithmetic on GF(3^{97}) through the Linux platform with reference to the software of literature[13], and did comparison as well as verification of the results with the same elliptic curve and input parameters, and finally the results obtained from the simulation verification in VIVADO 2018.3 were consistent with those obtained from Linux. The hardware platform we use is the XILINX-Xc7z020 device. Finally, we know through simulation that it takes only 0.335ms to calculate a scalar multiplication.



Name	Value
clk	1
reset	0
x1[193:0]	144162298554054820552a
y1[193:0]	1455955069089214054596
c[150:0]	7a46e0901f72546f8d3eba7
done	1
zero3	0
x3[193:0]	0824962a015a22550a9515
y3[193:0]	15a12425298242410a9641
zero1	0

0. 3350368 ms
0. 3350370 ms
0. 3350372 ms
0. 3350374 ms
0. 3350376 ms
0. 3350378 ms
144162298554054820552a05426081a1842886a58916a6249
1455955069089214054596a189a4420556589054140941695
7a46e0901f72546f8d3eba717e08644135de41
00000000000000000000000000000000
00000000000000000000000000000000
0824962a015a22550a9515464521926089841059890982949
15a12425298242410a9641481a92591261119406896852a10

Figure 9. Simulation results.

The main arithmetic in the scalar multiplication structure implemented is then simulated separately, the specific performance at a maximum frequency of 124Mhz and the resources occupied are shown in Table 1.

Table 1. GF(3^{97})- Major arithmetic performance and resource usage

GF(3^{97})-Arithmetic	**Time/us**	**Resources/slices**
multiplication	0.225	811
inversion	1.785	1373
Point addition	2.785	3201
Point doubling	2.325	2578

5.2. Performance Analysis and Comparison

Finally, in order to demonstrate that the scalar multiplication structure implemented in this paper has certain advantages in terms of both resource consumption and operational performance, the performance of scalar multiplication structures implemented in related conventional fields hardware is compared in this paper. Direct comparisons are difficult due to different platforms used, different algorithms defined and the variability of bit widths. Therefore, to ensure the relative fairness of the comparison, some conventional fields with similar bit widths are selected for it.

Table 2. Performance comparison with conventional fields of similar bit width

Design	Hardware platform	Bit width/ fields	Resources (Slices)	Frequency (MHz)	Time (Ms)	Number of Cycles(K)
[14]	Virtex-4	192	7080	21.55	15.87	342
[15]	Virtex-4	192	8590	48	2.3	110
[16]	XC4VLX	GF(2^{163})	9308	12.5	195.09	2438
[17]	Virtex-5	GF(2^{163})	9670	147.5	0.283	41.7
This paper	Xc7z020	194	7518	124	0.335	41.5

From Table 2, it can be seen that the scalar product structure designed and implemented in this paper consumes the least number of clock beats to compute one scalar result [14-16], which is slightly inferior to the computing performance of the design by Anissa[17], but saves about 22% in terms of resources. The comparison fully illustrates that the scalar product structure designed in this paper does have some advantages. However, the scalar multiplication structure implemented in this paper also has some shortcomings. If triple point arithmetic is used to implement the scalar multiplication algorithm, the performance will be better, but it will consume more resources, so how to improve the computing performance to a large extent without consuming too many resources is worthy of further study.

6. CONCLUSIONS

In this paper, we propose and implement a scalar product structure for elliptic curves on a base field of GF(3^m) based on the theory of elliptic curve cryptography methods. It is demonstrated by experimental comparison that the arithmetic performance of the implementation is sufficiently comparable to that of the conventional fields that have been extensively studied so far. This structure can be used as an IP core in some communication fields with higher security requirements. For finite fields of other bit widths, this can be achieved simply by modifying the design parameters, so it also has some generality. How to improve the inverse operation on finite field and scalar multiplication algorithm to increase the calculation speed is worthy of further study.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. U1631121). Thanks to all the contributors who have worked on this paper.

REFERENCES

[1] Jiang, J. Hou, J., Huang, H., Zhao, Y., Feng, X., & Science, S. O. (2019) Research on area-efficient low-entropy masking scheme for AES. Journal on Communications.

[2] RESCORLA E, MOZILLA. (2020) The transport layer security (TLS) proto-col version 1.3: RFC8446[S]. IETF, (2018-08).

[3] Yao T, Cao BW. (2018) An overview of cyberspace security[J]. China New Communications, 03(v.20): 170-170.

[4] Galbraith S D. (2001) Super singular Curves in Cryptography[C]. International Conference on the Theory and Application of Cryptology and Information Security.

[5] Shen Shao. (2015) Research on Fast Algorithms for ScalarMultiplication of Elliptic CurveCryptography over GF (3n)[J]. Computer Science & Application, 04(4): 390-399.

[6] Yeniaras, E., & Cenk, M. (2020). Faster Characteristic Three Polynomial Multiplication and Its Application to NTRU Prime Decapsulation.

[7] Kim, K. H., Kim, S. I., & Ju, S. C. (2007). New Fast Algorithms for Arithmetic on Elliptic Curves over Fields of Characteristic Three.

[8] Koblitz, N. (1987) Elliptic curve cryptosystems. Mathematics of Computation, 48(177), 203-209.

[9] El-Tantawy, S. A., Salas, A. H., Alharthi, M. R., & Engineering, M. (2021) On the analytical solutions of the forced damping duffing equation in the form of weierstrass elliptic function and its applications. Mathematical Problems in Engineering.

[10] Cohen H. (1998) Efficient Elliptic Curve Exponentiation Using Mixed Coordinates[C]. ASIACRYPTO'98.

[11] Li Fan, Li Yunfeng, Weng Tianheng, Zhang Junjie, (2020) The Rapid Parallel Realization of SM2 Point Operation Based on FPGA [J]. Electronic Measurement Technology,43(15):105-111.

[12] Salam, T., & Hossen, M. S. (2020). HECC (Hyperelliptic Curve Cryptography).

[13] Duanmu QF, Wang YB, Zhang KZ, (2009) GF(3^m)-ECC algorithm and its software implementation[J]. Computer Engineering, (14): 7-9.

[14] Hu, X., Zheng, X., Zhang, S., Cai, S., & Xiong, X. (2018). A low hardware consumption elliptic curve cryptographic architecture over gf(p) in embedded application. Electronics, 7(7), 104-.

[15] Javeed K, Wang X. (2016) FPGA Based High Speed SPA Resistant Elliptic Curve Scalar Multiplier Architecture[J]. International Journal of Reconfigurable Computing, (2016-7-10), 2016, 2016: 2.

[16] Imran, M., Kashif, M., & Rashid, M. (2018). Hardware Design and Implementation of Scalar Multiplication in Elliptic Curve Cryptography (ECC) over GF(2^163) on FPGA.

[17] Anissa, S., Medien, Z., Chiraz, M., & Mohsen, M. (2017). Design and implementation of low area/power elliptic curve digital signature hardware core. Electronics, 6(2), 46-.

AUTHORS

Tan Yongliang, Yunnan University, Main research on IoT security and cryptography

He Lesheng, Yunnan University, Associate Professor, Research on IoT security and embedded systems

Jin Haonan, Yunnan University, Main research on digital signal processing

Kong Qingyang, Yunnan University, China Main research on network security

PROOF OF RENEWABLE (POR) THE ROBE^2 PROTOCOL

Tom Davis[1] and Adel Elmessiry[2]

[1]Renewable Energy Alliance, Lichtenstein
[2]Cryptobloc Tech Universal, Canada

ABSTRACT

We are at a serious crossroads as it relates to carbon emissions and the condition of our planet. Global conditions are spiraling out of control. Climate change is widespread, occurring extremely fast, and intensifying. The consumption of nonrenewable energy sources is impacting both the environment and the economy in equal proportions. Up to this point society has tried to solve these problems with local solutions but we have fallen short. The missing component to solve the global problem is an alignment of individuals and organizations coming together, taking responsibility, and creating global solutions to meet the goal of being carbon negative by 2050.

In this paper, we propose the ROBe^2 protocol as the global solution that brings everyone together to solve these very important issues. Renewable Obligation Base energy economy (ROBe^2) is a protocol attempting to aggregate local renewable energy solutions into a global impact while providing an economically sound framework and allowing the creation of an economic incentive for using renewable energy in place of a fossil one [1].

KEYWORDS

Scientific Research, Researcher, Research, Knowledge, Learning, Applied Research, Decentralized Autonomous, Organization, DAO, Research Model, Research Activity, Blockchain, Emerging Technology, Incentive Design, Reputation Staking, Distributed Ledger Technology, Decentralized Infrastructure, Renewable, Renewable Energy.

1. INTRODUCTION

The current energy model is spiraling our world into an unsustainable future. If the world continues down the path of high fossil fuel emissions, given current knowledge of the consequences, it would be an act of extraordinary intergenerational injustice [2]. Not only does it negatively impact our earth and society, it will also have a devastating effect on our children and their health [3]. The urban poor are particularly vulnerable with overcrowded living conditions and inaccessibility to safe infrastructure, making them highly vulnerable to climate change impacts [4]. Children in developing nations will suffer tremendously as most of the mortality and morbidity rates related to climate change will come from the accessibility of drinkable water, shortage of food and the accelerated spread of vector-borne diseases [5]. In fact, the urgency of the issue has recently been emphasised by European scientists, who warn that action must be taken now to stabilize climate if a catastrophe is to be avoided [6]. For a sustainable future, society has no choice but to come together and use current technological breakthroughs to solve this crisis and create a better future now and for generations to come [7].

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 101-111, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111710

2. Problem Statement

The world's demands on the limited natural resources used to power industrial society are rapidly decreasing as demand for fossil fuels is rising. This accounts for more than 80% of the world's primary energy consumption [8]. In addition, the damage fossil fuels are causing in urban areas is causing hundreds of thousands of cases of premature deaths and respiratory illness [9]. Because of the seriousness of these issues, the Paris Climate Accord has declared a state of emergency and has called for a reduction in all carbon emissions with the goal of being carbon neutral by the year 2050 as well as declaring the objective of keeping the increase in global average temperature to well below 2°C above pre-industrial levels within this century and further to pursue efforts to limit the increase to 1.5°C [10].

In addition, the current IPCC (Intergovernmental Panel on Climate Change) report from August 9, 2021 has revealed new and disturbing data about the state of climate change and the effect of Co2 carbon emissions. Some of the most important empirical evidence includes: (1) It is unequivocal that human influence has warmed the atmosphere, ocean and land. Widespread and rapid changes in the atmosphere, ocean, cryosphere and biosphere have occurred. (2) Many changes due to past and future greenhouse gas emissions are irreversible for centuries to millennia, especially changes in the ocean, ice sheets and global sea level. (3) The scale of recent changes across the climate system as a whole and the present state of many aspects of the climate system are unprecedented over many centuries to many thousands of years, and (4) Human-induced climate change is already affecting many weather and climate extremes in every region across the globe [11].

Evidence of observed changes in extremes such as heatwaves, heavy precipitation, droughts, and tropical cyclones, and, in particular, their attribution to human influence, has strengthened since the Fifth Assessment Report (AR5) [12].

However, there is good news to come out of the report which is why we must act now. The report states that an immediate reduction in carbon emissions would have a swift and positive impact, "From a physical science perspective, limiting human-induced global warming to a specific level requires limiting cumulative CO2 emissions, reaching at least net zero CO2 emissions, along with strong reductions in other greenhouse gas emissions," would create strong, rapid and sustained reductions in CH4 emissions and would also limit the warming effect resulting from declining aerosol pollution and would improve air quality [12]. In other words, acting now makes an immediate difference in helping us return back to normal.

In fact, for the first time in history, all 195 member nations unanimously signed on to the accord and agreed with the findings [12]. These are the most critical issues of our time.

3. Current Solutions

Common solutions to help solve the climate crisis include, but are not limited to, (1) Keeping fossil fuels in the ground, (2) Switch to sustainable transport (3) Proper insulation of homes. (4) Improvisation of farming techniques, (5) Restore nature to absorb more carbon (6) Protect forests, (7) Protect the ocean, (8) Reduce plastic and (9) Invest in renewable energy. These are all important, but alone, they are not enough to make a significant impact.

The most popular form of fighting climate change and reducing carbon emissions is the deployment of the carbon credit system. In fact, carbon credits utilized for carbon load reductions could even be created through a private initiative to constitute a market that will complement regulatory-based initiatives such as national emissions trading systems [13].

Scientific advances could also improve the energy efficiency of existing technologies and develop newer, cheaper carbon-free technologies as discussed in detail by the "McKinsey abatement costs curve" [14] . We label this "the demand theory"— that is, the economy will stop demanding fossil fuels as alternatives become more cost-competitive [15].

But achieving net-zero CO2 emissions will require carbon capture and storage (CCS) to reduce current (Greenhouse Gas) GHG emission rates, and negative emissions technology (NET) to recapture previously emitted greenhouse gases but present NET examples are few, are at a small-scale and not deployable within a decade [16].

One of the greatest contributors to GHG is methane from mass producing agricultural farms around the world. Agriculture is a source for three primary GHGs: CO2, CH4, and N2O. It can also be a sink for CO2 through C sequestration into biomass products and soil organic matter [17]. These are responsible for a significant fraction of anthropogenic emissions, up to 30% according to the Intergovernmental Panel on Climate Change (IPCC) [18]. Without a global commitment to reducing GHG emissions from all sectors, including agriculture, no amount of agricultural adaptation will be sufficient under the destabilized climate of the future [19].

Many of these solutions can seem so overwhelming that smaller companies and individuals feel they aren't able to make much of an impact. While people may have a strong concern about the environment, the complexity of this particular social dilemma—its abstractness, time extendedness, and intergroup nature—tends to discourage actions that help reduce climate change [20]. Therefore, what is needed is a full-circle solution that everyone, and their networks, can be involved in that is making a statistical difference in reducing carbon emissions and creating renewable energies in each and every one of these areas.

4. THE EMERGENCE OF BLOCKCHAIN

Carbon credits, (Renewable Energy Credits) RECs and more advanced systems of those credits that are in development will soon be implemented via blockchain technology. These technologically driven forms of creating renewal makes the 2050 carbon-neutral goal a real possibility.

The requests for sustainable development technologies such as blockchain technology and smart contracts is extremely high. Blockchain and other distributed ledger technologies can enable global partnerships for open innovation that also help meet the goals of the EU Green Deal and the UN's Strategic Development Goals so blockchain is a forefront technology to solve the problem of sustainability and open innovation [21].

In this way, renewable energy sources, fuel cell systems, and other energy generating sources will be optimally combined and connected to the grid system using advanced energy transaction methods [22].

Blockchain technology often plays an essential role in the discussion about a base framework for new energy platform solutions or decentralized business models. Due to its unique ability to transparently document the common state of information within a network, it can provide trust between non-trusting parties in an increasingly granular energy system which has allowed the energy sector to be one of the most rapid adopters of blockchain technology [23].

5. Is Renewable a Valid Financial Option?

The real question is, “will renewable options work to sell back to the grid?” It’s clear that renewable energy will save money on high electricity costs, but will the utility company buy back the power being generated? “Will there be enough interest from utilities to justify the investment in specific renewables like wind and solar energy?

The answer is yes.

Utilities are under increasingly stringent federal and state mandates to develop renewable portfolios as part of the overall supply. In fact, 37 states have mandated renewable or alternative energy standards or goals, and many federal and state organizations are willing to help those working to achieve them” [24].

Therefore, a renewable energy solution connected to the blockchain that could prove ongoing and consistent renewables, would not only be financially beneficial to all parties involved, it would also create a new form of energy credit that would accelerate the UN’s SDG goals and 2050 carbon neutral goal.

6. The Proposed Solution

From our previous work presented in this paper, we clearly show that the biggest obstacle in adopting renewable energy is not the actual cost of the (conversion to) renewable energy but really the profitability of renewable energy usage. This effect requires us to think differently about how the energy is produced and how the energy producers are incentivized. The second point that we need to address in this proposed solution is that the solution should not be in the form of simple subsidization but rather needs to have means of continuing and being self-sufficient. This requires thinking about the system's sustainability over a long period of time. While there is no one solution that fits all, we can think about it in two stages. One would be the bootstrapping stage and the second is steady-state.

These are the factors that led us to suggest Renewable Obligation Base energy economy (ROBe^2). Similar to blockchain-based supply chains [25], the protocol brings forth a global economic incentive to produce renewable energy. The biggest factor in deciding how to produce energy is profitability. If producing energy through renewable sources is less profitable than the conventional methods, market dynamics will push energy producers towards not utilizing renewables [26].

To shift renewable energy production to profitability, the ROBe^2 protocol is going to compensate those committed to producing renewable energy so that producing renewable energy becomes more profitable than conventional ones [27].

7. Renewable Energy System

Standard renewable energy production can be viewed as a simple system that is composed of three stages: production, storage, and consumption [28] as outlined in Figure 1.

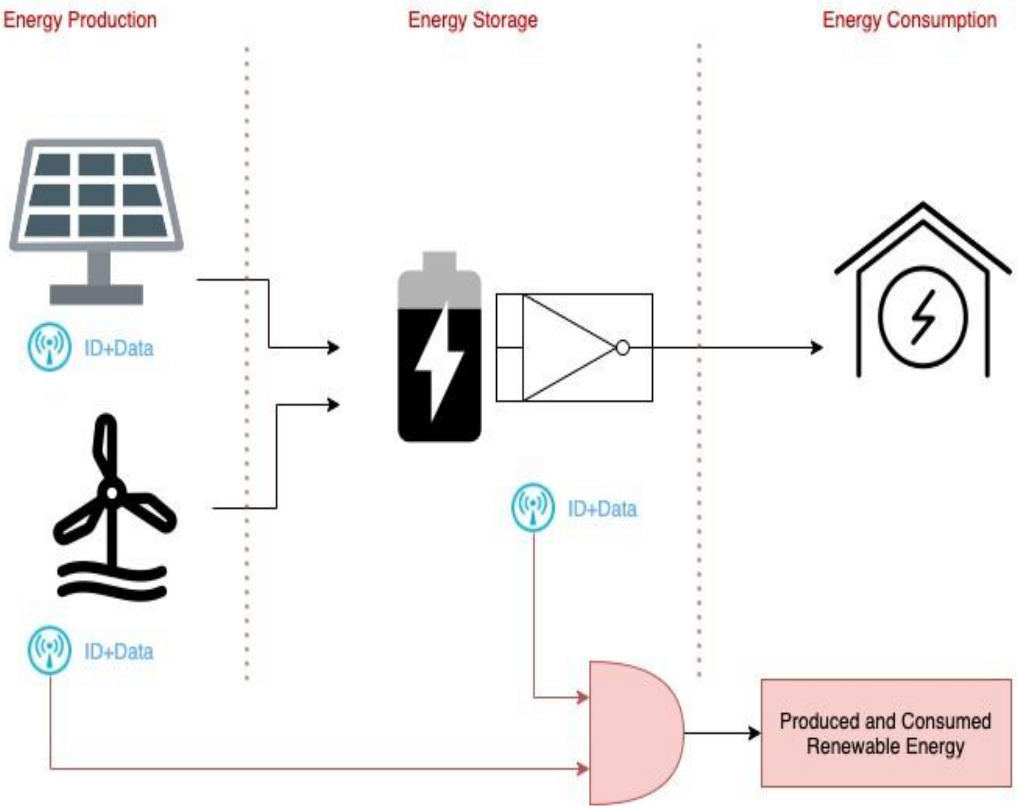


Figure 1. Renewable Energy System Depiction

Renewable Energy Production

There are many sources of renewable energy production, the most widely used are solar and wind production as mentioned earlier. In most cases, those sources utilize a form of charger unit that regulates the produced energy and delivers it to the storage unit [29].

In our proposal, those sources would be outfitted with a unique identifier by the manufacturer that is registered and authenticated. In advanced units, the control unit would also contain a method of communicating the produced energy through APIs to dashboards to measure the actual production. The measured production could be by a single unit or a whole farm of production units. The simplest format of the messages would be in the following format:

Device ID, Power Production. This is a critical part of validating the source of the energy.

Renewable Energy Storage

Since renewable energy is often generated in certain periods when the renewable source is ready to be harvested, that energy may not be ready to be consumed. The energy is then stored to be released at a later time. While many methods are used for energy storage, such as kinetic energy storage, hydraulic energy storage, the most prevailing method of storage is by utilizing a form of electric energy storage. This is normally done by utilizing a battery. Advanced batteries utilize a Battery Management System (BMS) [30]. The BMS’s main functionality is to regulate the battery charging and discharging. In the case of a multi-cell battery, the BMS assesses the health of each cell, the capacity to hold a charge, and control which cells are used in order to optimize the battery bank operations. Modern BMS has a serial number to identify the battery bank and may also have a sub serial number to identify the actual cell. In our proposal, the discharge of the battery is reported upstream using the same format for the APIs to allow measuring of the consumed renewable energy.

Renewable Energy Consumption

Renewable energy is not only consumed locally but also can be sold back to the grid to generate revenue for the producer. The key point is the profit difference between the energy produced from renewable sources and those produced from conventional sources.

For our purposes, we are concerned with the profit deferential rather than the full cost of energy production. The justification here is that producers will naturally gravitate to the production method yielding the highest profit. This is where the key difference is, we aim to incentivize renewable energy production and consumption by providing the financial incentive which turns renewable energy production to an equal or more profitable as compared to the other sources.

Proof of Renewable Mining

ROBe^2 protocol utilizes a novel approach to mining. The current popular mining systems are using Proof of Work (PoW) or Proof of Stake (PoS) [31]. While PoW is energy-intensive, PoS focuses on the ownership of the underlying asset rather than the contributions of the participants.

ROBe^2 adopts a novel approach that rewards the production of renewable energy. The main concept can be boiled down to two main points, namely: Verification of renewable energy, and Incentivization of renewable energy.

Verification of Renewable Energy

This is achieved through the two measuring points in the renewable energy production systems. Since the energy production sources will be reporting the produced energy, that part would verify that the energy is renewably produced. The consumption part is measured by the BMS which would report the discharged power. Both produced and discharged energy is recorded with the manufacturer's authenticated serial number and the data reported.

Incentivization of Renewable Energy

The second part of the protocol is to be able to incentivize renewable energy production and consumption. While most approaches are only focused on allocating a limited supply of incentivization, the problem we are addressing is unlimited. A solution that adopts a dual tokeneconomics would be better suited [32]. To address this problem, two mechanisms are set up to provide the incentivization. The first is protocol-assigned tokens, while the second depends on the protocol operations as a green contribution.

Governance Structure

To be able to autonomously run the operations of the ROBe^2 protocol, A decentralized autonomous organization, DAO, needs to be set up so that it can take charge of the protocol rollouts as well as governance [33]. The DAO will be in charge of how the technical aspects of the protocol are implemented and updated, as well as any new decisions that must be taken to adopt new devices, protocols or tokens that will exist in the future. The DAO membership will start with the developers of the protocol itself and then be expanded to include other contributors based on their participation in the protocol development. This community-based governance ensures that the decisions of the protocol will not be biased by one entity or the other and will be in the best interest of the entire community. As part of the incentivizations a DAO reward wallet will be established for the DAO members [34]. It is proposed that the DAO be a green DAO, so that its operations by definition will agree with the philosophy of the protocol as well.

Renewable Alliance Mining Pool

The mining pool represents the current ROBe^2 tokens available in the pool. The pool is used to distribute the rewards to the participants.

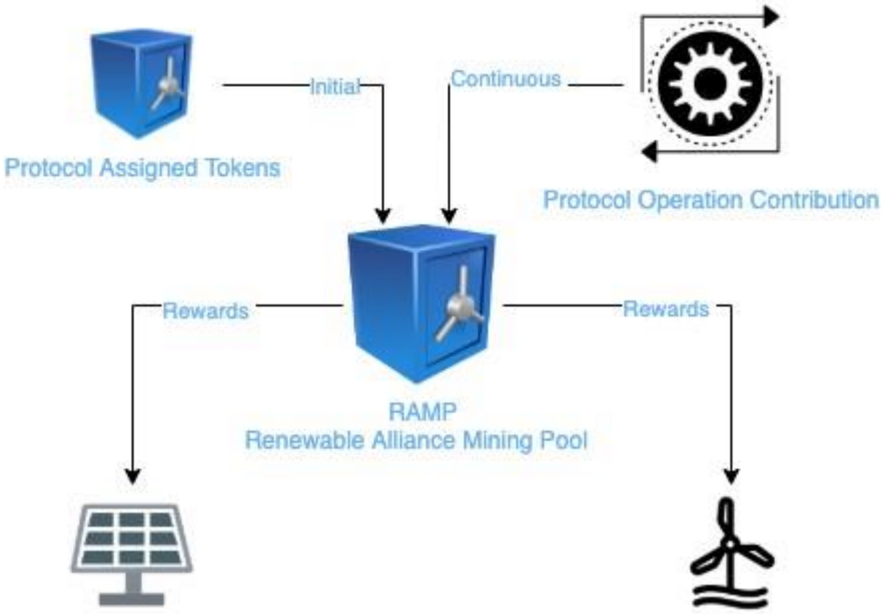

Initial
Continuous
Protocol Assigned Tokens
Protocol Operation Contribution
Rewards
Rewards
RAMP
Renewable Alliance Mining Pool

Figure 2. Renewable Alliance Mining Pool

The RAMP gets the tokens from two main sources:

Protocol Assigned Tokens

Those are the tokens assigned by the original protocol at inception. The tokens are locked in a mining contract and are released over time. Those tokens will eventually be exhausted.

Protocol Operation Contribution

To ensure the protocol's perpetual operations, the ROBe^2 protocol provides a way by which protocol users can contribute to the RAMP. Each time the ROBe^2 token is transacted, a **5%** contribution is deducted from the transaction and distributed as follows:

- **2%** to the GreenDAO development fund
- **3%** to the RAMP

The **3%** will provide the required revenue source for the perpetual rewards pool.

Mining Reward Distribution

The distribution depends on the size of the pool and the number of participants.
Rewards are calculated daily to provide a balance between the pool size and participation.
The equation used is shown below:

$$(EPi) = \frac{DR*REP\left(EP_i\right)*R(EP_i)}{\sum_{i=0}^{i=n} REP\left(EP_i\right)*R(EP_i)} \quad DMR$$

Where:
EP: Energy Producer
EP$_i$: The ith Energy Producer
REP(EP$_i$): The ith Energy Producer daily renewable producedenergy
RAMP: Renewable Alliance Mining Pool *n*: number of energy producers
R(EP$_i$): Reword rate for the ith Energy Producer, set by the renewable alliance
DMR(EP$_i$): Daily Mining Rewards for the ith Energy Producer
DR: daily rewards

ROBe^2 Protocol Algorithm

The algorithm basically is represented as follows:

1. Energy Producer (EP) must register with the Renewable Energy Alliance (REA)
2. Each energy producer is assigned a wallet
3. Each device is registered on the network
4. Each device is approved by the REA
5. The REA assigned the cost differentials for the energy producer in the zone
6. Each energy producer is granted a relative portion of the REA budget
7. The energy producer generates the renewable energy
8. The energy producer contributes it to the grid
9. The REA approved device will sign a transaction indicating the energy generation
10. The transaction is validated
11. Once the transaction is validated, the ROBe^2 reward tokens are sent from the REA wallet to the energy producer wallet

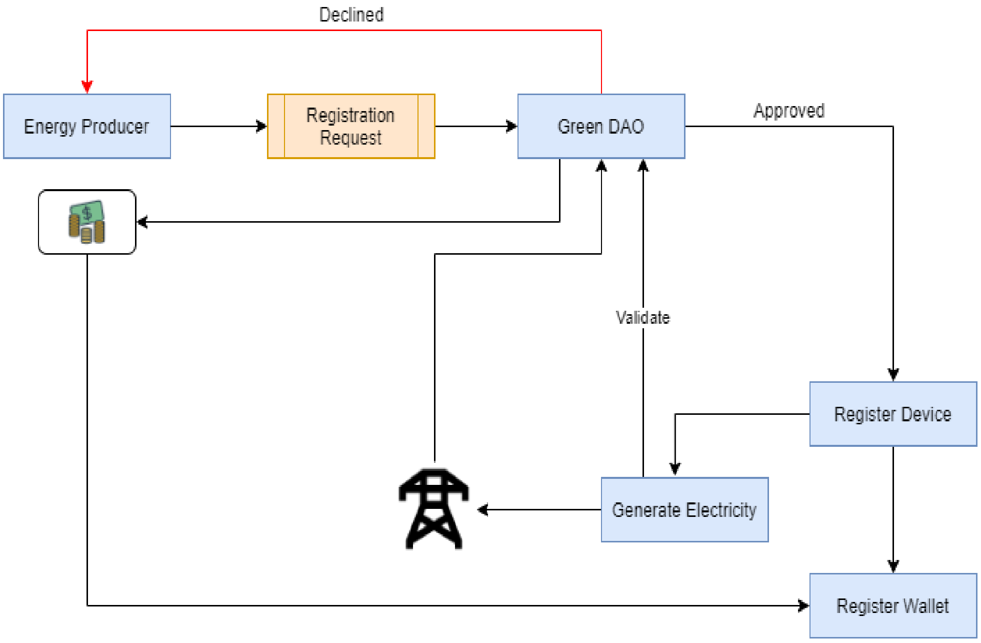


Figure 3. ROBe^2 Protocol Algorithm

CONCLUSION

In this work, we have explored the current status of global world energy production and how the current solutions fall short of addressing them. This is an unsustainable situation that leads to a spiral of consumption and poverty. Our proposed solution is not just an academic exercise, but rather provides a concrete way to measure proof of renewable energy consumption while subsequently utilizing the PoR as the basis of a blockchain-based incentivization protocol. We strongly believe that the outlined framework provides our best chance of changing the global trajectory from one way resource depletion into one of a renewable sustainable planet. We plan to call upon global thought leaders and industrial entrepreneurs to join our cause and adopt the protocol.

ACKNOWLEDGMENT

The Authors would like to express their appreciation to their respective organizations and for the use of IP 63/257,354.

REFERENCES

[1] Sen, Souvik&Ganguly, Sourav, 2017. "Opportunities,barriers and issues with renewable energy development – A discussion," Renewable and SustainableEnergy Reviews, Elsevier, vol. 69(C), pages 1170-1181.

[2] Assessing “Dangerous Climate Change”: Required Reduction of Carbon Emissions to Protect Young People, Future Generations and Nature**.** James Hansen ,PushkerKharecha, Makiko Sato, Valerie Masson-Delmotte, Frank Ackerman, David J. Beerling, Paul J. Hearty, Ove Hoegh-Guldberg, Shi-Ling Hsu, Camille Parmesan, Johan Rockstrom, Eelco J. Rohling, Jeffrey Sachs..

[3] Pollution from Fossil-Fuel Combustion is the Leading Environmental Threat to Global Pediatric Health and Equity: Solutions Exist, Frederica Perera. Columbia Center for Children’s Environmental Health, Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, 722 W. 168th Street, New York, NY 10032, USA*Int. J. Environ. Res. Public Health* 2018, *15*(1), 16; https://doi.org/10.3390/ijerph15010016.

[4] Baker, J. (2011c). Climate change, disaster risk, and the urban poor. The International Bank for Reconstruction and Development/The World Bank, 7-11.

[5] Bernstein, A. & Myers, S. (2011). Climate change and children's health. Center for Health and the Global Environment, Harvard Medical School, 1-6.

[6] Meinshausen, M. (2005). On the Risk to Overshoot 2 °C. Avoiding Dangerous Climate Change Symposium, Exeter, UK, 1-3 February.

[7] Technology Innovation and Climate Change Policy: an overview of issues and options. Grubb, M; (2004) Technology Innovation and Climate Change Policy: an overview of issues and options. Keio Economic Studies, 41 (2) pp. 103-132.

[8] S.H.Mohra, J Wang, G. Ellem, J Ward, D. Giurco. "**Projection of world fossil fuels by country.**" Fuel. Volume 141, 1 February 2015, Pages 120-135.

[9] "Lvovsky, Kseniya; Hughes, Gordon; Maddison, David; Ostro, Bart; Pearce, David. 2000. Environmental Costs of Fossil Fuels : A Rapid Assessment Method with Application to Six Cities. Environment Department papers;no. 78. Pollution Management series.. World Bank, Washington, DC. © World Bank, p. Xi.

[10] "Decarbonization in Complex Energy Systems: A Study on the Feasibility of Carbon Neutrality for Switzerland in 2050", Frontier in Energy Research, Xiang Li1*, Theodoros Damartzis1, Zoe Stadler, Stefano Moret, Boris Meier, Markus Friedl and François Maréchal, 16 November 2020 | https://doi.org/10.3389/fenrg.2020.549615.

[11] IPCC: (2021). Intergovernmental Panel on Climate Change: AR6 Reports.

[12] IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

[13] Matthew, John A. "How carbon credits could drive the emergence of renewable energies." *Energy Policy*, 2008, vol. 36, issue 10, 3633-3639 (2008).

[14] Frank Ackerman and Ramón Bueno. Use of McKinsey abatement cost curves for climate economics modeling. SEI-U.S. Working Paper WP-US-1102. January 25, 2011.

[15] Thomas Covert, Michael Greenstone, Christopher R. Knittel. "Will We Ever Stop Using Fossil Fuels?" MIT Center for Energy and Environmental Research Policy. February 2016.

[16] Negative emissions technologies and carbon capture and storage to achieve the Paris Agreement commitments. The Royal Society. R. Stuart Haszeldine, Stephanie Flude, Gareth Johnsonand Vivian Scott. April 2, 2018. https://royalsocietypublishing.org/doi/full/10.1098/rsta.2016.0447.

[17] Sharon Lachnicht Weyers, Donald C. Reicosky. "Agricultural opportunities to mitigate greenhouse gas emissions." Environmental Pollution. Volume 150, Issue1, November 2007, Pages 107-124.

[18] Francesco N Tubiello1, Mirella Salvatore1, Simone Rossi1,2, Alessandro Ferrara1, Nuala Fitton3 and Pete Smith, The FAOSTAT database of greenhouse gas emissions from agriculture. Environmental Research Letters, Volume 8, Number 1. 12 February 2013.

[19] Beddington, J., et al. (2012). Achieving food security in the face of climate change: Final report from the Commission on Sustainable Agriculture and Climate Change. CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). Copenhagen, Denmark.

[20] Climate Change: What Psychology Can Offer in Terms of Insights and Solutions. Current Directions inPsychological Science.Paul A. M. Van Lange, Jeff Joireman, Manfred Milinski. First Published July 12, 2018. https://journals.sagepub.com/doi/full/10.1177/0963721417753945..

[21] Fraga-Lamas, Paula and Fernandez-Carames, "Leveraging Blockchain for Sustainability and Open Innovation; Cyber-Resiliant Approach Toward EU Green Deal and UN Sustainable Development Goals. InTech Open Book Series. May 7, 2020.

[22] Blockchain of Carbon Trading for UN Sustainable Development Goals. Seong-Kyu Kim 1 andJun-HoHuh2, *Sustainability* 2020, *12*(10), 4021. 14 May 2020. https://doi.org/10.3390/su12104021.

[23] Zeiselmair, Andreas, et al. "Analysis and Application of Verifiable Computation Techniques in Blockchain Systems for the Energy Sector." *Frontiers in Blockchain*, vol. 4, 2021.

[24] Loftis, B. (2010). Untapped Resource: Converting Mine Sites into Renewable Energy Assets: Engineering, Geology, Mineralogy, Metallurgy, Chemistry, etc. *Engineering and Mining Journal, 211*(3), 50-55.

[25] ElMessiry, M., &ElMessiry, A. (2018, June). Blockchain framework for textile supply chain management. In International Conference on Blockchain (pp. 213-227). Springer, Cham.

[26] Leijon, M., Bernhoff, H., Berg, M., &Ågren, O. (2003). Economical considerations of renewable electric energy production—especially development of wave energy. Renewable Energy, 28(8), 1201-1209.

[27] Yang, D. X., Jing, Y. Q., Wang, C., Nie, P. Y., & Sun, P. (2021). Analysis of renewable energy subsidy in China under uncertainty: Feed-in tariff vs. renewable portfolio standard. Energy Strategy Reviews, 34, 100628.

[28] Heussen, K., Koch, S., Ulbig, A., &Andersson, G. (2011). Unified system-level modeling of intermittent renewable energy sources and energy storage for power system operation. IEEE Systems Journal, 6(1), 140-151.

[30] Cheng, K. W. E., Divakar, B. P., Wu, H., Ding, K., &Ho, H. F. (2010). Battery-management system (BMS) and SOC development for electrical vehicles. IEEE transactions on vehicular technology, 60(1), 76-88.

[31] Sriman, B., Kumar, S. G., &Shamili, P. (2021). Blockchain technology: Consensus protocol proof of work and proof of stake. In Intelligent Computing and Applications (pp. 395-406). Springer, Singapore.

[32] ElMessiry, M., ElMessiry, A., &ElMessiry, M. (2019, June). Dual token blockchain economy framework. In International Conference on Blockchain (pp. 157-170). Springer, Cham.

[33] Kaal, W. A. (2021). A Decentralized Autonomous Organization (DAO) of DAOs. Available at SSRN 3799320.

[34] Hassan, S., & De Filippi, P. (2021). Decentralized Autonomous Organization. Internet Policy Review, 10(2), 1-10.

AUTHORS

Dr. Tom Davis

Tom Davis is a visionary entrepreneur, CEO, and Leadership expert to Fortune 500 companies and non-profit organizations around the globe. He has worked in 40 countries and trained leaders from 80 nations. He is the Chief Renewable Officer of the Renewable Energy Alliance whose mission is to strategically bring people and organizations together to create a 100% renewable energy future to regenerate the world. Tom has worked with major companies across the globe including Ingersoll-Rand, Lear, Celsa Group, Marvell Technology, Ficosa, Grifols, Vueling Airlines, Intel, Chick-Fil-A, Coca-Cola, Pfizer, Bristol-Myers Squibb, and GlaxoSmithKline. He is focused on creating a renewable economic platform that can be accessed by anyone in the world.

Adel Elmessiry, Ph.D.

Tech entrepreneur, published expert on AI and Blockchain with 20+ years of Healthcare, Mentor, Advisors & Speaker. Adel is a serial entrepreneur with three successful technology companies taken from inception to acquisition. He has proven executive experience with a solid track record that includes over 10 years at HealthStream and 7 years at InVivoLink/ HealthTrust. Academically, he is holding a Ph.D. in Computer Science at NCSU Natural Language Processing. He serves as the president chief technology officer for AlphaFin, a Draper Goren Holm portfolio company. Adel is the founder of Crypto Bloc Tech, a cutting-edge blockchain technology company. Together we are on a mission to build the next financial technology ecosystem that will empower the global economy.

AN INTELLIGENT SYSTEM TO AUTOMATE HUMIDITY MONITORING AND HUMIDIFIER CONTROL USING INTERNET-OF-THINGS (IOT) AND ARTIFICIAL INTELLIGENCE

Qian Zhang[1] and Yu Sun[2]

[1]Jserra High School, San Juan Capistrano, CA 92675
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

Air conditioners are widely used in family homes all over the world. However, the side effects of using air conditioners and dehumidification can cause health problems if people remain in low-humidity environments. This paper traces the development of a software application and system to create an intelligent humidifier that automatically turns on or off for convenience or for those who cannot engage manual control. We applied our application to a humidifier for several days and conducted a qualitative evaluation of the approach. Results affirmed the usability and capacity of our automatic control system.

KEYWORDS

IoT, Machine Learning, Deep Learning, Artificial Intelligence.

1. INTRODUCTION

Since the first modern air conditioner was invented in 1902 by Willis Haviland Carrier, air conditioning has grown to become an indispensable part of people's lives. [1] It adjusts the room temperature, especially during scorching summers. However, problems can come with prolonged use of air conditioning, and a drop in indoor humidity levels is one of them. Although the air conditioner is not a dehumidifier, dehumidification is part of the cooling process since moisture is drawn out of the room together with heat. [2] This moisture collects on the evaporator coil, and is drained out of the home via the condensate pan and drain line. Heating a room during the winter has a similar effect since as the room is heated, the relative humidity drops.

People don't usually feel comfortable in dry air. Low humidity in a room can even cause health problems, such as frequent sore throats, chapped lips, and bloody noses. This is because as moisture evaporates, we lose a vital layer of protection that is effective at filtering bacteria. Also, when our noses and lips are split and irritated, the capillaries are more exposed, making it easier for microbes to pass directly into our bloodstream. In addition, some studies suggest that viruses such as influenza thrive in environments with low humidity. [3] Besides health problems, low humidity can also cause issues like splitting wooden floorboards or furniture.

To keep room humidity at comfortable levels, a humidifier can help. However, it is impossible to keep our attention on the internal environment all the time to decide whether to turn a humidifier

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 113-121, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111711

on or off, especially during work or sleep. Our appliance will solve this problem by evaluating the indoor environment and automatically turning the humidifier on or off.

In this paper, we trace the current humidifier market to find the best way to maintain the optimal humidity that can decrease when the AC system is cooling or heating a room. Our goal is to create a product that automatically monitors the indoor environment and works to maintain optimal humidity levels. Our method is inspired by Google Nest, which maintains a user's desired temperature by controlling the AC system using machine learning. There are some advantages of our product over a humidifier with machine learning. First, our system can predict the indoor environment, including humidity, by monitoring existing weather data and learning users' preferences. Second, it can automatically turn the humidifier on or off to maintain a certain humidity level, allowing users to remain comfortable while focusing on their work or sleeping. Third, our device can easily be applied to an existing humidifier in a user's home. There is also a mobile app to help users control our device. Therefore, our system has some advantages over the existing products.

We used support vector learning (SVM) and regression models to build the basic program needed for our device. [4] Through adjusting the regression model, polynomial features, and input data sets, we tested the accuracy of various machine learning models via the mean square error method to determine the most appropriate model to use.

The rest of this paper is organized as follows: Section 2 introduces the details of the challenges we encountered during the design and development of the sample; Section 3 focuses on our solutions in response to the challenges mentioned in Section 2; Section 4 presents relevant details on how we conducted the experiment, followed by related works in Section5. Finally, Section 6 provides concluding remarks, as well as the possible future applications of our device.

2. CHALLENGES

In order to develop a software application and system to create an intelligent humidifier that automatically turns on or off for convenience or for those who cannot engage manual control, a few challenges were identified as follows.

2.1. Challenge 1: AC systems reduce humidity when both heating and cooling

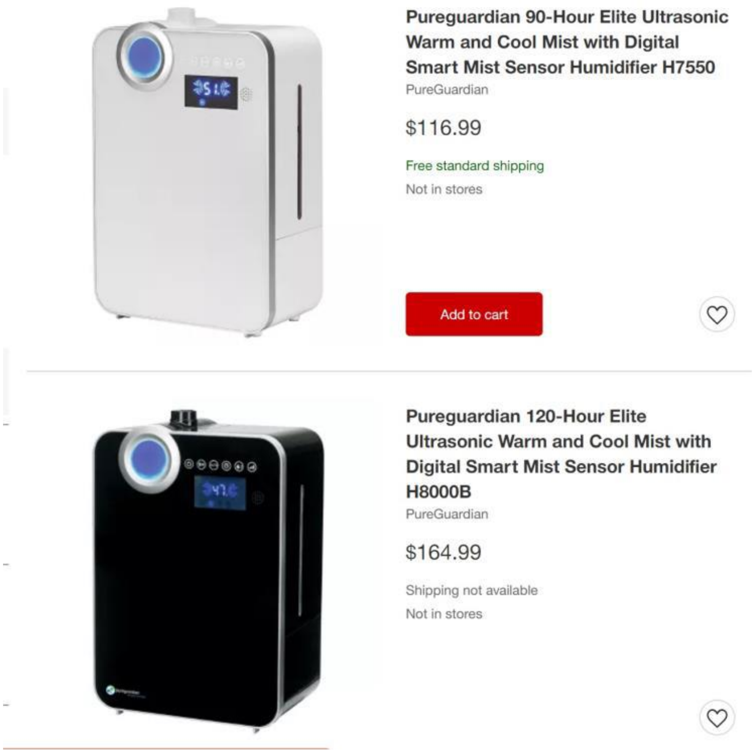
Air conditioners are widely used in people's homes as a necessary piece of equipment to help regulate the indoor environment. However, the only element most AC systems control is the temperature. Actually, other aspects such as the humidity should be considered to modify the indoor environment to meet optimal comfort levels. AC also plays the role of a dehumidifier since it removes water vapor in the air. [5] When air hits the cold evaporator coil inside the air handler, the AC makes the humidity condense on the oil and drain into the pipe that then exits outdoors. This also applies to the AC's heating system as well. Heating dries the air, causing the relative humidity to drop. For example, in the northern areas of China, people usually feel dry and uncomfortable when they get up in the morning during winter because the heat must stay on at night. They cannot turn it off because the heating pipes were buried underground by the government during urban planning. In short, existing devices that help modify indoor environments can have drawbacks, and are often unable to increase the humidity in the room.

2.2. Challenge 2: Health concerns of using humidifiers

Obviously, we can use humidifiers to increase humidity levels indoors. However, this is not a perfect solution. Popular portable humidifiers have drawbacks. First, people have to monitor them manually depending on their humidity preferences. While focusing on work or sleeping, they cannot always do this consistently, which means they may to keep the humidifier on or off and suffer an uncomfortable environment. The second disadvantage is that using a humidifier all the time wastes water and electric power. What is worse, overuse of humidifiers can also lead to health problems. If left on all the time, high humidity and misty conditions can have a negative effect on the lungs. [6] Also, humidifiers have the potential to release minerals and microorganisms into the air. These microorganisms might not be directly harmful, but can greatly affect people with asthma. Unclean humidifiers facilitate the growth of certain bacteria, which is associated with coughing and the common cold. Therefore, humidifiers require regular cleaning to reduce bacterial growth.

3. SOLUTION

To resolve the problem of having to turn a humidifier on or off or overuse it, some may choose to buy advanced humidifiers with automatic control systems. [7] While these options work, they can be expensive. For example, digital smart mist sensor humidifiers sell for up to $100 in the online store of Target™ (see Figure 1). These models usually only sense humidity and cannot adjust other indoor elements such as temperature or allergy management.



Pureguardian 90-Hour Elite Ultrasonic Warm and Cool Mist with Digital Smart Mist Sensor Humidifier H7550

PureGuardian

$116.99

Free standard shipping

Not in stores

Pureguardian 120-Hour Elite Ultrasonic Warm and Cool Mist with Digital Smart Mist Sensor Humidifier H8000B

PureGuardian

$164.99

Shipping not available

Not in stores

Figure 1. Examples of advanced humidifiers with automatic control systems

The name of our application is HC, which stands for Humidity Conditioner. The application was implemented using Thunkable. We used machine learning to make regular humidifiers into smart devices that serve as real-time humidity monitoring systems to offer automatic and remote humidity control. The application has two main pages: the login page and a page that features a humidity indicator, real-time humidity and temperature data, and a dynamic background that provides the weather and time (see Figure 3).

The HC application has a login page that allows users to sign in using multiple devices to control the humidifier. The home page displays real-time data in its and allows users to see the current humidity levels in percentages to manage their humidifier using an indicator/slider as a controller. There is also a button that allows users to choose from three modes: on/off/auto. On or off modes allow users to control the humidifier manually, while the auto mode uses machine learning to regulate the humidity according to users' specifications and preferences.

We used the raspberry pi 3 as the processor for our device. [9] The raspberry pi was selected over a microprocessor single-board microcontroller, such as an Arduino uno, for its Wi-Fi connectivity and storage capabilities to store data from the temperature-humidity sensor. [8]

The sensor used in our device is a DHT22 digital temperature-humidity sensor. [10] The DHT22 was chosen for several reasons including its compatibility with the raspberry pi, ease of use, low cost, and range and accuracy of readings. It can read humidity ranges from 0 to 100% with 2-5% accuracy and temperature ranges from -40 to 80°C with ±0.5°C accuracy with a sampling rate of 0.5 Hz.

The final component of our device is a power strip with a built-in power relay that provides power to both the raspberry pi and any humidifier. [15] The built-in power relay allows the raspberry pi to turn the outlet the humidifier is connected to on or off. The power strip has three outlet modes: always on, usually off, and usually on. The "off" is usually off by default unless a current is sent to the power relay. Similarly, the "usually on" mode is on by default unless a current is sent to the power relay.

The device requires that the raspberry pi be plugged into the power strip in the always on outlet and the humidifier in the usually off outlet, then a two-wire, signal and ground wire from the raspberry pi's gpio pins is connected to the power relay. Once powered, the raspberry pi reads the humidity levels from the DHT22 sensor. When the humidity level falls below the target threshold, the raspberry pi sends a signal to the power relay, turning on the humidifier plugged into the "usually off" outlet. Also, when the humidity level goes over the threshold, the raspberry pi stops the signal, turning off the humidifier.

In the HC application, we used machine learning for prediction and classification. [11] We used prediction at the bottom half of the home page where the predicted humidity levels are displayed (see Figure 2). Machine learning allows the system to predict patterns of humidity by studying existing data as well as users' manual control history. When users turn the humidifier on or off, the system uses these preferences to optimize its predictions.

In addition, the auto mode takes advantage of machine learning classification. The algorithm divides current and predicted humidity data to two categories, then decides whether to turn the humidifier on or off.

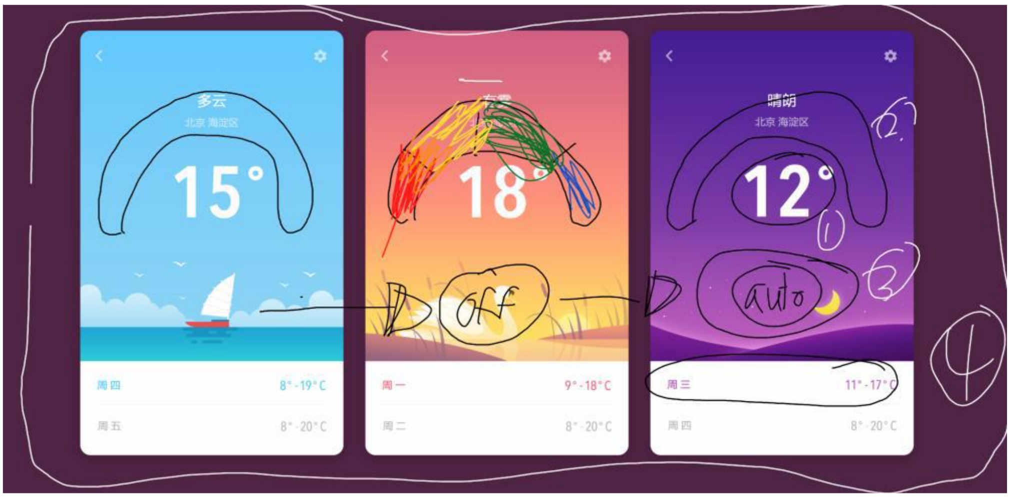


Figure 2. Humidity Conditioner UX

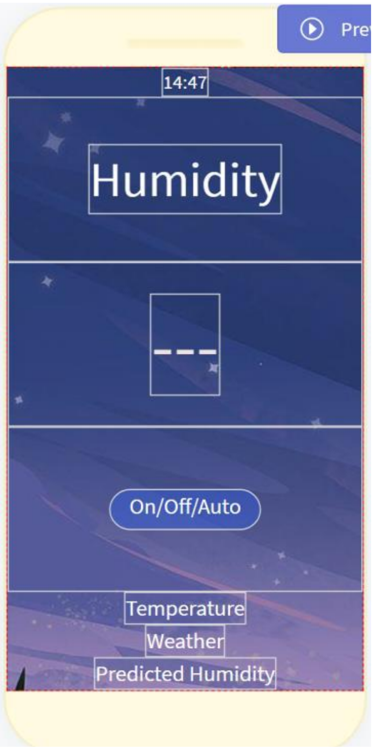
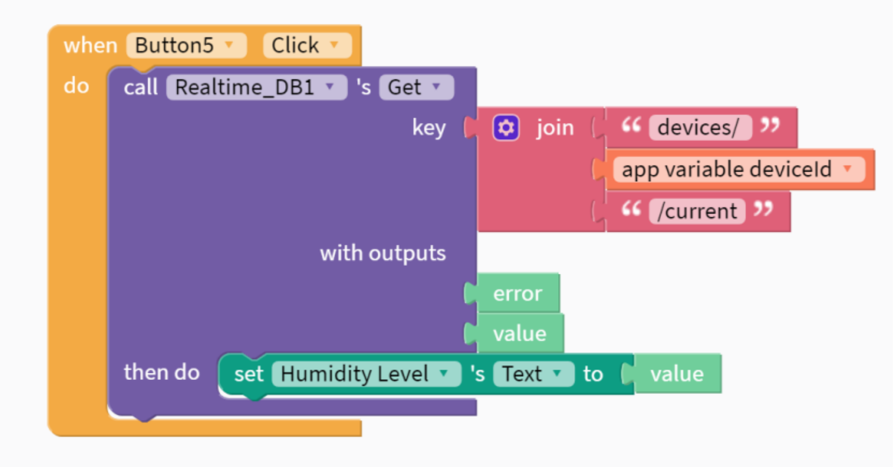


Figure 3. Display data including weather and time

For the design of the home page, we broke the page down into four rows (see Figure 4). In the first row, we added a “Humidity” title. In the second, real-time humidity levels are displayed. For the third row, we added a button that allows users to choose from three modes—on/off/auto. The fourth row displays other indoor-environment data, including the current temperature, weather, and predicted humidity levels. There is also a time-input that displays the present time at the top of the homepage.



```
when Button5 Click
do  call Realtime_DB1 's Get
                     key  join " devices/ "
                               app variable deviceId
                               " /current "
            with outputs
                     error
                     value
    then do  set Humidity Level 's Text to  value
```

Figure 4. Humidity Conditioner's homepage display rows

The block code of our home page operates as follows: users click on the button and the system retrieves their account and calls on the real-time database (firebase) for an output of either a percentage humidity level in or an error. The system then translates the real-time humidity to text and displays it in the second row of the homepage.

4. EXPERIMENT

We performed experiments to verify that our predictive models worked, and to ensure that we were selecting the optimal model, data, and parameters to produce the most accurate and satisfactory experience for users.

We conducted two experiments to evaluate the accuracy of two machine learning models—prediction and classification—by changing polynomial degrees and data set features. [12] The first experiment aimed to find a machine learning model that predicts indoor humidity based on other environmental elements such as temperature and wind speed. This experiment compared and evaluated the accuracy of three models created with sixteen data sets using polynomial degrees of 1, 2, and 3, respectively, to determine which was the most optimal one. The second experiment was conducted with the intake of sixteen datasets using different set features to find the best-fit machine learning model to determine whether or not to turn the humidifier on or off depending on environmental elements such as temperature and wind speed. We performed these experiments not only to verify that our predictive models work, but also to ensure that we were selecting the optimal model, data, and parameters to produce the most accurate and satisfactory experience for users.

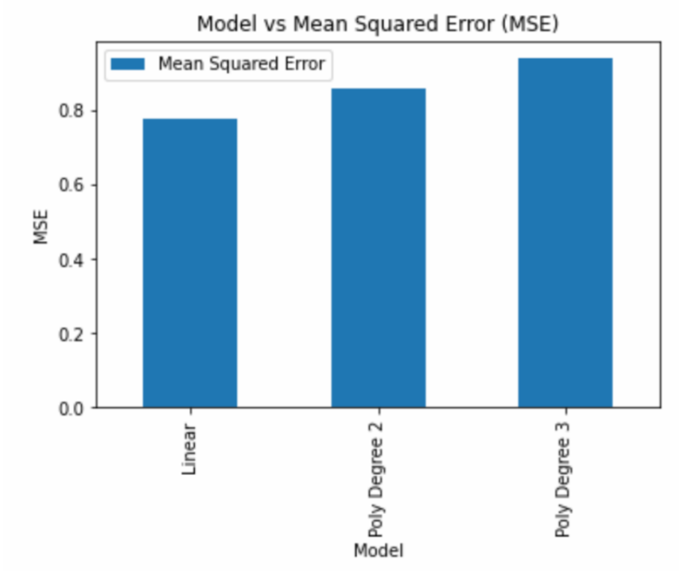


Figure 5. Experiment 1 data showing the best fit prediction model (polynomial degree of 3)

In experiment 1, we applied different polynomial parameters to the same data set to find out which model would produce the most accurate algorithm. The result of experiment 1 shows that, along with polynomial degrees of 1, 2, and 3, the accuracy of each is 0.787, 0.869, and 0.942 respectively. It revealed that the best fit prediction model is the one with a polynomial degree of 3 (see Figure 5).

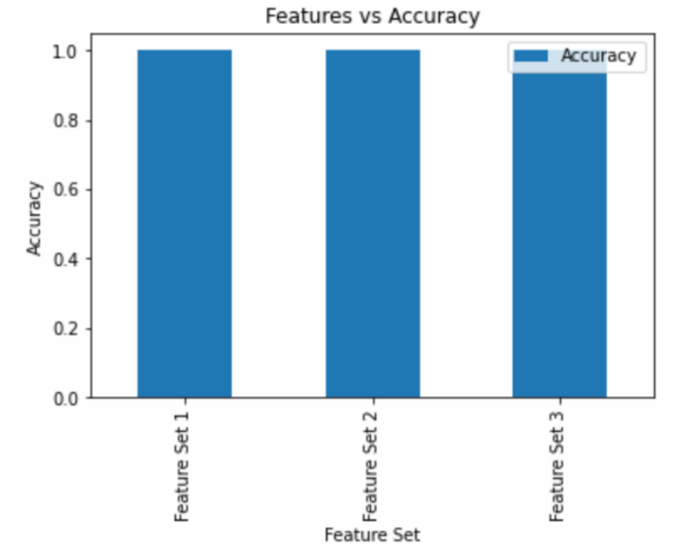


Figure 6. Experiment 2 data showing the accuracy of the three models were the same (1.0)

For experiment 2, we compared the SVM machine learning models that differ by their data sets to determine the best one for classification. We tried the three features of humidity, temperature, and air quality in varied combinations: the first was temperature and wind speed, the second was humidity and wind speed, and the third was temperature, humidity and wind speed. It turned out that the accuracy of these three models were all the same, which was 1.0 (see Figure 6).

5. RELATED WORK

Biqing, Li, et al. created an intelligent air humidifier. [13] Their design adopts STC89C52RC SCM control and connects via the auxiliary circuit to achieve automatic testing and sound-light alarm and to turn on the humidifier. They first set a desired level of humidity, D_0, then test the current indoor humidity. If the current value does not meet the required one, the humidifier turns on automatically. Our product is different in that we use machine learning, which allows us to predict the indoor environment based on temperature and humidity data instead of having to test the indoor environment.

Baughman, A., et al. studied the health risks brought on by high humidity. [6] They found that "[t]he primary influences of humidity on health are through biological pollutants." Most health issues are caused by pollutants such as fungi as well as viruses such as Streptococcus, Legionella, and the common cold and flu. In Section 2: Challenges, we review some of the negative effects of high humidity and misty conditions that can be brought about by keeping humidifiers on for too long. The releasing of minerals and microorganisms by humidifiers can also affect people with asthma. Therefore, machine learning is helpful in controlling humidifiers to reduce these associated health issues.

Ku, K. L., et al. addresses the creation of an automatic control system for thermal comfort within an entire building. [14] Their model utilizes "an adaptive neuro fuzzy inference system and a particle swarm algorithm" to create a nonlinear multivariable inverse PMV model so as to determine thermal comfort temperatures. Similarly, our design also makes use of machine learning. However, two main points make our design different from Ku's. The first is that our design aims to control the indoor humidity at optimal levels based on existing data including temperature. The second is that our model is made for one-room environments instead of entire buildings. Our product is therefore best for maintaining the moisture requirements of a single room.

6. CONCLUSION AND FUTURE WORK

Our application, HC, applies machine learning to help automatically control a humidifier to maintain optimal levels of indoor humidity. Two experiments were conducted to ensure the accuracy of two machine learning models—prediction and classification. In the first experiment, we determined the best prediction model by importing existing weather data and testing with different polynomial degrees. Our conclusion was that degree 3 is the best fit for the model. The second experiment was conducted with the intake of datasets with different set features to find the best model to classify whether to turn the humidifier on or off. The three SVM machine learning models performed at the same high accuracy level. As a result, our intelligent humidifier control application allows users to focus on their work without the necessity of manually maintaining or monitoring humidity levels.

Our application still has limitations, however. It currently does not learn preferences to improve its prediction and classification over time. In the future, we can optimize our application further by importing code that allows it to study users' preferences as they manually control the humidifier over time. Also, our design can only be applied to an existing humidifier, which may be inconvenient for some who don't already have one. To remedy this in future, we may design a more integrative model that includes both a basic humidifier and intelligent control system.

REFERENCES

[1] Yu, B. F., et al. "Review of research on air-conditioning systems and indoor air quality control for human health." *International journal of refrigeration* 32.1 (2009): 3-20.

[2] Yong, Li, et al. "Experimental study on a hybrid desiccant dehumidification and air conditioning system." (2006): 77-82.

[3] Wolkoff, Peder. "Indoor air humidity, air quality, and health–An overview." *International journal of hygiene and environmental health* 221.3 (2018): 376-390.

[4] Soentpiet, Rosanna. *Advances in kernel methods: support vector learning*. MIT press, 1999.

[5] Cai, Dehua, et al. "Performance analysis of a novel heat pump type air conditioner coupled with a liquid dehumidification/humidification cycle." *Energy Conversion and Management* 148 (2017): 1291-1305.

[6] Baughman, A., and Edward A. Arens. "Indoor humidity and human health--Part I: Literature review of health effects of humidity-influenced indoor pollutants." *ASHRAE transactions* 102 (1996): 192-211.

[7] Golnaraghi, Farid, and Benjamin C. Kuo. *Automatic control systems*. McGraw-Hill Education, 2017.

[8] Aloisio, Alberto, et al. "uSOP: A microprocessor-based service-oriented platform for control and monitoring." *IEEE Transactions on Nuclear Science* 64.6 (2017): 1185-1190.

[9] Pi, Raspberry. "Raspberry pi 3 model b." *online].(https://www. raspberrypi. org* (2015).

[10] Sikarwar, S., and B. C. Yadav. "Opto-electronic humidity sensor: A review." *Sensors and Actuators A: Physical* 233 (2015): 54-70.

[11] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349.6245 (2015): 255-260.

[12] Maulud, Dastan, and Adnan M. Abdulazeez. "A Review on Linear Regression Comprehensive in Machine Learning." *Journal of Applied Science and Technology Trends* 1.4 (2020): 140-147.

[13] Li, Biqing, et al. "Design of the intelligent air humidifier." (2016): 110-112.

[14] Ku, K. L., et al. "Automatic control system for thermal comfort based on predicted mean vote and energy saving." *IEEE Transactions on Automation Science and Engineering* 12.1 (2014): 378-383.

[15] Melo, Thallyson PS, Felipe M. Avila, and L. F. Avila. "Power Strip Automation with Internet of Things."

USING DIFFERENT ASSESSMENT INDICATORS IN SUPPORTING ONLINE LEARNING

Yew Kee Wong

School of Information Engineering, HuangHuai University,
Zhumadian, Henan, China

ABSTRACT

The assessment outcome for many online learning methods are based on the number of correct answers and than convert it into one final mark or grade. We discovered that when using online learning, we can extract more detail information from the learning process and these information are useful for the assessor to plan an effective and efficient learning model for the learner. Statistical analysis is an important part of an assessment when performing the online learning outcome. The assessment indicators include the difficulty level of the question, time spend in answering and the variation in choosing answer. In this paper we will present the findings of these assessment indicators and how it can improve the way the learner being assessed when using online learning system. We developed a statistical analysis algorithm which can assess the online learning outcomes more effectively using quantifiable measurements. A number of examples of using this statistical analysis algorithm arepresented.

KEYWORDS

Artificial Intelligence, Assessment Indicator, Online Learning, Statistical Analysis Algorithm.

1. INTRODUCTION

Many online learning assessment systems that use multiple choice approach is based on the correct answers to judge a learner on their understanding of what they have learned [1]. We have carry out various experiments on these quantifiable measurements (assessment indicators) on Mathematics and English subject on different learner groups. With these assessment indicators, assessors and learners can easily assess the online learning performances [2].

We carry out experiments from 100 online voluntary learners from 8 different countries: USA, Canada, United Kingdom, Malaysia, Singapore, Philippines, Thailand and Indonesia. The experiments involved learners from the age between 9 and 10. The experiment questions we have used is the UK National Mathematics Curriculum, Year 5, Geometry topic, multiple choice questions. All 100 online voluntary learners, speak and write fluent English and have knowledge in this Geometry topic.

2. ASSESSMENT INDICATORS AND STATISTICAL ANALYSIS

In this research, we have identified 3 critical assessment indicators which can influence the learners' learning progress and understanding. These 3 assessment indicators have inter-relationships with the underlying final mark from the assessment [3]. At the end of the assessment, the artificial intelligence engine will analyse all the statistical information from these 3 indicators and provide a recommendation for the assessor and learner.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 123-128, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111712

(1). **Difficulty Level** (measure by the complexity of the questions [4]).

Each question will have the difficulty level embedded. For example, we take a topic in Addition from Mathematics subject. For an Addition topic, we can assign the Difficulty level to these 3 questions depending on the complexity, i.e. 4 + 3 = ? (Low Difficulty), 755 + 958 = ? (Medium Difficulty) and 7,431,398,214 + 32,883,295 = ? (High Difficulty).

Level Terms	Quantitative Measurement
High (hardest)	*3*
Medium	*2*
Low (easiest)	*1*

(2). **Understanding Level** (measure by the time from the question appear to submission).

Each question will have the understanding level embedded. For example, assuming there is a question with Level Term - High (from a to b) where “a = 3 seconds” and “b = 5 seconds”. In this example, if the learner can submit the answer between 3 to 5 seconds after the question appeared than the answer will be assigned 2 points for the Understanding indicator.

Level Terms	Quantitative Measurement
High (from a to b) fastest	*2*
Medium (from b to c)	*1*
Low (from c onwards) slowest	*0*

(3). **Confident Level** (variation in choosing an answer before submission).

For each question, we will capture the behaviour of the learner when choosing an answer before submission [5]. For example, for most learners if they are confident and prudent on choosing the correct answer, they will submit the answer once decided without making any changes. If the learner didn’t make any changes when answer this example question, than this answer will be assigned 2 points for the Confident indicator.

Level Terms	Quantitative Measurement
High (no change on first pick)	*2*
Medium (one change)	*1*
Low (two changes or more)	*0*

In this research, we have carry out multiple experiments to evaluate the use of assessment indicators and the statistical information generated when the learner performing the assessment [6][7]. We have conducted 3 detail experiments and the outcomes generated show promising result on the learners’ overall learning performance. Below are the 3 experiments summary which we have conducted on 100 online voluntary learners. All marks and indicators have been converted to percentage (%) prior for further analysis by the artificial intelligence engine.

2.1. Experiments

Experiment 1:

20 Multiple Choice Questions (Difficulty level : 1)

Type of questions : Year 5 - UK National Mathematics Curriculum (Topic: Geometry)Standard marking scheme : Lowest (0 / 20) and Highest (20 / 20) {marks} Understanding level : Lowest (0 / 40) and Highest (40 / 40) {20 questions x 2 points}

Confident level : Lowest (0 / 40) and Highest (40 / 40) {20 questions x 2 points}Number of learners involved in this experiment : 100

Summary Results :

- 10% of the learners who cannot move on to Experiment 2 in the 1st attempt,majority of their Understanding level and/or Confident level are under 50%. And the learners are required to have a 2nd attempt.
- In the 2nd attempt, all the remaining 10% learners have successfully move on to Experiment 2 and both indicators show a major improvement and all indicators above 50%.
- 90% of the learners who achieve moving to Experiment 2 in the 1st attempt, have shown promising result of over 50% in both Understanding level and Confident level.

Experiment 2:

20 Multiple Choice Questions (Difficulty level : 2)
Type of questions : Year 5 - UK National Mathematics Curriculum (Topic: Geometry)Standard marking scheme : Lowest (0 / 20) and Highest (20 / 20) {marks} Understanding level : Lowest (0 / 40) and Highest (40 / 40) {20 questions x 2 points}

Confidence level : Lowest (0 / 40) and Highest (40 / 40) {20 questions x 2 points}Number of learners involved in this experiment : 100

Summary Results :

- 23% of the learners who cannot move on to Experiment 3 in the 1st attempt, majority of their Understanding level and/or Confident level are under 50%. And the learners are required to have a 2nd attempt.
- In the 2nd attempt, 17% out of 23% learners have successfully move on to Experiment 3 and both indicators show a major improvement and all indicators above 50%.
- In the 3rd attempt, all the remaining 5% out of 23% learners have successfully move on to Experiment 3 and both indicators show a major improvement and all indicators above 50%.
- 77% of the learners who achieve moving to Experiment 3 in the 1st attempt, have shown promising result of over 50% in both Understanding level and Confident level.

Experiment 3:

20 Multiple Choice Questions (Difficulty level : 3)

Type of questions : Year 5 - UK National Mathematics Curriculum (Topic: Geometry)Standard marking scheme : Lowest (0 / 20) and Highest (20 / 20) {marks} Understanding level : Lowest (0 / 40) and Highest (40 / 40) {20 questions x 2 points}
Confidence level : Lowest (0 / 40) and Highest (40 / 40) {20 questions x 2 points}Number of learners involved in this experiment : 100

Summary Results :

- 56% of the learners who cannot complete the Experiment 3 in the 1st attempt, majority of their Understanding level and/or Confident level are under 50%. And the learners are required to have a 2nd attempt.

- In the 2nd attempt, 19% out of 56% learners have successfully completed the Experiment 3 and both indicators show a major improvement and all indicators above 75%.
- In the 3rd attempt, all the remaining 15% out of 37% (56% - 19%) learners have successfully completed the Experiment 3 and both indicators show a major improvement and all indicators above 75%.
- In the 4th attempt, all the remaining 22% (37% - 15%) learners have successfully completed the Experiment 3 and both indicators show a major improvement and all indicators above 75%.
- 44% of the learners who completed the Experiment 3 in the 1st attempt, have shown promising result of over 75% in both Understanding level and Confident level.

Both the assessment indicators and statistical information stand alone do not have any representations and it is meaningless without others being analysed altogether [8]. Furthermore,in order to have an effective and efficient online learning outcome for the learner, the assessor requires to design and develop the curriculum, learning materials and Q&A using a hybrid integrated model [9]. The curriculum needs to be an all rounded learning blueprint, where learner can improve their understanding in a progressive manner and user-friendly approach.

3. ARTIFICIAL INTELLIGENCE RULES

The artificial intelligence rules define the way the online learning system assigned learning materials and exercises for the learner to follow. These are the basic rules which we have carry out in our experiments, in which we find it effective in improving the learners understanding.

Rule number	Difficulty level	Correct answers(%)	Understanding level (%)	Confident level (%)	Recommendation (Response)
1	1	< 50	Nil	Nil	Repeat the same difficulty level = 1 exercise
2	1	≥ 50	< 50	< 50	Repeat the same difficulty level = 1 exercise
3	1	≥ 50	< 50	≥ 50	Repeat the same difficulty level = 1 exercise
4	1	≥ 50	≥ 50	< 50	Repeat the same difficulty level = 1 exercise
5	1	≥ 50	≥ 50	≥ 50	Move to next difficulty level = 2 exercise
6	2	< 50	Nil	Nil	Repeat the same difficulty level = 2 exercise
7	2	≥ 50	< 50	< 50	Repeat the same difficulty level = 2 exercise
8	2	≥ 50	< 50	≥ 50	Repeat the same difficulty level = 2 exercise
9	2	≥ 50	≥ 50	< 50	Repeat the same difficulty level = 2 exercise
10	2	≥ 50	≥ 50	≥ 50	Move to next difficulty level = 3 exercise
11	3	< 75	Nil	Nil	Repeat the same difficulty level = 3 exercise
12	3	≥ 75	< 50	< 50	Repeat the same difficulty level = 3 exercise
13	3	≥ 75	< 50	≥ 50	Repeat the same difficulty level = 3 exercise
14	3	≥ 75	≥ 50	< 50	Repeat the same difficulty level = 3 exercise
15	3	≥ 75	≥ 50	≥ 50	Move to next topic exercise

Figure 1. The artificial intelligence rules applied in the experiments.

Online learning assessor and learner can modify all the assessment indicators accordingly (depending on various conditions and overall standard requirements) [10].

4. Online Learning using Artificial Intelligence System

The online learning using artificial intelligence system includes several components which can be integrated as one complete artificial intelligence online learning system [11]. These are the standard components:-

1. Reasoning − It is the set of processes that empowers us to provide basis for judgement, making decisions, and prediction.
2. Learning − It is the activity of gaining information or skill by studying, practising, being educated, or experiencing something. Learning improves the awareness of the subjects of the study.
3. Problem Solving − It is the procedure in which one perceives and tries to arrive at a desired solution from a current situation by taking some path, which is obstructed by known or unknown hurdles.
4. Perception − It is the way of acquiring, interpreting, selecting, and organizing sensory information.
5. Linguistic Intelligence − It is one's ability to use, comprehend, talk, and compose the verbal and written language. It is significant in interpersonal communication.

The potential of online learning system include 4 factors of accessibility, flexibility, interactivity, and collaboration of online learning afforded by the technology. In terms of the challenges to online learning, 6 are identified: defining online learning; proposing a new legacy of epistemology-social constructivism for all; quality assurance and standards; commitment versus innovation; copyright and intellectual property; and personal learning in social constructivism [12].

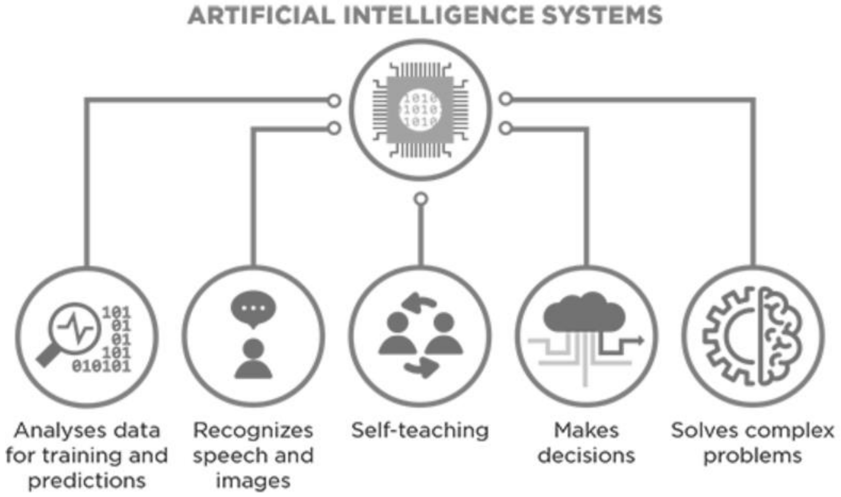

ARTIFICIAL INTELLIGENCE SYSTEMS
1010
1010
1010
101
01
01
101
010101
Analyses data for training and predictions
Recognizes speech and images
Self-teaching
Makes decisions
Solves complex problems

Figure 2. The artificial intelligence online learning system components.

5. Conclusions

This paper proposed artificial intelligence online learning involves 3 stages. Stage One involves design and development of the curriculum with learning materials, Q&A and other assessment indicators. Stage Two involves the implementation of a creative artificial intelligence rules. And Stage Three involves the user-friendly learning process and analysis operation. This model can generates flexibility when designing and developing the online learning system. The new statistical analysis algorithm with various assessment indicators show promising results in artificial intelligence online learning and further evaluation and research is in progress.

REFERENCES

[1] Bill Joy, (2000). Why the Future Doesn't Need Us, WIRED-IDEAS, USA.

[2] William Li, (2018). Prediction Distortion in Monte Carlo Tree Search and an Improved Algorithm, Journal of Intelligent Learning Systems and Applications, Vol. 10, No. 2.

[3] Ofra Walter, Vered Shenaar-Golan and Zeevik Greenberg, (2015). Effect of Short-Term Intervention Program on Academic Self-Efficacy in Higher Education, Psychology, Vol. 6, No. 10.

[4] Calum Chace, (2019). Artificial Intelligence and the Two Singularities, Chapman & Hall/CRC.

[5] Piero Mella, (2017). Intelligence and Stupidity – The Educational Power of Cipolla's Test and ofthe "Social Wheel", Creative Education, Vol. 8, No. 15.

[6] Zhongzhi Shi, (2019). Cognitive Machine Learning, International Journal of Intelligence Science, Vol. 9, No. 4.

[7] Crescenzio Gallo and Vito Capozzi, (2019). Feature Selection with Non Linear PCA: A Neural Network Approach, Journal of Applied Mathematics and Physics, Vol. 7, No. 10.

[8] Charles Kivunja, (2015). Creative Engagement of Digital Learners with Gardner's Bodily-Kinesthetic Intelligence to Enhance Their Critical Thinking, Creative Education, Vol. 6, No. 6.

[9] Evangelia Foutsitzi, Georgia Papantoniou, Evangelia Karagiannopoulou, Harilaos Zaragas and Despina Moraitou, (2019). The Factor Structure of the Tacit Knowledge Inventory for High School Teachers in a Greek Context.

[10] Nick Bostrom and Eliezer Yudkowsky, (2011). The Ethics of Artificial Intelligence, Cambridge University Press.

[11] Gus Bekdash, (2019). Using Human History, Psychology, and Biology to Make AI Safe for Humans, Chapman & Hall/CRC.

[12] The Student Circles.com, Artificial Intelligence Study Notes https://www.thestudentcircle.com/quickguide.php?url=artificial-intelligence

AUTHOR

Prof. Yew Kee Wong (Eric) is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, Hewlett- Packard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.

AN INTELLIGENT DATA-DRIVEN ANALYTICS SYSTEM TO ASSIST SPORTS PLAYER TRAINING AND IMPROVEMENT USING INTERNET-OF-THINGS (IOT) AND BIG DATA ANALYSIS

Julius Wu[1], Jerry Wang[2], Jonathan Sahagun[3] and Yu Sun[4]

[1]Irvine High School, Irvine, USA
[2]SMIC Private School, Shanghai, China
[3] California State University, Los Angeles, CA, 91706
[4]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

Our product is a very unique tracking tool that not only tracks the movement of players on a map, but also the velocity of each player. We have an application that coaches usually hold onto during a game or a practice. It shows coachesan accurate data sample of where each player is and what they are doing on the field whether it be grinding or fooling around. It also helps coaches see accurate gameplay during a game if the recording is not available. When coaches select elite players, they also get a presentation of each players' skills and how accurate they are when running different routes.

KEYWORDS

IoT, Machine learning, Data Mining.

1. INTRODUCTION

The application we develop is to track football players' data like numbers of steps, average velocity when training and use this data to provide summary and recommendation for football players. We expect this program will be loved by many American football players around the world as this program provides useful tips for them [1]. Compared to other traditional ways of training, our program has several benefits. First, the cost of training using this application is relatively low. We only need a kit that has an accelerator, GPS, and cellular network to track your data and report it back to your application. Second, football players can get access to detailed information through your application that can't be found in the traditional way. On the application, football players can not only see their average velocity and steps they run during matches, but also the trajectory of matches and the graph relating velocity and time. From these data, the application can automatically give a summary of your performance and possibly a suggestion for your training. Combined with detailed information, football players can have customized training, save time, and maximize their performance in big matches [2]. This application also has low barriers and it's universal to many football players, whether professional or not.

Existing method and tool use mobile phones to track the steps and locations that we need for the football tracker. This method seems good and low cost, while there's several problems within it.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 129-143, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111713

The first problem is the accuracy. Mobile phones usually have dozens of meters' variation to the actual location. This inaccuracy will cause inaccurate positioning and has a very negative impact on sportsmen's training. People will not know the close speed they run during the game and can hardly improve their training effect. The second problem is that we need to keep the mobile screen on in order to gain accurate locations. Many smartphones have a power-saving mechanism that makes applications stop tracking if the screens go off [3]. As soccer players play their sports, they don't care about their phones so their screens will inevitably go out. When the screen goes off, the application will stop tracking users' location, steps, etc. Thus, when the game is over, the athlete can't get an accurate result for their games.

Our goal in this research paper is to present their speed, location, and trajectory plot so they can use them to improve their training effect. Our method is inspired by the increasing iot networks around the world. These iot networks can connect items like bicycles together. After you ride that bicycle, the mobile phone will automatically show some basic information like speed, distance you traveled, and the map which shows you trajectory. These features are provided by dedicated logic boards. We believe that this dedicated logic board has advantages in several aspects. First, it can track athletes' information all the time without disconnecting like a mobile phone. Football games usually last about one and half hours, perhaps more. By using our tool kit to track location and speed information, athletes will not worry about being unable to collect information because it's always connected to the cloud through cellular network and sends you speed and location [4]. In this way, athletes can focus only on the game instead of checking their phones to assure that it's connected to the cloud. The second benefit is that this kit provides more accurate information thana mobile phone. Compared to mobile phones, the kit provides us more accurate location information. Typically, the error of locations of mobile phones can be up to dozens of meters. On the other hand, this kit provides you the error of location as small as several meters. The smaller the error, the more accurate the location is. If we have a more accurate location, we can track the athletes locations more precisely thus can present a more accurate trajectory plot. The kit can also count steps more accurately since it carries professional sensors [5]. The more accurate steps count can provide a more accurate distance and average speed values. With all this more accurate information, athletes can use this information to improve their strategy to play football games.

In two application scenarios, we demonstrate how the above combination of techniques increases the accuracy of tracking players' speed and location [6]. First, we show the usefulness of our approach: a case study in which one player carries the device and crosses the line by 50 times. The device will detect whether the player crosses the line or not. The player will run at a variety of speeds to make the result we test more authentic. In the second experiment, we will do the same set of tests on 5 different players. These players will carry our devices and cross the line. They will also run at a variety of speeds. Each player will cross the line for 10 times. The purpose is to test whether different players will affect the accuracy of our device to track speed and location.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. How to integrate the hardware with software at real-time

The first challenge we encountered is to learn how the programming software works. As we only had a little foundation about programming, we need to learn how the android studio works [7]. The first problem is basically installing the software. As it is a development tool, the android studio offers us with a variety of customization options. We need to figure out what plug-in we need for this software. During this process, the software runs errors so we need to spend time using Windows command processor (CMD) to solve the problem in order to let it run [8]. The second problem is learning the programming language. As we know, different programming software relies on different languages. The language we rely on is Dart. In order to program our app, we need to learn the Dart language. We need to learn proper terms, arrangements, and logic. Studying this language takes us time since we not only need to know these terms but also need to operate and try them to make them familiar and make the program work.

2.2. How to debug and update the system program remotely

When we write our program, we will inevitably encounter problems. In these months we wrote the applications, this happened many times. I remembered that when we wrote the login function, we couldn't log in despite finding nothing wrong with our program written to support this function and the firebase which authenticates the username and the password. This problem puzzled us for quite a long time as we inspected the code android studio and information in the firebase again and again. After checking it over and over again, we find out that it’s actually the problem of connecting firebase to android studio. Although codes in android studio and the user information in firebase had setted up seperately, it’s not connected with one another. It means that when we input username and password in android studio, it can’t check with the firebase, thus we can’t login it. To solve this problem, we inserted a firebase sdk to let the cloud connect with the program so that the problem of login and registering is solved. Following that, we can build more features like map, tracking information, etc.

2.3. How to collect and persist information at large-scale

In the beginning, we thought something that we could just do was to zoom in on the maps enough to scope in on a field and have our data collected there, so we did not think much further than that. Later while we were testing, we could see the tracker collect the data on a larger scale but as soon as we zoomed in on the map, everything got blurred out. We later realized that due to some restrictions, zooming in too much on a map would blur out everything rendering the field unclear. With that in the way, we could not see our players on the map clearly and we couldn’t tell where on the field they would be at. We solved this problem by projecting an imaginary field onto the map by plugging in the coordinates of 2 opposite corners of the football field onto the map. With the field correctly positioned on the map, we could now graph out where players would be at.

3. SOLUTION

The football tracking program shows trajectory map, duration, distance, average speed, and start and end time. The program is designed to visualize the performance of athletes in games by using data. By analyzing the data combined with the trajectory map, athletes can find out their problemsand make customized plans for their training in order to perform better in the next game or next season [9].

The football tracking application works with three parts: the application, the cloud(firebase), and the lot kit. The lot kit basically collects critical information we need for the application. The information includes distance, duration time, locations tracked by GPS sensor. These locations are then sent to the firebase from the device using cellular network. Thisinformation will be sent to the corresponding account. If we open the application and log in the corresponding account, we can see data we collected. By clicking different dates, we can see information we collected like distance, duration time, and locations. By using this information, the application created a trajectory map that let you see the trajectory during the game.

This system is implemented with the following components:

1) Mobile app development
2) Server/Firebase
3) Sensors and hardware

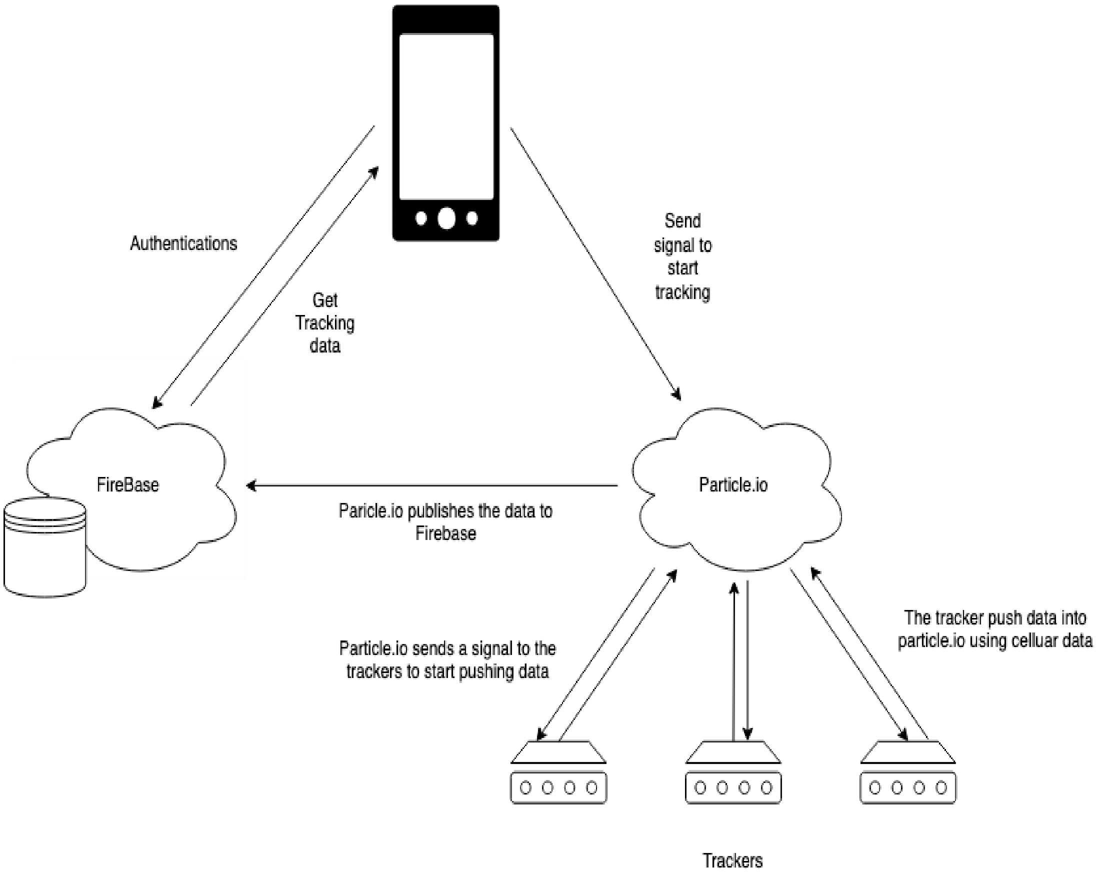

Authentications
Get Tracking data
Send signal to start tracking
FireBase
Paricle.io publishes the data to Firebase
Particle.io
Particle.io sends a signal to the trackers to start pushing data
The tracker push data into particle.io using celluar data
Trackers

Figure 1. The overview of the solution

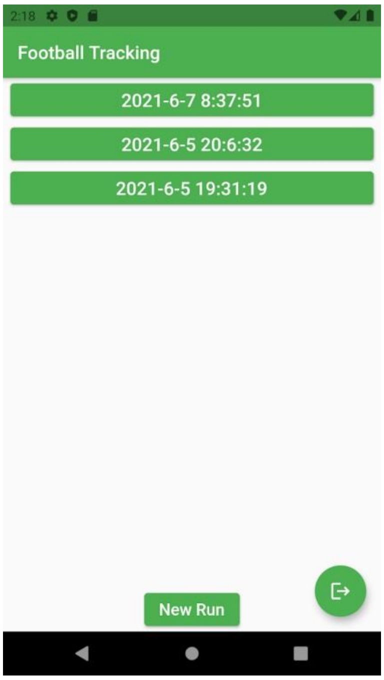
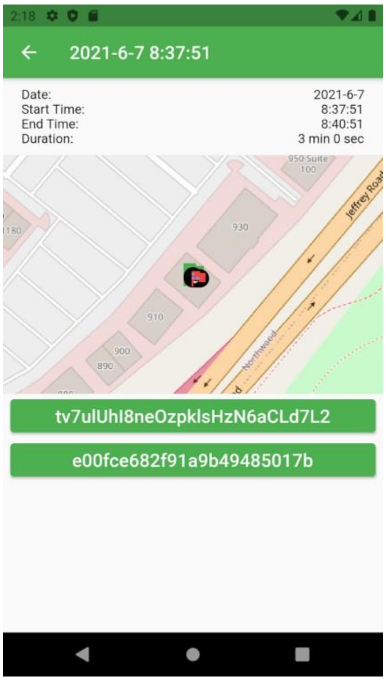
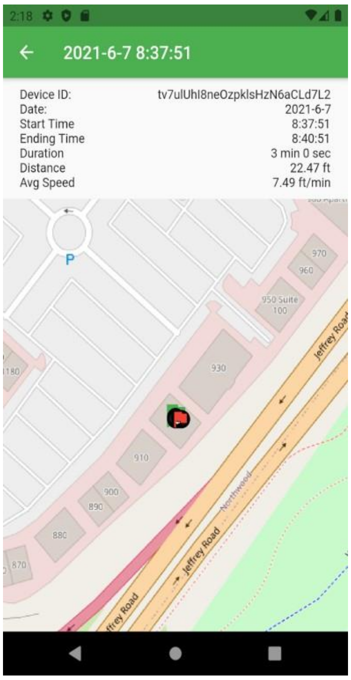


Figure 2. List of dates the tracker was running



Date:	2021-6-7
Start Time:	8:37:51
End Time:	8:40:51
Duration:	3 min 0 sec

Figure 3. Example of one the dates



2:18

2021-6-7 8:37:51

Device ID:	tv7ulUhl8neOzpklsHzN6aCLd7L2
Date:	2021-6-7
Start Time	8:37:51
Ending Time	8:40:51
Duration	3 min 0 sec
Distance	22.47 ft
Avg Speed	7.49 ft/min

Figure 4. Example of one of the trackers details

This function sends a message to the particle.io to tell the device to start publishing

```
static Future<Response> turnOn(int durationMinutes) async{
_baseURL =
"https://api.particle.io/v1/devices/${deviceID}/${turnOnFunctionName}";
print(_baseURL);

Response response = await http.post(Uri.encodeFull(_baseURL),
body: {
"arg": durationMinutes.toString()
},
headers: {
"Content-Type": "application/x-www-form-urlencoded","Accept":
"application/json",
"Authorization": "Bearer $token"
}
);
return response;
}
```

This function sends a message to the particle.io to tell the device to stop publishing.

```
static Future<Response> turnoff() async{
_baseURL =
"https://api.particle.io/v1/devices/${deviceID}/${turnOffFunctionName}";
print(_baseURL);
Response response = await http.post(Uri.encodeFull(_baseURL), headers: {
"Content-Type": "application/x-www-form-urlencoded","Accept":
"application/json",
"Authorization": "Bearer $token"
}
);
return response;
}
```

The next three functions calculate the speed of a player/device by its GPS coordinates and its timestamps.

```
// calculated in meters per hour
double calculateSpeed(){
double distance = calculateTotalDistance();
int seconds = gpsCoordinates.last.timestamp - starttimestamp;double hours = 1.0
* seconds / 60.0 / 60.0;


return distance/hours;
}

// calculated in meters
double calculateTotalDistance(){double totalGPSDistance = 0;
if (gpsCoordinates.length >= 2){
for(int i = 0; i < gpsCoordinates.length-1; i++){
var temp = calculateDistance(gpsCoordinates[i], gpsCoordinates[i+1]);
totalGPSDistance +=  temp;
}
}


return totalGPSDistance;
}
```

```
// Returns answer in meters
double calculateDistance(GPS coord1, GPS coord2){
// Radius of the Earth in meters
double r = 6371000.0;


double phi_1 = coord1.latitude * pi / 180;double phi_2 = coord2.latitude * pi /
180;


double delta_phi = (coord2.latitude - coord1.latitude) * pi / 180; double
delta_lambda = (coord2.longitude - coord1.longitude) * pi / 180;


 double a = pow(sin(delta_phi / 2.0), 2) + cos(phi_1) * cos(phi_2) *
pow(sin(delta_lambda / 2.0), 2);
double c = 2 * atan2(sqrt(a), sqrt(1-a));


return r * c ;
}
```

We used a framework called Flutter and it helped us create the inner design of the app and the tracking equipment. It helped us simulate the app on a phone so that we could just make an app without publishing it and using it on the phone.

We later brought in the use of a firebase as well. Firebase helps us keep all the data we use that we get from our hardware. The hardware comes from a company called Particle and it helps us send the data from the tracker with a cellular signal.

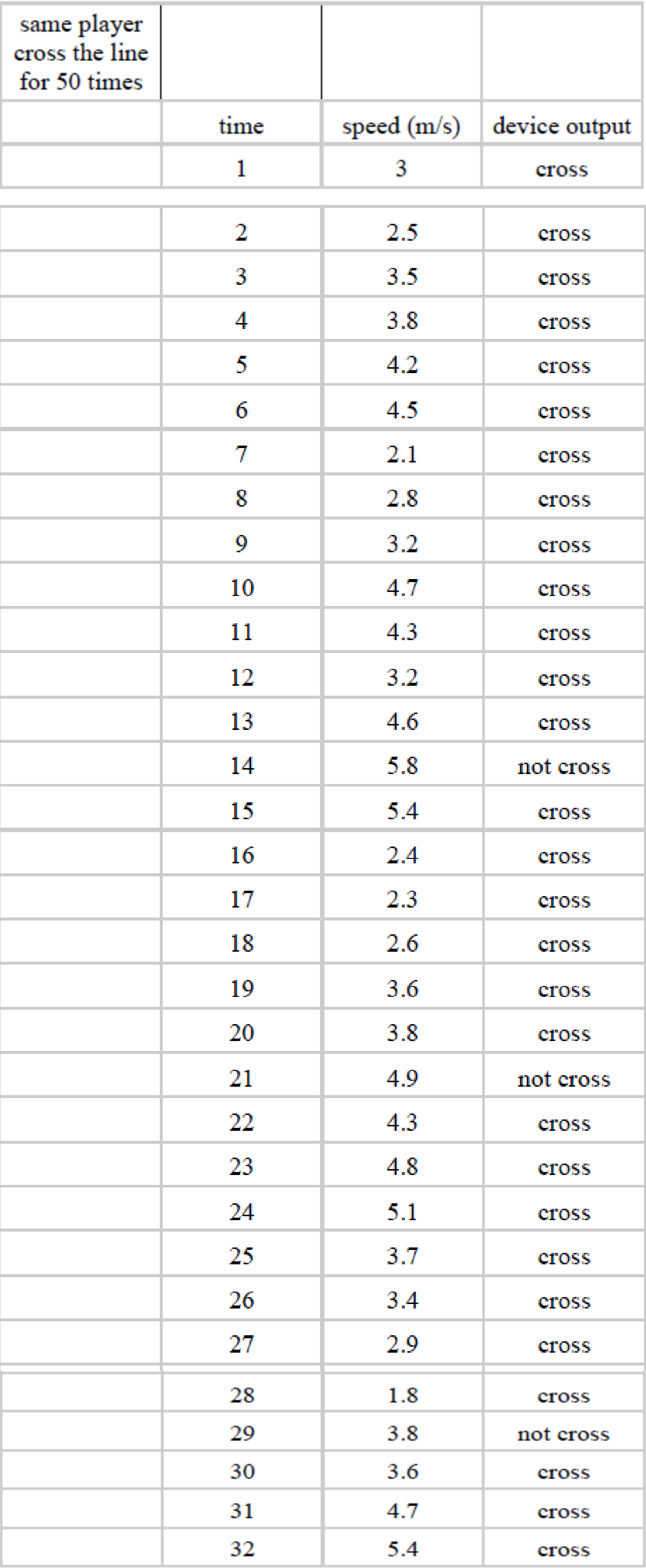
4. EXPERIMENT

Two experiments have been conducted in order to measure the accuracy and scalability.

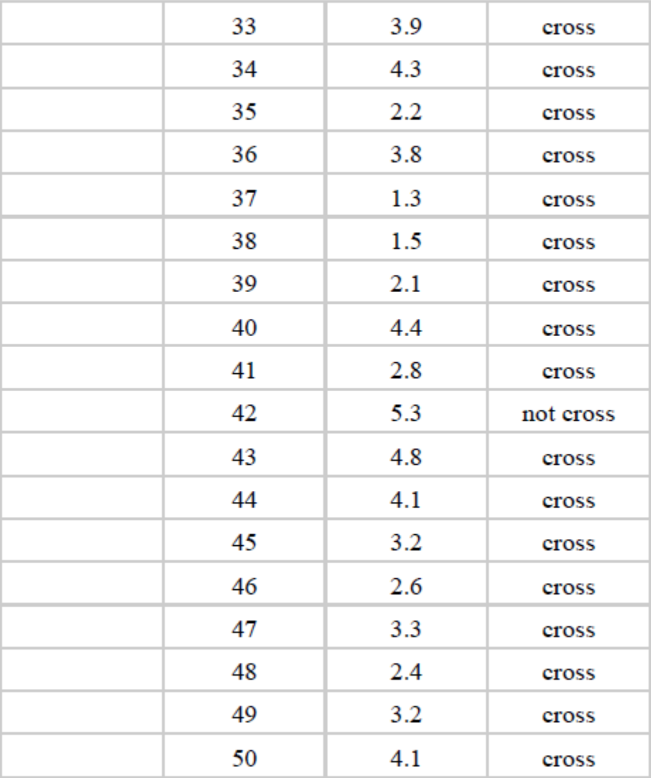
4.1. Experiment 1. The accuracy of detecting if the player crosses the line or not

The experiment is conducted as follows: one player carries the device and crosses the line 50 times. We will use whether the device detects whether the player crosses the line or not. We use the speed algorithms to help devices determine whether a player crosses the line or not.

Table 1. same player cross the line for 50 times



same player cross the line for 50 times			
	time	speed (m/s)	device output
	1	3	cross
	2	2.5	cross
	3	3.5	cross
	4	3.8	cross
	5	4.2	cross
	6	4.5	cross
	7	2.1	cross
	8	2.8	cross
	9	3.2	cross
	10	4.7	cross
	11	4.3	cross
	12	3.2	cross
	13	4.6	cross
	14	5.8	not cross
	15	5.4	cross
	16	2.4	cross
	17	2.3	cross
	18	2.6	cross
	19	3.6	cross
	20	3.8	cross
	21	4.9	not cross
	22	4.3	cross
	23	4.8	cross
	24	5.1	cross
	25	3.7	cross
	26	3.4	cross
	27	2.9	cross
	28	1.8	cross
	29	3.8	not cross
	30	3.6	cross
	31	4.7	cross
	32	5.4	cross



	33	3.9	cross
	34	4.3	cross
	35	2.2	cross
	36	3.8	cross
	37	1.3	cross
	38	1.5	cross
	39	2.1	cross
	40	4.4	cross
	41	2.8	cross
	42	5.3	not cross
	43	4.8	cross
	44	4.1	cross
	45	3.2	cross
	46	2.6	cross
	47	3.3	cross
	48	2.4	cross
	49	3.2	cross
	50	4.1	cross

Summary: from this data, we can see that the device generally captures whether the player passes the line or not with an accuracy of 92%. However, sometimes it does not capture the data because the player’s speed is probably high or the device has margin of error.

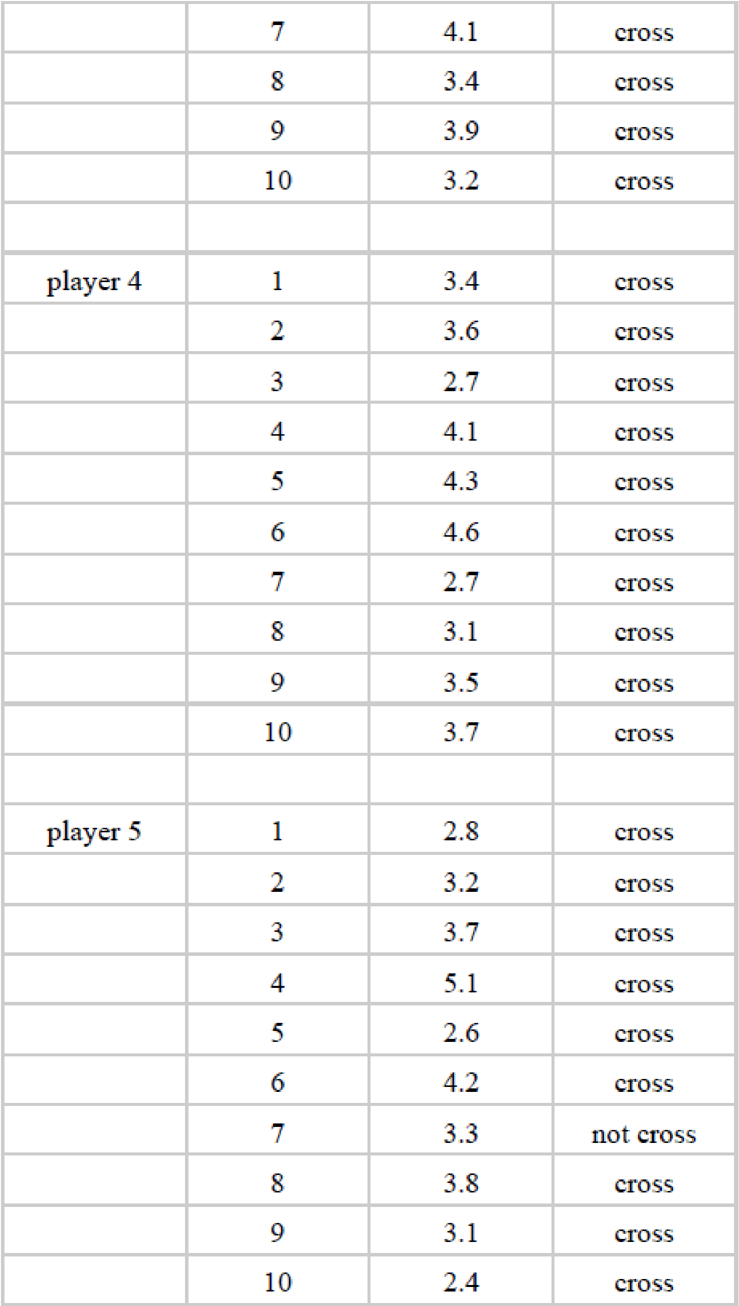
Experiment 2. The accuracy of detecting multiple football players crossing line activityconcurrently

We try to let different football players carry this device and cross lines to find out whether this device works for various situations.

In this experiment, 5 different players will carry this device and cross the line. Each player will use different speeds to do this for 10 times.

Table 2. different players cross the line

player1	time	speed (m/s)	device output
	1	3.6	cross
	2	2.8	cross
	3	2.5	cross
	4	3.9	cross
	5	4.3	cross
	6	4.9	cross
	7	5.2	cross
	8	3.3	cross
	9	4.6	not cross
	10	3.7	cross
player 2	1	3.2	cross
	2	3.9	cross
	3	3.6	cross
	4	2.8	cross
	5	3.4	cross
	6	4.3	cross
	7	4.7	cross
	8	5.1	cross
	9	2.4	cross
	10	3.4	cross
player3			
	1	2.3	cross
	2	2.8	cross
	3	3.3	cross
	4	3.5	cross
	5	3.1	cross
	6	3.8	not cross



	7	4.1	cross
	8	3.4	cross
	9	3.9	cross
	10	3.2	cross
player 4	1	3.4	cross
	2	3.6	cross
	3	2.7	cross
	4	4.1	cross
	5	4.3	cross
	6	4.6	cross
	7	2.7	cross
	8	3.1	cross
	9	3.5	cross
	10	3.7	cross
player 5	1	2.8	cross
	2	3.2	cross
	3	3.7	cross
	4	5.1	cross
	5	2.6	cross
	6	4.2	cross
	7	3.3	not cross
	8	3.8	cross
	9	3.1	cross
	10	2.4	cross

From this result, we can find out that the device generally captures players crossing the line precisely with an accuracy of 94%. Few errors (results that show players didn't cross the line) are shown.

From these results above, we can find out that in two experiments, the device successfully detects players crossing the line most times. This shows the device can detect the location properly no matter which player is carrying the device and the speed players run. This shows that the device has relatively high accuracy. The result also meets our expectation that the device will detect most situations in which players cross the line. The evidence is that among hundreds of times players cross the line, there's only fewer than 10 times that the device didn't detect players crossing the line.

5. RELATED WORK

This work titled "Architecture of an IoT-based system for football supervision (IoT Football)" is to use communication technology like ZigBee (one type of personal area networks) and embed sensing devices (e.g. sensors and RFID) [16]. By collecting footballers' information through sensors and sending them to the cloud, this device can help monitor the health of footballers and reduce the occurrence of adverse health conditions like hypoglycemia, swallowing the tongue and shortness of breath.

These two works share some similarities and have quite a difference. The similarity between our program and their program is that both use sensing devices to collect necessary data for certain purposes. Both programs also use communication technology that lets the information be sent from the device to the cloud. One difference between them is that two programs collect different sets of data. Their program collects footballers' health data while our program collects footballers' distance and location data. Their program uses communication technology like Zigbee while our program uses communication technology 3g cellular [10].

The clear advantage of our solution is that it can be used easier in a variety of locations. This advantage is caused by 3g cellular. Their communication solution using Zigbee is limited in a small scale and needs professionals to debug them. So, our solution with a 3g network can be used in numerous locations around the United States and can be used in large areas like football fields.

This work titled "IoT for Next-Generation Racket Sports Training" is to collect information by using wireless wearable sensing device (WSD) [11]. The system of WSD is capable of recognizing three different actions, i.e., smashes , clears , and drops , with an accuracy rate of 97% [12]. With this information, they can differentiate racket sports players between professional, subelite, and amateur players from their stroke performance. This IoT framework aims to change the way of racket sports training from experience-driven (subjective) to data-driven (objective).

Our work has several similarities and differences between this work. The similarity is that both works use data to provide advice for sports. Same as the WSD of this work, our system is low-cost, easy-to-use, and computationally efficient. The difference is that their solution shows users their level of performance (professional, subelite, and amateur players), while our solution presents distance, speed, and duration time.

Our solution provides users with more detailed information in an intuitive way. By carrying our devices in the football game, users can get detailed information like speed, distance and duration right after the game, they can also see the map so that they know their trajectory in the game. This gives footballers the opportunity to improve their performance with collected information.

This work titled "Continuous health monitoring of sportspeople using IoT devices based wearable technology" is to be used in wearable tracking devices to collect the health details and track the exercise records for reducing the risk factors. Machine learning techniques are introduced to analyze and monitor sportspersons' health.

Our work has several similarities and differences. The similarity is that both works use tracking devices in order to monitor and analyze important data of sportsmen. The difference is that this work uses wearable devices while our work uses dedicated devices that sportsmen carry on. The second difference is that our device tracks data like speed, velocity, and location while this device tracks health-related data.

Our solution can show you your trajectory during matches and football players can find out the data easily on their mobile phone [15]. They can use this data to further improve their performance and know what they did during matches. Their solution, however, emphasizes on health which sportsmen need professional doctors in order to make the data meaningful. So the threshold for football players to use the data of their device is quite high.

6. CONCLUSIONS

We will have several people run the same routes on the same field. If everyone runs at a decent pace, we should be getting different samples of data per person. If the tracker can show us the difference between each player and if it is accurate enough to tell us where each player is where the coach needs to know where they are, we can tell that the application is accurate.

Accuracy- Although we use more professional sensors to collect crucial information, it’s inevitable that there’s error in locations and distance [13]. The problem is that the GPS and the distance counting sensor do not exactly reflect this information. The location and distance we get from the sensor may have a little bit of variation compared to the true location and distance.

Practicability- The price of tracking equipment (Particles tracking one) is a little expensive. It costs about 160 dollars per tracker so there’s still room for price reduction to make it more universal, helping more people.

Optimization- We can still improve the process of collecting data so that this process is easier.

In the future, if possible, we could upgrade our current hardware equipment to 4G devices instead of staying on 3G. I feel like 4G could improve the signal and perhaps send out more accurate data instead of being just an estimated location. 4G could also reach out to more rural locations so the technology is more available to everyone. We can also have the tracker run at smaller interval ticks so that the data given will be more accurate.

Currently the tracking device is quite expensive, so in the future if one of the companies decide to sponsor us, we could use that money to make the device a lot less expensive [14]. Thus more teams could use our equipment at a lower price and not have to worry about it being worse in quality.

REFERENCES

[1] Lee, In, and Kyoochun Lee. "The Internet of Things (IoT): Applications, investments, andchallenges for enterprises." Business Horizons 58.4 (2015): 431-440.

[2] Ishida, Kazunari. "IoT application in sports to support skill acquisition and improvement." 2019IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA). IEEE, 2019.

[3] Wang, Yufan, et al. "IoT for next-generation racket sports training." IEEE Internet of ThingsJournal 5.6 (2018): 4558-4566.

[4] Huifeng, Wang, Seifedine Nimer Kadry, and Ebin Deni Raj. "Continuous health monitoring of sportsperson using IoT devices based wearable technology." Computer Communications 160 (2020): 588-595.

[5] Madakam, Somayya, et al. "Internet of Things (IoT): A literature review." Journal of Computer and Communications 3.05 (2015): 164.

[6] Ikram, Mohammed Abdulaziz, Mohammad Dahman Alshehri, and Farookh Khadeer Hussain. "Architecture of an IoT-based system for football supervision (IoT Football)." 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT). IEEE, 2015.

[7] Nastic, Stefan, et al. "Provisioning software-defined IoT cloud systems." 2014 internationalconference on future internet of things and cloud. IEEE, 2014.

[8] Gerhana, Yana Aditia, et al. "Decision support system for football player's position with tsukamoto fuzzy inference system." MATEC Web of Conferences. Vol. 197. EDP Sciences, 2018.

[9] Bi, Zhuming, et al. "IoT-based system for communication and coordination of football robot team." Internet Research (2017).

[10] Markov, Marko, et al. "Application of Firebase Cloud Service for Storing and Analyzing Data from IoT Mobile Devices." J. Mech. Autom. Identif. Technol 3 (2018): 17-20.

[11] Xu, Yetong, and Wenwu Hu. "Load evaluation of campus football match based on microprocessorand IoT wearable equipment." Microprocessors and Microsystems 81 (2021): 103778.

[12] Yu, F. U. "IoT Application in Monitoring System of Competitive Sports." Communications Technology (2012)

[13] Farooq, M. Umar, et al. "A review on internet of things (IoT)." International journal of computer applications 113.1 (2015): 1-7.

[14] Wilkerson, Gary B., Ashish Gupta, and Marisa A. Colston. "Mitigating sports injury risks using internet of things and analytics approaches." Risk analysis 38.7 (2018): 1348-1360.

[15] Elijah, Olakunle, et al. "An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges." IEEE Internet of Things Journal 5.5 (2018): 3758-3773.

[16] Stoyanova, Maria, et al. "A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues." IEEE Communications Surveys & Tutorials 22.2 (2020): 1191-1221.

PLAYGUESSR: COMMERCIAL APPLICATION OF MACHINE LEARNING IN FOOTBALL PLAY PREDICTION

Jason Wu[1], Evan Gunnell[2] and Yu Sun[2]

[1]Barrington High School, 616 W Main St, Barrington, IL 60010
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

The offensive strategy in American football strives to be enigmatic. A strong offense has a well rounded offensive playbook, rotating offensive plays in attempts to disrupt any predictive patterns. Therefore, it has always been in theinterest of defensive coordinators to offer accurate predictions of the upcoming play to minimize offensive yardage gain. A well advised defense can change its positioning and coverage schemes, given solely whether the next play will be a run or a pass. Although coaches have developed traditional heuristics for tendency-based play prediction, they are limited to patterns discerned by human consciousness.

This paper aims to take advantage of recent professional football databases in an attempt to develop a machine learning classification model for predicting opponent play-calling, as well as implement said model into a novel application aimed at deployment on all levels of play.

We conducted research on various classification models and features. Utilizing 10 past seasons from the National Football League (NFL), we devised random forest classification models with optimally chosen features. We also investigated the importance of Synthetic Minority Oversampling Technique (SMOTE) in training with inherently imbalanced datasets. The final model was able to achieve an NFL league average accuracy of 89.52%, with 91.69% as the highest team-specific accuracy. This accuracy is substantially higher than past projects of similar goals.

KEYWORDS

Football Analytics, Machine Learning, Classification, Play Prediction.

1. INTRODUCTION

Football is distinctive in its compartmentalized, or "discrete", plays. Hence, football plays naturally contain features, such as the score differential, down, and time remaining. Similarly, most football plays can be classified into two: running—moving the ball upfield from the line of scrimmage—and passing—throwing the ball from behind the line of scrimmage to a receiver positioned downfield. All remaining plays are "special", such as those when the ball is not in play (kickoffs and points-after-touchdown) and those when the ball remains in play (punts, field-goals, and quarterback kneels) [7].

American football, like the majority of competitive sports, is based on ambiguity [2]. From a defensive standpoint, the offensive advantage relies heavily on the unpredictable nature of play-calling [3]. Defensive coordinators rely on traditional heuristics, usually on past opponent tendencies, to predict offensive plays. Simple predictions can lead to shifts in defensive

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 145-164, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111714

positioning and coverage schemes that greatly hinder offensive capabilities.

Given the readily accessible professional football datasets, as well as the inherent need for improved defensive prediction tools for in-game coaching assistance, this paper seeks to devise an accurate machine learning model for said predictions, as well as fully implement an accompanying mobile app for ready deployment at all levels of play [4].

Traditional prediction heuristics are indispensable to football [12]. Opponent behaviors are usually inferred through careful analysis of opponent game footage. However, such a method is somewhat ineffective, since much of the underlying patterns are indiscernible to humans, due to inability to consider a wide set of potential influential features. Furthermore, human bias and retroactive interference limit the extent to which team-specific tendencies are discovered. Moreover, tendency-based metrics, such as run/pass ratios, have wildly varying performances. For example, recent years saw the National Football League (NFL) shift to adopt a pass-heavy offense. However, teams with balanced playbooks will have run/pass ratios that approach 1, which, intuitively, lowers prediction accuracy.

Past papers have considered solely the prediction of two types of targets: run plays and pass plays, with varying success [6]. Although special offensive plays, such as punts, field-goals, and quarterback kneels are rare and self-revealing, they can still reveal certain attitudes from opposing teams that are not discernible by simple run/pass target sets. For example, an offensive coordinator who favors run/pass plays over punting on 4th down can be inferred to be riskier and prone to turnovers. Even more, run/pass-only target sets significantly reduce data for certain situations. For example, 4th down data is greatly reduced when punting plays are neglected, which skews the data-by-down distribution.

Past papers have also focused mainly on professional level deployment. The extensive datasets that result from NFL coverage allows for papers to implement models not easily generalizable to other levels of play. For example, one popular input factor commonly featured by past papers is Madden player ratings. Video game statistics, aside from the fact that they are tailored to serve diverging purposes that confound model development, are not easily replicable on lower levels, such as high-school and collegiate teams.

Current years saw an uprising in publicly maintained professional sports databases [7]. In this paper, we aim to accomplish the goal of accurately predicting opponent play-calling. We were inspired by past papers to utilize the expansive Kaggle dataset developed by Horowitz et al, detailing 2009-2018 NFL play-by-play statistics. Past papers have attempted to solely predict whether the upcoming offensive play will result in a run or a pass. However, by removing special plays, such as punts, a small yet significant portion of the dataset is lost [8]. Our paper is the first to recognize the usefulness of offensive special plays in the target set, since offensive special plays do not disrupt game continuity and function heavily similar to other offensive plays. Furthermore, we predicted that including special plays offers relevant information to both offenses and defenses alike regarding an opponent’s situational behaviors. More elaboration is provided in the subsequent section.

Furthermore, we prioritized usability factors, such as time expenses, practicability, and functionality. Moreover, our application is aimed towards deployment on all levels of play, whereas previous papers have majorly focused on experimentation and development on the professional level with the NFL.

Much of our motives behind our evaluation metrics and experiments chosen are described in the following section, but our priorities are generally targeted towards responding to the challenges described in Section 2. We chose 5-fold cross validation for evaluation, which allows a model trained on 8 past NFL seasons to be evaluated for 2 NFL seasons. This metric allows for confirmation that play-calling tendencies are persistent and independent of specific matchups. Similarly, models were trained with team-specific data, and the league average—across 32 teams—is recorded. This is done to ensure that our experiments are not skewed by teams with varying tendency strengths.

Our experiments target specific questions and challenges that guide the development of our machine learning model. First, we analyzed the potential application of Synthetic Minority Oversampling Technique (SMOTE) on the dataset, due to an inherent imbalance. We evaluated simple classification models trained on both pre-SMOTE and post-SMOTE data. Secondly, we explored various popular classification models and chose one with optimal accuracy and time expenses, again weighted with regard to cross validation scores. Third, we studied the influence of our feature set and greatly reduced the final feature set to maximize model accuracy.

The paper's structure is as follows: Section 2 describes a small subset of significant challenges that we encountered and overcame during the design and experiments; Section 3 dissects our solution into details by analyzing our implementation methods and the motives behind our choices; Section 4 presents our experiments, as well as an abundant collection of recorded data, that were targeted to correspond to the various challenges that we discuss in Section 2; Section 5 gives a brief analysis of recent related works, including a critique and a comparison with our paper. Finally, Section 6 gives a brief concluding summary, as well as defining areas of future work.

2. CHALLENGES

In order to build the prediction model, a few challenges have been identified as follows.

2.1. Choosing optimal dataset features & targets

The dataset that we utilized to conduct research featured an extensive list of features, 255 in total [9]. However, we distinguished only 15 features to have potential relevance to influence play calling. Althoughthere was an overabundance of potential targets, ranging from total WPA (win percentage added) to two-point conversion probability, we ultimately decided on play-type [10]. That is, whether the play would result in a pass attempt or a run attempt. Our rationale behind this being that run/pass play predictions are actionable data for the defense. In addition, we added all possible offensive plays, such as punts, field-goals, and quarterback kneels (only excluding special plays that take place after scoring, such as extra points and kickoffs). Past papers have erred in the ruling that only run/pass predictions are constructive to the defense. These erroneous perspectives limit the dataset, especially on 4th down, where punts are most likely and most frequent. In the perspective of commercial potentials, a client should be able to input arbitrary factors for predictions beyond the next play. For example, if a team is predicted to attempt a field-goal later in an offensive drive, an offensive coordinator can determine the intensity of the next offensive drive, based on the score differential. Another example considers the end of the game (or when the score differential is very high), when the leading team is likely to run out the play clock with the quarterback kneeling when in possession of the ball. However, this practice varies with teams, and some will favor offensive drives in an attempt to score until the end of the game, which comes with associated risks of turnovers. Coordinators equipped with hypothetical predictions about opponent endgame behaviors can then adjust the offensive drive intensities, as well as defensive prevention schemes, among many other controllable factors, to take advantage.

A greater challenge arises from choosing relevant features. An overabundance of factors can not only confound the model and decrease the predictive accuracy, but can also significantly slow down the client-side processes, hence diminishing commercial practicability. Unnecessary factors not only make training the model more timely, but can also decrease the quality of client-side usage. More features meansmore inputs during in-game scenarios. Time is a significantly limiting variable, given that NFL regulationsonly allow 40 seconds in between each play. Variables that consume time yet do not yield comparable increases in prediction accuracy must not be equated into the final product.

Our solution is a multi-layered analysis using feature importances [11]. All 15 potential features were firstgauged for predictive influence, and although some factors distinguished themselves on the extremes of the spectrums, others were prone to confoundment due to the mixing of variables. This warranted a simplemanual grid-search using a small subset of potential factors similar in influence. In the end, the 15 potential features were minimized into a small subset of 7 showing maximum accuracy; each feature was considered simple for client-side entry, hence increasing efficiency.

2.2. Competing for user attention with other professional products/past papers

We have found several past papers, alongside numerous community projects, with similar goals of predicting football play-calling, with the oldest dating from 2015. However, they were primarily concerned with implementation solely on the professional level. Hence, they incorporated numerous NFL-specific features, such as Madden player ratings, that cannot be easily generalized to other levels of play. Our paper is distinguishable from past papers in that we seek to bring play prediction to all levels of play, using simple yet sufficient models [12].
Furthermore, we did not find any similar commercial product that shares a common target. Some challenges that may arise from the app-implementation include the creation of a streamlined and efficient UI, capable of importing large amounts of training data, as well as providing a high amount of customizability for the consumer’s preferences. Similarly, the former challenge warrants compatibility with existing popular game-film sharing systems, and the latter would require the ability of choosing specific datasets and features to predict from. An example of the latter would be using the dataset of all past opponent games versus using solely the dataset of past opponent games against a specific team, since play-calling variability varies drastically between play-levels. Furthermore, organizations may have ethical regulations that may conflict with the use of said technologies on the field.

2.3. Determining how to measure model accuracy

The dataset we utilized was expansive, providing data from 10 NFL seasons. This allowed a high variability of metrics for measuring the model’s accuracy. The challenge arises from the consideration of real-life scenarios. For a commercial product, a possible utilization is providing predictions during an ongoing game. Hence, a possible accuracy metric is comparing model predictions versus actual results in aspecific game. However, another consideration arises from the potential high variability of a team’s playbook throughout the season, or, on a greater scope, throughout the franchise’s short-term history. Hence, we decided that an appropriate accuracy metric would be cross validation with 5 folds. Roughly speaking, 5 fold cross validation allows a model trained with solely 80% of the data to be tested across 2 NFL seasons, which amounts to around 32-46 games.

We believe that using such a metric, in conjunction with the resulting model accuracies, has led to discoveries of generalized play-calling trends and tendencies throughout the NFL. For example, the fact that our model achieved a high validity through 2 NFL suggests that play-calling tendencies are rather innate, and not significantly influenced by a specific opponent, hence reducing the perceived variability in NFL play-calling.

2.4. Finding data and SMOTE

Traditionally, professional sports data has been long privatized, with only a small fraction of relevant data making its way to the public through media outlets. However, recent years saw the rise of publicly maintained sites documenting a wide range of professional sports data, from player drafts and trades to detailed post-game performance reports. This, in conjunction with recently developed web-scraping libraries targeting professional sports databases, form the premise that made this paper possible.

Due to extensive testing by previous papers on past NFL seasons, our original intention was to switch focus to a different level of play. We settled upon high-school game films as the most appropriate candidate, since college football is heavily similar to the professional level, and levels below high-school are hard to find due to the scarcity and obscurity of their datasets. However, high-school data being the candidate would result in significant limitations in research capabilities. Due to the constrictions of high-school level data, we are provided with less features, as well as less data. Moreover, the level of dynamism in high-school level play is, intuitively, higher than those at the professional levels. Personnel are almost completely replaced each year, and the skill and performance vary significantly per team. With our main intention being the development of a sound model, we chose to avoid these self-imposed limitations, and instead chose past NFL seasons as our research dataset.

Another challenge also arises from the fact that the dataset is intuitively skewed. Visualization of the types of data reveals that the play types are heavily skewed. Running and passing plays More details can be found in later sections describing methodologies.

2.5.Choosing optimal model & optimizing hyperparameters (accuracy and speed)

Once the metric for measuring accuracy was established, we turned to evaluating different metrics using said accuracy metric. This took place before the features were optimized. We chose several popular classifier models, since there was a high variability in performance and efficiency. Models quick to train with data usually performed less adequately against those with longer training times. Due to the fact that model retraining is infrequent and not required, we prioritized accuracy over efficiency. In the end, the model with the best accuracy was a random forest classifier. Interestingly, the traditional heuristics that coaches rely on are heavily similar with the model’s decision trees. However, the decision tree is limited inthe expansiveness of the feature set that it can utilize.

Following work was done on optimizing hyperparameters using a grid search cross validation. There were several popular and significantly influential settings that we considered, ranging from the number of trees in an ensemble to the max depth reached by each tree. The grid search sets were limited with respect to time restraints. For example, even though an increase in n-estimators may boost model accuracy, its tradeoff is significantly slower training times and potentially renders the possibility of in-game model retraining futile, not to mention slowing down prediction processes. Similarly, hyperparameters, such as max depth, have the potential to increase efficiency at the cost of model accuracy. We found that default hyperparameters kept the time costs sufficiently low, while not sacrificing accuracy. More of our optimization methodologies can be found in the next section.

3. SOLUTION

A mobile app [PlayGuessr] trains a machine learning classification model using the play conditions (eg. down, time, score differential) as inputs and play type as outputs (eg. run, pass, punt). Firstly, we implemented the app to receive the client's desired training dataset, as well as any prediction inputs. For ease of customizability, the training data is categorized by games. This can allow the client to select, with specificity, the data that the model will train with.

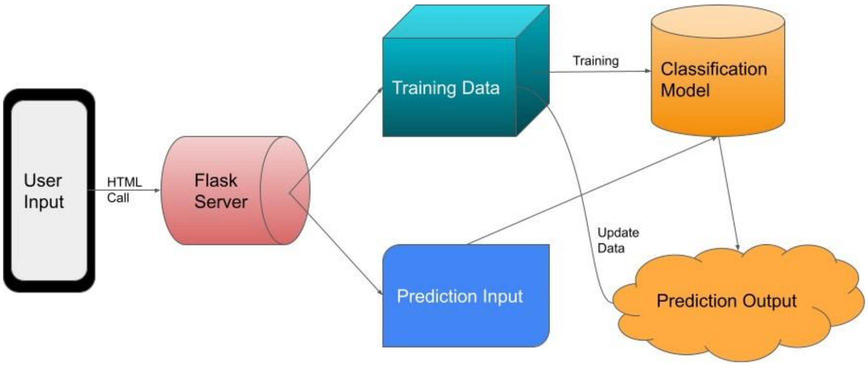


Figure 1. Overview of the app

As Figure 1 demonstrates, the model trains with the client-selected datasets and can readily take inputs from the client. The client's inputs are sent to the model, which returns a play type prediction from the client' s inputs. After the play, the client is prompted to record the actual play type in the app. The app maintains a dataset of the past plays during an ongoing game and gives the client the option to retrain the model with said plays.

PlayGuessr is implemented as follows. Firstly, the backend is divided into three main components. The training set is produced from user-provided data. We used the Pandas library for Python to parse our intended features and convert Data Frames into training sets. SMOTE was performed on the training data, with a K-Neighbors value of 2. The second backend component is our machine learning model. After obtaining our training data, we fitted a random forest classifier model using said data. The model is now ready to make predictions. Thirdly, the Flask module for Python web framework was utilized to set up a server for our predictive algorithm.

A design prototype (Figure 3) is provided below, demonstrating the chronological relationships between each screen.

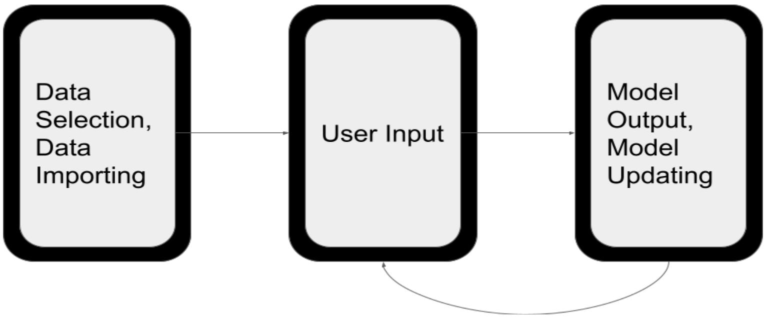
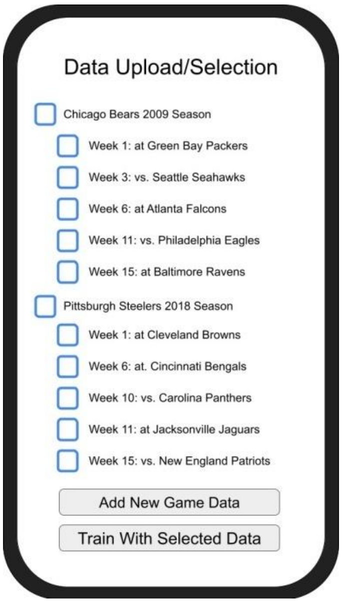


Figure 2. Design prototype



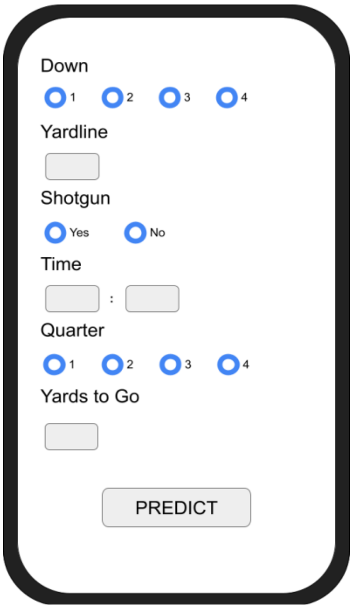
Data Upload/Selection

□ Chicago Bears 2009 Season
- □ Week 1: at Green Bay Packers
- □ Week 3: vs. Seattle Seahawks
- □ Week 6: at Atlanta Falcons
- □ Week 11: vs. Philadelphia Eagles
- □ Week 15: at Baltimore Ravens

□ Pittsburgh Steelers 2018 Season
- □ Week 1: at Cleveland Browns
- □ Week 6: at. Cincinnati Bengals
- □ Week 10: vs. Carolina Panthers
- □ Week 11: at Jacksonville Jaguars
- □ Week 15: vs. New England Patriots

Figure 3. Uploading and Selecting Training Set

The app will first prompt the user to choose their datasets to use for model training (Figure 3). After the training set is established, the backend will fit the model accordingly, and predictions are ready to be made. The user will be taken to the input screen (Figure 4).



Down
○ 1 ○ 2 ○ 3 ○ 4

Yardline

Shotgun
○ Yes ○ No

Time
:

Quarter
○ 1 ○ 2 ○ 3 ○ 4

Yards to Go

Figure 4. Input Screen

The input screen consists of six out of the seven features, due to net yards per offensive drive hard to keep track by the user. Instead of traditional textboxes, we have implemented buttons for features with discrete values. After the features are selected, the user will prompt the app to send a server command to run our predictive algorithm. The results will be returned in the final output screen (Figure 5).

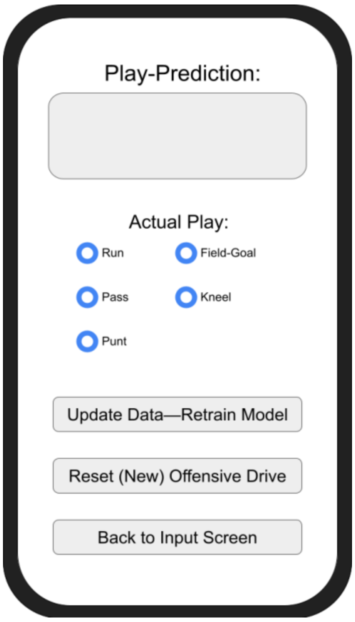
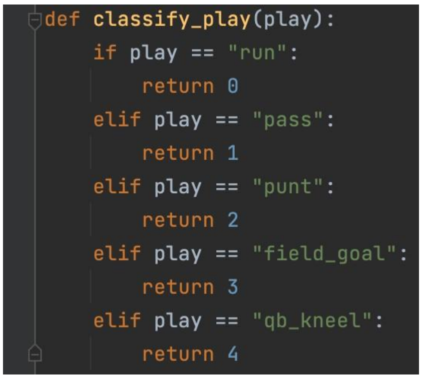


Figure 5. Output Screen

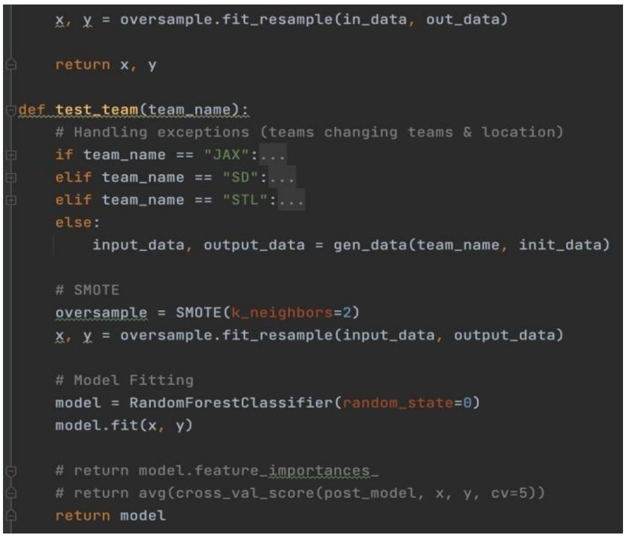
After the play, the user will be prompted to provide the actual resulting play. The actual output is kept in a temporary list in the backend, alongside the provided input features. After every prediction, the user will be provided the ability to update the training dataset with new information from the ongoing game. Due to the time expenses associated with retraining a random forest classification model, it will be up to the user to update the model at a time of convenience. Moreover, the net yards feature will be implemented in the output screen, allowing the user to end an offensive drive to effectively reset the net yardage to zero. The user will be returned to the input screen, where the next offensive prediction can be made again.

The following figures (Figures 6, 7, 8) demonstrate the various functions utilized in the processes described above, primarily to parse and construct training data, as well as evaluate models.



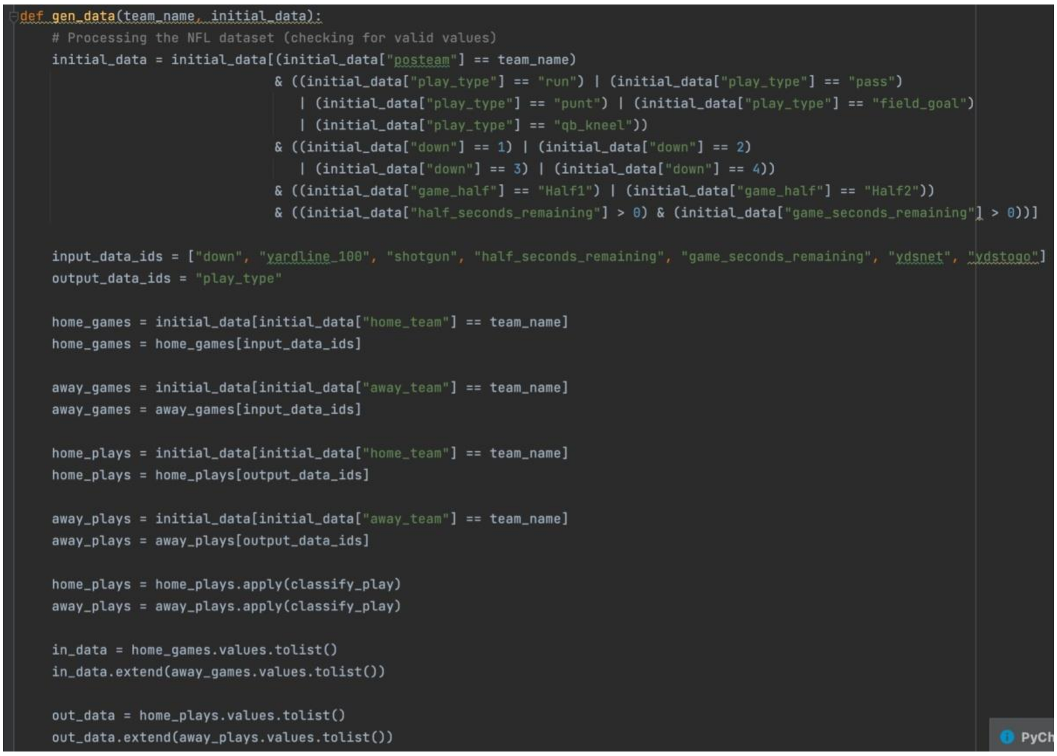
```
def classify_play(play):
    if play == "run":
        return 0
    elif play == "pass":
        return 1
    elif play == "punt":
        return 2
    elif play == "field_goal":
        return 3
    elif play == "qb_kneel":
        return 4
```

Figure 6. Code for Play Type Vectorization



```
    x, y = oversample.fit_resample(in_data, out_data)

    return x, y

def test_team(team_name):
    # Handling exceptions (teams changing teams & location)
    if team_name == "JAX":...
    elif team_name == "SD":...
    elif team_name == "STL":...
    else:
        input_data, output_data = gen_data(team_name, init_data)

    # SMOTE
    oversample = SMOTE(k_neighbors=2)
    x, y = oversample.fit_resample(input_data, output_data)

    # Model Fitting
    model = RandomForestClassifier(random_state=0)
    model.fit(x, y)

    # return model.feature_importances_
    # return avg(cross_val_score(post_model, x, y, cv=5))
    return model
```

Figure 7. Code for Experimentation



```
def gen_data(team_name, initial_data):
    # Processing the NFL dataset (checking for valid values)
    initial_data = initial_data[(initial_data["posteam"] == team_name)
                                & ((initial_data["play_type"] == "run") | (initial_data["play_type"] == "pass")
                                   | (initial_data["play_type"] == "punt") | (initial_data["play_type"] == "field_goal")
                                   | (initial_data["play_type"] == "qb_kneel"))
                                & ((initial_data["down"] == 1) | (initial_data["down"] == 2)
                                   | (initial_data["down"] == 3) | (initial_data["down"] == 4))
                                & ((initial_data["game_half"] == "Half1") | (initial_data["game_half"] == "Half2"))
                                & ((initial_data["half_seconds_remaining"] > 0) & (initial_data["game_seconds_remaining"] > 0))]

    input_data_ids = ["down", "yardline_100", "shotgun", "half_seconds_remaining", "game_seconds_remaining", "ydsnet", "ydstogo"]
    output_data_ids = "play_type"

    home_games = initial_data[initial_data["home_team"] == team_name]
    home_games = home_games[input_data_ids]

    away_games = initial_data[initial_data["away_team"] == team_name]
    away_games = away_games[input_data_ids]

    home_plays = initial_data[initial_data["home_team"] == team_name]
    home_plays = home_plays[output_data_ids]

    away_plays = initial_data[initial_data["away_team"] == team_name]
    away_plays = away_plays[output_data_ids]

    home_plays = home_plays.apply(classify_play)
    away_plays = away_plays.apply(classify_play)

    in_data = home_games.values.tolist()
    in_data.extend(away_games.values.tolist())

    out_data = home_plays.values.tolist()
    out_data.extend(away_plays.values.tolist())
```

Figure 8. Code for Parsing Data

4. EXPERIMENTATION

Experiment 1

Note: Code snippets will be provided to detail the exact experiments taking place. We created two functions, gen_data (used to generate training data, given a specific team, see Figure 9), and test_team (performing SMOTE [when necessary] and fitting model with training data, able to return the model, CV scores, and feature importances). Please refer to Section 3 for the implementation of said programs. These two functions may be slightly altered with our following experiments.

The first experiment seeks to determine the relevancy of SMOTE, or Synthetic Minority Oversampling Technique. Intuitively, running and passing plays constitute the majority of plays called, significantly overshadowing situational plays (punts, field-goals, and quarterback kneels). The following table (Table 1) and PyPlot graph (Figure 9) describes the total occurrences of each play type over the 2009-2018 seasons. Another PyPlot Graph is an average of each play type's occurrences averaged per NFL team. Given an imbalanced dataset, we wanted to test the effectiveness of SMOTE in model accuracy, using a basic SVM classification model as our metric (using all available features).We applied SMOTE to our training data using a baseline K-Neighbors value of 2. We trained one model with the SMOTE undergone data, also keeping a control model trained with non-SMOTE data. We then compared average 5-fold cross validation results from the two models. This was done using the Chicago Bears, but results can be generalized to other teams as well, and the degree of efficacy is predicted to be dependent upon the varying play type ratios of each team. Those with higher disparities are predicted to benefit more from SMOTE. A Pyplot of the post-SMOTE training data for play types can be found below as well.

Table 1. Pyplot of the post-SMOTE training data

Play Type	**Total Occurrences**	**Average Occurrences**	**Percent of Total Plays**
Run	126,663	3,958.22	37.22%
Pass	177,933	5,560.41	52.29%
Punt	22,832	713.50	6.71%
Field-Goal	9,249	289.03	2.72%
Quarterback Kneel	3,603	112.59	1.06%

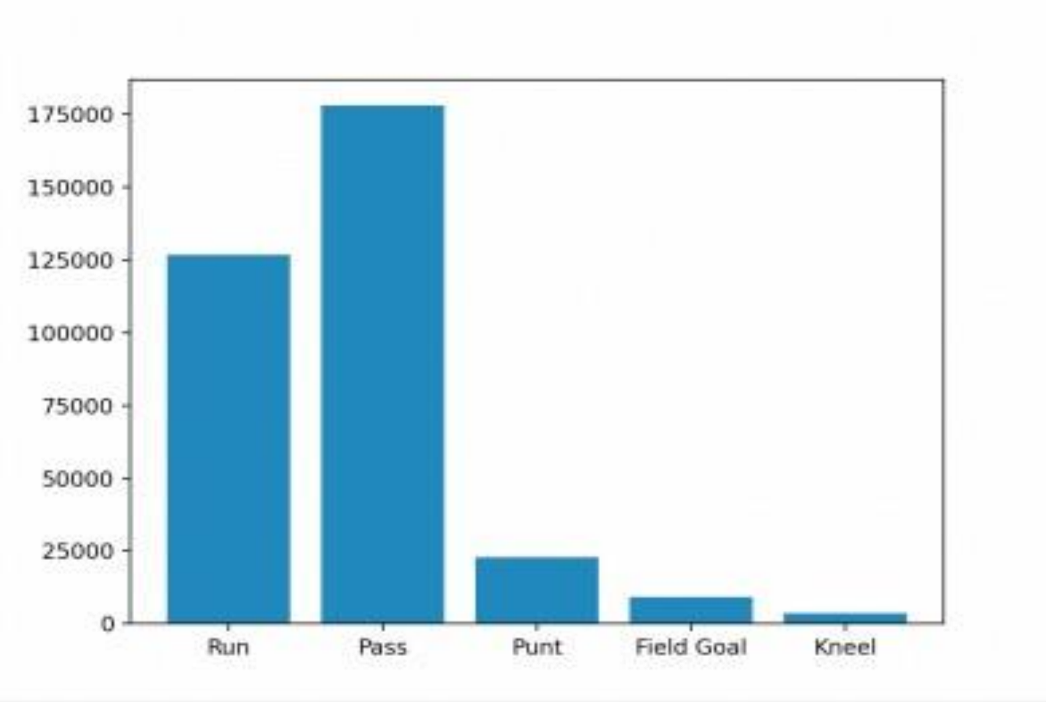


Figure 9. Total Occurrences in the NFL Per Play Type

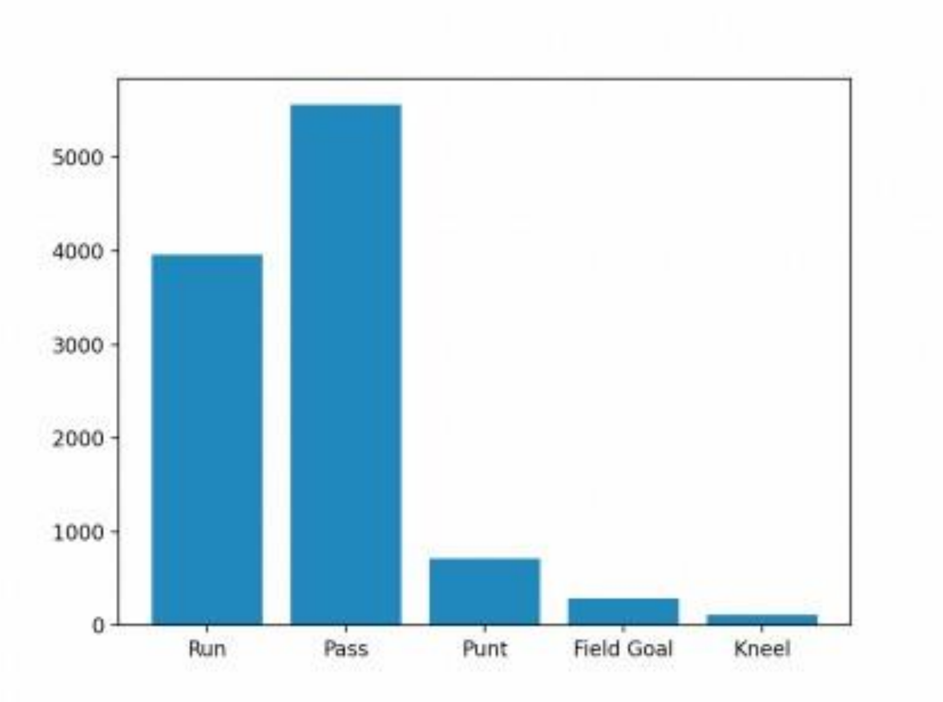


Figure 10. Average Occurrences Per Team Per Play Type

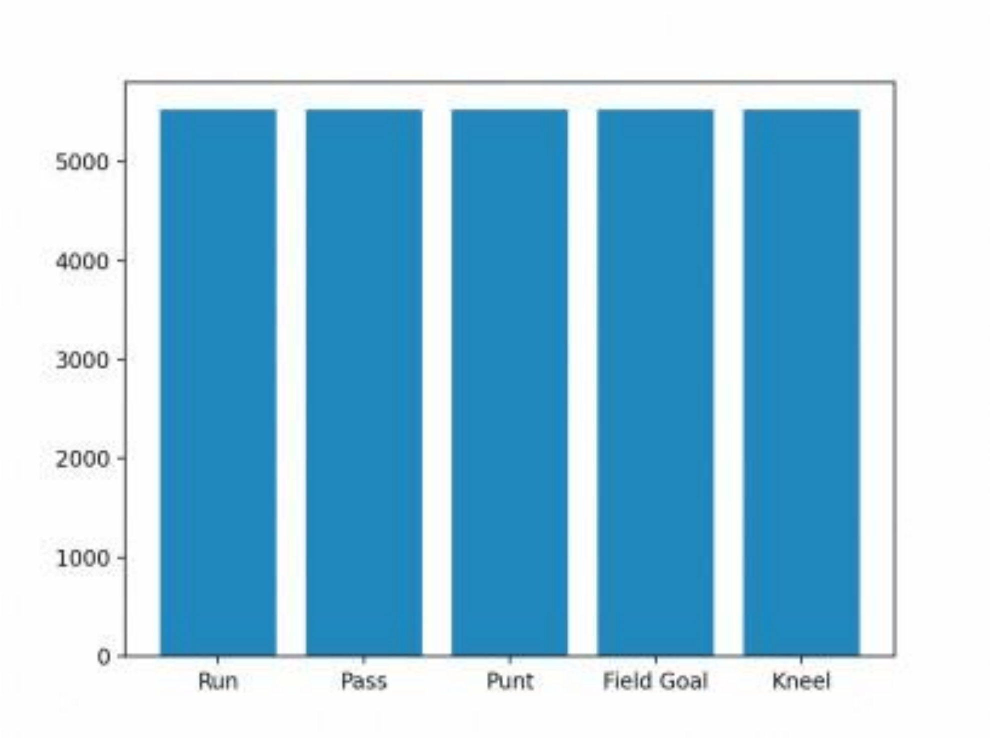


Figure 11. Post-SMOTE Distribution for the Chicago Bears

Table 2. Table of accuracy

Control SVM Accuracy	Experimental SVM Accuracy (SMOTE)
51.73%	57.41%

The results are as followed: Above, Figure 11 demonstrates the resulting distribution for the Chicago Bears after SMOTE was performed. Furthermore, Table 2 (above) describes the resulting model accuracies for both pre-SMOTE and post-SMOTE models. Code for the experiment can be found in Figure 12.

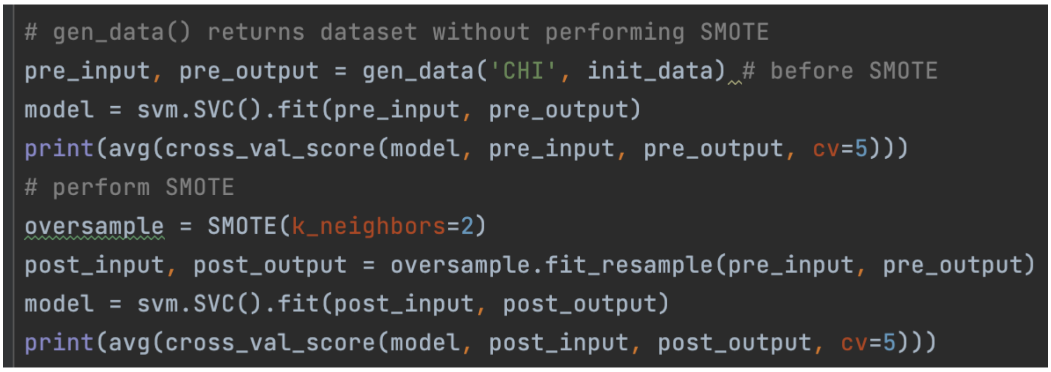

```
# gen_data() returns dataset without performing SMOTE
pre_input, pre_output = gen_data('CHI', init_data) # before SMOTE
model = svm.SVC().fit(pre_input, pre_output)
print(avg(cross_val_score(model, pre_input, pre_output, cv=5)))
# perform SMOTE
oversample = SMOTE(k_neighbors=2)
post_input, post_output = oversample.fit_resample(pre_input, pre_output)
model = svm.SVC().fit(post_input, post_output)
print(avg(cross_val_score(model, post_input, post_output, cv=5)))
```

Figure 12. Code for Experiment 1

As the results demonstrate, SMOTE was capable of elevating a basic SVM classification model’s accuracyby 10.97%. We discovered that SMOTE is essential for boosting accuracy due to the imbalance nature of our datasets. Further experimentation with increasing K-Neighbor values did not yield any significant nor constant upwards trends in model accuracy, and the K-Neighbor value was determined to be at two in order to compensate for future time expenses. The following experiment explores the SMOTE-induced accuracy enhancements with different popular classification models.

Experiment 2

We have discovered that SMOTE significantly boosts a model's accuracy due to the imbalance in our datasets. However, different classification models respond differently to SMOTE, and their increases in accuracy also vary. This experiment aims to find the most optimal post-SMOTE model. We considered 5 popular classification models: SVM, Random Forest, Naive Bayes (Gaussian NB), Passive Aggressive, andGradient Boosting. We chose to remain with the Chicago Bears for our testing data. For features, we have included all available features as a baseline. All hyperparameters are set at default values, as provided by Scikit-Learn, except for a constant and uniform random state. The following table shows results for all 5 models. Each model is trained on both pre-SMOTE and post-SMOTE data (with K-neighbors remaining at 2) and evaluated by a 5-fold cross validation metric. The percent increase in performance is also recorded in Table 3. Note: the code from Experiment 1 was reused with slight modifications by substituting various models in place of SVM.

Table 3. Results for 5 models

Model	Pre-SMOTE	Post-SMOTE	% Increase
SVM	51.73%	57.41%	10.97%
Random Forest	71.67%	88.82%	23.94%
Naive Bayes	65.75%	81.30%	23.66%
Passive Aggressive	48.74%	57.50%	17.98%
Gradient Boosting	74.04%	87.95%	18.79%

Our data demonstrates that there is a high variability between the effects of SMOTE on final model accuracy. We decided upon Random Forest as our model of choice, as it yielded the highest post SMOTE accuracy. This experiment further supports our previous experiment demonstrating the pivotal role of SMOTE. Gradient boosting is another highly favorable candidate with the second highest post-SMOTE accuracy. Had our experiment been done solely on non-SMOTE datasets, we would have chosen gradient boosting as the most optimal model, since it has the highest pre-SMOTE accuracy.

Experiment 3

The previous experiments have yielded a desirable data set and a competent model. The final experiment seeks to define the optimal subset of features with the highest accuracy and intuitive convenience. We predict that due to different teams having varying priorities, an optimal feature subset forthe Chicago Bears might not be optimal for teams differing in values and priorities. To mitigate team-specific feature selection, we performed feature analysis on all 32 NFL teams with all available features. We used random forest classifiers along with post-SMOTE datasets. The team-specific data is then averaged to obtain an appropriate generalized feature list for the entire league.

The below table (Table 4) and PyPlot (Figure 13) demonstrates the relative average importance of ourfeature set. Note: the function test_team was modified to return the feature importances of the random forest classification model trained from the team's data.

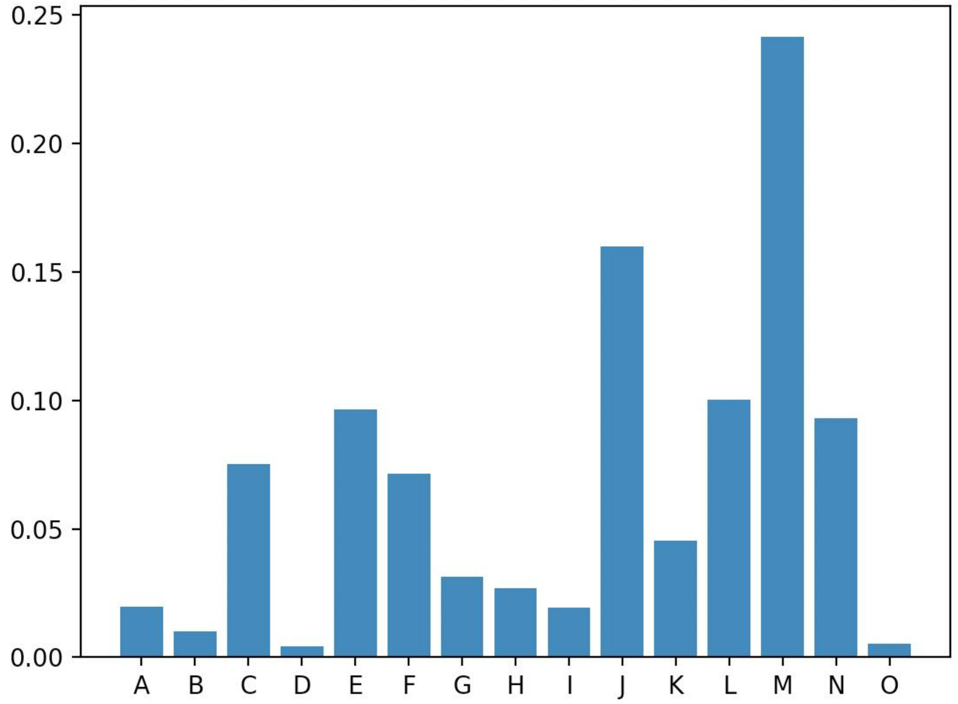


Figure 13. The relative average importance of our feature set chart

Table 4. Average Importances for All Features

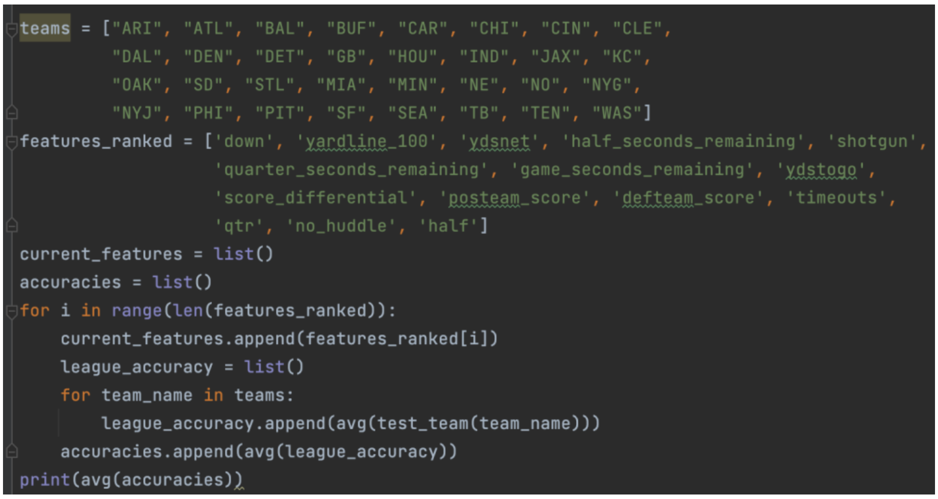
Feature	Importance (%)
A. Timeouts Remaining	1.96%
B. Quarter	1.02%
C. Quarter Seconds Remaining	7.52%
D. Half	0.42%
E. Half Seconds Remaining	9.65%
F. Game Seconds Remaining	7.16%
G. Score Differential	3.12%

H. Offensive Team Score	2.69%
I. Defensive Team Score	2.69%
J. Yardline	15.99%
K. Yards to Go	4.52%
L. Net Yards (On Drive)	10.03%
M. Down	24.14%
N. Shotgun (Formation)	9.31%
O. No Huddle	0.52%

From the average feature importances, our set of 15 features can be ranked by their percentages. The ordergoes as follows: down, yardline, net yards on a drive, half seconds remaining, shotgun, quarter seconds remaining, game seconds remaining, yards to go, score differential, offensive/defensive team score, timeouts remaining, quarter, no huddle, half.

We then trained the model 15 times, starting with solely the most influential feature, adding the next most influential feature every iteration, as well as recording the league average accuracy for each iteration. We seek to set a pruning point in our features set to discard any irrelevant features bringing insufficient boosts in accuracy in return for added inconvenience for client-side management.

The following table (Table 5) describes our findings, as well as a PyPlot graph (Figure 17) that charts theprogress in accuracy boosting per feature added. Code for the experiment can be found in Figure 14.



```
teams = ["ARI", "ATL", "BAL", "BUF", "CAR", "CHI", "CIN", "CLE",
         "DAL", "DEN", "DET", "GB", "HOU", "IND", "JAX", "KC",
         "OAK", "SD", "STL", "MIA", "MIN", "NE", "NO", "NYG",
         "NYJ", "PHI", "PIT", "SF", "SEA", "TB", "TEN", "WAS"]
features_ranked = ['down', 'yardline_100', 'ydsnet', 'half_seconds_remaining', 'shotgun',
                   'quarter_seconds_remaining', 'game_seconds_remaining', 'ydstogo',
                   'score_differential', 'posteam_score', 'defteam_score', 'timeouts',
                   'qtr', 'no_huddle', 'half']
current_features = list()
accuracies = list()
for i in range(len(features_ranked)):
    current_features.append(features_ranked[i])
    league_accuracy = list()
    for team_name in teams:
        league_accuracy.append(avg(test_team(team_name)))
    accuracies.append(avg(league_accuracy))
print(avg(accuracies))
```

Figure 14. Experiment 3 code

Table 5. The progress in accuracy boosting per feature added

Feature Added	Cumulative Model Accuracy	Accuracy Difference	Increase In Percentage
A. Down	44.56%	N/A	N/A
B. Yardline	79.30%	+34.74%	+77.97%
C. Net Yards	84.71%	+5.42%	+6.83%
D. Half Seconds Remaining	86.34%	+1.62%	+1.92%
E. Shotgun	88.81%	+2.47%	+2.86%
F. Quarter SecondsRemaining	88.96%	+0.14%	+0.16%
G. Game Seconds Remaining	89.04%	+0.082%	+0.092%
H. Yards to Go	89.62%	+0.58%	+0.65%
I. Score Differential	89.59%	-0.028%	-0.031%
J. Offensive Team Score	89.40%	-0.19%	-0.22%
K. Defensive Team Score	89.16%	-0.24%	-0.27%
L. Timeouts Remaining	89.06%	-0.10%	-0.12%
M. Quarter	89.02%	-0.036%	-0.041%
N. No Huddle	88.94%	-0.082%	-0.092%
O. Half	88.92%	-0.023%	-0.026%

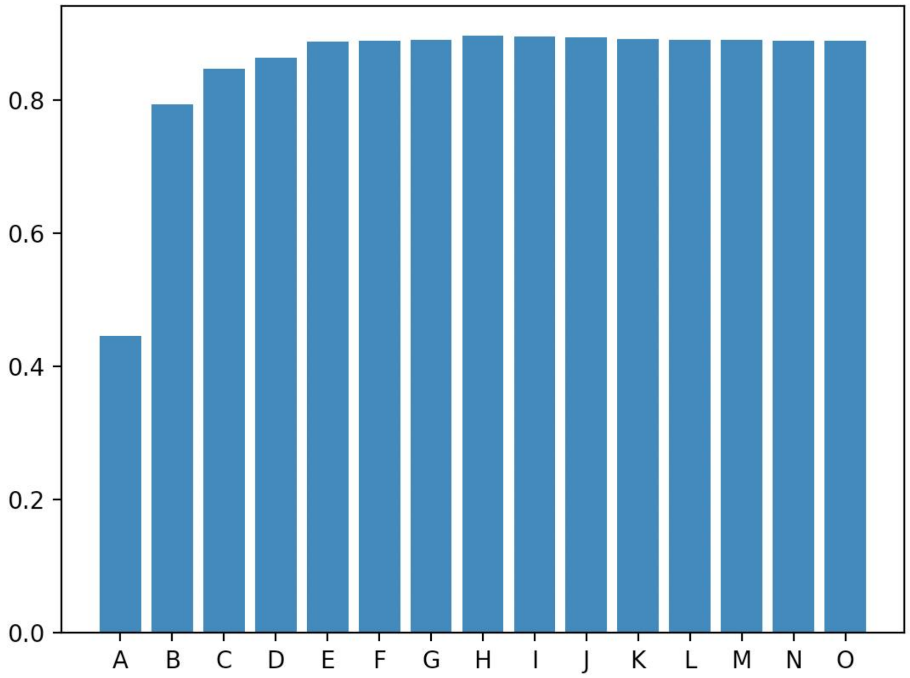


Figure 15. The progress in accuracy boosting per feature added chart

The graph (Figure 19) shows that model accuracy quickly rose and plateaued. The graph also reveals that model accuracy was maximized when the feature set consisted of factors A-H. The table confirms the absolute maxima of 89.62%. Furthermore, each successive addition of features resulted in a decrease in model accuracy. The results have defined our final feature set as follows: down, yardline, net yards, half seconds remaining, shotgun, quarter seconds remaining, game seconds remaining, yards to go.

Discussion

The first experiment was designed in response to the imbalanced nature of our dataset. As previously discussed, the imbalance in the dataset arises from the comparatively rare occurrences of offensive special plays, involving punts, field-goals, and quarterback kneels. Hence, our challenge was whether advanced techniques that remedy imbalanced classification is practicable and desirable in our project. The experiment validated our hypothesis, and the subsequent experiment further supported our predictions, with all models showing significant yet varying degrees of improvement after the training data had undergone SMOTE.

Our second experiment was tailored towards the challenge of choosing the optimal model. Past papershave been exceptionally widespread in their classification models of choice, and we evaluated a small set of popular classification models for their accuracy. The result demonstrated that random forest and gradient boosting classifiers were the strongest candidates, with gradient boosting holding the highest accuracy with pre-SMOTE training data, and random forest with post-SMOTE training data. Both classification models are good candidates, yet time complexities, as well as our utilization of SMOTE, drove us to choose random forest as our model of choice.

Our final experiment targeted the challenge of optimal feature selection in an effort to maximize model accuracy as well as minimize potential confounding usage. The accuracy vs feature set graph defined a clear absolute extremum. Although alternative evaluating metrics, such as analysis of variance with multiple factors, may be better for feature selection, our method of evaluation is less computationally expensive and also does not lie far from the optimal feature set.

5. RELATED WORKS

Joash Fernandes et al (2020) sought to design an intuitive and interpretable model for predicting run/pass plays in the NFL. The authors utilized a significantly smaller dataset compared to ours, consisting of the 2013-2016 NFL seasons. The authors considered the following features: year, quarter, minute, second, down, yards to go, yardline, and offensive formation; these are heavily similar to our feature set. The authors also derived many features, including the previous play, home/away game, point differential, passing and completion proportions, average yards per play type, and average position group scores (weighted by the position player count on the field). The paper experimented with complex models, assessing classification trees, K-nearest neighbors, random forests, and neural networks, which constitutes a more expansive list than our models. Similar to our model, the authors chose a decision tree model, however, with limited splits.

The finalized model achieved a lower accuracy of 65.3% with only three variables—down, point differential, and yards to go—and ten splits. Although the model was trained from league-wide data, team-specific trees scored accuracies between 64.7% to 82.5%, which had significantly higher variance than our results.

Lee et al (2016) set out to build upon past research to develop a model featuring more diverse features, primarily formation and player ratings. The paper utilized past NFL data from the 2011-2014 seasons, which is significantly smaller than the size of our training set, although featuring more details. The authors chose the following features from their dataset: score difference, quarter, time left in quarter, down, yards to go, player counter per position, formation, out of position players, turnovers, and home/away games. The authors settled on team-on-league type data. Moreover, the feature set contains many derived features, such as positional group rankings.

The authors experimented with four different models: logistic regression, linear discriminant analysis, gradient boosting machine, and random forest, most of which were assessed in our paper as well. Gradientboosting scored the highest, with a 75.7% accuracy, which is in agreement with our findings. However, the authors seeked to combine models in order to achieve an optimal accuracy, and they reported that a combined model with gradient boosting having a 60% and random forest a 40% weight showed the best performance, with an accuracy of 75.9%. The team was also able to achieve a game-specific accuracy ranging from 47.2% to 94.6%, as well as a season-specific accuracy ranging from 67.6% to 86%, which isalso more wildly variant than our results.

Ota (2017) presents a unique viewpoint by advising against using raw statistics, due to lack of specificscenarios and resulting discontinuous models. Instead, Ota advocates for a more ideal “aggregate” model.

Ota’s study utilized solely one NFL season, from 2016, which is significantly and concerningly smaller. The author included the following features: down, yards to go, score differential, game time remaining, and yardline. The list is marginally smaller than ours, as well as containing all static and observable variables.

The study performed experimentation on a neural network with customizable hyperparameters, which we did not assess. It must be clarified that models were trained separately per down. The average accuracy, 68.9%, was higher than the baseline run/pass ratio driven model accuracy, which scored 61.3%. It was noted that as the downs incremented, the model accuracy rose considerably, with 3rd down scoring the highest, at 86.8%.

The study was extended into research of situational-based models. Logistic regressions, linear regressions, support vector machines, and random forests were tested for accuracy, all of which were repeated in our study. Ota chose to reduce the feature set even more, with only team, down, distance, and yardline remaining. Holistic average accuracy proved to be at 65.67%, only a slight improvement over the naïve tendency-based model, which scored 62.86%. Both experiments yielded lesser accuracies than our results.

While Ota put forth two perspectives—aggregate and situational—both suggest that their practicality is not fully realized. The aggregate model is too broad in its reach to offer specific insight, and much of the general trends revealed are too intuitive to offer insight. Perhaps if the aggregate model was more specified towards a certain area, it would serve more value to coordinators [13].

6. CONCLUSIONS

In this paper, we addressed the potential of football in-game predictions. We experimented with dataset engineering techniques, such as SMOTE, to satisfy our eccentric target set that deviates from previous studies in its inclusion of offensive special plays. We also experimented with different machine learning classification models, and found that random forest and gradient boosting both have high-performances. We chose to advance with random forest due to time complexities, and in an effort to further minimize potential confounding variables during deployment, we sought to reduce our feature set using feature importances.

The experiments proved to be highly successful in responding to the encountered challenges. The experiment design was focused around both team-specific and league aggregate evaluations to minimize potential bias with outlying teams. The experimental results show that our model had a clear advantage over previous models, with a 89.52% accuracy across the league. This confirmed that offensive play-calling tendencies are independent of the defense and deeply rooted in the offensive team itself.

Limitations

Overall, we were unable to predict and evaluate the effects of deployment in other levels of play. The NFL dataset is expansive and may not be attainable in the same manner, whether it be personnel longevity or league expansiveness, in college and high-school level play.

Furthermore, hardware limitations prevented experiments from reaching further complexity. Hyperparameter tuning and analysis of variance were excluded for time expenses.

Even more, the dataset presented limiting feature sets. We also noticed that the dataset featured erroneous data, primarily for score differential, although the error is marginal and rare.

The choice to incorporate more plays into our label set meant that certain models, such as those that govern two-target predictions (logistic regression, for example) were made unavailable.

Future Works

A major area remains on the practicability of our model's deployment in lower level play. Further research will be done on whether the conditions of high-school and college football, such as higher player turnover rates and smaller datasets, warrants adjustments to our model.

Furthermore, more features can be evaluated for their influences. Some appealing features include gameday conditions, and specific offensive formations. The label set can also experiment with specificity. For example, expanding the target set to include all possible offensive plays, or providing multi-class outputs for different levels of specificity. Even more, deep learning models, especially those of recurrent neural networks, have yet to be tested.

REFERENCES

[1] Lee, Peter, Ryan Chen, and Vihan Lakshman. "Predicting offensive play types in the National Football League."(2016): 1-5.

[2] Pincivero, Danny M., and Tudor O. Bompa. "A physiological review of American football." Sports Medicine23.4 (1997): 247-260.

[3] Ota, Karson L. Football play type prediction and tendency analysis. Diss. Massachusetts Institute of Technology, 2017.

[4] Ötting, Marius. "Predicting play calls in the National Football League using hidden Markov models." arXiv preprint arXiv:2003.10791 (2020).

[5] Czaczkes, Benjamin, and Yoav Ganzach. "The natural selection of prediction heuristics: Anchoring andadjustment versus representativeness." Journal of Behavioral Decision Making 9.2 (1996): 125-139.

[6] Yashiro, Kotaro, and Yohei Nakada. "Computational Method for Optimal Attack Play Consisting of Run Plays and Hand-pass Plays for Seven-a-side Rugby." 2020 IEEE International Symposium on Multimedia (ISM). IEEE, 2020.

[7] Joash Fernandes, Craig, et al. "Predicting plays in the National Football League." Journal of Sports Analytics 6.1 (2020): 35-43.

[8] Cook, Sally. How to Speak Football: From Ankle Breaker to Zebra: An Illustrated Guide to Gridiron Gab.Flatiron Books, 2016.

[9] Ferryman, James, and Ali Shahrokni. "Pets2009: Dataset and challenge." 2009 Twelfth IEEE internationalworkshop on performance evaluation of tracking and surveillance. IEEE, 2009.

[10] Shen, Ting, et al. "Decision supporting model for one-year conversion probability from MCI to AD using CNN and SVM." 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018.

[11] Maina, James, and Kunihito Matsui. "Elastic multi-layered analysis using DE-integration." Publications of the Research institute for Mathematical Sciences 41.4 (2005): 853-867.

[12] Goyal, Udgam. Leveraging machine learning to predict playcalling tendencies in the NFL. Diss. Massachusetts Institute of Technology, 2020.

[13] Gray, Philip K., and Stephen F. Gray. "Testing market efficiency: Evidence from the NFL sports betting market." The Journal of Finance 52.4 (1997): 1725-1737.

A MOBILE PLATFORM FOR FOOD DONATION AND DELIVERY SYSTEM USING AI AND MACHINE LEARNING

George Zhou[1], Marisabel Chang[2] and Yu Sun[2]

[1]Santa Margarita Catholic High school,
22062 Antonio PkwyRancho Santa Margarita, CA 92688
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

Within the last year through the turmoil of the Covid-19 pandemic, an increasing number of families and individuals are experiencing food insecurity due to a loss of job, illnesses, or other financial struggles [4]. Many families in the Orange County area and abroad are turning to free food sources such as community food pantries or banks. Using specified surveys to food insecure families, we discovered a need for a solution to enhance the accessibility and usability of food pantries [5]. Therefore, we created a software application that uses artificial intelligence to locate specific items for users to request, and allow volunteers to see those requests and pick up the resources from food pantries, and deliver them directly to the homes of individuals. This paper shows the process in which this idea was created and how it was applied, along with the conduction of the qualitative evaluation of the approach. The results show that the software application allowed families and individuals to receive quality groceries at a much higher frequency, regardless of multiple constraints.

KEYWORDS

Mobile Platform, machine learning, data mining.

1. INTRODUCTION

A growing number of people are out of employment [6]. Hunger and food insecurity has grown dramatically due to healthcare and the economic crisis caused by the Corona virus [7]. Before the pandemic, more than 35 million Americans lived in households that struggled against hunger, and one in ten (10.5 percent) of households in the U.S. will experience food insecurity in 2019. Due to the effects of the corona virus pandemic, this number has grown to more than 42 million. People who may experience food insecurity, including a potential 13 million children according to USDA and Feeding America, may reside in surprisingly wealthy communities or demographics, but because of the recent pandemic, have recently been in financial strules. In Orange County, the average household experiencing food insecurity grew from 8.9% to 12.1% in 2019 to 2020, a total of 427,058 households. Many households that experience food insecurity do not qualify for federal nutrition programs and need to rely on their local food banks and other hunger relief organizations for support [14]. Therefore, there has been an increased need for free resources not coming directly from the government, such as food pantries, food bands, food drives, and or local distributions. Food pantries across the nation, specifically in Orange county, have seen drastic increases in visits and food distributed. Taking our Crossline Church food pantry as an example, we distributed 2.1 million lbs in 2021, growing from $125,000 lbs in 2019, that is 1,600% growth. In the first quarter of 2021, we've already outpaced the first quarter of

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 165-176, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111715

2020, distributing 586,754 lbs of food in Q1 of 2021 versus 95,510 lbs from Q1 of 2020, another 500% growth.

There is no direct competition that exactly matches our business model. However there are few related products and services that are inline with our mission, including Food Bank’s mobile food pantries are set up at sites that are close to the communities that have a higher concentration of vulnerable families [8].However it differs from our business model in that the mobile food pantries are still limited and can not cover an unlimited geographical area like our application can. Additionally, the mobile food pantries do not have a digital platform that connects to the Clients directly [9]. Therefore, the supply and demand is not connected and therefore has the same limitation like traditional food pantries [10]. Another application that has similar methodologies include the MealConnect App, which connects food banks with restaurants and grocery stores. The app matches the excess food at the restaurants and grocery stores with local food banks to reduce food waste. However the App does not connect directly with Clients, nor deliver to them directly. Delivery with Dignity Orange County program connects the restaurants with the volunteers through non-profit agencies to help deliver food to people who need the program based on “Triple Threat” criteria. This includes those that are at the highest risk for COVID-19 per CDC guidelines, as well as those who are ineligible or have not served by any community organization for the provision of food to their homes. It also includes families or individuals that are financially unable to meet their food needs without leaving their home, and who do not have a reliable support system of friends or family to assist. You could see that this effort does not have the digital platform and resources like that of our application, and it also does not connect directly with those individuals and catering to their needs.

PantryGo is a data driven, easy to use, smart delivery app that responds to general food insecurity as well as the healthcare and economic crisis created by the novel coronavirus. It works to assist our “Clients”, the most vulnerable and isolated individuals in OC who are not able to reach quality food sources due to various constraints through partnerships with “Services providers”, including local food pantries, grocery stores, and leading nonprofit agencies [11]. The program recruits "Volunteers" to deliver meals, pantry food, and excess groceries directly to the doorsteps of Clients. Simultaneous to feeding those in need, PantryGo is reducing the food waste in food pantries and grocery stores by providing estimated demand, food shelf time, food availability, resources allocation and other Artificial Intelligent information through the app. The strengths of PantryGo compared to its competitors include its reliability, which means that clients can count on PantryGo to quickly deliver fresh food and do not have to question whether proper food supply is available. Service Providers can count on PantryGo to know what and how much supplies are needed to meet the needs of those families. Next is its operational efficiency, meaning that PantryGo is the fastest service or delivery among food pantries. Technology is not widely used in the food pantry realm, or any service providing free resources to the most vulnerable. The uniqueness of this product will help food pantries, grocery stores, and all Service Providers to serve their Clients better, by saving their time, saving gas, saving resources, through one platform. Another strength is the quality of PantryGo, in which it serves all types of people in any circumstance. Regardless of physical, mental, or geographical constraints, the app allows any person to receive quality food and nutrition. Instead of quick meals from fast food restaurants or cheap, unhealthy snacks, the app allows for simple, quick, free deliveries of fresh produce. Additionally, the data that PantryGo provides is a SnapShot of hunger and food insecurity in any given community, that helps identify the numerous informational data regarding the behavior of the clients. This includes the most frequently requested items, as well as a spread chart on the locations of these clients and average delivery time for volunteers. This information can be used by food pantries and other agencies to better understand the tendencies and behavior of these families, so they can mold their programs or services to these findings.

Lastly, PantryGo allows for a stronger community, in which it helps our churches, government agencies, and policy makers to allocate the resources and tackle the root cause of food insecurity. Management teams can oversee, perform and solve problems better than before.

To prove our results, we can look at the percentage of which food waste has been reduced within the food bank system and individual food pantries, as well as the number of food received by food insecure families and individuals. In finding the decrease in the amount of food waste within the food bank system and individual food pantries, we can compare the current amount of food waste left at food banks, pantries, restaurants, and grocery stores compared to the amount of food waste left over after the implementation of the app. To find the percentage decrease of the food waste, we can divide the final amount by the original amount. This percentage will show the effectiveness of our application, furthermore proving our results in finding a useful solution. PantryGo, as stated before, hopes to greatly reduce food waste and surplus amounts of food that go to waste, as it is negatively impacting the environment, while simultaneously out of reach for families and individuals who desperately need those resources. Therefore, in measurement of the decrease in food waste that was able to reach these families and individuals, we can determine the success of our solution and furthermore prove our results. Additionally, in finding the number of food insecure families and individuals who are receiving more food after the launch of PantryGo, we can distribute online surveys to question whether their food supply has increased or decreased [12]. Because of the easy to use nature of the application, and the convenience in which it offers, families will be able to receive food at a much higher rate.

The rest of the paper is organized as follows: Section 2 explains the details and description of the challenges that were faced during the experiment, as well as the development and design of the solution. Next is section 3, which discusses the solution that was developed in wake of the topic and the methodology of that solution. Following that is section 4, which focuses on the experiment itself and an evaluation of that experiment in its process, followed by section 5, which discusses related works within the same field or topic that will be presented and evaluated. Finally, section 6 presents the concluding remarks and ideas, as well as future work intended for this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Finding a Solution

One of the first and biggest challenges that we faced in solving the problems that re-occurred at the food pantry was finding the solution itself. The food bank system along with the idea of a food pantry have existed for a few decades already, and since its creation, no one has come up with a solution to improve the reach of the food pantry and bring it to a technological level. Therefore, in the early stages of PantryGo, it was difficult finding the right structure and system for the application and the process in which the food would be delivered. Furthermore, in finding a solution, we did not want to alter the system on which the food pantry is already operating, as many volunteers have spent countless hours and days perfecting the system in place. After many weeks of brainstorming and adjustment, we were able to keep the current food bank system while moving the system to a digital platform and expanding the reach and accessibility of the food pantry system.

2.2. Creating the App

Most likely the biggest challenge of the experiment/project is the actual designing, development, and engineering of the mobile application within PantryGo. As a coder who only had a few weeks of experiment under the belt, developing a professional application was extremely hard from the get go. The implementation of code and constant refinement of the design as well as the UI/UX meant hundreds of alterations and changes to the code of the app. Furthermore, the difficulty in finding the balance between a functioning app with basic UI/UX and a professional grade app with very developed UI/UX led to many periods of distress and difficulty finding the right balance. As a team, we want to be able to put out a product that is easy to use, pleasing to the eye, and professional grade, but because of the limited time frame, as well as minimal funding, it is difficult to find the right balance between professional and functional.

2.3. Promotion

A challenge that is considered an ongoing challenge and one that may never be resolved is promotion or the advertisement of the mobile app and PantryGo as a whole. First off, the creation of the promotion assets, including logo design, website design, social media, flyer design, and graphic art design all require a long process of revision and perfection until satisfaction. After the theoretical creation and design of these assets are finished, there is a long process in which the small details and tedious development takes place. Additionally, many of the assets went through multiple changes and rounds of dissatisfaction, where they were completely thrown away, and created from scratch. After these assets were completed, they needed to be distributed to the target consumers and the right people who would spread the word of the application and the project even further. Difficulty in convincing people of the project and the applicability of the project was one of the hardest parts in utilizing the promotional assets and making sure it reaches the greatest number of people.

3. SOLUTION

PantryGo users will be able to sign up by choosing from 3 different user categories. The first category is the Foodbank profile. This profile will be used by Foodbank managers or anyone in charge of the operations at the specific food pantry. Foodbank profile users will be able to display details regarding their operational times, inventory, location, and other details that other users can view, saving time, energy, and confusion for clients and volunteers. The second category is the customer or client profile. This profile will be used by individuals or family members looking to receive food from the food pantry. Users can insert information regarding location and other personal information if desired. They will also be able to choose a grocery list with desired items, and then choose a food pantry that will automatically appear with the available items of the grocery list [13]. The third user is the volunteer profile. This user will be able to put in personal information so that other users can see the specifics of their volunteer for liability and safety reasons. The volunteers will be able to select from delivery requests by clients on that day, and will be able to see the details of the delivery, including location and pick up times. Additionally, the foodbank profiles will also be able to see customer requests that come in, and as those customers select items, the inventory within those food banks automatically change, and can be altered by the food bank manager, depending on the available inventory. Figure 1 shows the main components and flow of the app.

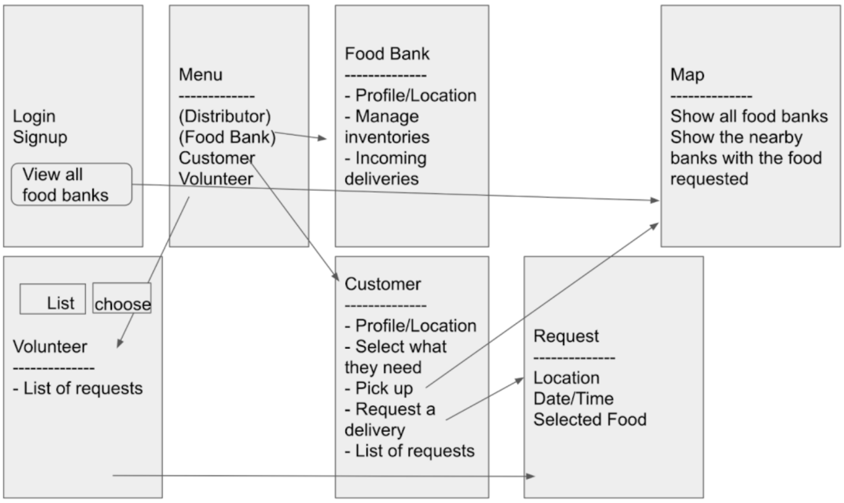
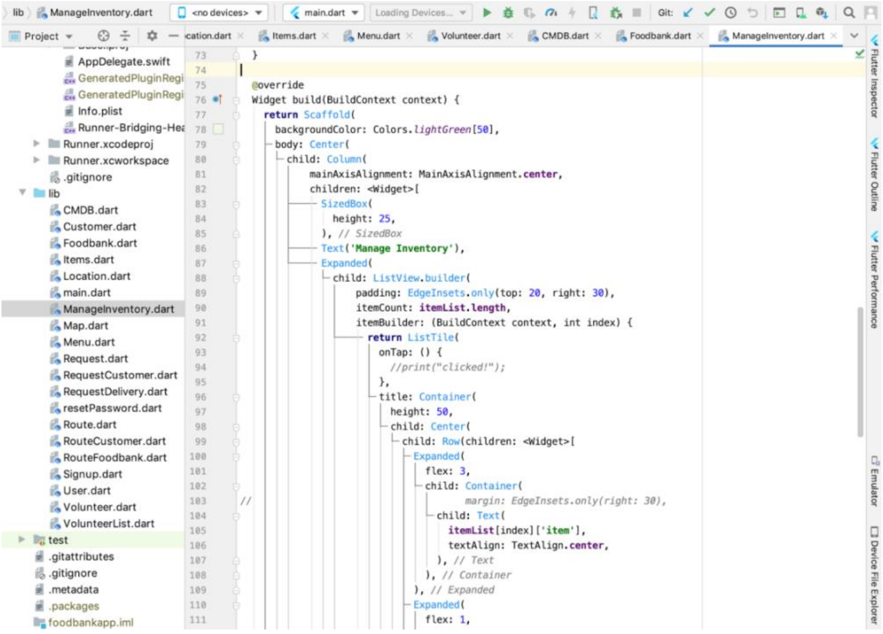


Figure 1. Overview of the app

Component 1: Food Bank User

The first component is the food bank user, in which we implemented the code to where the main functions of the food bank user is to be able to manage the inventory of the food pantry. In other words, the food bank or pantry manager will be able to change all of the food items and boxes that are available for pickup or delivery on a daily basis. Food bank users will also be able to change their locations and upload their information for display to other users such as customers and volunteers. They will also be able to see all of the food requests from customers, as well as all of the details of that request, including whether the request was completed by the volunteer or not. Figure 2 shows a segment of the code for the food bank users ability to change inventory. Figure 3 shows screenshots of the app from the food bank users perspective.



```
  }

  @override
  Widget build(BuildContext context) {
    return Scaffold(
      backgroundColor: Colors.lightGreen[50],
      body: Center(
        child: Column(
            mainAxisAlignment: MainAxisAlignment.center,
            children: <Widget>[
              SizedBox(
                height: 25,
              ), // SizedBox
              Text('Manage Inventory'),
              Expanded(
                child: ListView.builder(
                    padding: EdgeInsets.only(top: 20, right: 30),
                    itemCount: itemList.length,
                    itemBuilder: (BuildContext context, int index) {
                      return ListTile(
                        onTap: () {
                          //print("clicked!");
                        },
                        title: Container(
                          height: 50,
                          child: Center(
                            child: Row(children: <Widget>[
                              Expanded(
                                flex: 3,
                                child: Container(
//                                      margin: EdgeInsets.only(right: 30),
                                  child: Text(
                                    itemList[index]['item'],
                                    textAlign: TextAlign.center,
                                  ), // Text
                                ), // Container
                              ), // Expanded
                              Expanded(
                                flex: 1,
```

Figure 2. Food bank users ability to change inventory

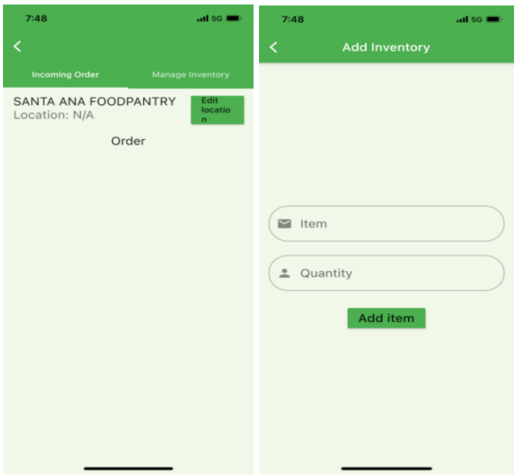
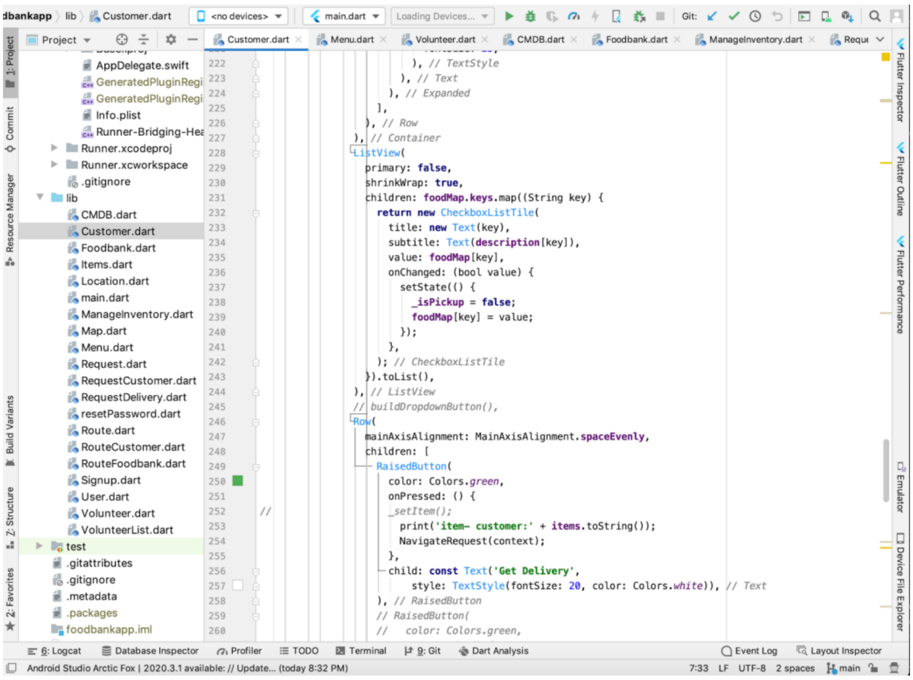


Figure 3. Food bank users perspective

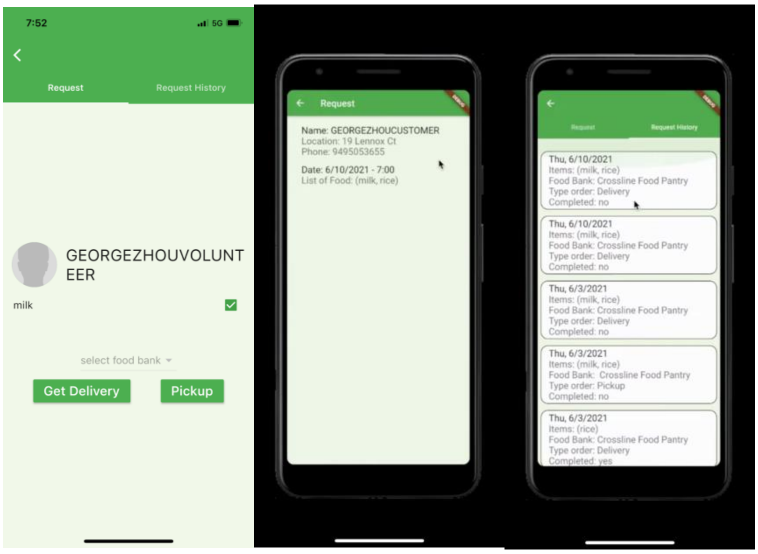
Component 2: Customer User

The second component is the customer user, which has a few main functions. The customer will be able to request a food list based on the available items from a food pantry within their area, and through artificial intelligence, the food items which they selected will automatically link with a food pantry with the specified items. Customers will also be able to see all of their past requests and the details of the request, and when a volunteer selects that customers request, it will automatically move into request history as completed. Figure 4 shows a segment of the code that allows customers to request specific food items, while Figure 5 shows screenshots from the viewpoint of the customer user.



```
          ), // TextStyle
        ), // Text
      ), // Expanded
    ],
  ), // Row
), // Container
ListView(
  primary: false,
  shrinkWrap: true,
  children: foodMap.keys.map((String key) {
    return new CheckboxListTile(
      title: new Text(key),
      subtitle: Text(description[key]),
      value: foodMap[key],
      onChanged: (bool value) {
        setState(() {
          _isPickup = false;
          foodMap[key] = value;
        });
      },
    ); // CheckboxListTile
  }).toList(),
), // ListView
// buildDropdownButton(),
Row(
  mainAxisAlignment: MainAxisAlignment.spaceEvenly,
  children: [
    RaisedButton(
      color: Colors.green,
      onPressed: () {
//    _setItem();
        print('item- customer:' + items.toString());
        NavigateRequest(context);
      },
      child: const Text('Get Delivery',
          style: TextStyle(fontSize: 20, color: Colors.white)), // Text
    ), // RaisedButton
    // RaisedButton(
    //   color: Colors.green,
```

Figure 4. Customers request specific food items



Request

Name: GEORGEZHOUCUSTOMER
Location: 19 Lennox Ct
Phone: 9495053655

Date: 6/10/2021 - 7:00
List of Food: (milk, rice)

Thu, 6/10/2021
Items: (milk, rice)
Food Bank: Crossline Food Pantry
Type order: Delivery
Completed: no

Thu, 6/10/2021
Items: (milk, rice)
Food Bank: Crossline Food Pantry
Type order: Delivery
Completed: no

Thu, 6/3/2021
Items: (milk, rice)
Food Bank: Crossline Food Pantry
Type order: Delivery
Completed: no

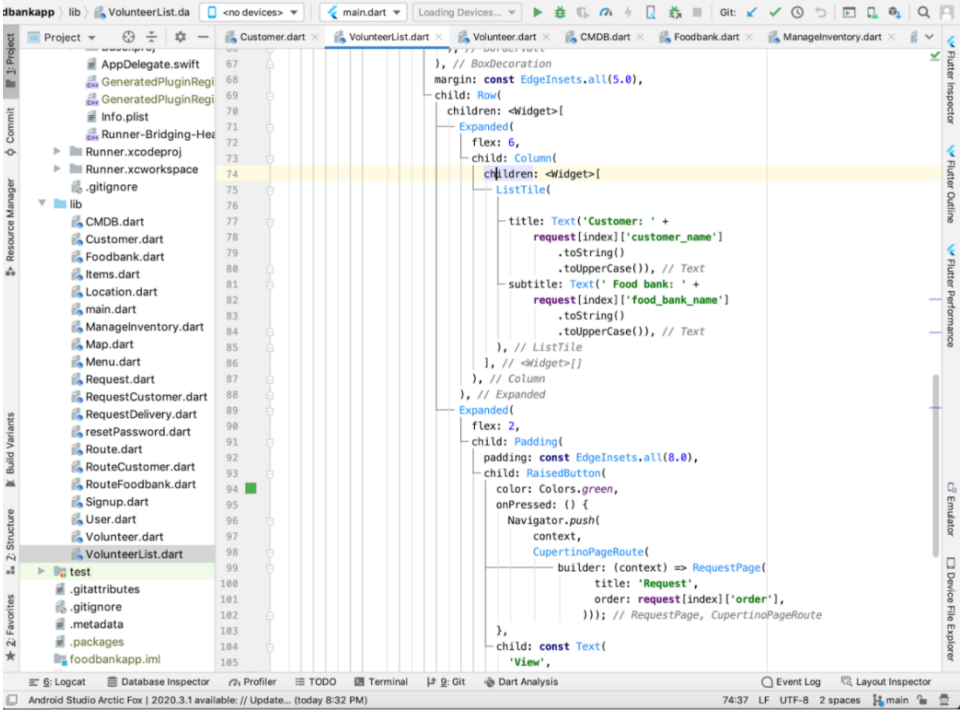
Thu, 6/3/2021
Items: (milk, rice)
Food Bank: Crossline Food Pantry
Type order: Pickup
Completed: no

Thu, 6/3/2021
Items: (rice)
Food Bank: Crossline Food Pantry
Type order: Delivery
Completed: yes

Figure 5. Screenshots from the viewpoint of the customer user

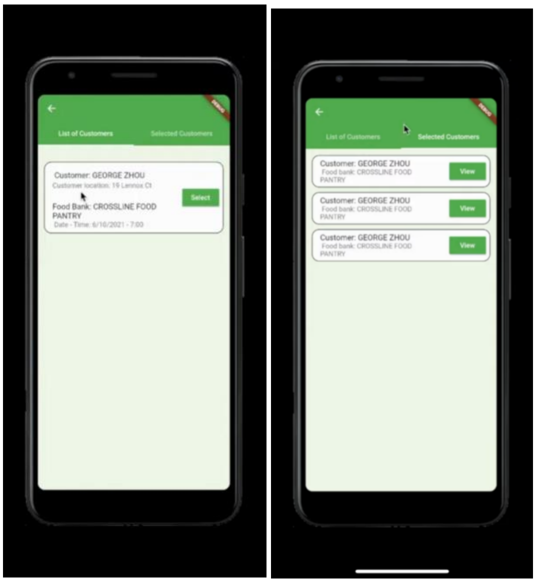
Component 3: Volunteer User

The third component is the volunteer user, which has the main functions of selecting customer food requests. The volunteer will automatically receive all food requests made by any customer at any time, including all the details retained to the request, such as location, time, date, and food items. Additionally, once the volunteer selects the request, it will go to the list of customers, and once completed, will update the screens on the other two users. In Figure 6, it will show the code that allows volunteers to select customer requests, and in Figure 7, it shows screenshots from the view of the volunteer user.



```
), // BoxDecoration
margin: const EdgeInsets.all(5.0),
child: Row(
  children: <Widget>[
    Expanded(
      flex: 6,
      child: Column(
        children: <Widget>[
          ListTile(

            title: Text('Customer: ' +
                request[index]['customer_name']
                    .toString()
                    .toUpperCase()), // Text
            subtitle: Text(' Food bank: ' +
                request[index]['food_bank_name']
                    .toString()
                    .toUpperCase()), // Text
          ), // ListTile
        ], // <Widget>[]
      ), // Column
    ), // Expanded
    Expanded(
      flex: 2,
      child: Padding(
        padding: const EdgeInsets.all(8.0),
        child: RaisedButton(
          color: Colors.green,
          onPressed: () {
            Navigator.push(
                context,
                CupertinoPageRoute(
                    builder: (context) => RequestPage(
                          title: 'Request',
                          order: request[index]['order'],
                        ))); // RequestPage, CupertinoPageRoute
          },
          child: const Text(
            'View',
```

Figure 6. Volunteers select customer requests



Customer: GEORGE ZHOU
Customer location: 19 Lennox Ct
Food Bank: CROSSLINE FOOD PANTRY
Date - Time: 6/10/2021 - 7:00

Customer: GEORGE ZHOU
Food bank: CROSSLINE FOOD PANTRY

Customer: GEORGE ZHOU
Food bank: CROSSLINE FOOD PANTRY

Customer: GEORGE ZHOU
Food bank: CROSSLINE FOOD PANTRY

Figure 7. Screenshots from the view of the volunteer user

The connection between the 3 components comes in the form of steps. The first step involves the food bank user, who will insert the inventory and the quantity of each specific item. Next, the customer user will be able to see the food items from the selective food pantry, and will then select the items they desire, and select the date and time on which they would like the request delivered [15]. Then, that request will be sent to both the food pantry and the volunteer, where the volunteer will select that food request and pick up the food from the food pantry, and then deliver it to the home of the individual or family.

4. Experiment

To be able to evaluate the efficiency of our approach, we gathered data and information over a 2 week span from the Crossline Church food pantry through the tallying of requests, weighing of food waste, and surveying of customers and volunteers. Two experiments were conducted to test the solution and its effectiveness. The first experiment explores the effects of the application on food waste, and the second experiment evaluates the applications effect on volunteer involvement.

Experiment 1: PantryGo's effect on Food Waste

This experiment involves the collection of data of the amount of food waste left over after each week. The first week was monitored without the usage of our solution, and the second week is monitored with the implementation of the application. Food waste has become a global crisis, especially in the United States, where surpluses of food has led to detrimental effects to the environment as a carbon footprint. The application attracts more customers to the food pantry and expands its reach, effectively reducing the food waste that piles up at locations such as food pantries. In Figure 8, it shows the pounds of food waste left over after week 1 and week 2. Clearly, the amount of food waste was reduced by more than 50% after the implementation of the application, showing its effectiveness in greater food pantry usage.

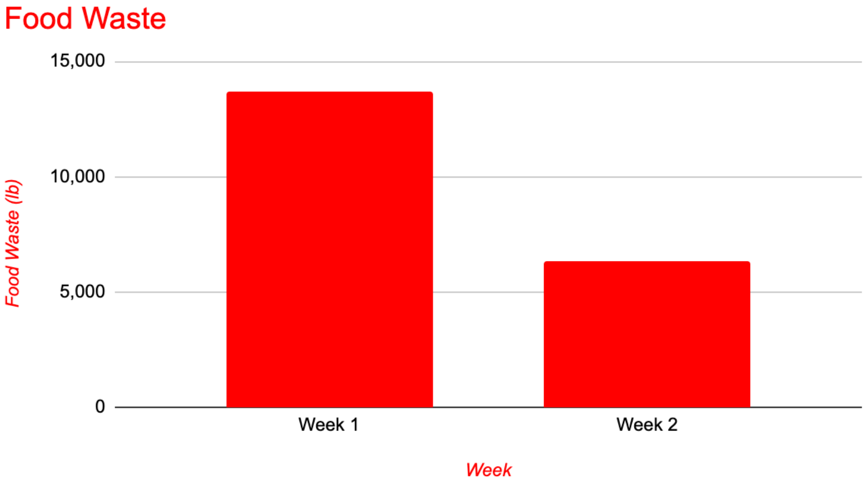


Figure 8. Food waste

Experiment 2: PantryGo's effect on Volunteer Involvement

This experiment involves the gathering of data and information based on surveys given to different youth volunteer groups asking for whether they were interested in volunteering for the food pantry. The first survey (during the first week) was given before the control group was notified of the opportunity to serve as delivery drivers, while the second survey (during the second week) was given right after the group was notified of the existence of the app. Figure 9, week 1, shows the number of volunteers who committed to volunteer for the food pantry that week in multiple positions except for food delivery driver, while week 2 shows the number of volunteers committed to work in multiple positions including food delivery driver. The figure evidently shows that there was an increase in volunteers in the second week, having linear correlation to the introduction of the application or solution. Therefore, teenage volunteers will have more involvement and interest in volunteering when given the opportunity to be a delivery driver, as these teenagers tend to enjoy car rides over physical labor.

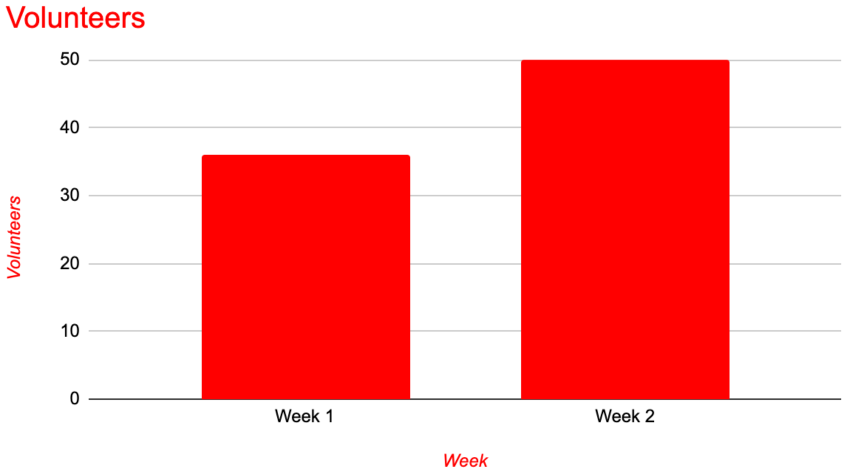


Figure 9. Volunteers

Both experiments show the effect on which the application or solution has addressed the problems faced within food banks and pantries, ranging from excess food waste or limited volunteers. The experiments showed that both of theses aspects were benefited after the introduction and implementation of PantryGo, both decreasing food waste and increasing volunteer involvement. These results reach my expectation in the positive benefits the solution brings to the food pantry system and current condition.

5. RELATED WORK

Leah R. Kicinski in her publication “Characteristics of Short and Long-Term Food Pantry Users”[1] discusses the modern research findings of the eating habits and overall health of frequent food pantry visitors. Kickiski and her team of researchers at Grand State Valley University used an interview based data collections strategy within food insecure residents in the Michigan area. Her findings indicate that food pantries are catering to two distinct group of people, one consisting of recently unemployed individuals, or families who have been under the poverty line for over a decade. Similar to my work, this paper discusses the flaws within the food pantry system, including its lack of accessibility, but does not give a solution to these problems, as Kicinski states, “Pantry users are not all identical in their characteristic, reasons for food need, or their food pantry experiences” [2]. Kicinski’s generalized statement only mentions the culprit behind the reason for the food pantries’ lack of accessibility, while my paper discusses a solution that can cater to all of the differences that cause different constraints.

“The Cost of Free Assistance: Why Low Income Individuals Do Not Access Food Pantries”, published by Kelley Fong, Rachel A. Wright, and Christopher Wimer discusses the possible reasons why many food insecure families do not utilize the free resources provided by food pantries. For most of these families, constraints including minimal transportation, distance, working multiple jobs, illnesses, or fast food options often lead them to ignore the resources provided by food pantries. This publication released by the Department of Sociology at Harvard University goes into great detail regarding certain research findings and data regarding the status of these low income families, and similar to my paper, give possible solutions for food pantries to meet these families in the middle.

In a very recently published paper titled “A Community Partnership for Home Delivery of Food Boxes to Covid-19 Quarantined and Isolated Families" led by a number of authors headed by Emily English examines the partnership between a community group that delivers food boxes to families who are isolated due to illnesses or COVID-19 cases, barring them from shopping for groceries. The community group, already connected to this community of residents, were able to deliver to over 531 families within the area. The publication, in its summary of the project's accomplishments and mission, is very similar to some of the main motives of PantryGo and its target market. The article also goes into great detail regarding the specific transfer strategies utilized, as they state, “Addresses were then sent to the regional transit partner to create routes and secure the appropriate number of buses and drivers to be utilized for deliveries" [3]. Similar to my publication, the addresses of these families were mentioned in ensuring the process of delivery is efficient and safe.

6. CONCLUSIONS

Because of the increasing demand for food pantry resources and distribution due to the ongoing pandemic, causing unemployment rates to skyrocket, I have proposed an application that can be used to expand the reach and accessibility of food pantries. The application uses artificial intelligence to fluidly offer users to choose their desired food items, linking these families or individuals with food pantries and volunteers, who will deliver their specific food requests directly to their door. This allows for families who experience constraints, barring them from reaching their local food pantry, including illnesses, limited transportation, large distances, multiple jobs, or time constraints to be able to still access the free resources provided by their local food pantry. To evaluate the effectiveness of this application, two separate experiments were conducted using gathered information through the form of food waste and volunteer involvement. The first experiment measured the change in food waste left over after a week of food pantry operation with and without the implementation of the application. The amount of food waste decreased drastically after the use of the application, showing its immediate effect in increasing the usage of the food pantry and its resources. Furthermore, the application can also benefit restaurants and grocery stores in reducing its food waste and carbon footprint by bringing its surplus of food to food pantries, and then to families who are experiencing food insecurity. The second experiment consisted of taking surveys of volunteers before and after the introduction of the application. The results showed that volunteers were much more likely to commit to volunteering when the opportunity of becoming a delivery driver was offered, furthermore showing the applications ability to attract teenage volunteers and community collaboration. The overall results of these two experiments validate the applicability of the solution, as well as its estimated success and efficiency.

The current limitations of the application is expanding the project to other food pantries and banks in different regions or states. Currently, the application is running in Southern Orange County, where the food pantry serves many residents in Orange County. PantryGo hopes to spread to other states and regions where the need to this application is at an even greater level, but because of its relatively early stages in operation, it is difficult to sell the idea to other food pantry managers or systems. These food pantries often already have a system of distributing their food to their community, and have worked hard to make that system as efficient as possible.

PantryGo hopes to find other collaborators or leaders within those different states or regions to represent the application and promote it similar to that of which is occurring in Orange County. These representatives in each region can be regional managers or board members of the organization, and can collaborate with other board members or regional leaders to grow as a family within PantryGo and spread the application nationwide.

REFERENCES

[1] Kicinski, Leah R. “CHARACTERISTICS OF SHORT AND LONG-TERM FOOD PANTRY USERS.” Michigan Sociological Review, vol. 26, 2012, pp. 58–74. JSTOR, www.jstor.org/stable/23292651. Accessed 5 Aug. 2021.

[2] Fong, K.; Wright, R.; Wimer, C. The cost of free assistance: Why low-income individuals do not access food pantries. J. Sociol. Soc. Welf. 2016, 43, 71–93.

[3] Emily English, Christopher R. Long, Krista Langston, Bonnie Faitak, April L. Brown, Amanda Echegoyen, Joel Gardner, Casey Cowan, Debbie Rambo, Brenda Perritt, Barb Laubenstein, Alyssa Snyder, Pat Bourke, Melisa Lelan& Pearl A. McElfish (2021) A Community Partnership for Home Delivery of Food Boxes to COVID-19 Quarantined and Isolated Families, Journal of Hunger & Environmental Nutrition, 16:1, 19-28, DOI: 10.1080/19320248.2020.1863284

[4] Gundersen, Craig, and James P. Ziliak. "Food insecurity and health outcomes." Health affairs 34.11 (2015): 1830-1839.

[5] Bhattarai, Gandhi Raj, Patricia A. Duffy, and Jennie Raymond. "Use of food pantries and food stamps in low‐income households in the United States." Journal of Consumer Affairs 39.2 (2005): 276-298.

[6] Layard, Richard, Stephen Nickell, and Richard Jackman. "The unemployment crisis." (1994).

[7] He, Feng, Yu Deng, and Weina Li. "Coronavirus disease 2019: What we know?." Journal of medical virology 92.7 (2020): 719-725.

[8] Arney, Fiona, and Dorothy Scott, eds. Working with vulnerable families: A partnership approach. Cambridge University Press, 2013.

[9] De Reuver, Mark, Carsten Sørensen, and Rahul C. Basole. "The digital platform: a research agenda." Journal of Information Technology 33.2 (2018): 124-135.

[10] Gale, David. "The law of supply and demand." Mathematica scandinavica (1955): 155-169.

[11] Burkhart, Patrick J., and Suzanne Reuss. Successful strategic planning: A guide for nonprofit agencies and organizations. Sage, 1993.

[12] Gordon, Jeffry S., and Ryan McNew. "Developing the online survey." Nursing Clinics of North America 43.4 (2008): 605-619.

[13] Bassett, Raewyn, Brenda Beagan, and Gwen E. Chapman. "Grocery lists: connecting family, household and grocery store." British Food Journal (2008).

[14] Holmes, Eleanor, et al. " “Nothing is going to change three months from now” : A mixed methods characterization of food bank use in Greater Vancouver." Social Science & Medicine 200 (2018): 129-136.

[15] Greenberg, Michael, Gwendolyn Greenberg, and Lauren Mazza. "Food pantries, poverty, and social justice." (2010): 2021-2022.

FAST CONVOLUTION BASED ON WINOGRAD MINIMUM FILTERING: INTRODUCTION AND DEVELOPMENT

Gan Tong and Libo Huang

School of Computer, National University of
Defense Technology, Changsha, China

ABSTRACT

Convolutional Neural Network (CNN) has been widely used in various fields and played an important role. Convolution operators are the fundamental component of convolutional neural networks, and it is also the most time-consuming part of network training and inference. In recent years, researchers have proposed several fast convolution algorithms including FFT and Winograd. Among them, Winograd convolution significantly reduces the multiplication operations in convolution, and it also takes up less memory space than FFT convolution. Therefore, Winograd convolution has quickly become the first choice for fast convolution implementation within a few years. At present, there is no systematic summary of the convolution algorithm. This article aims to fill this gap and provide detailed references for follow-up researchers. This article summarizes the development of Winograd convolution from the three aspects of algorithm expansion, algorithm optimization, implementation, and application, and finally makes a simple outlook on the possible future directions.

KEYWORDS

Winograd Minimum Filtering, Winograd Convolution, Fast Convolution, Convolution Optimization.

1. INTRODUCTION

Convolutional Neural Networks (CNN) are widely used in tasks such as computer vision and natural language processing. CNN with deep learning has reached or even surpassed the level of human experts in some fields by constructing deeper and more complex networks. At the same time, deeper CNN also brings more parameters and greater computing power requirements. Therefore, more and more research attempts to accelerate the training and inference of CNN and the use of fast convolution operators is an important method among them.

Fast convolution operator uses fast convolution algorithm to achieve convolution, including FFT convolution [1] and Winograd convolution [2]. This type of convolution converts the matrix multiplication in the original convolution into the corresponding element-wise multiplication (Hadamard product) by linearly transforming the input feature map and convolution kernel of the convolution operator to the corresponding domain. The result of the corresponding element-wise multiplication can be restored to the original feature mapping domain after the corresponding inverse linear transformation. In this "transform-calculation-inverse transformation" process, the number of multiplication operations is considerably reduced compared to direct convolution, and the cost is an increase in the number of addition operations.

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 177-191, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111716

On most modern processors, the execution efficiency of addition is much higher than that of multiplication. Therefore, we can replace the convolution implementation in the model with a fast convolution operator and use the reduced multiplication operations to improve the execution efficiency of the model. The same is to reduce the multiplication operation. The linear transformation in the Winograd convolution is to map the real number to the real number domain, and the FFT convolution is to map to the complex number domain. Therefore, the memory usage during the Winograd convolution operation only needs half of the FFT convolution, making the Winograd convolution the most popular fast convolution operator.

However, there are many challenges in directly applying Winograd convolution. First, the earliest proposed Winograd convolution has a limited scope of application. It can only be applied to two-dimensional convolutions with unit stride size and small convolution kernels. When applied to large convolution kernels, there will be numerical instability [3]. Secondly, due to the complexity of linear transformation and inverse linear transformation, the optimization of fast convolution operators on a specific platform is difficult to achieve, such as the use of parallelism and data locality [4]. In addition, Winograd convolution and network compression technology represented by pruning and quantization are difficult to directly combine, so it is not easy to be deployed on platforms with insufficient computing power and energy consumption restrictions [5].

In response to these problems, researchers have done a lot of work, but so far there is no published article that systematically summarizes related work. In order to make it easier for the follow-up researchers to understand and master the previous work, this article summarizes the development of Winograd from the three aspects of algorithm generalization, algorithm optimization, implementation and application, and looks forward to the possible future directions. The structure of this paper is as follows: Section 2 introduces the introduction and algorithm development of Winograd convolution; Section 3 introduces the optimization of Winograd convolution algorithm in three aspects; Section 4 introduces the realization and practical application of Winograd convolution on several types of platforms; Chapter Five summarizes this article and looks forward to possible future research directions.

2. THE INTRODUCTION AND DEVELOPMENT OF WINOGRAD CONVOLUTION

2.1. Winograd Minimum Filtering Algorithm

Winograd proposed the minimum filtering algorithm of finite impulse response (FIR) filtering in 1980 [6]. The minimum filtering algorithm gives m output generated by the FIR filter of r taps, namely $\mathrm{F}(\mathrm{m}, \mathrm{r})$, the minimum number of multiplications needed $\mu(\mathrm{F}(\mathrm{m},, \mathrm{r}))$ is $\mathrm{m} + \mathrm{r} - 1$. Taking $\mathrm{F}(2,3)$ as an example, the input $\mathrm{d} = [\mathrm{d}_0,\ \mathrm{d}_1,\ \mathrm{d}_2,\ \mathrm{d}_3]$ is a vector of size 4, filter $\mathrm{g} = [\mathrm{g}_0,\ \mathrm{g}_1,\ \mathrm{g}_3]^{\mathrm{T}}$, then:

$$F(2,3) = \begin{bmatrix} d_0 & d_1 & d_2 \\ d_1 & d_2 & d_3 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} m_1 + m_2 + m_3 \\ m_2 - m_3 - m_4 \end{bmatrix}$$

where:

$$m_1 = (d_0 - d_2) g_0 \quad m_2 = (d_1 + d_2) \frac{g_0 + g_1 + g_2}{2}$$

$$m_4 = (d_1 - d_3) g_2 \quad m_3 = (d_2 - d_1) \frac{g_0 - g_1 + g_2}{2}$$

When calculating $[\mathrm{m}_1,\ \mathrm{m}_2,\ \mathrm{m}_3,\ \mathrm{m}_4]$, the number of multiplications involved in the algorithm is $\mu(\mathrm{F}(2{,}3)) = 2 + 3 - 1 = 4$, the number of addition operations that need to be performed on d is 4, and the number of addition operations that need to be performed on g is 3 (the value of $\mathrm{g}_0 + \mathrm{g}_2$ can be calculated only once); using$[\mathrm{m}_1,\ \mathrm{m}_2,\ \mathrm{m}_3,\ \mathrm{m}_4]$ to get the result of $\mathrm{F}(2{,}3)$ requires 4 additions. The number of multiplications required by the algorithm has been reduced from 6 to 4.

2.2. The Introduction of Winograd convolution

Winograd minimum filtering algorithm can be expressed in the form of a matrix:

$$Y = A^T[(Gg) \odot (B^T d)]$$

where$g$ is the filter vector, d is the input data vector, Y is the output data vector, G is the filter transformation matrix, B^T is the data transformation matrix, $\odot$ is the corresponding bit multiplication of the matrix (Hadamard product), A^T represents the output transformation matrix. For $F(2{,}3)$, the matrices are:

$$B = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix},$$

$$A^T = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & -1 & -1 \end{bmatrix}, g = [g_0 \quad g_1 \quad g_2]^T, d = [d_0 \quad d_1 \quad d_2 \quad d_3]^T$$

By nesting the one-dimensional minimum filtering algorithm $F(m, r)$, we can get the two-dimensional minimum filtering algorithm $F(m \times m, r \times r)$:

$$Y = A^T[(GgG^T) \odot (B^T dB)]A$$

Now the size of the filter g is $r \times r$, the size of output Y is $m \times m$, and the size of input d is $(m + r - 1) \times (m + r - 1)$. The number of multiplications required by the two-dimensional minimum filtering algorithm is $(m + r - 1)^2$, while the number of multiplications required by the original convolution algorithm is $m \times m \times r \times r$. For $F(2 \times 2{,}3 \times 3)$, the number of multiplications is reduced from 36 to 16, which is a reduction of 2.25 times. Even if the additional addition operations are included, there are great benefits. We can naturally split Winograd convolution into four separate stages (as shown in Figure 1):

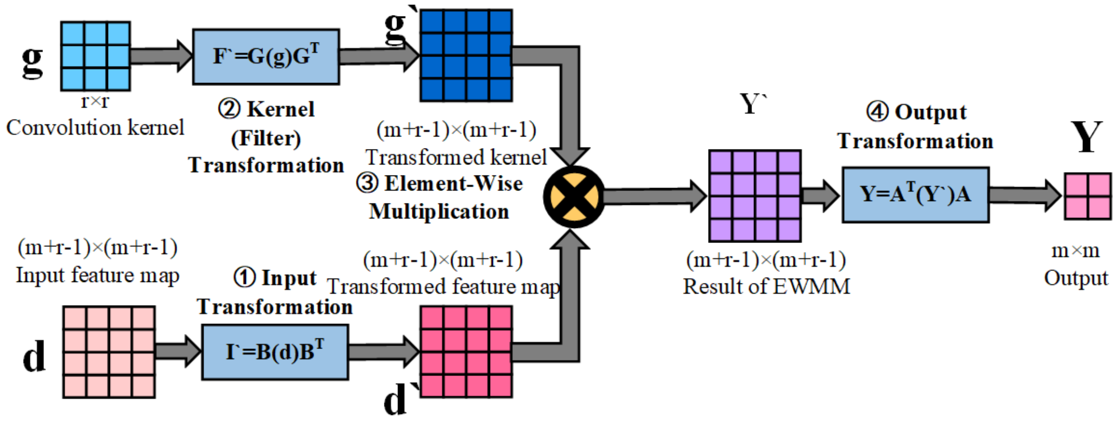


Figure 1. Four stages of Winograd convolution

- **Input Transformation** (**ITrans**): $d' = B^T dB$, transform the input tensor to Winograd domain, the size of d' is $(m + r - 1) \times (m + r - 1)$;
- **Kernel Transformation** (**KTrans**): $g' = GgG^T$, transform the convolution kernel to Winograd domain, the size of g' is $(m + r - 1) \times (m + r - 1)$;
- **Element-Wise Matrix Multiplication** (**EWMM**): $Y' = g' \odot d'$, which is the calculation stage of Winograd convolution, the size of Y' is $(m + r - 1) \times (m + r - 1)$;
- **Output Transformation** (**OTrans**): $Y = A^T Y' A$, inversely transform the result of EWMM from Winograd domain to feature map tensor domain, the size of Y is $m \times m$.
-

[2] was the first that applied the Winograd minimum filtering algorithm to CNN, and the performance of the convolution operator is improved by reducing the number of multiplications. For a two-dimensional convolution operator, the output needs to be divided into $m \times m$ tiles. The input corresponding to the convolution is the input slices of $(m + r - 1) \times (m + r - 1)$ that overlap each other. There is an overlap of $r - 1$ between the input slice and the adjacent slice. According to the analysis of [7], a large slice size and a small convolution kernel size can make the overlap area of repeated calculations less, but it will also bring greater numerical errors. Experiments show that the performance of $F(2 \times 2,3 \times 3)$ on multiple convolutions exceeds cuDNN, and the memory size used is much lower than FFT convolution. As a result, the Winograd convolution was introduced into CNN.

2.3. Generalization and Extension of Winograd Convolution

2.3.1. Generalization

The introduction of Winograd convolution [2] is a milestone, but the convolution only supports two-dimensional convolution operators with $r = 3$ and $r = 2$, and the tilesize does not exceed 6. But this is far from satisfying the various types of convolution operators in modern CNNs, so there are many subsequent studies that generalize Winograd convolution to various types.

[8], [9], [10] realized Winograd convolution to support larger arbitrary convolution kernel size and slice size, [11], [12], [13], [14] proposed a decomposition method to decompose the large convolution kernel size and tile size into several kind of small Winograd convolution. Three-dimensional convolution is used to process time and space feature information and is the main component of 3D-CNN. [14], [15], [16], [17], [18]nested one-dimensional Winograd convolution

and two-dimensional Winograd convolution to obtained three-dimensional Winograd convolution. [8] generalized the Winograd convolution to N-dimensions. Down-sampling often uses a stride convolution operator to reduce the size of the feature map through non-unit step convolution. [19] extended the algorithm to three dimensions while achieving a step size of 2. [10], [11], [12], [13], [14] used matrix decomposition to extend the algorithm to any stride size. [20] combined the decomposition algorithm and nesting to better solve the problem of arbitrary stridesize. Dilated convolution and transposed convolution are often used in image segmentation, super-resolution and other fields. [21] proposed a dilated Winograd convolution to support the dilation of 2 and 4. [22] converted the transposed convolution into multiple basic convolutions through predefined decomposition and interleaving operations and implemented the support of the Winograd convolution for the transposed convolution.

2.3.2. Extension

In addition to generalizing to various convolutions, there are many attempts to extend the linear transformation of Winograd convolution itself. The Winograd algorithm family linearly transforms the input tiles and convolution kernel into the Winograd domain, performs the Hadamard product, and then inversely transforms back to the feature map domain. For the specified convolution kernel and tile size, the linear transformation matrices A, G, and B are known before calculation. Convolution can be expressed as polynomial multiplication. Map the elements of the convolution kernel $g(x)$ and the input vector $d(x)$ to the coefficients of the polynomials $g(x)$ and $d(x)$ respectively, then the elements of the output vector y (convolution of g and d) are equal to the coefficient of the polynomial $y(x) = g(x)d(x)$. The Winograd convolution algorithm family is based on the Chinese Remainder Theorem (CRT) on polynomials. The convolution output can be obtained by taking the remainder of the polynomial in the irreducible and coprime polynomial congruence system. Solving the congruence equations is to obtain the specific solution of the linear transformation matrix according to the coefficients of the polynomial [6].

[23] extended the convolution polynomial used in the Winograd convolution algorithm to a higher-order polynomial. Experiments have shown that second-order polynomials will significantly reduce the error, but will also increase the number of multiplications, so there is a trade-off between the number of multiplications and the accuracy of floating-point numbers. [24] extended polynomial multiplication to the complex number domain and used the symmetry of conjugate complex multiplication to further reduce the number of multiplications. [25] proposed to introduce Winograd convolution into the remainder system (RNS) to realize the low-precision quantization operation of Winograd convolution and support larger input slice size. [26] Innovatively introduced the Fermat Number Transformation (FNT). On the one hand, using this transformation can ensure that the intermediate calculation results are all unsigned numbers, and on the other hand, all calculations are simplified to shift and addition operations.

In addition, [27] combined Winograd convolution with Strassen's algorithm. Strassen algorithm [28] is an algorithm to reduce the number of matrix operations. [2] pointed out that the operation reduction of Strassen algorithm is much smaller than Winograd algorithm, but [27] replaced the convolution used in Strassen algorithm with Winograd convolution and combined the reduction of operations brought by the two to achieve further optimization. [29] applied Winograd convolution to the additive neural network, replacing multiplication with addition, maintaining considerable performance and reducing power consumption.

3. Optimization of Winograd Convolution

3.1. Pruning

Pruning is an effective technique commonly used in CNN optimization. Pruning is mainly used to prune the weights of the convolution operator in CNN, and the weights that have little effect on the output will be set to zero. The convolution kernel after pruning becomes a sparse tensor, which brings two advantages. One is that storing sparse convolution kernel tensor weights in a specific compression format can reduce memory usage, and the other is that many elements in the sparse tensor are 0, so the amount of calculation for convolution can be reduced. For the convolution of the convolutional layer and the fully connected layer, the parameters can be reduced by more than 90%. However, it is difficult to directly apply pruning on Winograd convolution, because the sparse convolution kernel will become a dense matrix after transforming to the Winograd domain, which violates the original intention of pruning.

[5], [30], [31], [32] propose to apply pruning on Winograd convolution and FFT convolution. They applied a linear rectification unit (ReLU) after the input transformation to obtain a sparse Winograd domain tensor. At the same time, the transformed convolution kernel is pruned to obtain a sparse Winograd domain convolution kernel. At this point, the two tensors in the calculation phase have become sparse tensors. [33] applied pooling after the input transformation, the principle is the same as the application of ReLU. [34], [35] designed a new memory data layout for sparse Winograd convolution. [36] proposed to learn the pruning coefficient of Winograd convolution locally and reached a sparse rate of more than 90%. [37] pointed out that the use of ReLU method changed the network layout, they proposed to apply spatial structure pruning on the transformed feature map tensor, and then transfer its sparsity to the convolution kernel of the Winograd domain. [10], [38] proposed that the results of the pruning of the above methods are irregular, which is not conducive to the performance of hardware, so the position-sensitive sub-row balance coefficient pruning mode and sparse row balance compression are respectively proposed. [39] proposed a new coding format to solve the coding overhead caused by the sparsity of the active area. [40] introduced the Zero-Skip hardware mechanism, skipped the calculation of 0 weight, and provided hardware support for the sparse matrix operation after pruning.

3.2. Low Precision and Quantization

It is another common method in CNN to sacrifice precision and reduce memory footprint and computational efficiency. Changing the parameters of the CNN model from a 32-bit floating-point number to a 16-bit floating-point number or quantizing it to an 8-bit fixed-point number or even lower precision, can compress the model and improve computing efficiency without losing the accuracy of the model. When Winograd convolution was first introduced, single-precision and half-precision floating-point numbers were tested at the same time, but experiments have shown that using half-precision floating-point numbers will lead to larger absolute errors [2]. Winograd convolution can also be combined with quantization. [24] proposed a uniform affine quantization to generate a quantized convolution kernel and expressed in 8-bit unsigned integer and dynamic range. [41] proposed dynamic layered application of different convolution implementation and quantization on CNN to reduce computational complexity, including the quantization of Winograd convolution. [42] proposed to apply Winograd convolution to an 8-bit network and use learning to solve the problem of accuracy loss. [25] The introduction of RNS transformation also enables quantization to operate at low precision. [43] proposed to model the accuracy loss and use different quantization levels for feature maps and convolution kernels. [44] replaced the canonical basis polynomials in the Winograd transform with Legendre basis polynomials, and proposed quantization based on basis transformation techniques. [45] embed linear quantization directly into the Winograd domain to achieve low-precision quantization. [46] further explored the use of the Winograd algorithm to optimize the convolution kernel with 4-6-bit precision. [47] applied quantization on feature map slices, and applied particle swarm optimization technology to find the threshold of quantization. In addition, [10], [48], [49], [50] also applied 8-bit quantization technique on Winograd convolution.

3.3. Numerical Stability

Winograd convolution has only been applied to the 3×3 convolution kernel and small input tiles for a long time, because of the inherent numerical instability in the Winograd convolution calculation. When applied to larger convolution kernels or input tiles, the polynomial coefficients of the Winograd transform increase exponentially. This imbalance will be reflected in the elements of the transformation matrix, resulting in large relative errors. [7] studied that the source of this numerical instability is the large-scale Vandermonde matrix in the transformation [3] and proposed carefully selecting the corresponding polynomials that exhibit the smallest exponential growth. They also proposed scaling the transformation matrix to alleviate numerical instability. [23] used higher-order polynomials to reduce the error of Winograd convolution, but the cost was an increase in the number of multiplications. [42] handed over the processing of numerical errors to training to learn better convolution kernel weights and quantization in Winograd convolution. [51] proved mathematically that large convolution kernels can be solved by overlap and addition. [20], [52] solved large-size convolution kernel and non-unit step convolution into small convolution kernels to solve the numerical accuracy problem. [53] selected the appropriate output tile size based on symbolic calculation and meta-programming automation to balance numerical stability and efficiency. [54] proved that the floating-point calculation order in linear transformation affects accuracy, rearranged the calculation order in linear transformation based on Huffman coding, and proposed a mixed-precision algorithm.

4. Implementations and Applications of Winograd Convolution

4.1. Implementation

The high performance brought by Winograd convolution allows researchers to quickly deploy it to various platforms, in addition to CPUs, GPUs, etc., but also FPGA platforms, mobile terminals, and edge computing devices that have strict requirements for efficiency and power consumption.

4.1.1. CPU

[5], [36] implemented pruning and retraining of Winograd convolution on the CPU. [55], [56] compared the performance of FFT and Winograd on CPU, and their performance characteristics were analysed. [15] implemented three-dimensional Winograd convolution on a multi-core CPU by using the specific large memory of the CPU platform. [57] use JIT optimization technology to accelerate the realization of the direct convolution kernel Winograd convolution on the small convolution kernel on the x86 CPU architecture. [58] relied on the automatic vectorization of the compiler, the calculation stage was converted to batched GEMM to achieve performance improvement. [8] proposed a custom data layout on the CPU, using vectorized instructions to achieve efficient memory access. [59] used the L3-Cache of the CPU to reuse the convolution kernel, but it cannot support convolution with an excessively large number of channels. [60] utilized the similarity in Winograd convolution to achieve deep data reuse on the CPU. In

addition, [61], [62] implemented Winograd convolution and efficient inference libraries on the mobile ARM platform CPU.

4.1.2. GPU

[63] performed similar performance comparisons with [55], [56] on the GPU. [16], [17] implemented the three-dimensional Winograd convolution on the GPU, but the calculation stage in [16] directly called the matrix multiplication implementation of the cuBLAS library, and [17] manually wrote a specific implementation. The data parallelism and intra-slice parallelism of Winograd convolution are used in [64] to achieve multi-dimensional parallel training on large-scale GPU clusters. [65] used MegaKernel technique to fuse the four stages of Winograd convolution and used a well-designed task mapping algorithm to achieve a significant performance improvement on the GPU. [66] used SASS assembler to optimize Winograd convolution, merge global memory access and make shared memory access conflict-free, used cache to design pipeline and to improve calculation intensity, and used conventional registers to fill the shortcomings of insufficient predicate registers. [32] realized the pruning technology on the GPU and proposed a dynamic batch size algorithm to improve the training speed. [53] also implemented Winograd convolution for mobile GPUs.

4.1.3. FPGA

[67] described in detail the minimum requirements for Winograd convolution hardware implementation and implemented the basic modules of Winograd convolution on FPGA. [68] cached all intermediate feature mappings in the stream buffer to achieve a high energy consumption ratio FPGA implementation. [69] designed a line buffer structure to cache feature maps and reuse data from different slices, and designed an efficient parallel execution unit for Winograd convolution, and dealt with the sparse case in [34]. [14], [70], [71] unified the two-dimensional and three-dimensional Winograd, and built a unified template on the FPGA. [72] implemented hybrid convolution on FPGA and analysed the occasions suitable for FFT and Winograd convolution. [35], [73], [74], [75] unified the realization of the Winograd convolution kernel matrix multiplication and maximize the reusability of the module. [76], [77] conducted a comprehensive design space exploration on the realization of Winograd convolution on FPGA. [11] proposed a decomposition method for the convolution of various parameters, which simplifies the hardware implementation. [78] designed an open-source back-end framework on the CPU-FPGA heterogeneous platform, but only supports Winograd convolution with unit stride size, while [43] implements fine-grained scheduling and supports more general convolution. In addition, [79], [80], [81], [82], [83] and [84] also implemented Winograd convolution on FPGA and explored the design space, and [85] fully evaluated the complete CNN implementation on FPGA.

4.1.4. Other

[40], [86] utilized the SIMT architecture of GPGPU to process CNN using Winograd convolution in parallel, and [40] also added support for Zero-Skip. [87], [88] used high-efficiency Winograd convolution on IoT devices to achieve high performance. [41], [89], [90] used random calculation and approximate calculation to complete the implementation. [91] is implemented on ReRAM, which improves data reuse based on tiles, and [48] implemented 8-bit quantized convolution based on DRAM architecture. [18] realized the three-dimensional Winograd convolution on the vector DSP. By expanding a $F(2 \times 2,3 \times 3)$ convolution instruction and adding a new calculation module, [92] implemented Winograd convolution on an open source RISC-V framework.

Many frameworks integrate Winograd convolution to improve model execution efficiency. Cltorch [4] is a hardware-independent back-end platform based on OpenCL. [93] implemented a tool for mapping the Caffe model to FPGA and chose whether to apply Winograd convolution based on dynamic programming. [94] implemented a software stack supporting Winograd convolution, generating high-efficiency load for Intel hardware including CPU, integrated display, and neural computing stick. In addition, Winograd convolution has been integrated into popular deep learning frameworks and neural network libraries.

4.2. Applications

The fast convolution represented by Winograd aims to accelerate the convolution to improve the execution efficiency of the CNN model. It can be used in scenarios that require real-time performance. [72] implemented a face recognition system using hybrid convolution. [95] Implemented an accelerator for action recognition based on three-dimensional Winograd convolution. [22,96] introduced Winograd convolution into real-time super-resolution, but there is a difference in upsampling. [22] uses the Winograd implementation of transposed convolution, while [96] uses shuffle layer instead of transposed convolution. [49] implemented a speech recognition accelerator on wearable devices and applied an 8-bit integer Winograd convolutional network.

5. CONCLUSION

Winograd convolution is currently the most widely used fast convolution operator. Since the introduction of CNN, its scope of use has gradually covered all types of convolutions in modern CNN with the in-depth research of researchers, and the combination with pruning, quantization and other technologies has also matured. Winograd convolution has been integrated in various platform deep learning frameworks and neural network libraries, which can generate efficient workloads for various hardware platforms. On the FPGA platform, it is possible to customize the implementation of software and hardware coordination for Winograd convolution, but how to make good use of computing power and memory levels on general computing platforms such as CPU and GPU remains to be further studied. For example, the four-stage integration of Winograd, the optimization of data flow, the trade-off between computational intensity and memory access efficiency, may all be a breakthrough in optimization on a general-purpose computing platform.

REFERENCES

[1] M. Mathieu, M. Henaff, and Y. LeCun, “Fast Training of Convolutional Networks through FFTs,” ArXiv13125851 Cs, Mar. 2014, Accessed: Sep. 29, 2021. [Online]. Available: http://arxiv.org/abs/1312.5851

[2] A. Lavin and S. Gray, “Fast Algorithms for Convolutional Neural Networks,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 4013–4021. doi: 10.1109/CVPR.2016.435.

[3] V. Y. Pan, “How Bad Are VandermondeMatrices?,” SIAM J. Matrix Anal. Appl., vol. 37, no. 2, pp. 676–694, Jan. 2016, doi: 10.1137/15M1030170.

[4] H. Perkins, “cltorch: a Hardware-Agnostic Backend for the Torch Deep Neural Network Library, Based on OpenCL,” ArXiv160604884 Cs, Jun. 2016, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1606.04884

[5] X. Liu and Y. Turakhia, “Pruning of Winograd and FFT Based Convolution Algorithm,” p. 7, Jun. 2016.

[6] S. Winograd, Arithmetic Complexity of Computations. Society for Industrial and Applied Mathematics, 1980. doi: 10.1137/1.9781611970364.

[7] K. Vincent, K. Stephano, M. Frumkin, B. Ginsburg, and J. Demouth, "ON IMPROVING THE NUMERICAL STABILITY OF WINOGRAD CONVOLUTIONS," p. 4, 2017.

[8] Z. Jia, A. Zlateski, F. Durand, and K. Li, "Optimizing N-dimensional, winograd-based convolution for manycore CPUs," in Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, Vienna Austria, Feb. 2018, pp. 109 – 123. doi: 10.1145/3178487.3178496.

[9] A. Cariow and G. Cariowa, "Minimal Filtering Algorithms for Convolutional Neural Networks," ArXiv, p. 11, 2020.

[10] D. Wu, X. Fan, W. Cao, and L. Wang, "SWM: A High-Performance Sparse-Winograd Matrix Multiplication CNN Accelerator," IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 29, no. 5, pp. 936–949, May 2021, doi: 10.1109/TVLSI.2021.3060041.

[11] C. Yang, Y. Wang, X. Wang, and L. Geng, "WRA: A 2.2-to-6.3 TOPS Highly Unified Dynamically Reconfigurable Accelerator Using a Novel Winograd Decomposition Algorithm for Convolutional Neural Networks," IEEE Trans. Circuits Syst. Regul. Pap., vol. 66, no. 9, pp. 3480–3493, Sep. 2019, doi: 10.1109/TCSI.2019.2928682.

[12] C. Yang, Y. Wang, X. Wang, and L. Geng, "A Stride-Based Convolution Decomposition Method to Stretch CNN Acceleration Algorithms for Efficient and Flexible Hardware Implementation," IEEE Trans. Circuits Syst. Regul. Pap., vol. 67, no. 9, pp. 3007–3020, Sep. 2020, doi: 10.1109/TCSI.2020.2985727.

[13] J. Pan and D. Chen, "Accelerate Non-unit Stride Convolutions with Winograd Algorithms," in Proceedings of the 26th Asia and South Pacific Design Automation Conference, Tokyo Japan, Jan. 2021, pp. 358–364. doi: 10.1145/3394885.3431534.

[14] H. Deng, J. Wang, H. Ye, S. Xiao, X. Meng, and Z. Yu, "3D-VNPU: A Flexible Accelerator for 2D/3D CNNs on FPGA," in 2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), Orlando, FL, USA, May 2021, pp. 181–185. doi: 10.1109/FCCM51124.2021.00029.

[15] D. Budden, A. Matveev, S. Santurkar, S. R. Chaudhuri, and N. Shavit, "Deep Tensor Convolution on Multicores," Proc. 34 Th Int. Conf. Mach. Learn., p. 10, 2017.

[16] Q. Lan, Z. Wang, M. Wen, C. Zhang, and Y. Wang, "High Performance Implementation of 3D Convolutional Neural Networks on a GPU," Comput. Intell. Neurosci., vol. 2017, pp. 1–8, 2017, doi: 10.1155/2017/8348671.

[17] Z. Wang, Q. Lan, H. He, and C. Zhang, "Winograd Algorithm for 3D Convolution Neural Networks," in Artificial Neural Networks and Machine Learning – ICANN 2017, vol. 10614, A. Lintas, S. Rovetta, P. F. M. J. Verschure, and A. E. P. Villa, Eds. Cham: Springer International Publishing, 2017, pp. 609–616. doi: 10.1007/978-3-319-68612-7_69.

[18] W. Chen, Y. Wang, C. Yang, and Y. Li, "Hardware Acceleration Implementation of Three-Dimensional Convolutional Neural Network on Vector Digital Signal Processors," in 2020 4th International Conference on Robotics and Automation Sciences (ICRAS), Wuhan, China, Jun. 2020, pp. 122–129. doi: 10.1109/ICRAS49812.2020.9135062.

[19] J. Yepez and S.-B. Ko, "Stride 2 1-D, 2-D, and 3-D Winograd for Convolutional Neural Networks," IEEE Trans. Very Large Scale Integr. VLSI Syst., p. 11, 2020.

[20] J. Jiang, X. Chen, and C.-Y. Tsui, "A Reconfigurable Winograd CNN Accelerator with Nesting Decomposition Algorithm for Computing Convolution with Large Filters," ArXiv, p. 6, 2021.

[21] M. Kim, C. Park, S. Kim, T. Hong, and W. W. Ro, "Efficient Dilated-Winograd Convolutional Neural Networks," in 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, Sep. 2019, pp. 2711–2715. doi: 10.1109/ICIP.2019.8803277.

[22] B. Shi, Z. Tang, G. Luo, and M. Jiang, "Winograd-Based Real-Time Super-Resolution System on FPGA," in 2019 International Conference on Field-Programmable Technology (ICFPT), Tianjin, China, Dec. 2019, pp. 423–426. doi: 10.1109/ICFPT47387.2019.00083.

[23] B. Barabasz and D. Gregg, Winograd Convolution for DNNs - Beyond Linear Polynomials, vol. 11946. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-35166-3.

[24] L. Meng and J. Brothers, "Efficient Winograd Convolution via Integer Arithmetic," ArXiv190101965 Cs, Jan. 2019, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1901.01965

[25] Z.-G. Liu and M. Mattina, "Efficient Residue Number System Based Winograd Convolution," in Computer Vision – ECCV 2020, vol. 12364, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 53–68. doi: 10.1007/978-3-030-58529-7_4.

[26] W. Xu, Z. Zhang, X. You, and C. Zhang, “Reconfigurable and Low-Complexity Accelerator for Convolutional and Generative Networks Over Finite Fields,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 39, no. 12, pp. 4894–4907, Dec. 2020, doi: 10.1109/TCAD.2020.2973355.

[27] Y. Zhao, D. Wang, and L. Wang, “Convolution Accelerator Designs Using Fast Algorithms,” Algorithms, vol. 12, no. 5, p. 112, May 2019, doi: 10.3390/a12050112.

[28] V. Trassen, “Gaussian elimination is not optimal,” p. 3.

[29] W. Li, H. Chen, M. Huang, X. Chen, C. Xu, and Y. Wang, “Winograd Algorithm for AdderNet,” ArXiv210505530 Cs, May 2021, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/2105.05530

[30] X. Liu, J. Pool, S. Han, and W. J. Dally, “Efficient Sparse-Winograd Convolutional Neural Networks,” ArXiv180206367 Cs, Feb. 2018, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1802.06367

[31] H. Wang, W. Liu, T. Xu, J. Lin, and Z. Wang, “A Low-latency Sparse-Winograd Accelerator for Convolutional Neural Networks,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, May 2019, pp. 1448–1452. doi: 10.1109/ICASSP.2019.8683512.

[32] S. Zheng, L. Wang, and G. Gupta, “Efficient Ensemble Sparse Convolutional Neural Networks with Dynamic Batch Size,” in Computer Vision and Image Processing, vol. 1378, S. K. Singh, P. Roy, B. Raman, and P. Nagabhushan, Eds. Singapore: Springer Singapore, 2021, pp. 262–277. doi: 10.1007/978-981-16-1103-2_23.

[33] Y. Choi, M. El-Khamy, and J. Lee, “Jointly Sparse Convolutional Neural Networks in Dual Spatial-winograd Domains,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, May 2019, pp. 2792–2796. doi: 10.1109/ICASSP.2019.8682922.

[34] L. Lu and Y. Liang, “SpWA: an efficient sparse winograd convolutional neural networks accelerator on FPGAs,” in Proceedings of the 55th Annual Design Automation Conference, San Francisco California, Jun. 2018, pp. 1–6. doi: 10.1145/3195970.3196120.

[35] F. Shi, H. Li, Y. Gao, B. Kuschner, and S.-C. Zhu, “Sparse Winograd Convolutional neural networks on small-scale systolic arrays,” ArXiv181001973 Cs, Oct. 2018, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1810.01973

[36] S. Li, J. Park, and P. T. P. Tang, “Enabling Sparse Winograd Convolution by Native Pruning,” ArXiv170208597 Cs, Oct. 2017, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1702.08597

[37] J. Yu, J. Park, and M. Naumov, “Spatial-Winograd Pruning Enabling Sparse Winograd Convolution,” ArXiv190102132 Cs, Jan. 2019, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1901.02132

[38] T. Yang, Y. Liao, J. Shi, Y. Liang, N. Jing, and L. Jiang, “A Winograd-Based CNN Accelerator with a Fine-Grained Regular Sparsity Pattern,” in 2020 30th International Conference on Field-Programmable Logic and Applications (FPL), Gothenburg, Sweden, Aug. 2020, pp. 254–261. doi: 10.1109/FPL50879.2020.00050.

[39] X. Wang, C. Wang, J. Cao, L. Gong, and X. Zhou, “WinoNN: Optimizing FPGA-Based Convolutional Neural Network Accelerators Using Sparse Winograd Algorithm,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 39, no. 11, pp. 4290–4302, Nov. 2020, doi: 10.1109/TCAD.2020.3012323.

[40] H. Park, D. Kim, J. Ahn, and S. Yoo, “Zero and data reuse-aware fast convolution for deep neural networks on GPU,” in Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis, Pittsburgh Pennsylvania, Oct. 2016, pp. 1–10. doi: 10.1145/2968456.2968476.

[41] Y. Gong, B. Liu, W. Ge, and L. Shi, “ARA: Cross-Layer approximate computing framework based reconfigurable architecture for CNNs,” Microelectron. J., vol. 87, pp. 33–44, May 2019, doi: 10.1016/j.mejo.2019.03.011.

[42] J. Fernandez-Marques, P. N. Whatmough, A. Mundy, and M. Mattina, “Searching for Winograd-aware Quantized Networks,” MLSys, p. 16, 2020.

[43] W. Zhang, X. Liao, and H. Jin, “Fine-grained Scheduling in FPGA-Based Convolutional Neural Networks,” in 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, Apr. 2020, pp. 120–128. doi: 10.1109/ICCCBDA49378.2020.9095680.

[44] B. Barabasz, “Quantaized Winograd/Toom-Cook Convolution for DNNs: Beyond Canonical Polynomials Base,” ArXiv200411077 Cs Math Stat, Apr. 2020, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/2004.11077

[45] G. Li, L. Liu, X. Wang, X. Ma, and X. Feng, “Lance: efficient low-precision quantized winograd convolution for neural networks based on graphics processing units,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, May 2020, pp. 3842–3846. doi: 10.1109/ICASSP40776.2020.9054562.

[46] Q. Han et al., “Extremely Low-bit Convolution Optimization for Quantized Neural Network on Modern Computer Architectures,” in 49th International Conference on Parallel Processing - ICPP, Edmonton AB Canada, Aug. 2020, pp. 1–12. doi: 10.1145/3404397.3404407.

[47] D. Sabir, M. A. Hanif, A. Hassan, S. Rehman, and M. Shafique, “TiQSA: Workload Minimization in Convolutional Neural Networks Using Tile Quantization and Symmetry Approximation,” IEEE Access, vol. 9, pp. 53647–53668, 2021, doi: 10.1109/ACCESS.2021.3069906.

[48] M. M. Ghaffar, C. Sudarshan, C. Weis, M. Jung, and N. Wehn, “A Low Power In-DRAM Architecture for Quantized CNNs using Fast Winograd Convolutions,” in The International Symposium on Memory Systems, Washington DC USA, Sep. 2020, pp. 158–168. doi: 10.1145/3422575.3422790.

[49] Y. Yao et al., “INT8 Winograd Acceleration for Conv1D Equipped ASR Models Deployed on Mobile Devices,” ArXiv201014841 Cs Eess, Oct. 2020, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/2010.14841

[50] Y. Cao, C. Song, and Y. Tang, “Efficient LUT-based FPGA Accelerator Design for Universal Quantized CNN Inference,” in 2021 2nd Asia Service Sciences and Software Engineering Conference, Macau Macao, Feb. 2021, pp. 108–115. doi: 10.1145/3456126.3456140.

[51] C. Ju and E. Solomonik, “Derivation and Analysis of Fast Bilinear Algorithms for Convolution,” SIAM Rev., vol. 62, no. 4, pp. 743–777, Jan. 2020, doi: 10.1137/19M1301059.

[52] D. Huang et al., “DWM: A Decomposable Winograd Method for Convolution Acceleration,” Proc. AAAI Conf. Artif. Intell., vol. 34, no. 04, pp. 4174–4181, Apr. 2020, doi: 10.1609/aaai.v34i04.5838.

[53] A. Mazaheri, T. Beringer, M. Moskewicz, F. Wolf, and A. Jannesari, “Accelerating winograd convolutions using symbolic computation and meta-programming,” in Proceedings of the Fifteenth European Conference on Computer Systems, Heraklion Greece, Apr. 2020, pp. 1–14. doi: 10.1145/3342195.3387549.

[54] B. Barabasz, A. Anderson, K. M. Soodhalter, and D. Gregg, “Error Analysis and Improving the Accuracy of Winograd Convolution for Deep Neural Networks,” ACM Trans. Math. Softw., vol. 46, no. 4, pp. 1–33, Nov. 2020, doi: 10.1145/3412380.

[55] A. Zlateski, Z. Jia, K. Li, and F. Durand, “FFT Convolutions are Faster than Winograd on Modern CPUs, Here is Why,” ArXiv180907851 Cs, Sep. 2018, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1809.07851

[56] A. Zlateski, Z. Jia, K. Li, and F. Durand, “The anatomy of efficient FFT and winograd convolutions on modern CPUs,” in Proceedings of the ACM International Conference on Supercomputing, Phoenix Arizona, Jun. 2019, pp. 414–424. doi: 10.1145/3330345.3330382.

[57] A. Heinecke et al., “Understanding the Performance of Small Convolution Operations for CNN on Intel Architecture,” SC’17, p. 2, 2017.

[58] S. N. Ragate, “Optimization of Spatial Convolution in ConvNets on Intel KNL,” 2017.

[59] R. Gelashvili, N. Shavit, and A. Zlateski, “L3 Fusion: Fast Transformed Convolutions on CPUs,” ArXiv191202165 Cs, Dec. 2019, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/1912.02165

[60] R. Wu, F. Zhang, Z. Zheng, X. Du, and X. Shen, “Exploring deep reuse in winograd CNN inference,” in Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, Virtual Event Republic of Korea, Feb. 2021, pp. 483–484. doi: 10.1145/3437801.3441588.

[61] P. Maji, A. Mundy, G. Dasika, J. Beu, M. Mattina, and R. Mullins, “Efficient Winograd or Cook-Toom Convolution Kernel Implementation on Widely Used Mobile CPUs,” in 2019 2nd Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2), Washington, DC, USA, Feb. 2019, pp. 1–5. doi: 10.1109/EMC249363.2019.00008.

[62] H. Lan et al., “FeatherCNN: Fast Inference Computation with TensorGEMM on ARM Architectures,” IEEE Trans. Parallel Distrib. Syst., vol. 31, no. 3, pp. 580–594, Mar. 2020, doi: 10.1109/TPDS.2019.2939785.

[63] H. Kim, H. Nam, W. Jung, and J. Lee, “Performance Analysis of CNN Frameworks for GPUs,” p. 10, 2017.

[64] B. Hong, Y. Ro, and J. Kim, “Multi-dimensional Parallel Training of Winograd Layer on Memory-Centric Architecture,” in 2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Fukuoka, Oct. 2018, pp. 682–695. doi: 10.1109/MICRO.2018.00061.

[65] L. Jia, Y. Liang, X. Li, L. Lu, and S. Yan, “Enabling Efficient Fast Convolution Algorithms on GPUs via MegaKernels,” IEEE Trans. Comput., pp. 1–1, 2020, doi: 10.1109/TC.2020.2973144.

[66] D. Yan, W. Wang, and X. Chu, “Optimizing batched winograd convolution on GPUs,” in Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, San Diego California, Feb. 2020, pp. 32–44. doi: 10.1145/3332466.3374520.

[67] A. CARIOW and G. CARIOWA, “Hardware-Efficient Structure of the Accelerating Module for Implementation of Convolutional Neural Network Basic Operation.pdf,” Meas. Autom. Monit., 2017.

[68] U. Aydonat, S. O’Connell, D. Capalija, A. C. Ling, and G. R. Chiu, “An OpenCLTM Deep Learning Accelerator on Arria 10,” in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey California USA, Feb. 2017, pp. 55–64. doi: 10.1145/3020078.3021738.

[69] L. Lu, Y. Liang, Q. Xiao, and S. Yan, “Evaluating Fast Algorithms for Convolutional Neural Networks on FPGAs,” presented at the 2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines, Salt Lake City, UT, May 2017.

[70] J. Shen, Y. Huang, Z. Wang, Y. Qiao, M. Wen, and C. Zhang, “Towards a Uniform Template-based Architecture for Accelerating 2D and 3D CNNs on FPGA,” in Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey CALIFORNIA USA, Feb. 2018, pp. 97–106. doi: 10.1145/3174243.3174257.

[71] J. Shen, Y. Huang, M. Wen, and C. Zhang, “Toward an Efficient Deep Pipelined Template-Based Architecture for Accelerating the Entire 2-D and 3-D CNNs on FPGA,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 39, no. 7, pp. 1442–1455, Jul. 2020, doi: 10.1109/TCAD.2019.2912894.

[72] C. Zhuge, X. Liu, X. Zhang, S. Gummadi, J. Xiong, and D. Chen, “Face Recognition with Hybrid Efficient Convolution Algorithms on FPGAs,” in Proceedings of the 2018 on Great Lakes Symposium on VLSI, Chicago IL USA, May 2018, pp. 123–128. doi: 10.1145/3194554.3194597.

[73] S. Kala, J. Mathew, B. R. Jose, and S. Nalesh, “UniWiG: Unified Winograd-GEMM Architecture for Accelerating CNN on FPGAs,” in 2019 32nd International Conference on VLSI Design and 2019 18th International Conference on Embedded Systems (VLSID), Delhi, NCR, India, Jan. 2019, pp. 209–214. doi: 10.1109/VLSID.2019.00055.

[74] S. Kala, B. R. Jose, J. Mathew, and S. Nalesh, “High-Performance CNN Accelerator on FPGA Using Unified Winograd-GEMM Architecture,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 27, no. 12, pp. 2816–2828, Dec. 2019, doi: 10.1109/TVLSI.2019.2941250.

[75] S. Kala and S. Nalesh, “Efficient CNN Accelerator on FPGA,” IETE J. Res., vol. 66, no. 6, pp. 733–740, Nov. 2020, doi: 10.1080/03772063.2020.1821797.

[76] A. Ahmad and M. A. Pasha, “Towards Design Space Exploration and Optimization of Fast Algorithms for Convolutional Neural Networks (CNNs) on FPGAs,” in 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, Mar. 2019, pp. 1106–1111. doi: 10.23919/DATE.2019.8715272.

[77] X. Liu, Y. Chen, C. Hao, A. Dhar, and D. Chen, “WinoCNN: Kernel Sharing Winograd Systolic Array for Efficient Convolutional Neural Network Acceleration on FPGAs,” ArXiv210704244 Cs, Jul. 2021, Accessed: Sep. 19, 2021. [Online]. Available: http://arxiv.org/abs/2107.04244

[78] R. DiCecco, G. Lacey, J. Vasiljevic, P. Chow, G. Taylor, and S. Areibi, “Caffeinated FPGAs: FPGA framework For Convolutional Neural Networks,” in 2016 International Conference on Field-Programmable Technology (FPT), Xi’an, China, Dec. 2016, pp. 265–268. doi: 10.1109/FPT.2016.7929549.

[79] A. Podili, C. Zhang, and V. Prasanna, “Fast and efficient implementation of Convolutional Neural Networks on FPGA,” in 2017 IEEE 28th International Conference on Application-specific Systems, Architectures and Processors (ASAP), Seattle, WA, USA, Jul. 2017, pp. 11–18. doi: 10.1109/ASAP.2017.7995253.

[80] Y. Huang, J. Shen, Z. Wang, M. Wen, and C. Zhang, “A High-efficiency FPGA-based Accelerator for Convolutional Neural Networks using Winograd Algorithm,” J. Phys. Conf. Ser., vol. 1026, p. 012019, May 2018, doi: 10.1088/1742-6596/1026/1/012019.

[81] M. R. Vemparala, A. Frickenstein, and W. Stechele, “An Efficient FPGA Accelerator Design for Optimized CNNs Using OpenCL,” in Architecture of Computing Systems – ARCS 2019, vol. 11479, M. Schoeberl, C. Hochberger, S. Uhrig, J. Brehm, and T. Pionteck, Eds. Cham: Springer International Publishing, 2019, pp. 236–249. doi: 10.1007/978-3-030-18656-2_18.

[82] Z. Bai, H. Fan, L. Liu, L. Liu, and D. Wang, “An OpenCL-Based FPGA Accelerator with the Winograd’s Minimal Filtering Algorithm for Convolution Neuron Networks,” in 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, Dec. 2019, pp. 277–282. doi: 10.1109/ICCC47050.2019.9064413.

[83] A. Ahmad and M. A. Pasha, “FFConv: An FPGA-based Accelerator for Fast Convolution Layers in Convolutional Neural Networks,” ACM Trans. Embed. Comput. Syst., vol. 19, no. 2, pp. 1–24, Mar. 2020, doi: 10.1145/3380548.

[84] H. Ye, X. Zhang, Z. Huang, G. Chen, and D. Chen, “HybridDNN: A Framework for High-Performance Hybrid DNN Accelerator Design and Implementation,” in 2020 57th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jul. 2020, pp. 1–6. doi: 10.1109/DAC18072.2020.9218684.

[85] P. Xiyuan, Y. Jinxiang, Y. Bowen, L. Liansheng, and P. Yu, ″A Review of FPGA ‑ Based Custom Computing Architecture for Convolutional Neural Network Inference,″ Chin. J. Electron., vol. 30, no. 1, pp. 1‒17, Jan. 2021, doi: 10.1049/cje.2020.11.002.

[86] H. Jeon, K. Lee, S. Han, and K. Lee, “The parallelization of convolution on a CNN using a SIMT based GPGPU,” in 2016 International SoC Design Conference (ISOCC), Jeju, South Korea, Oct. 2016, pp. 333–334. doi: 10.1109/ISOCC.2016.7799813.

[87] A. Xygkis, L. Papadopoulos, D. Moloney, D. Soudris, and S. Yous, “Efficient winograd-based convolution kernel implementation on edge devices,” in Proceedings of the 55th Annual Design Automation Conference, San Francisco California, Jun. 2018, pp. 1–6. doi: 10.1145/3195970.3196041.

[88] G. Mahale, P. Udupa, K. K. Chandrasekharan, and S. Lee, “WinDConv: A Fused Datapath CNN Accelerator for Power-Efficient Edge Devices,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 39, no. 11, pp. 4278–4289, Nov. 2020, doi: 10.1109/TCAD.2020.3013096.

[89] H. Wang, Z. Zhang, X. You, and C. Zhang, “Low-Complexity Winograd Convolution Architecture Based on Stochastic Computing,” in 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, Nov. 2018, pp. 1–5. doi: 10.1109/ICDSP.2018.8631556.

[90] G. Lentaris, G. Chatzitsompanis, V. Leon, K. Pekmestzi, and D. Soudris, “Combining Arithmetic Approximation Techniques for Improved CNN Circuit Design,” in 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Glasgow, UK, Nov. 2020, pp. 1–4. doi: 10.1109/ICECS49266.2020.9294869.

[91] J. Lin, S. Li, X. Hu, L. Deng, and Y. Xie, “CNNWire: Boosting Convolutional Neural Network with Winograd on ReRAM based Accelerators,” in Proceedings of the 2019 on Great Lakes Symposium on VLSI, Tysons Corner VA USA, May 2019, pp. 283–286. doi: 10.1145/3299874.3318018.

[92] S. Wang, J. Zhu, Q. Wang, C. He, and T. T. Ye, “Customized Instruction on RISC-V for Winograd-Based Convolution Acceleration,” in 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP), NJ, USA, Jul. 2021, pp. 65–68. doi: 10.1109/ASAP52443.2021.00018.

[93] Q. Xiao, Y. Liang, L. Lu, S. Yan, and Y.-W. Tai, “Exploring Heterogeneous Algorithms for Accelerating Deep Convolutional Neural Networks on FPGAs,” in Proceedings of the 54th Annual Design Automation Conference 2017, Austin TX USA, Jun. 2017, pp. 1–6. doi: 10.1145/3061639.3062244.

[94] A. Demidovskij et al., “OpenVINO Deep Learning Workbench: Comprehensive Analysis and Tuning of Neural Networks Inference,” in 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), Oct. 2019, pp. 783–787. doi: 10.1109/ICCVW.2019.00104.

[95] M. Lou, J. Li, G. Wang, and G. He, “AR-C3D: Action Recognition Accelerator for Human-Computer Interaction on FPGA,” in 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, May 2019, pp. 1–4. doi: 10.1109/ISCAS.2019.8702353.

[96] P.-W. Yen, Y.-S. Lin, C.-Y. Chang, and S.-Y. Chien, “Real-time Super Resolution CNN Accelerator with Constant Kernel Size Winograd Convolution,” 2020 2nd IEEE Int. Conf. Artif. Intell. Circuits Syst. AICAS, p. 5, 2020.

AUTHORS

GanTong (1995-) is a postgraduate of National University of Defense Technology. His main research area is CNN acceleration on different computer architectures.

Libo Huang (1983-) is a master supervisor of National University of Defense Technology. He is an associate researcher and mainly specializes in computer architecture.

A CONTEXT-AWARE AND IMMERSIVE PUZZLE GAME USING MACHINE LEARNING AND BIG DATA ANALYSIS

Peiyi Li[1], Peilin Li[2], John Morris[3] and Yu Sun[3]

[1]University of California, Irvine, Irvine, CA 92697
[2]Coding Minds Academy, Irvine, CA 92620
[3]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

Recent years, video games have become one of the main forms of entertainment for people of all ages, in which millions of members publicly show their screenshots while playing games or share their experience of playing games [4]. Puzzle game is a popular game genre among various video games, it challenges players to find the correct solution by providing them with different logic/conceptual problems. However, designing a good puzzle game is not an easy task [5]. This paper designs a puzzle game for players of all age ranges with proper difficulty level, various puzzle mechanics and attractive background setting stories. We applied our games to different players to test play and conducted a qualitative evaluation of the approach. The results show that the pace of puzzle games affects play experience a lot and the difficulty level of the puzzles affects players' feelings to the game.

KEYWORDS

Puzzle game, game design, video games, adventure game.

1. INTRODUCTION

Puzzle game is a game genre that requires players to find at least one correct solution in order to solve the challenge faced [6]. Among all the game genres, puzzle games mainly concentrated on logical and conceptual challenges [7]. Puzzle games can not only practice players' abilities to use their brains but also improve players' visual-spatial reasoning. By solving complicated puzzles, players will also gain a huge sense of accomplishment. The topic we're going to discuss in this paper is the pace of puzzle games and how the difficulty level of the puzzles affects users' game experience while playing puzzle games. As game developers, we have to be responsible for not only those basic game elements like scripts and game mechanics but also consider the pace of game flow and the difficulty level that will affect game time and users' game experience. The proper difficulty level can make players sandwiched between the desire of conquering difficulties and trying to give up when seeing no hope of solving the problem. This topic will also discuss how to use plots to connect the entire game among different scenes and choose different puzzles in different scenes that make the scene look reasonable. Players will feel weird if seeing a pipe puzzle inside the bedroom but feel reasonable if seeing a number lock. That's why choosing proper puzzles is important while designing a puzzle game.

There are some difficulty estimating techniques and systems that have been proposed to estimate the difficulty level of puzzles in puzzle games, which allows the user to choose proper difficulty

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 193-202, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111717

level of puzzles that is not too easy to make players feel boring or too hard to make players feel frustrated, these proposals assume there's a difficulty function that can combine different aspects of the levels of these puzzle games, for example level size, and provide difficulty ratings, which is rarely the case in practice. Their implementations are also limited in scale, with samples given for only estimating specific game type like Flow, Lazors and Move, which is only a small part of puzzles that can be used when making puzzle games. Other techniques, such as rating the difficulty level of Sudoku problems with human oriented, general difficulty criteria, are also not comprehensive when designing puzzle games. Because the rating methods are limited to Sudoku or at most, constraint satisfaction problems (CSP), the method used cannot be used by game developers effectively while designing puzzle games [8].

In this paper, we follow the same line of research by first building an abstract for the game, navigating the problem while building, then finding possible solutions, playtesting those solutions and improving the solutions [9]. Our goal is to find the balance between difficulty level of challenges and playability of those puzzles. Our method is inspired by unit testing, which validated that each unit of the software code performs as expected. There are some good features of the method we used. First, the game method can be improved while the designing process is limited by availability and playability. Second, the background story of the game can perfectly match the game scenes that gives players an immersive game experience. Therefore, we believe that using the method we chose to design puzzle games can give players good play experience with both challenging and interesting feelings.

We're going to prove the results by collecting survey results from players who've played the puzzle game we've made [10]. The survey is going to collect data like difficulty level of challenges inside the game rating from one to five and playability of these puzzles also rating from one to five. Since there isn't a standardized and convincing rating method for puzzle games' puzzles worldwide, the data we collected about the difficulty level of challenges inside the game and the playability of the game will be pretty subjective. In order to minimize the possible bias that occurred when analyzing the data, we're going to use data from players in different age ranges and maximize the number of players being researched as much as possible. By analyzing the data we collected, we're going to find the balance between difficulty and playability of games. Not only provide game designers a good method to make a better game, but also give players better play experience with both challenging and interesting feelings [15].

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Game design and game logic

Among various topics that puzzle games might use, room escaping is one of the most common topics that fit puzzle games properly. So we decided to design a room escaping game at the very beginning of the design process, but how to make the game stand out among all the other room escaping games is a crucial problem. Since the puzzle mechanics are similar among all puzzle games, we chose to make the game unique by setting unique multi-thread endings. There are in total five different endings that the player might face. These endings are not displayed randomly, based on the explore degree of the players towards the story, they'll face different endings that fit what they've got recognized to the story.

2.2. Context-aware 2D game scene development

Since the whole game is designed by our team including developing the scripts and organizing artworks, there are multiple strange errors that occur while developing the scripts. For example, the scaling script that is attached to the object is not working, or the code conflicts with another method in another script. We've used a lot of time fixing those errors and managing the project to make it work as expected. We wrote thousands of lines of codes for this project.

2.3. Deciding puzzle difficulty level

The pace of a puzzle game is mainly managed by the logic flow and time used by the players to solve the puzzle. Therefore, deciding the puzzle difficulty level is one of the most important aspects that we need to consider while developing our puzzle game. If the puzzle is too easy, the gameplay time is too short and players didn't get enough challenges while playing the game. If the puzzle is too complicated and hard, players are easy to feel frustrated and they tend to give up playing the game instead of trying to find solutions.

3. SOLUTION

In order to change the pace of a puzzle game, we have to limit the playtime of each scene, and the playability of each puzzle. To control the pace of a puzzle game, we chose to make the difficulty level of the puzzles changeable by making parameters of the puzzle handling function editable and using the parent/child or neighbors methods in scripts to make the objects free to set instead of preset every parameter inside scripts. By editing these values, the difficulty level of each puzzle is easy to change and we decide the difficulty level by collecting the players average time used to solve each puzzle. The total playtime of this game is estimated to be less than an hour, if most players are stuck on one scene for more than 15 minutes, that means the difficulty level of the puzzle in this scene need to adjust and the hints or logic of solving the puzzle needs to be clearer.

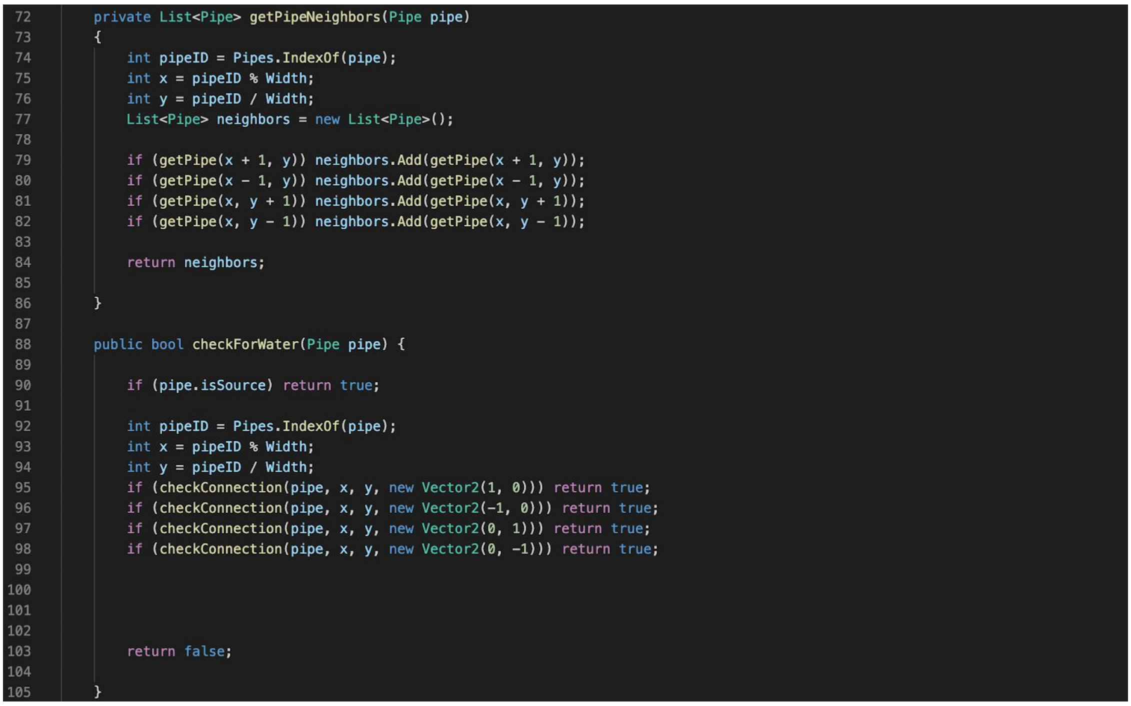
The first scene is a bedroom with a locked drawer inside. The number lock is obvious enough so every player who enters this scene will notice the locked drawer and know their goal is to unlock the drawer. By interacting with the objects in the scene carefully, players are able to find the hints to the only lock except for the locked door inside the room which is the number lock. The hint for this lock is a mathematical problem with proper difficulty, players just have to list down different solutions to find the correct one, and they will realize the correct answer is three numbers, also the number lock requires three digits as the solution. After unlocking the drawer, the player will get the key to unlock the door to get to the next scene.

The second scene is a corridor with puzzles and pieces that are scattered around the space. By seeing the blank frame and scattered pieces, players will realize they have to find all pieces that are hidden inside the scene. Although they didn't know what would happen after they finished the puzzle, at least they knew their goal. After completing the piece puzzle, a key will show up from the painting and the player will collect the key to unlock the locked door.

The third scene is a bathroom. At first, players will be a little bit confused about what they should do, but after they interact with the mirror inside the bathroom, a pipe puzzle shows up and we assume that everyone knows how to play the pipe puzzle. Just simply let water flow from the start pipe to the end pipe. After solving the pipe puzzle, when players click the sink, the sink will be shown as filled with water status and a key is lying at the bottom of the sink. By collecting the key, players are able to enter the next scene by unlocking the locked door.

The fourth scene is a living room. It is pretty obvious that a clock dial is displaying at the center of the scene but missing clock hands. Players will know their goal is to find the clock hands and the correct time that these clock hands should point to. After solving the clock puzzle, the TV inside the living room will turn on and players should realize how to interact with the TV and will reach the next scene. There's only one object inside the final scene so the player will know their goal is to click that object.

All the values that related to the difficulty of each puzzle, for example the hints that decide solutions of the number lock or the time in the clock puzzle that clock hands should point to are editable, which means the difficulty level of all the puzzles could be improved after play testing.



```
    private List<Pipe> getPipeNeighbors(Pipe pipe)
    {
        int pipeID = Pipes.IndexOf(pipe);
        int x = pipeID % Width;
        int y = pipeID / Width;
        List<Pipe> neighbors = new List<Pipe>();

        if (getPipe(x + 1, y)) neighbors.Add(getPipe(x + 1, y));
        if (getPipe(x - 1, y)) neighbors.Add(getPipe(x - 1, y));
        if (getPipe(x, y + 1)) neighbors.Add(getPipe(x, y + 1));
        if (getPipe(x, y - 1)) neighbors.Add(getPipe(x, y - 1));

        return neighbors;

    }

    public bool checkForWater(Pipe pipe) {

        if (pipe.isSource) return true;

        int pipeID = Pipes.IndexOf(pipe);
        int x = pipeID % Width;
        int y = pipeID / Width;
        if (checkConnection(pipe, x, y, new Vector2(1, 0))) return true;
        if (checkConnection(pipe, x, y, new Vector2(-1, 0))) return true;
        if (checkConnection(pipe, x, y, new Vector2(0, 1))) return true;
        if (checkConnection(pipe, x, y, new Vector2(0, -1))) return true;




        return false;

    }
```

Figure 1. The code excerpt of how to check pipe neighbors

4. Experiment

4.1. Experiment 1

To test the difficulty level and playability of our game, we sent a survey to players who've played this game. The survey includes each puzzles' difficulty level rating from 1-5 and playability of each puzzle's rating from 1-5. After rating difficulty level and playability, the player will also provide a word or two about their play experience and any improvements recommended. After receiving those survey results, we'll use data analysis methods to make a table and see the result analysis.

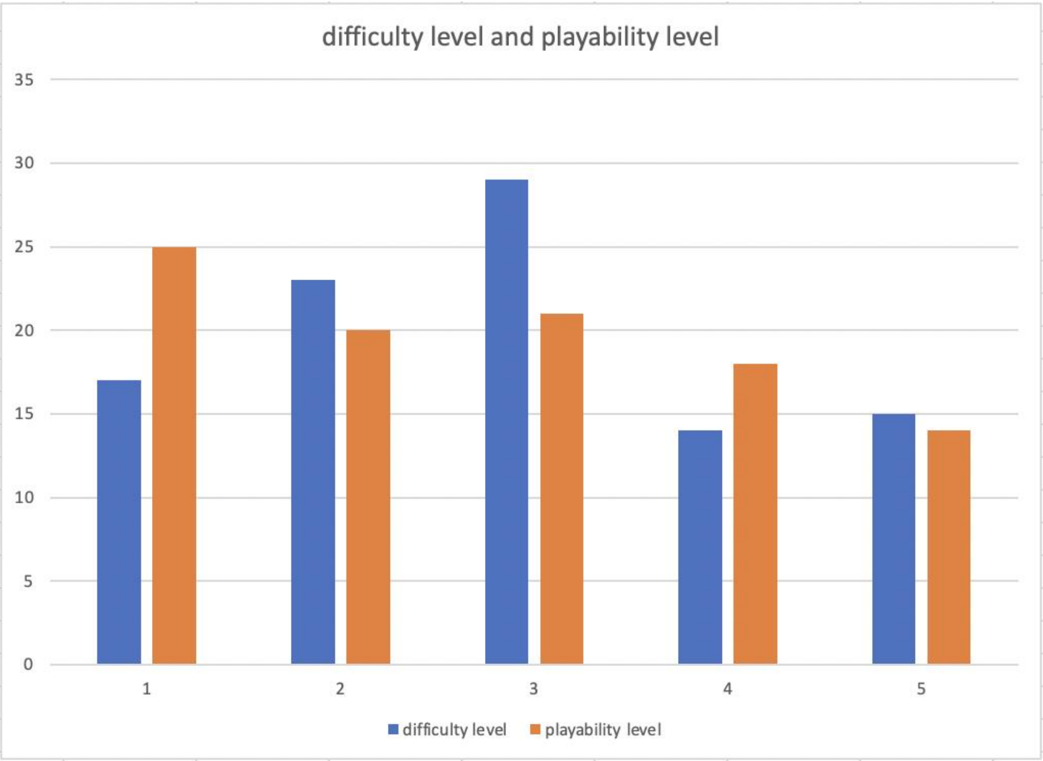


Figure 2. The difficulty and playability levels

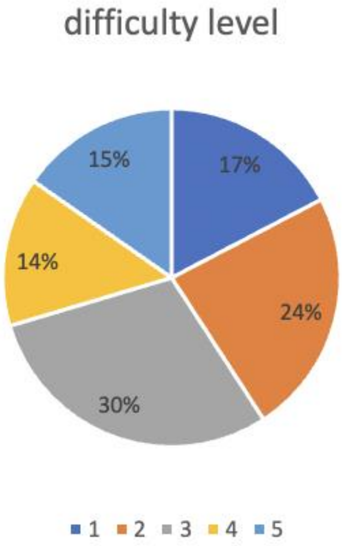


Figure 3. The specific difficulty level analysis

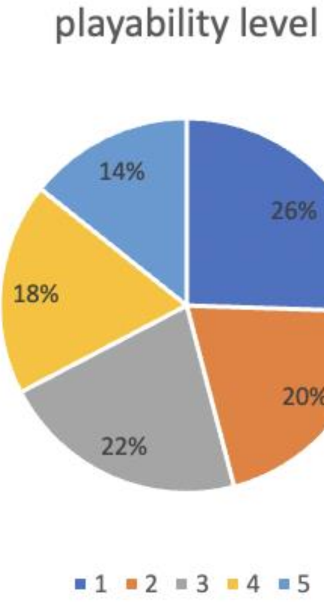


Figure 4. The specific playability level analysis

The result of experiment 1 shows players' difficulty level and playability level of the game that was collected from 100 surveys. We can see from the bar chart that peoples' opinion about the difficulty level of the game is pretty diverse, but the majority of people who took the survey thought the difficulty level of the game is 3 which is a median number among 1 to 5. This value shows the game designer that the difficulty level of the game is proper and acceptable by most people. From the pie chart we can know that most people think the playability level of the game is only 1, which informs the game designer that their game needs to be more playable. From the improvements recommended in the survey that we collected, the most frequently shown suggestion is that puzzles need to cooperate more with objects inside the space.

4.2. Experiment 2

To test the players' age range, average time used to play the game and the attractiveness of multiple endings, we sent a survey to players who've played this game. The survey includes players' age, the total time they played the game. After filling out the survey, the player will also answer their preferences to play this game again to unlock different endings rating from 1-5. After receiving those survey results, we'll use data analysis methods to make a table and see the result analysis.

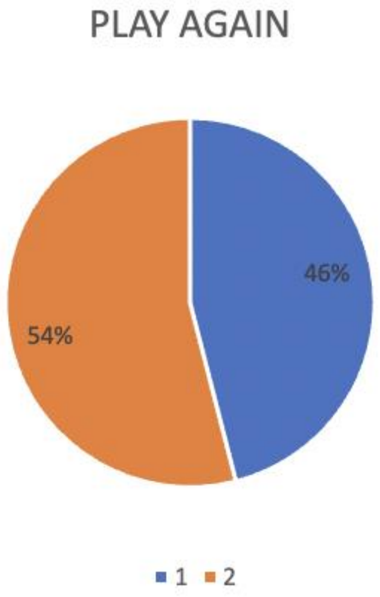


Figure 5. The analysis of continuing playing

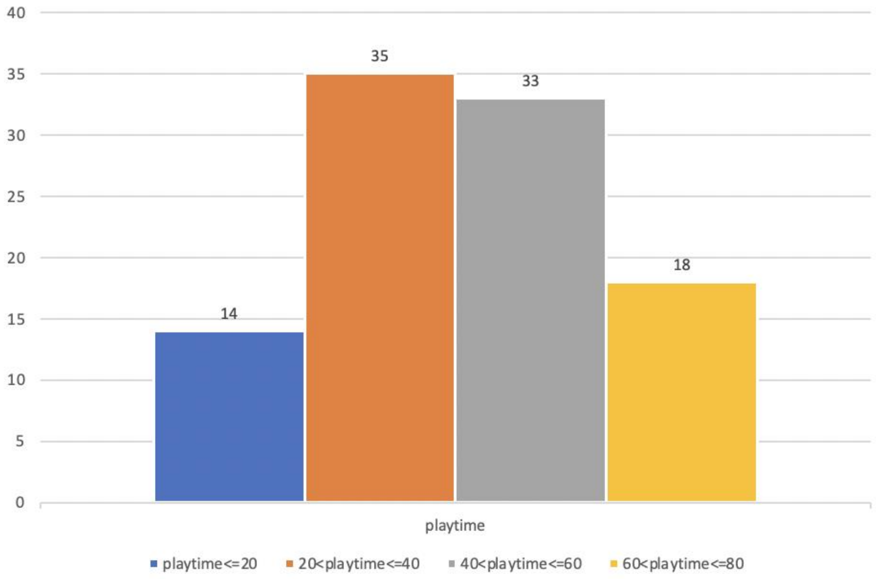


Figure 6. The comparison of game play time

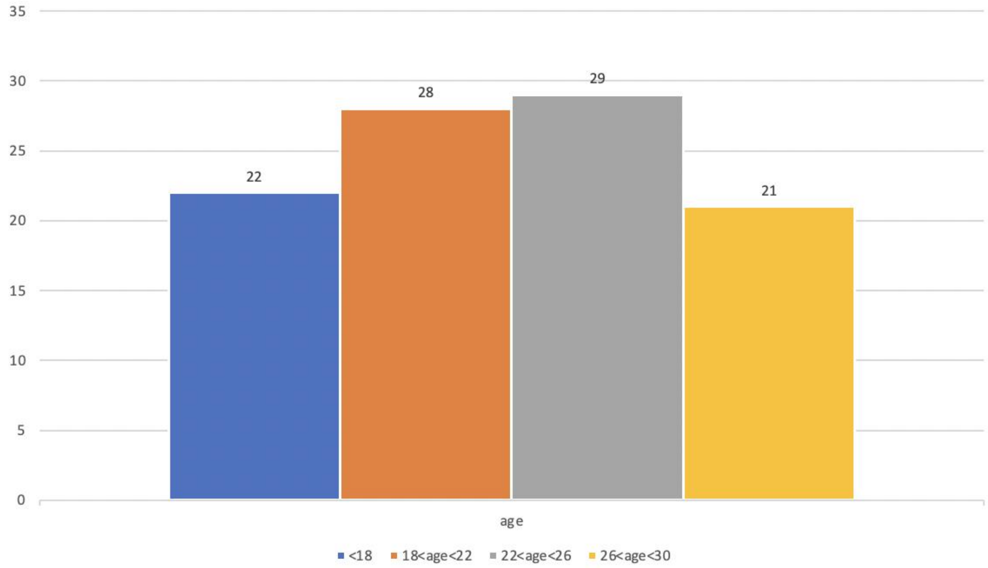


Figure 7. The comparison of player age groups

The result of experiment 2 shows players' age range, average time used to play the game and whether players would like to play the game again to unlock different endings. We can see from the pie chart that 46% of players who finish the survey choose option 1 which is marked as "Yes" while 54% of players who finish the survey choose option 2 which is marked as "No". The reason they chose not to play again is that they think repeated game experience is boring. This feedback notified game designers to make more changes to the game when it's being played again so that players would be more likely to play again. From the playtime bar chart we know that most people can complete playing the game in 60 minutes which is the expected playtime of the game. This table shows that the expected playtime of the game can be achieved. The age range bar chart tells the game designer ages of his players, which help the game designer better analyze the needs of his target audiences.

By giving surveys to players who've played the game and analyzing the data, issues, and suggestions collected, game designers are able to modify the game and precisely target the needs of those players. Surveys can help game designers communicate better with their players. The rating system of difficulty level and playability level can help game designers better shape their game and cater to the needs of the public. What's more, Data analysis is an intuitive and effective tool for game designers to improve their work.

5. Related Work

Marc van Kreveld, et al provided a method that can automatically rating the difficulty of puzzle game levels. They used a difficulty function to calculate the difficulty level of puzzle games and choose variables while playing the game or watching others playing the game to measure the final results. Marc van Kreveld's difficulty function can measure the difficulty level of most puzzle games automatically, but finding variables from the game takes a lot of time and the standard they found for one puzzle may not fit similar puzzles due to changes and complicated settings of that puzzle. Compared to Marc van Kreveld's method, our method is more precise since it's only serving for one specific game and it's more likely a useful tool for game designers instead of a standard rating method for all puzzles.

Conor Linehan, et al talked about the relationship between the pace of challenges, players' enjoyment and difficulty experience, and players' ability to learn from game play. Compared to our paper which discussed an effective tool that can help game designers improve their work, Conor Linehan's paper found the relationship between pace of challenges and the learning curve of challenges introduced to players. Conor Linehan's method helps us a lot when designing puzzles and manipulates the balance between difficulty level and players' game experience while designing.

GaëlleGuigon, et al presented a creation tool for designing serious games with riddles like escape room games. GaëlleGuigon's method is quite similar to our method since the purpose of GaëlleGuigon's and our method is the same: helping game designers to build a successful game. GaëlleGuigon's creation tool focuses more on how to develop the game while our rating system method focuses more on how to improve the game based on the game we've made. GaëlleGuigon's tool is more friendly to people who haven't started their game yet. However, the method mentioned in this paper can give more useful suggestions to game designers to improve their game.

6. CONCLUSIONS

In summary, in order to help game designers manipulate the pace of puzzles and give players better game experience, we provide an effective method which is a rating system that rates the difficulty level and playability level of the game by players who have tested this game [11]. We've collected 100 pieces of surveys from players' who've played our game and analyzed data from the surveys. By analyzing variables like difficulty level, playability level, playtime and whether the player is willing to play the game again to unlock different endings, we found out many problems from the game that could be improved. Effectiveness of this method is that it can help game designers target issues occurring inside the game precisely, and suggestions provided by people who've finished the surveys give game designers a possible direction to improve the game and make the game fits the needs of majority players.

However, there are limitations of the method we provided [12]. First of all, collecting surveys from multiple players' who've to playtest the game may be a long-term process. People might not playtest your game, or they choose not to finish the survey after playing the game, or players don't even finish playing the whole game. Waiting for enough data to analyze the difficulty level and playability level of the game takes too much time. Secondly, this method only fits one game at a time. If you change the game that needs to be analyzed for its user feedback, all the data collected for the previous game makes no sense to the current game, a new survey must be formed and a new waiting process of collecting enough data starts [13].

In order to solve these limitations, we'll try to build a platform for game designers to playtest their games and collect data needed [14]. In order to absorb enough players who would like to playtest and give feedback, the platform will provide enough rewards to these players. The rewards are provided by game designers, it may be coupons of other games or special gifts inside the game after the game is published.

REFERENCES

[1] M. van Kreveld, M. Löffler and P. Mutser, "Automated puzzle difficulty estimation," 2015 IEEE Conference on Computational Intelligence and Games (CIG), 2015, pp. 415-422, doi: 10.1109/CIG.2015.7317913.

[2] Lincoln, Conor Linehan University of, et al. "Learning Curves: Analysing Pace and Challenge in Four Successful Puzzle Games." Learning Curves | Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play, 1 Oct. 2014, dl.acm.org/doi/abs/10.1145/2658537.2658695.

[3] GaëlleGuigon, Mathieu Vermeulen, JérémieHumeau. A Creation Tool for Serious Puzzle Games. CSEDU 2019, May 2019, Heraklion, Greece. pp.556-561, 10.5220/0007796405560561 . hal-02132554

[4] Funk, Jeanne B. "Reevaluating the impact of video games." Clinical pediatrics 32.2 (1993): 86-90.

[5] Lin, Chien-Heng, and Chien-Min Chen. "Developing spatial visualization and mental rotation with a digital puzzle game at primary school level." Computers in Human Behavior 57 (2016): 23-30.

[6] Grace, Lindsay. "Game type and game genre." Retrieved February 22.2009 (2005): 8.

[7] Toulmin, Stephen. "From logical systems to conceptual populations." PSA 1970. Springer, Dordrecht, 1971. 552-564.

[8] Pang, Shanchen, et al. "Notice of Retraction: Rating and Generating Sudoku Puzzles." 2010 Second International Workshop on Education Technology and Computer Science. Vol. 3. IEEE, 2010.

[9] Henzinger, Thomas A., et al. "Abstract interpretation of game properties." International Static Analysis Symposium. Springer, Berlin, Heidelberg, 2000.

[10] Scheuren, Fritz. "What is a Survey?." Alexandria: American Statistical Association, 2004.

[11] IJsselsteijn, Wijnand, et al. "Measuring the experience of digital game enjoyment." Proceedings of measuring behavior. Vol. 2008. No. 2008. Maastricht, the Netherlands: Noldus, 2008.

[12] Simon, Marilyn K., and Jim Goes. "Scope, limitations, and delimitations." (2013).

[13] Taylor-Powell, Ellen, and Carol Hermann. "Collecting evaluation data: surveys." Washington, DC: University of Wisconsin-Extension (2000).

[14] Hamel, Gary, and Michele Zanini. "Build a change platform, not a change program." Retrieved November 12 (2014): 2014.

[15] Brathwaite, Brenda, and Ian Schreiber. Challenges for game designers. Boston, Massachusetts: Course Technology/Cengage Learning, 2009.

INTELLIGENT SPEED ADAPTIVE SYSTEM USING IMAGE REGRESSION METHOD FOR HIGHWAY AND URBAN ROADS

Bhavesh Sharma[1] and Junaid Ali[2]

[1]Department of Electrical, Electronics and Communication Engineering, Engineering College, Ajmer, India
[2]Department of Mechanical Engineering, Indian Institute of Technology, Madras, India

ABSTRACT

Intelligent Speed Adaptive System (ISAS) is an emerging technology in the field of autonomous vehicles. However, the public acceptance rate of ISAS is drastically low because of several downfalls i.e. reliability and low accuracy. Various researchers have contributed methodologies to enhance the traffic prediction scores and algorithms to improve the overall adaptability of ISAS. The literature is scarce for Image Regression in this range of application. Computer vision has proved its iota in stream of object detection in self-driving technology in which most of the models are assisted through the complex web of neural nets and live imaging systems. In this article, some major issues related to the present technology of the ISAS and discussed new methodologies to get higher prediction accuracy to control the speed of vehicle through Image Regression technique to develop a computer vision model to predict the speed of vehicle with each frame of live images.

KEYWORDS

Intelligent Systems, Self-Driving Vehicle, Image Processing, Image Regression, Computer Vision, Automotive.

1. INTRODUCTION

With the rapid growth of automobile users across the world and the surplus adaption of state-of-art technology in the latest automobiles, the automotive industries have turned themselves from batch-type producers to mass vehicle producers. The growing user density in the region also increases the chances of jammed traffics, accidents, and nonetheless environmental pollution due to idling of hundreds of vehicles on the signal crossings. Researchers across the globe are working judiciously on each subject of traffic control systems, accident prevention systems, and pollution control systems. So many efforts have already been done in the direction of developing an Intelligent Speed Adaption System to control the speed of the automobile to prevent traffic and ultimately accidents due to high vehicle speeds, lack of control during cornering, and overtaking. With each passing generation of humans as well as automobiles, the dire need for high horsepower, more speed, cutting edge aerodynamic design, and intelligent behaviour of the vehicle in response of the driver is in demand and with Artificial Intelligence taking over the course of almost everything, automotive industries are also developing state-of-art intelligent vehicle systems to give premium experience to the customers [1].

ISA is based on a speed limiter incorporated within each vehicle that can take into account speed limit restrictions, that can adjust the maximum driving speed to the speed limit specified by the

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 203-213, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111718

roadside infrastructure, and that can provide feedback to the driver or take autonomous action when that speed limit is exceeded. ISA systems could use fixed or dynamic speed limits. In the fixed case, the driver is informed about the speed limit, which could be obtained from a static database. Dynamic speed limits take into account the current road conditions such as bad weather, slippery roads, or major incidents before prescribing the speed limit. If we assume Road Speed Limit as the range of speeds with a minimum and maximum value rather than a single absolute value. Then the highlighted benefit of such a system is that driver will have a low speed as well as high speed for an individual road to drive on. The selection of speed between the ranges will be defined as per mathematical algorithm fed on live traffic data in form of live image frames. In simpler words, the vehicle will choose to go slow if the traffic is dense on the road whereas the vehicle will choose to go fast but within the speed limit if the traffic is open on the road. However, the underlying problem with this method is limiting vehicle dynamics considerations in determining the road safety speed limits. The speeds assigned on the highways are determined based on vehicle dynamics of average automobile models which is not possible to be altered immediately. Such limitations are frustrating for drivers of the latest technology automobiles which can perform better because of advanced design and dynamics. We understand each vehicle has its defined performance capability a 1200 cc engine is going to over-take 800 cc engines for plausible engineering reasons. But it would be arrantly injustice for a 1200 cc engine vehicle to lag in traffic due to speed limits based on 800 cc engine capabilities. Revolutionizing the road speed limit data across the globe will be a difficult task. But we can revolutionize the way our vehicles respond to traffic and speed limits with the help of computer vision powered by image regression tools [2] [3].

Moreover, the vehicle speed prediction is not just limited to the above-mentioned factors but also the un-registered obstacles on the road e.g. wild animals crossing the road, construction work, pitfalls on road, and any unidentified object lying on road. To tackle this problem Computer Vision came to our rescue and with a dataset of 10,000 high-quality images provided by the Indian Transport Department. The dataset contains images of all possible obstacles a vehicle could face while driving on urban or rural roads. Computer Vision assisted with Convolution Neural Networks model envisages the vehicle to process live images from roads and make decisions backed by the mathematical algorithm to prevent a collision or slows down the speed of the vehicle on the road.

Most of the speed adaption system depends on GPS which are highly inaccurate and don’t update from time to time, create lags while driving. A robust system is required that works offline and thus adapts the speed within seconds. Most of the GPS-based models are slow and thus cannot provide good results in urban areas. Here we have aimed to predict the speed of the vehicle using the images taken from the car dashboard. The neural net is been trained to take an input image segment it and thus provide a speed estimation. Each training image or frames speed is been annotated at a given time frame that is used to train the neural net. Thus, improving the overall accuracy of the computer vision model to detect obstacles and adjust speed based on algorithm [4].

2. Data Pre-Processing

At first, the image data is converted into a proper R, G, B channel with a height, width of 120x120x3. There are two types of data on which experiments is been done, Highway data and Urban data

For neural nets, the images are not been converted into arrays but for other models like SVR, Random Forest, and linear regression models the image is been converted into the NumPy arrays. At each frame, the speed of the vehicle is been recorded and thus annotated with frames. CNN-based image regression is mainly used to detect bounding boxes when there is a task of classification but if proper features is been extracted it can be used to predict the continuous values.

2.1. Background Subtraction in Image Processing

Background subtraction is a technique used to recognize the moving object in a video. The recognition of the images is done by using static cameras. The fundamental principle behind using background subtraction is recognizing the image from a difference between the reference frame of the image and the present frame of the image. The reference frame in this technique is known as the background image. The background model must be static while extracting the front objects from the video. The generic name of the front object in the video is known as the foreground object. Background subtraction is divided into two categories:

i. Parametric Background Subtraction
ii. Non-Parametric Background Subtraction

Two major techniques that are frequently used for background subtraction are pixel-based and block-based. In the case of statistical representation of the image non-parametric pixel technique is used. Shadow detection and illuminations are two major factors that impact the quality of the background subtraction. While capturing the foreground object, the background in the image must be static. The major application of background subtraction is in video surveillance and video analysis. The quality of the image extracted after the background subtraction can be enhanced by using various methods such as phosphine dots. These images as shown in figure 1 and figure 2 are then passed through neural nets and VGG 16.

Figure 1. Parametric Background Subtraction

Figure 2. Non-Parametric Background Subtraction

3. Proposed Solution

When a proper CNN is constructed and while tuning the last layer to predict the continuous value rather than probabilities. While compiling the neural net MAE (Mean Absolute Error) with Adam optimizer is used. At here we proposed a solution to map a regression function using CNN for getting a prediction of speed from the images. For that, we have used the ISA^2 dataset [5]. Humans can't detect the speed of the vehicle from a single image but CNN can help to predict the same. That's why we proposed a novel solution to detect the speed of the vehicle using a single image. We proposed four algorithms for the same. The deep learning solution and three classical machine learning solutions.

3.1. Neural Network Regressor Model

The CNN regressor model directly takes an input image which is been rescaled in RGB format. The number of channels is not been reduced which passing the image from the CNN model. The input shape of the images been 120x120x3. The CNN regressor network is been trained to learn from the mapping from the input images with their labelled speeds.

$$\hat{s}_W = f(W, I_i) \tag{1}$$

The above equation hold for the mapping for the image to speed by CNN, where $f(W, I_i): I_i \to \mathbb{R}$ represents the mapping of the input image to speed. The loss function for the trainable model is been a mean square error (MSE).

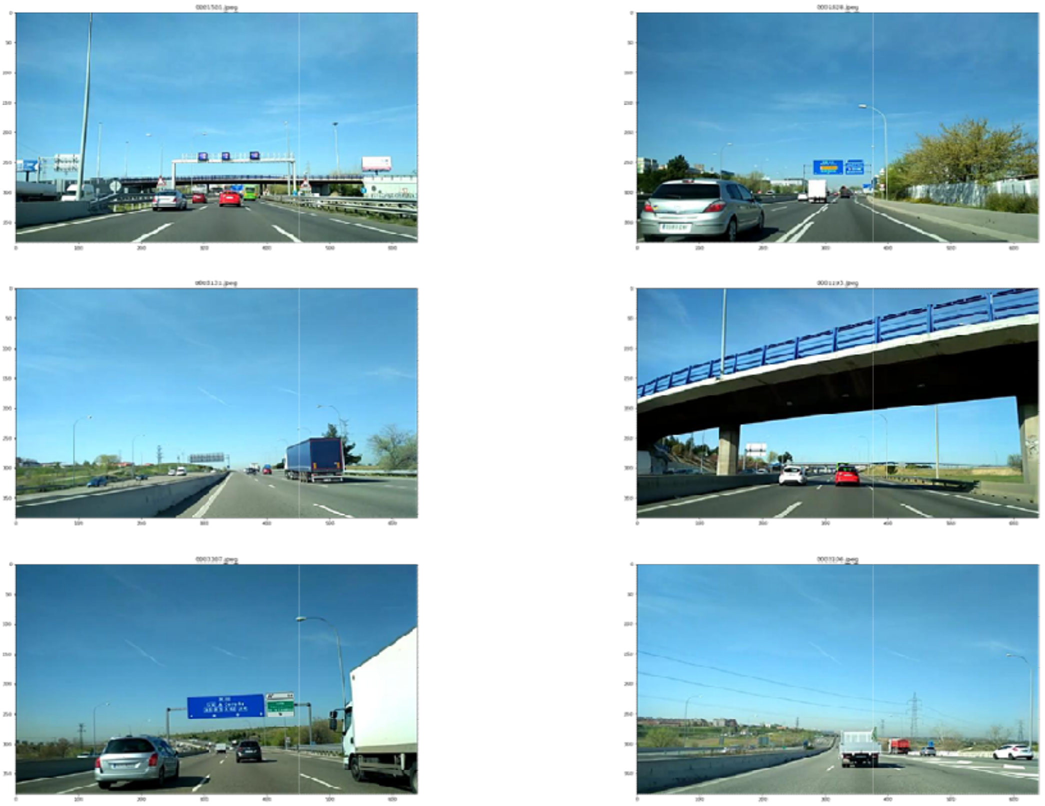


Figure 3. Sample Dataset – Highway Traffic Images

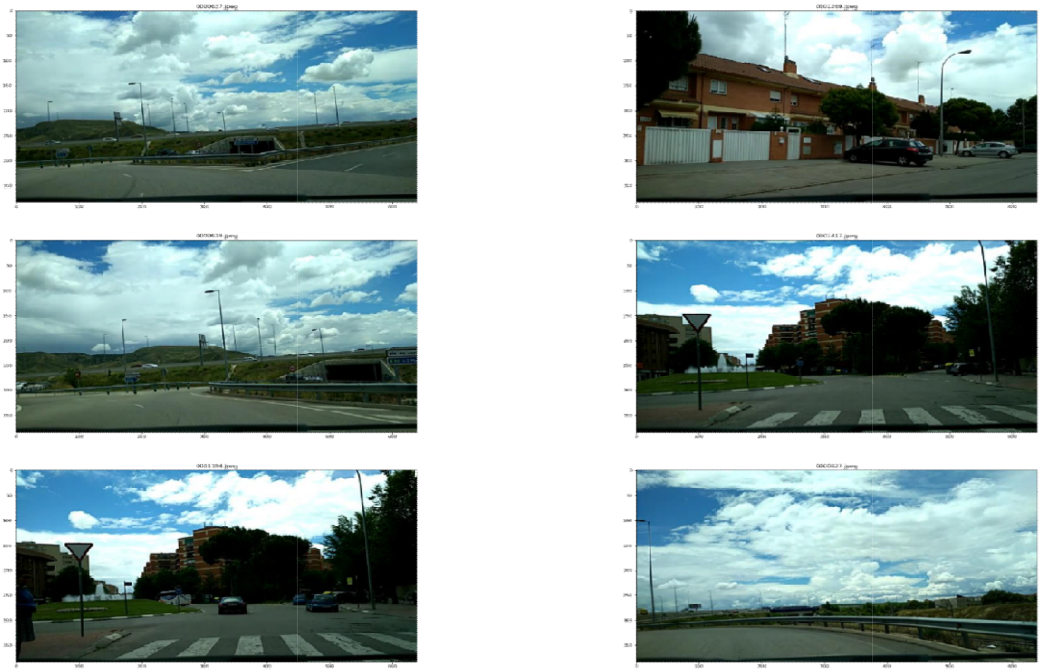


Figure 4. Sample Dataset of Urban Traffic Images

These images as shown in figure 3 and figure 4 for the Highways and Urban data is been passed through CNN with the following properties.

The proposed neural net has an input layer with dimensions 120x120x3. The second layer consists of CONV2D with the size of 118x118x16 with a max-pooling layer of 59x59x16. Then second CONV2D layer is been inserted with the size of 28x28x32 with an additional global polling average and the dense layers. The output layer is accompanied by the MSE loss function. The model is been trained with an iteration of 1000 and validation error is been noted at each epoch. To evaluate the model MAE or mean absolute error has been used for both the image sets of Highway and Urban.

For K images in training or testing set, the MAE is given by equation 2 below,

$$\frac{1}{K}\sum_{i=1}^{K} |s_{r_i} - \hat{s}_i| \qquad (2)$$

3.2. SVM Regressor Model

In SVM, the images can't directly feed onto the model, that why it needs to break down in array format, for this the images are been flatten in the dimension of 2800x43200 for training sets. The same dimension is been used for the urban image datasets. The images are been a break down in the arrays which contains the features of any given images. After that, the images are been normalize and passed through noise filters. The support vector regression model is trained using these images with labelled speed for each image set as shown in figure 5.

The SVR has the following constraints as shown in equations 3 to 7,

$$\underset{2}{\overset{1}{min}} \parallel w \parallel^2 + C\sum_{i=1}^{N} \xi_i + \xi_i^* (3)$$

$$y_i - w^T x_i \leq \varepsilon + \xi_i' i = 1 \ldots N \qquad (4)$$

$$w^T x_i - y_i \leq \varepsilon + \xi_i i = 1 \ldots N \qquad (5)$$

$$\xi_i \xi_i \geq 0 i = 1 \ldots N \qquad (6)$$

$$\mathcal{L}(w, \xi', \xi, \lambda, \lambda^\circ, \alpha, \alpha^*) = \frac{1}{2} \parallel w \parallel^2 + C\sum_{i=1}^{N} \xi_i + \xi_i^* + \sum_{*1}^{N} \alpha_i^* (y_i - w^T x_i - \varepsilon - \xi_i^*) + \sum_{i=1}^{N} \alpha_i(-y_i + w^T x_i - \varepsilon - \xi_i) - \sum_{i=1}^{N} \lambda \xi_i + \lambda_i \xi_i' (7)$$

Where the slack variable or the cost variable ξ_i is introduced.

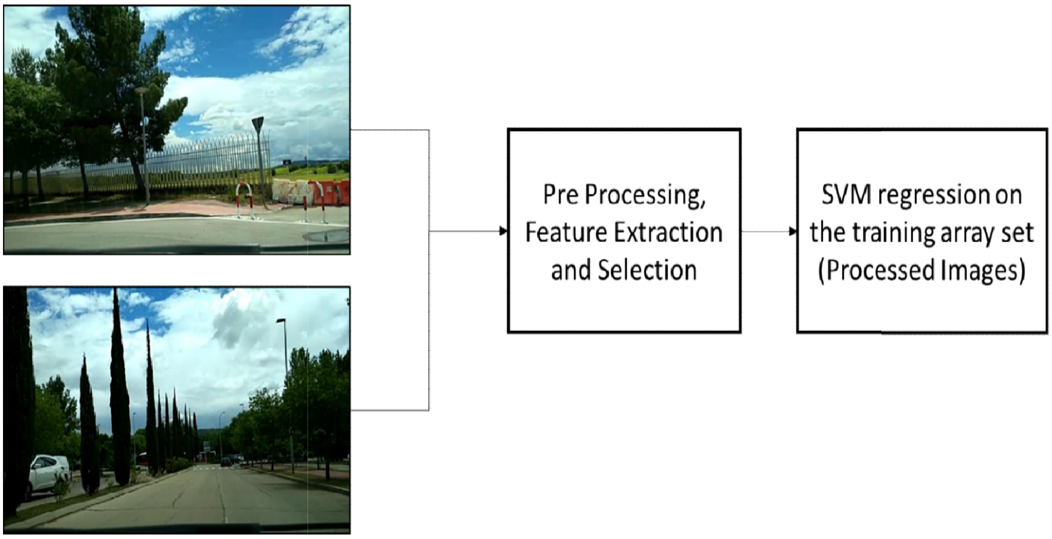

Pre Processing, Feature Extraction and Selection
SVM regression on the training array set (Processed Images)

Figure 5. Algorithm of Feature Extraction and Speed Prediction

3.3. Random Forest Regression

The random forest regression model is a part of the supervised learning used to get higher prediction accuracy of traffic to control the speed of the vehicle. Numerous image training sets are assembled to improvise the higher prediction score. Random forest regression is a part of supervised learning based on the ensemble technique. The random forest technique can be used for both regression and the classification of the images. By taking the decision tree as a base of the model, column sampling and the rows sampling are done in the random forest technique. In the case of the random forest regression, if more image training sets are used then the variance of the model decreases, and hence the stability of the model increases. The random forest algorithm works by developing a decision tree multitude at the time of image datasets and provides the mean or mode value of all individual trees prediction values. Lower the value of the variance shows the higher stability of the model. When the image training sets are decreases then the value of the variance increases so, the stability of the model decreases. The mathematics behind the random forest regression is as shown from equation 8 to 12:

3.3.1. Splitting Criterion

$$RSS = \sum left\,(Y_i - Y_l)^2 + \sum right(Y_i - Y_r)^2 \tag{8}$$

Where,

$$Y_l = Left\ node\ Y - mean\ Value \tag{9}$$

$$Y_r = right\ node\ Y - mean\ value \tag{10}$$

3.3.2. Gini Criterion

$$Gini = n_L \sum_{n=1}^{k} p_{kl}\,(1 - p_{kl}) + n_R \sum_{n=1}^{k} p_{kr}\,(1 - p_{kr}) \qquad (11)$$

Where,

$$p_{kl} = Left\ node\ proportional\ of\ class\ k$$
$$p_{kr} = right\ node\ proportion\ of\ class\ k$$

The Gini index value is used to determine the frequency at which any individual element among the dataset is mislabelled. The range of the Gini index is 0 to 1.

3.3.3. Mean Absolute Error (MAE)

The MAE in the case of the random forest regression model is given by

$$MAE = \frac{\sum_{I=1}^{N} abs(Y_I - \lambda(X_I))}{N} \qquad (12)$$

3.4. VGG 16 (Oxford Net)

VGG 16 is a CNN model used for image recognition at a very large scale. The dimension of the image training dataset used in this report is 2800 X 43200. The dimension of the input RGB image training dataset to convolution layer 1 is 120 X120 X3. In a VGG 16 architecture model, the input set of image datasets is allowed to pass through a series of convolutional layers. This stack of the convolutional layers comprises filters with tiny receptive fields. Receptive fields are used to capture the notions present in the images.VGG16 is part of the convolutional neural network used to solve the problem related to computer vision. Common computer vision problems are the classification of image datasets and regression problems. VGG model is divided into two categories based on the number of layers present in the model (VGG16, and VGG19). VGG is still a powerful classifier used for the classification of the image dataset. VGG16 algorithms can be used for regression and classification tasks. Mean squared value and the mean squared error in the case of VGG16 is given by the following mathematical relationship as shown in equation 13 and 14.

$$mae = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - X_i^{GT}\right| \qquad (13)$$
$$mse = \left(\frac{1}{n}\sum_{i=1}^{n}\left|X_i - X_i^{GT}\right|\right)2 \qquad (14)$$

4. Results & Discussion

To evaluate models and their performance the standard Mean Absolute Error metric is been used which computes the difference between the actual speed and predicted speed averaged over the K value of images. The results for the Highway as well as Urban areas is been calculated separately. The following tables is been constructed for the Highway as well as for Urban images.

Table I: Model Summary

Method	Highway (MAE) km/hr	Urban (MAE) km/hr
CNN based approach	10.45	9.640
SVM Regressor	8.473	6.460
Random Forest Regression	8.428	8.465
VGG16	9.668	8.279

From the above Table I, can be interpreted that the VGG16 has performed very well on the Highway data. Similarly, the SVM regressor has shown very good results in the urban image data. The tuned neural net has performed well on both datasets but didn't produce better results as compared to SVM or Boosted Trees.

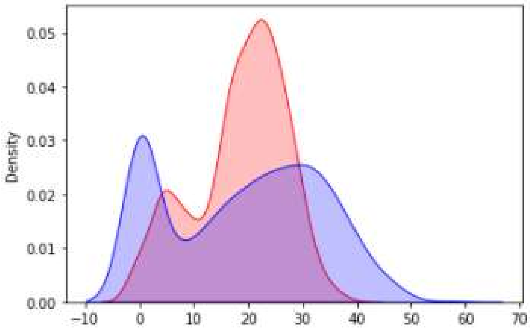


Figure 6. Density plot for predicted and actual values

The above plot shown in figure 6 describes the density plots for predicted and actual values where the red ones are of the predicted and the blue ones are for actual values. The predicted values show the same KDE pattern as the true images speed. This helps to determine how close the distribution of the predicted and the actual speed data. In figure 7 and 8, epochs are plotted against the losses and the decreasing trend is showing the increasing accuracy of the model which is desirable for accurate speed prediction.

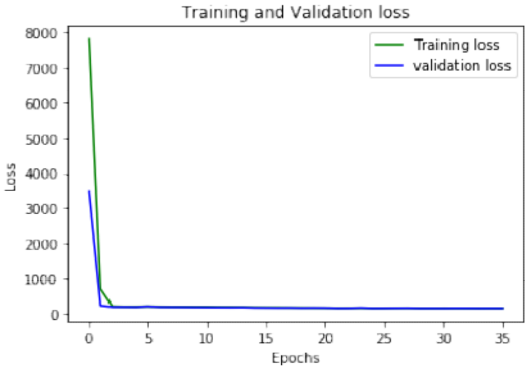


Figure 7. Loss -Epoch Curve - Highway Image Dataset

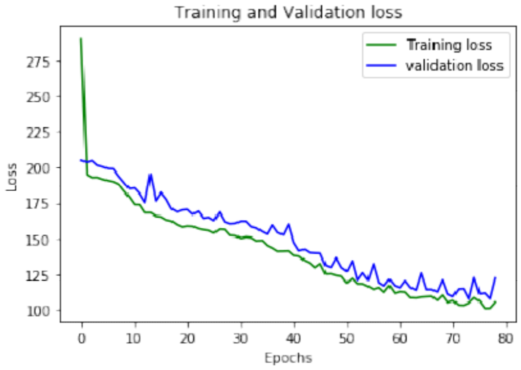


Figure 8. Loss- Epoch - Urban Image Dataset

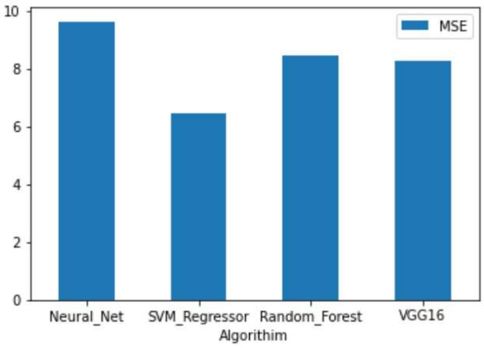


Figure 9. MSE comparison for Urban Data

The above results shown in figure 9and 10 the lowest MAE is been produced by the SVM regressor on the Urban data. The SVM regressor has performed better than the VGG16 and Neural Net as the MSE of these two is very high. The Urban data don’t have many features as the roads and traffic density is very less as compared to Highway Data.

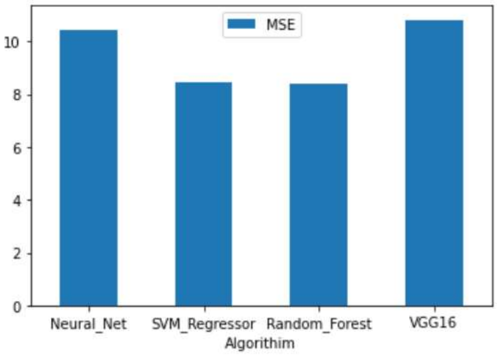


Figure 10. Comparison of MSE in Highway data

The number of features in the High way data is much more as compared to the Urban Data as the number of cars and traffic density increases in Highway data. SVM regression and Random forest have worked very well on the Highway data as compared to VGG16 and Designed Neural Net.

5. CONCLUSION & FUTURE SCOPE

In this article we have performed computing experiments to conclude that through the given image frame from vehicles while driving, the speed of the vehicle can be detected using machine learning algorithms. Our objective was to demonstrate the capability and prowess of Image Regression in computer vision techniques to detect the speed of running vehicle which is potent as compared to the conventional GPS tracking APIs and vehicle platooning as previous works claims. We have achieved our objective by comparing different machine learning and neural networks models on same dataset. Among the Highway Image data and Urban Image data theleast MAE has been shown by the SVM and Random Forest Regressor whereas VGG 16 and NN have shown comparatively poor results. The ML algorithms were able to capture the features and thus able to make a speed predictions with minimal error. In future a segmentation technique and optical flow can be used enhance the model performances through reducing the computing time through hyper parameters tuning and advance algorithms. Image regression for speed detection of vehicle can be interlocked with automobile ECU to control braking and speed control in dense traffic areas and accident prone areas. Computer vision assisted machines have better visibility range and reaction timing than a human beings, wide angle and multipoint focus is far superior then human retina. In countries like India, Pakistan and Bangladesh where roads are flooded with wildlife and carefree human beings, intelligent decision making can save many human and animal lives by preventing road accidents.

6. ACKNOWLEDGEMENT

We would like to show our gratitude towards IROS 2018 Workshop, 10th Planning, Perception and Navigation for Intelligent Vehicles for providing us with valuable dataset for our research.

REFERENCES

[1]. L Baskar, B De Schutter, and H. Hellendoorn, "Intelligent speed adaptation in intelligent vehicle highway systems – A model predictive control approach," Proceedings of the 10th TRAIL Congress 2008 – TRAIL in Perspective – CD-ROM, Rotterdam, The Netherlands, 13 pp., Oct. 2008.

[2]. D Shukla and E Patel, "Speed determination of moving vehicles using lucas-kanade algorithm," IJCATR, vol. 2, pp. 32–36, 01 2012.

[3]. I Sreedevi, M. Gupta, and P. Asok Bhattacharyya, "Vehicle tracking and speed estimation using optical flow method," International Journal of Engineering Science and Technology, vol. 3, 01 2011.

[4]. A.J.E Atkociunas, R. Blake and M. Kazimianec, "Image processing in road traffic analysis," in Image Processing in Road Traffic Analysis, vol. 10, 2005, pp. 315–332.

[5]. Herranz-Perdiguero, C. and López-Sastre, "ISA2 : Intelligent Speed Adaptation from Appearance", IROS 2018 Workshop, 10th Planning, Perception and Navigation for Intelligent Vehicles (PPNIV' 18)

AN ALGORITHM-ADAPTIVE SOURCE CODE CONVERTER TO AUTOMATE THE TRANSLATION FROM PYTHON TO JAVA

Eric Jin[1] and Yu Sun[2]

[1]Northwood High School, 4515 Portola Parkway, Irvine, CA 92620, USA
[2]California State Polytechnic University, Pomona, CA, 91768, USA

ABSTRACT

In the fields of computer science, there exist hundreds of different programming languages. They often have different usage and strength but also have a huge number of overlapping abilities [1]. Especially the kind of general-purpose coding language that is widely used by people, for example Java, Python and C++ [2]. However, there is a lack of comprehensive methods for the conversion for codes from one language to another [3], making the task of converting a program in between multiple coding languages hard and inconvenient. This paper thoroughly explained how my team designs a tool that converts Python source code into Java which has the exact same function and features. We applied this converter, or transpiler, to many Python codes, and successfully turned them into Java codes. Two qualitative experiments were conducted to test the effectiveness of the converter. 1. Converting Python solutions of 5 United States Computer Science Olympic (USACO) problems into Java solutions and conducting a qualitative evaluation of the correctness of the produced solution; 2. converting codes of various lengths from 10 different users to test the adaptability of this converter with randomized input. The results show that this converter is capable of an error rate less than 10% out of the entire code, and the translated code can perform the exact same function as the original code.

KEYWORDS

Algorithm, programing language translation, Python, Java.

1. INTRODUCTION

There are nearly as many programming languages in this world as human languages, but the conversion between these languages of computers is still a developing field [1]. There are comprehensive translations between all human languages, but there aren't many for programming languages [3]. The solution to this problem is hidden in coding itself, algorithms can be built to convert code between coding languages. This paper is about an algorithm we built that does this task: an algorithm that can translate Python code into Java code. Why choose Python and Java? Because they are among the list of the most used programming languages in the current world [4].

The converter is able to take in a file containing Python source codes and convert it to Java source codes that have the same performance. However, perfection in translation is impossible to achieve due to some of the fundamental differences between the two languages [5]. The converter we built is able to achieve more than 90% of correct translation. This program is useful in many aspects. First it will be a helpful tool to beginners learning these languages, it is convenient with a tool where one can just type in a line of code one already knows and receive the exact same code

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 215-229, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111719

in the language one is learning. Especially in situations when multiple coding languages of the same code might be needed. For example, the United States of America Computer Science Olympics (USACO), a national competition open to all high school students, often have problems that are only doable with certain languages. It saves time and works to avoid code the same program again in another language. Also, actual programming projects, like building an application, might need to have different versions in different languages to meet the needs of the users of different platforms [6].

There has been existing algorithms aiming to transpile one programing language to another [7]. Google's Google Web Toolkit (GWT) turns Java to JavaScript, Facebook's hiphop compiler compiles PHP into C [8]. What our converter is doing is parallel to the transpilers of the two great tech giants: turning one programming language to another on the source code level. Our converter is unique since it is doing transpilation between Java and Python, which is different from the other existing transpilers. However, google and Facebook's transpiler optimize the original source code during the transpilation, this is something that our algorithm is not able to do [8].

Just within the field of transpiling Java and Python, there also exists a method called Jython [9]. It is a plugin of Java which allows users to "freely mix the two languages both during development and in shipping products." according to their website. While looking similar to our converter, this is not the same thing as transpiling. With Jython developers can switch part of a Java code to Python code, having the same ability. but it does not have the functionality of turning one source code to another source code, which is the main purpose of our algorithm.

The approach we took to solve the problem of converting one coding language to another is a method similar to the enumeration method [10]. Which means, using if statements to list out and detect every possible structure that exists in the Python code, and convert each part of the code into its Java version. First of all, after a Python file is inputted, the converter breaks it line by line, for each line it quickly converts the simple parts of the spaces for indentation and comments, so only the meaningful code is carried into enumeration. In a certain order, the algorithm checks for unique words which represent a list of commonly used Python code structures, for example a line containing an isolated "=" sign is dealing with a variable: the right side value needs to be stored into the left side. The sutures are the following: "defining a function/methods", "calling a function", "using a list to store values", "interactions like for loop and while loop", "if statements and booleans", "casting variables to another type", and finally "creating or updating a variable". The order is needed since multiple structures can occur in the same line, for example an if statement which checks some value in a list using a method. Plus all kinds of edge cases not included in the common used code structure listed above, like "open and reading files" "for each loop" "the in function".Each of these code structures were taken out and convert into Java codes, while the other parts of the Python line remain the same. After the Python line comes out of all these checking cases, almost every single part will be converted into Java code, thus it is written to the Java file as output.

This algorithm is focusing on the transpilation of Python source code to Java should contain the following abilities: the width of convertible Python code, the correctness of the output Java code, and the accommodation to any user with different coding habits. We designed two experiments to measure these activities. For Experiment 1, First, we decided to convert solutions written in Python for five United States of American Computer Science Olympic (USACO) algorithm problems [13]. We will measure how much of the converted code, which is in Java, is needed to fix before it can run properly, since the transpilation cannot be perfect. Then we check if the converted Java code is able to output the correct results to the problem just as their Python version, in order to prove the relation actually works.

For the second experiment, we gathered raw Python source code from 10 random coders that have length from short to long. Similar analysis from experiment 1 is applied: the percentage of error is calculated by the amount of incorrectly translated characters out of the total length. Also, a graph of length vs error is plotted to show if there's any correlation between the length of the code and the effectiveness of the converter.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the construction of the converter. Section 3 takes an overall go through of all the codes within thealgorithms, and some close look at the details of the code. Section 4 presents the structure of the two experiments conducted along with analysis of the data generated, followed by the introduction of related similar works in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Arrays

One fundamental difference between the language of Java and Python is that Python has a data structure called list, which can store any kind of elements with any length in one single list structure [11]. In simple words, the type and length of the list is modifiable. However, this feature brings convenience at the cost of using extra memory spaces and slowing the speed of the code. Meanwhile, Java has two similar structures: array and ArrayLists. A Java array has the similar syntax as a Python list (e.g., list[0]) is a code that gets the first element in both a Python list and a Java array. However, the array must be declared with a fixed length and a fixed type, which does not fit with the flexible features of the Python list. On the other hand, the ArrayList needs a fixed type for the element it contains, while having the modifiable length of the list, but its syntax is very different. In the end, after successfully handling the syntax, We had chosen to use ArrayList in the output Java code to replace every list created in the input Python code. So, a (list = [ ]) will be converted into (e.g., ArrayList list = new ArrayList<>();). To solve the fixed data type, we simply declared them all as Objects, which is the fundamental data type in Java. However, this triggers a more complicated issue: casting the elements saved as objects in the ArrayLists to their designated data type whenever they are used.

2.2. Variable

In Python, you can declare a variable as an integer, but later on change its value to a string, the Python syntax generally disregards data types and will not generate a compiling error for any non matching data type mistakes. However, Java syntax is strict with data types. Once a new variable is declared, it requires a fixed data type. Thus, there's no useful information for the type of the variable in the input Python line except the value of the variable itself. The algorithm keeps track of every variable created, both the variable name and its value, by detecting the lines of codes that contain an isolated single "=". If the variable name is not the record, it launches a series of complicated case work to categorize the value of the variable and grants the corresponding Java variables their proper type. This also includes the special case of initializing an ArrayList. With this method, the algorithm is able to cast the elements with type objects in the ArrayListonce they need to be accessed.

2.3. Functions

The ultimate challenge is that, no matter how comprehensive the converter is, it can never cover the entirety of Python to Java conversion. Because each language has their own libraries and functions, the number is countles and increasing everyday. Our converter did not solve this challenge but offers a fairly clean fix to the problem, which is to find the matching pair of functions. For example, both languages contain a math library and can call functions to do mathematical operations. Python's pow(a,b) is the same thing as Java's Math.pow(a,b). Therefore, our converter contains a dictionary of corresponding functions. Once a typical structure of calling functions in the input Python code is identified, the dictionary returns the corresponding Java function. Now it only contains the common functions of Java and Python, but it is easy to add to the dictionary if a new matching pair function is needed.

3. Solution

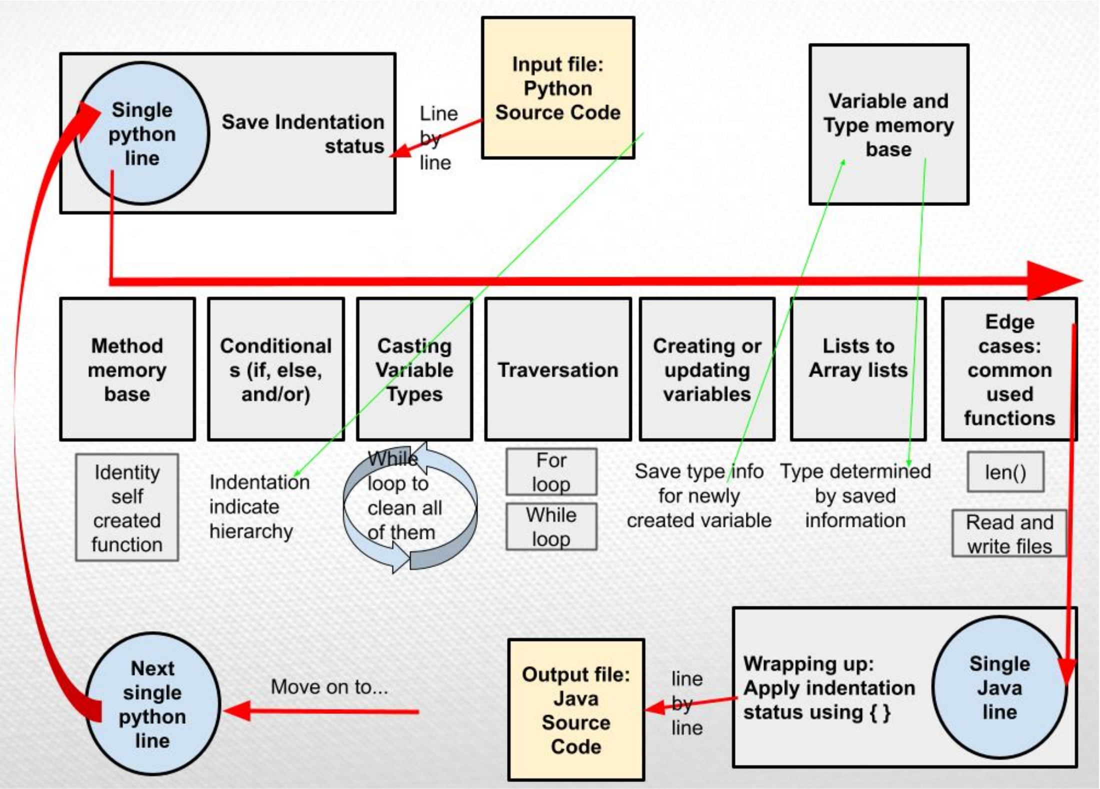

Single python line
Save Indentation status
Line by line
Input file: Python Source Code
Variable and Type memory base
Method memory base
Conditionals (if, else, and/or)
Casting Variable Types
Traversation
Creating or updating variables
Lists to Array lists
Edge cases: common used functions
Identity self created function
Indentation indicate hierarchy
While loop to clean all of them
For loop
While loop
Save type info for newly created variable
Type determined by saved information
len()
Read and write files
Next single python line
Move on to...
Output file: Java Source Code
line by line
Wrapping up: Apply indentation status using { }
Single Java line

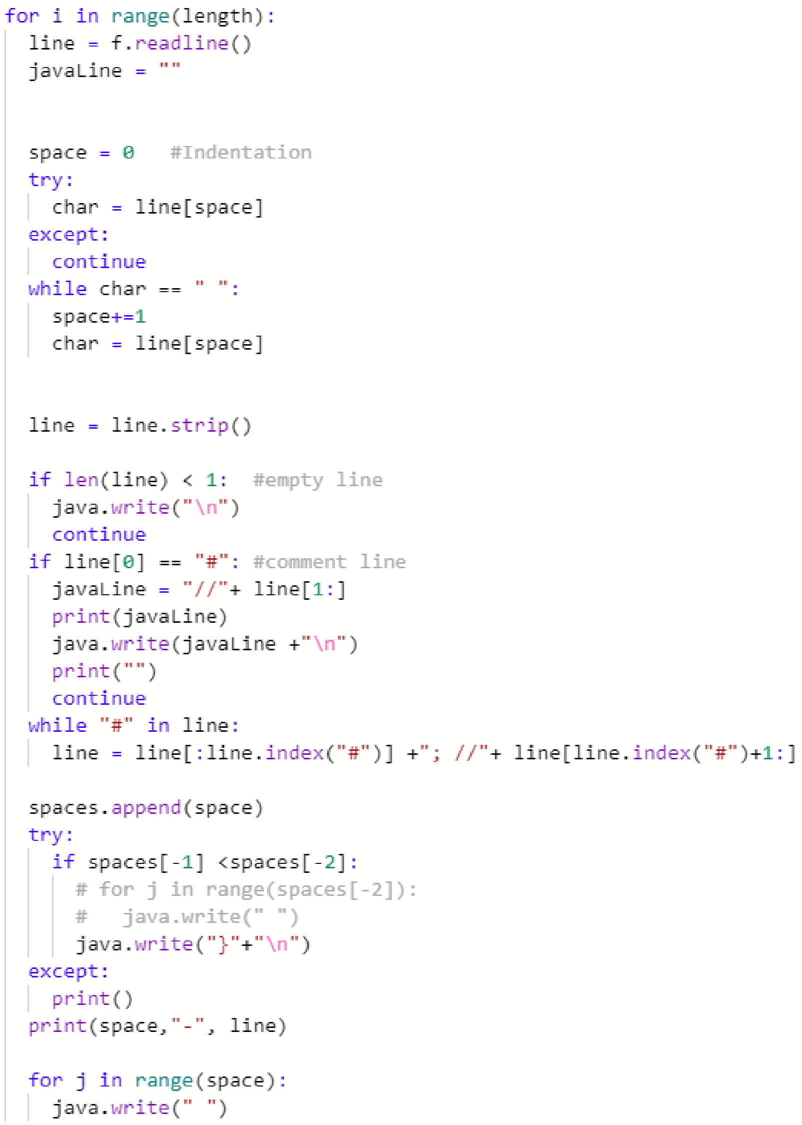
Figure 1. Overview of the solution

The converter is an algorithm coded in Python that converts Python source code to Java source code with the same function [3]. It does it in a line-by-line process, in a way similar to the enumeration method. Usually each line is analyzed separately, meaning the algorithms treat each line as converting the same thing as if it is the only imputed Python line. Focusing on the actual process, it uses string parsing to identify certain structures in the code and convert it to Java code. The structure it is looking for are the following in this order: calling or creating methods/functions; conditional; casting variables; for loop and while loop; dealing with variables; Python list; and other edge cases of commonly used functions. Besides the line-by-line process, there is certain information that is meaningful to the entirety of the code, such as variables and methods used throughout code. The algorithm observes variables and functions when they are first created and saves them with their parameters and type for later use. Reading one Python line, writing one Java line after all Python code is converted, the Java code is outputted as a Java file.

3.2. Handling with indentation:

Example:

```
list = []                                         ArrayList list = new ArrayList<>();
#   a for loop adding value to list     →         //   a for loop adding value to list
for i in range(5):                                for(int i = 0; i < 5; i+=1 ) {
list.append(i)                                 list.add(i);
print(list)                                         }
                                                System.out.println(list);
```



```
for i in range(length):
  line = f.readline()
  javaLine = ""


  space = 0   #Indentation
  try:
    char = line[space]
  except:
    continue
  while char == " ":
    space+=1
    char = line[space]


  line = line.strip()

  if len(line) < 1:  #empty line
    java.write("\n")
    continue
  if line[0] == "#": #comment line
    javaLine = "//"+ line[1:]
    print(javaLine)
    java.write(javaLine +"\n")
    print("")
    continue
  while "#" in line:
    line = line[:line.index("#")] +"; //"+ line[line.index("#")+1:]

  spaces.append(space)
  try:
    if spaces[-1] <spaces[-2]:
      # for j in range(spaces[-2]):
      #   java.write(" ")
      java.write("}"+"\n")
  except:
    print()
  print(space,"-", line)

  for j in range(space):
    java.write(" ")
```

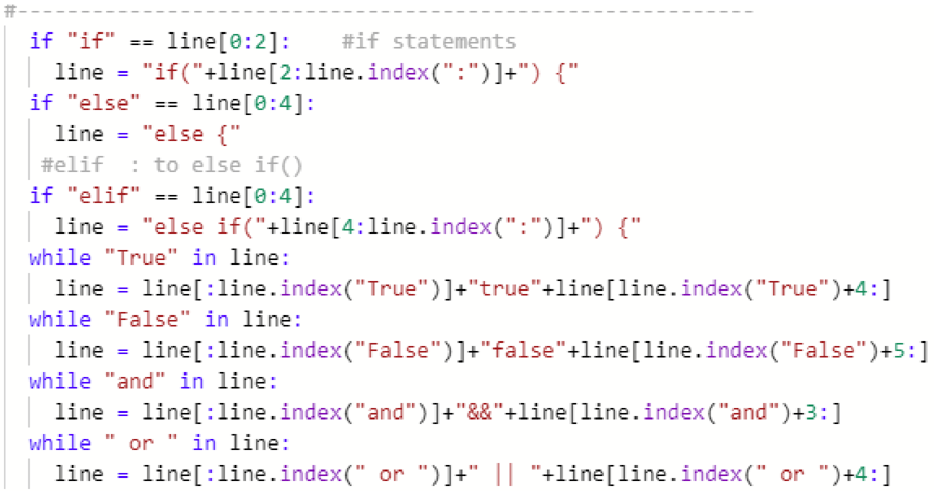
Figure 2. Code for handling with indentation

The body of this algorithm is a for loop, which reads the file containing the Python source code line by line. the “JavaLine” variable is created so save the Java version of the currentline. First the situation of an empty line is detected, so the code can go on counting how many spaces are in front of the first character in this line. This number indicating the indentation is recorded in a list, then all the spaces are stripped off the line so only the meaningful code is passed on to the following analysis. This section of the code focuses on the important information revealed by indentation. Since the Python indentation determines the hierarchy of the code, it needs to be converted into pairs of “{ “ and “}”, the corresponding structure in Java. The opening “{" will be taken care of later in the algorithm when a “:” is see at the end of the code, so the current section only needs to add closing “}” when needed. To do this, the algorithm traverses through the list of spaces, whenever a line has less space than its previous one, it means the closure of a prior level. Thus, a closing “}” will be added to the javelin. However, certain indentation structures cannot be identified by this method, so sometimes the user needs to manually fix the “{“ and “}” in the Java code.

3.2.1. If statements and logical operations:

Example:

```
if a == 2 and (True or False):              if( a == 2 && (true || false)) {
        print("case 1")              →              System.out.println("case 1");
                                             }
else:                                       else {
        print("case 2")                             System.out.println("case 2");
                                            }
```



```
#------------------------------------------------------------
if "if" == line[0:2]:    #if statements
  line = "if("+line[2:line.index(":")]+") {"
if "else" == line[0:4]:
  line = "else {"
 #elif  : to else if()
if "elif" == line[0:4]:
  line = "else if("+line[4:line.index(":")]+") {"
while "True" in line:
  line = line[:line.index("True")]+"true"+line[line.index("True")+4:]
while "False" in line:
  line = line[:line.index("False")]+"false"+line[line.index("False")+5:]
while "and" in line:
  line = line[:line.index("and")]+"&&"+line[line.index("and")+3:]
while " or " in line:
  line = line[:line.index(" or ")]+" || "+line[line.index(" or ")+4:]
```

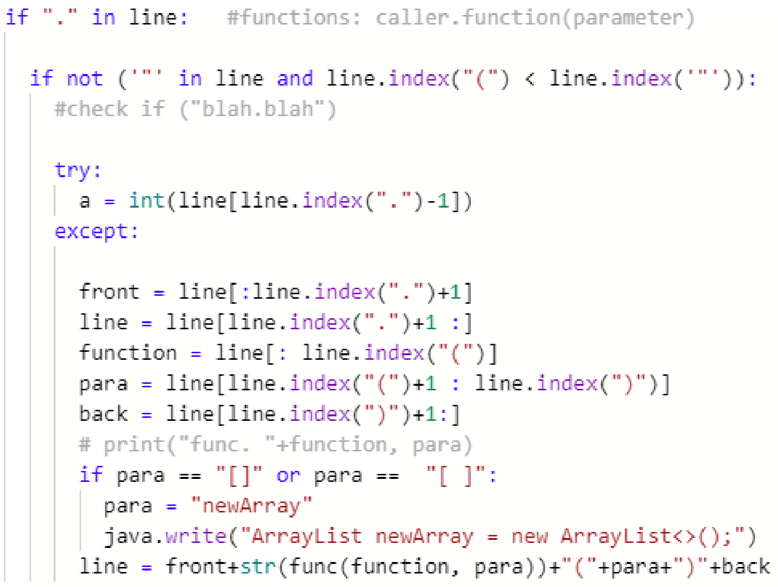
Figure 3. Code for statements and logical operations

How the algorithm handles if statements work like this: picks out the part of the code where Java and Python is different and turns it into the Java version. This specific segment is simple, which makes it a typical example. First, if “if” , “else” or “else if” is seen in the front of the line, it immediately makes this line part of an if-else structure with no exception. Therefore, the only thing needed to do is to do some small editing of syntax, adding parentheses and braces. Another thing different between Python and Java is how they write “true” “false” “and” “else”, while loops are needed to find all of them since multiple can exist in the same line.

3.2.2. Functions:

Example:

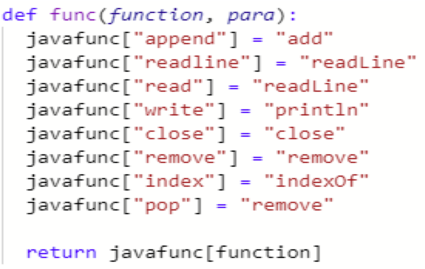
list1.append([ ]) → ArrayListnewArray = new ArrayList<>(); list1.add(newArray);
list2.pop(0) → list2.remove(0)



```
if "." in line:   #functions: caller.function(parameter)

  if not ('"' in line and line.index("(") < line.index('"')):
    #check if ("blah.blah")

    try:
      a = int(line[line.index(".")-1])
    except:

      front = line[:line.index(".")+1]
      line = line[line.index(".")+1 :]
      function = line[: line.index("(")]
      para = line[line.index("(")+1 : line.index(")")]
      back = line[line.index(")")+1:]
      # print("func. "+function, para)
      if para == "[]" or para ==  "[ ]":
        para = "newArray"
        java.write("ArrayList newArray = new ArrayList<>();")
      line = front+str(func(function, para))+"("+para+")"+back
```

Figure 4. Code for functions

The first line of code detects if a dot”.” is presented in the line variable, which is a string containing the Python line. The second line checks if the dot exists as part of a string by checking if the quotation mark is also in the line. However, even if it is not in a string, another unwanted situation is when a dot is used as a decimal point between numbers like “1.25”, the try conditional excludes it by checking if the character before the dot is a number. If those are not the case, the algorithms got what it is looking for: calling functions. Normally a dot in the Python syntax has the usage of calling a function from an imported dependency, which usually has the structure of “caller.function(parameter)” . The algorithm uses some string slicing to separate each part, among these parts the function is replaced by the Java version of the function by calling the func() dictionary containing this info, which will be explained in delta below. The parameter is also important, when the parameter is a list, Java needs a new ArrayList to be created, which is what the last chunk of the code is doing. After getting all the parts in Java, the algorithm reassembles them into one Java line calling a Java function.



```
def func(function, para):
  javafunc["append"] = "add"
  javafunc["readline"] = "readLine"
  javafunc["read"] = "readLine"
  javafunc["write"] = "println"
  javafunc["close"] = "close"
  javafunc["remove"] = "remove"
  javafunc["index"] = "indexOf"
  javafunc["pop"] = "remove"

  return javafunc[function]
```

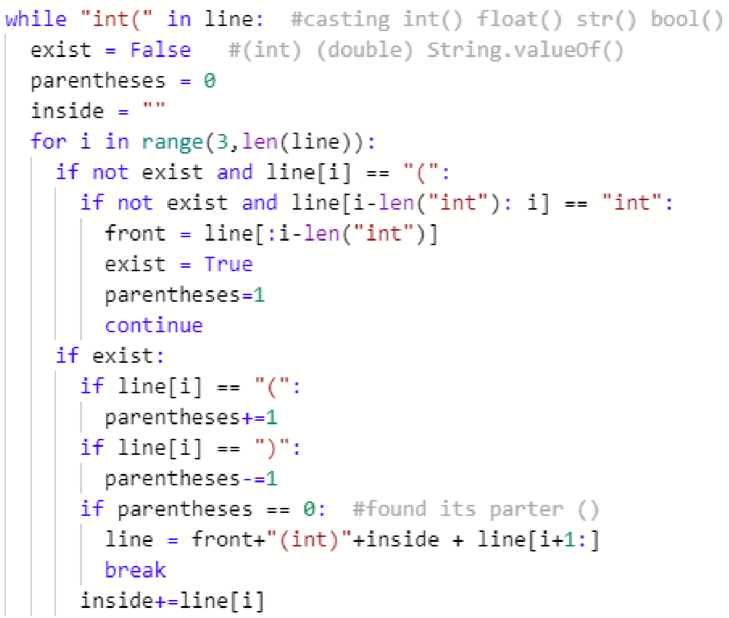
Figure 5. Part code for the function memory base

This is part of the code for the function memory base, which is used for saving the Python and Java functions that are interchangeable. It can be updated to include more pair as times goes on, depending on the dependencies used in the Python code.

3.2.3. Casting variable types:

Example:

int(number1)	→	(int) number1;
float(number2)	→	(double) number2;
str(x)	→	String.valueOf(x);
float (str (int (a)))	→	(double) String.valueOf((int) a) ;



```
while "int(" in line:  #casting int() float() str() bool()
    exist = False   #(int) (double) String.valueOf()
    parentheses = 0
    inside = ""
    for i in range(3,len(line)):
        if not exist and line[i] == "(":
            if not exist and line[i-len("int"): i] == "int":
                front = line[:i-len("int")]
                exist = True
                parentheses=1
                continue
        if exist:
            if line[i] == "(":
                parentheses+=1
            if line[i] == ")":
                parentheses-=1
            if parentheses == 0:  #found its parter ()
                line = front+"(int)"+inside + line[i+1:]
                break
            inside+=line[i]
```

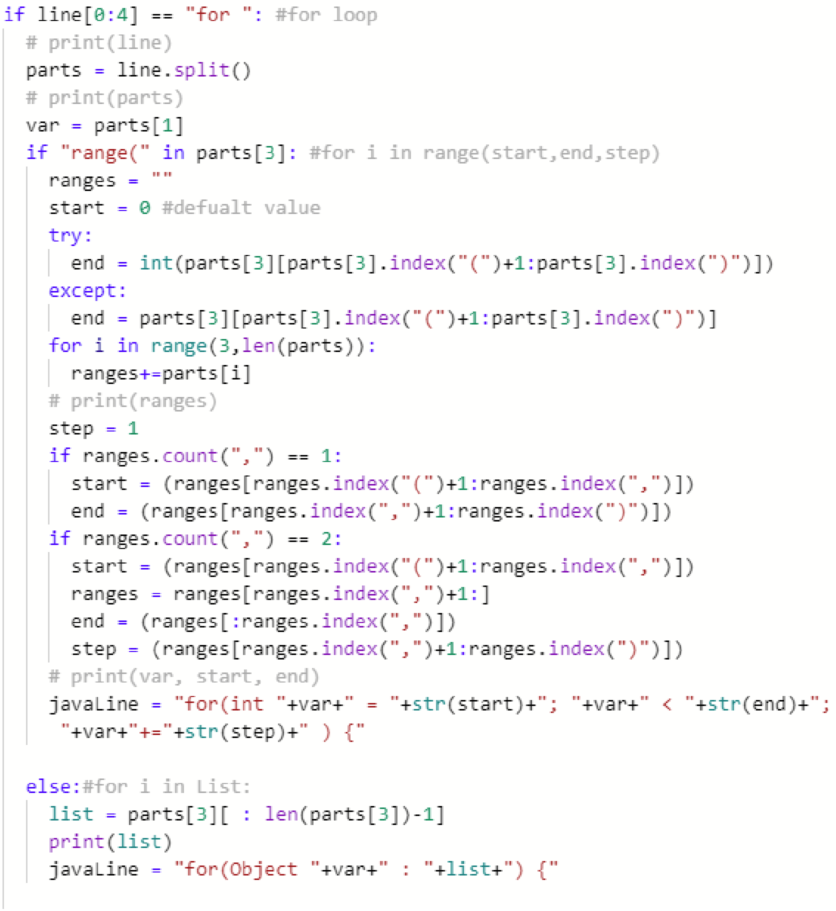
Figure 6. Code for casting variable types

The code section above it for converting the casting of integers which means turning int(a) into (int) a, for other variable types (string, boolean float) the code is nearly the exact same as this one so it is enough to look closely at just this one for integer. Recognizing the casting of variables is different from other structures since it may appear many times in one single Python line, making the while loop a fit tool to find and change all of them. This also means the algorithm would need to identify the pair of parentheses belonging to this specific cast action. After identifying any “int(“ still existing in the current line, the algorithm starts to count the parenthesis after it. Only if the counter (variable parenthesis) notices that the number of open parenthesis and close parenthesis matches, it is certain that which part of the code is the cast actually casting. Thus, the same process of breaking up the line and resembling it into Java syntax is applied, and this specific cast has been successfully converted. However, the algorithm will not move on until all the casting in the current line is finished, making it capable of handling nested casting operations.

3.2.4. For loop

Example:

for i in range(10):	→	for(int i = 0; i < 10; i+=1) {
for a in range(0,50,2):	→	for(int a = 0; a < 50; a+=2) {
for b in range(n):	→	for(int b = 0; b < n; b+=1) {
for value in elements:	→	for(Object value : elements) {



```
if line[0:4] == "for ": #for loop
    # print(line)
    parts = line.split()
    # print(parts)
    var = parts[1]
    if "range(" in parts[3]: #for i in range(start,end,step)
        ranges = ""
        start = 0 #defualt value
        try:
            end = int(parts[3][parts[3].index("(")+1:parts[3].index(")")])
        except:
            end = parts[3][parts[3].index("(")+1:parts[3].index(")")]
        for i in range(3,len(parts)):
            ranges+=parts[i]
        # print(ranges)
        step = 1
        if ranges.count(",") == 1:
            start = (ranges[ranges.index("(")+1:ranges.index(",")])
            end = (ranges[ranges.index(",")+1:ranges.index(")")])
        if ranges.count(",") == 2:
            start = (ranges[ranges.index("(")+1:ranges.index(",")])
            ranges = ranges[ranges.index(",")+1:]
            end = (ranges[:ranges.index(",")])
            step = (ranges[ranges.index(",")+1:ranges.index(")")])
        # print(var, start, end)
        javaLine = "for(int "+var+" = "+str(start)+"; "+var+" < "+str(end)+";
         "+var+"+="+str(step)+" ) {"

    else:#for i in List:
        list = parts[3][ : len(parts[3])-1]
        print(list)
        javaLine = "for(Object "+var+" : "+list+") {"
```

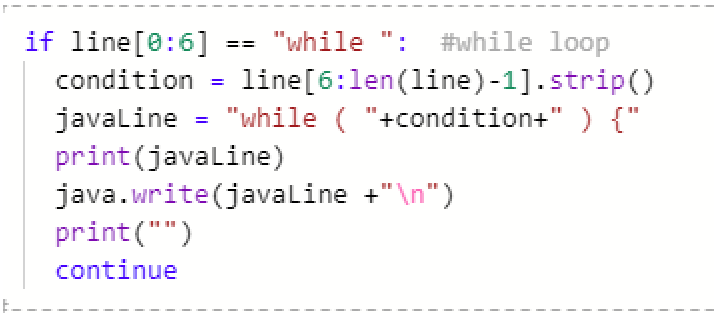
Figure 7. Code for loop

If the Python line starts with “for” it is definitely a for loop (or a for each loop). A Python for loop consists of up to 3 parts: start, end, step, in the form of “for i in range (start, end, step). The Java version will be for (int i = start; i< end; i+= step). However, usually the only parameter used by programmers is only the end, to repeat the loop for x times. So the algorithm recognizes the different situations by splitting the Python line and counting how many commas within the range () string. if either step or start is not used, they will be set to the default value 1 and 0. Then, all the parts including the variable (var), start, end and step are resembled back in Java syntax. However, the string slicing has the shortcomes of being unable to deal with extra spaces in the Python line, so it is dependent on the assumption of the syntax of the imputed Python code.

Similarly, a for each loop is converted by grabbing the single element and the group of elements from the Python line and putting it in Java format of “for (Object element: elements) {“. notice that the type is settled to be Object since it cannot be determined from the original Python code.

3.2.5. While loop

while a== 2: → while (a == 2) {



```
if line[0:6] == "while ":  #while loop
    condition = line[6:len(line)-1].strip()
    javaLine = "while ( "+condition+" ) {"
    print(javaLine)
    java.write(javaLine +"\n")
    print("")
    continue
```

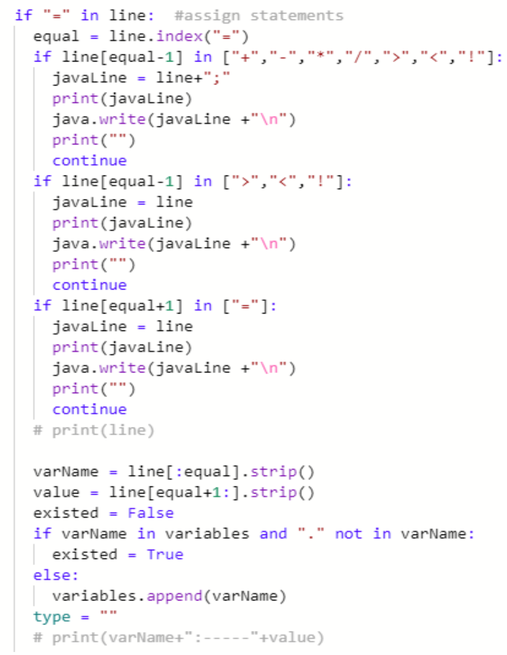
Figure 8. Code for while loop

On the other hand, the while loop is much simpler than the for loop. Just detecting the line starting with “while” will eliminate other possibilities. The process of conversion is extracting the boolean statement and putting it in parentheses.

3.2.6. Variables

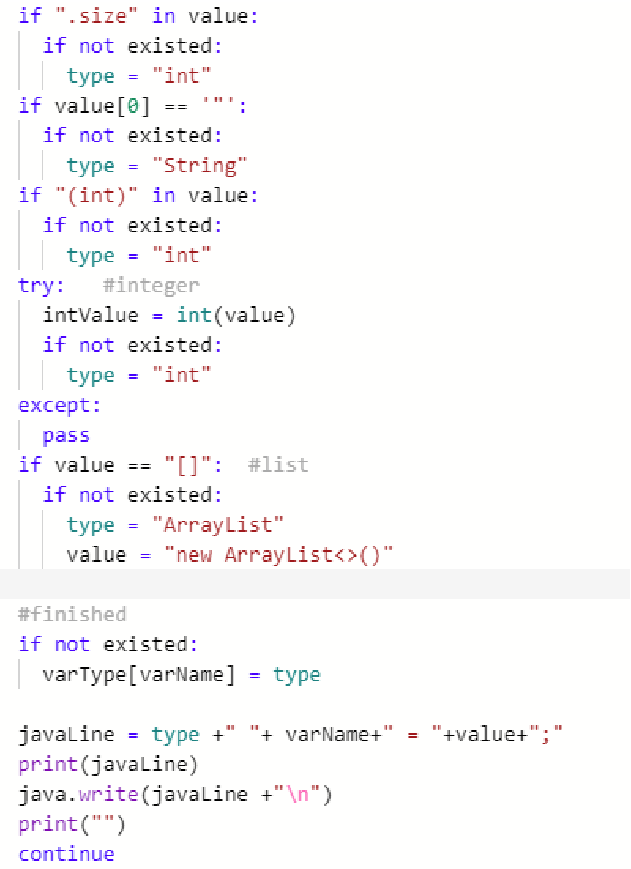
Example:

a = 2		int a = 2;
a = a * 3		a = a * 3;
name = “Eric”	→	String name = “Eric”;
name = name + “Jin”		name = name + “Jin”;
list = []		ArrayList list = new ArrayList<>();



```
if "=" in line:  #assign statements
    equal = line.index("=")
    if line[equal-1] in ["+","-","*","/",">","<","!"]:
        javaLine = line+";"
        print(javaLine)
        java.write(javaLine +"\n")
        print("")
        continue
    if line[equal-1] in [">","<","!"]:
        javaLine = line
        print(javaLine)
        java.write(javaLine +"\n")
        print("")
        continue
    if line[equal+1] in ["="]:
        javaLine = line
        print(javaLine)
        java.write(javaLine +"\n")
        print("")
        continue
    # print(line)

    varName = line[:equal].strip()
    value = line[equal+1:].strip()
    existed = False
    if varName in variables and "." not in varName:
        existed = True
    else:
        variables.append(varName)
    type = ""
    # print(varName+":-----"+value)
```

Figure 9. Code for variables (1)



```
if ".size" in value:
    if not existed:
        type = "int"
if value[0] == '"':
    if not existed:
        type = "String"
if "(int)" in value:
    if not existed:
        type = "int"
try:   #integer
    intValue = int(value)
    if not existed:
        type = "int"
except:
    pass
if value == "[]":  #list
    if not existed:
        type = "ArrayList"
        value = "new ArrayList<>()"

#finished
if not existed:
    varType[varName] = type

javaLine = type +" "+ varName+" = "+value+";"
print(javaLine)
java.write(javaLine +"\n")
print("")
continue
```

Figure 10. Code for variables (2)

The algorithm identifies equal signs "=" in the current line, if the equal sign is preceded with a mathematical operation sign, that means the Java version is the same as the Python version. So the line simply gets written to the Java file, since at this point of the algorithm other Python parts in the current line are already converted. However, the case of an isolated equal sign is where a variable is created, or the value of a variable is changed. Here a major difference between Python and Java occurs: Java requires a fixed type to be assigned to variables that are first created. The algorithm counters this issue by saving created variables to a dictionary(varName : type), so excited variables can be extinguished from newly created variables. A line involving creating a variable has two components, the variable name on the left side, and whatever the value is on the right side. By detecting certain features for the value, like quotation marks for string and "[]" for ArrayList, in most cases a fixed type for the can be determined. Then this information is saved to the dictionary mentioned earlier. However, due to the major difference that Python treats everything as an object, in some case the variable type cannot be determined by just looking at the Python code, thus the user needs to manually provide this information.

4. EXPERIMENT

Experiment 1: Correctness

The convertor takes in Python code and converts it line by line to Java code. In order to determine the correctness of the translated code, we designed an experiment using Python code samples from United States of America Computer Science Olympic (USACO) questions. The USACO is a competition consisting of coding problems of different difficulties about algorithms, and different numbers of test cases required to pass for each question [13]. We first randomly selected 5 problems from the USACO, 2 bronze level and 3 silver level, and wrote Python solutions to each of them which successfully passed all the cases. Then, put the 8 Python code into the converter one by one, to generate 8 Javatranslations of the code. Therefore, since the converter cannot produce 100% perfect Java codes, we are able to count the number of characters that needed to be manfully fixed in the Java code before the code executed with no trouble. This number indicates the correctness of the translated code. Finally, the ability of the Java code is supported by observing if it can pass the same problem just like the original Python code does. Premise: the 8 Python codes are able to pass the 5 bronze problem and 3 silver problems.

Table 1. Data of correctness

correctness (changed characters / total characters)	proof of functionality
79 / 917 = 8.6 %	pass (bronze)
28 / 883 = 3.2%	pass (bronze)
132 / 1131 = 11.7%	pass (silver)
75 / 725 = 10.3%	pass (silver)
158 / 1263 = 12.5%	pass (silver)
average: 9.26%	

The percentage on the left indicated the ratio of characters was manually fixed in the converted code due to unavoidable syntax and logical error, until the code executed without mistake. This is the measurement of the correctness of the translation process. The average percentage of error is 9.26%.

After fixing the code, the experiment demonstrates it works just as fine as the original Python code by seeing if the Java code is able to pass for each USACO problem. Fortunately, the result is indicating that after fixing a certain numberof mistakes. The converted code has the exact abilities and functions as the original code. Overall, the average percent of characters needed to be fixed in the translated code is 9.26%. The conclusion can be drawn from this experiment is: if the Python code is in the field of algorithms building, after fixing about 9% of the converted Java code, it will have the exact same function as the original Python code.

Experiment 2: Adaptability

Experiment one uses Python code inputs from only one coder, who is the builder of this convertor. However, different coders have different coding habits, which could influence the function of the convertor. To test the effectiveness of this convertor on different coders, we gathered 10 Python code samples from 10 different coders, ranging from 10 lines to 200 lines, and used the converter to transpile it into Java code. Analysis done on the 10 outputs is shown below:

Table 2. Data of Adaptability

total characters	117	225	102	32	605	289	473	632	1154	1141
changed characters	9	13	6	2	51	19	23	16	45	73
percentage of changed	7.7%	5.8%	5.9%	6.3%	8.4%	6.5%	4.8%	2.5%	3.9%	5.1%

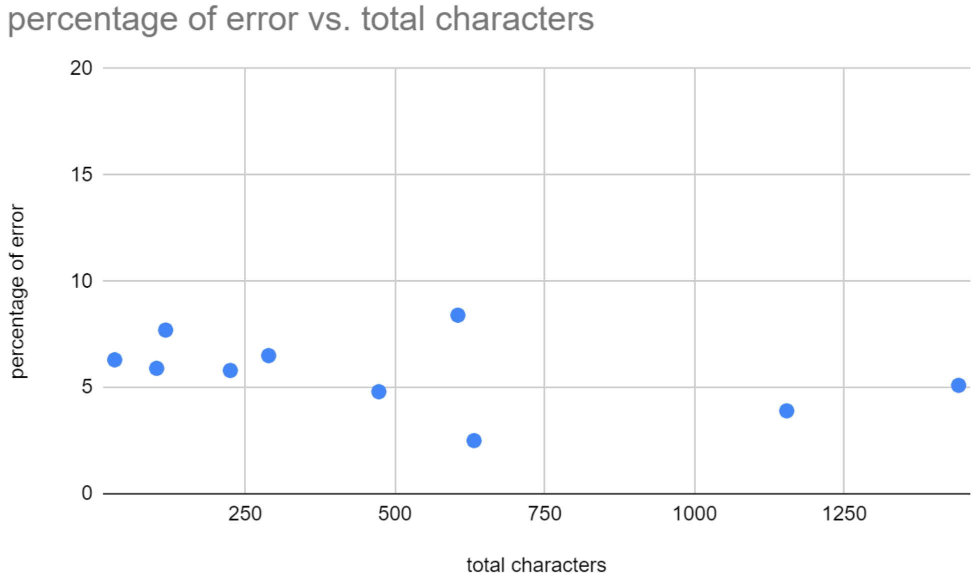


Figure 11. Percentage vs. Characters

The 10 translated Java codes each have length from 60 characters to more than 1000 characters. Using the same calculation method as experiment 1: the number of characters that did not get converted successfully out of the total characters = the percentage of error. As can be seen, for different coders, the percentage of error of translation fluctuates between 2% - 9%. Thus, we can say that the converter has good adaptability to different coders. However, more question remains: could the length of the code be an issue? would the converter be less effective if the code is longer? To answer this, a graph of percentage of error vs total characters is constructed showing there's no obvious correlation between the length of the code and the correctness of translation. Thus, the conclusion can be made is this: the converter has a 2% -9% error rate in the translation of Python to Java code, depending on the coder and length of the input code.

The goal of the two above experiments is to prove the trustability of the converter in terms of measuring its correctness of translation and adaptability to various kinds of inputs. Experiment 1 shows that the complicated algorithmic Python codes from USACO have only 9.2% of error rate, and the converted code can perform the extent function as the original code. Experiment 2 is testing how well the converter performs under different situations. Observation supporters that the error rate is 2% - 9%, and the converter works effectively with different coders and various input lengths. Although the sample size used in the two experiments cannot be judged as large data, overall, the error rate never exceeds 10% of the entire code piece. Thus, the converter is trustable with a translation rate more than 90% under any situation.

5. Related Work

Lachaux et al. presented a programming code translator using machine learning [14]. They selected a number of source code GitHub repositories and trained a machine learning model to automatically translate the code from one language to another. Although machine learning allows a more automated approach to perform the translation, it cannot capture all the possibilities or rules, and there is always an accuracy issue. Our work is totally based on the rules, so we have a more reliable foundation to guarantee the accuracy. In addition, our work specifically focuses on Java to Python translation in order to improve the accuracy, while their work targets a more generalized language transition.

Abazyan et al. demonstrated their version of an interlanguage translation for Python and Java, it is an algorithm that translates from source to source, and it uses machine learning to correct the translation whenever the accuracy is lower than 60% [15]. Their algorithm is able to conduct back translating, which means it can perform both Python to Java and Java to Python translation. Although their method does have a more general usage, the reliance of machine learning to fix low accuracy translations shows the instability of translation. Our work focuses on only handwriting algorithms and involves no AI, which generates a more stable accuracy among the translation from Python to Java.

Aggarwal et al. had constructed a tool to convert Python 2 codes into the higher version Python 3, it uses statistical machine translation, which is a technique used in natural language translation [16]. Their score for evaluating the accuracy is as high as 99.37. The conversion of code from a previous version to a newer version of the same coding language looks similar to cross language translation, but they are fundamentally different. While our converter is doing the task of Python to Java translation, similar challenges exist in the analysis of Python source code are also mentioned in Aggarwal‘s article. However our work can not achieve that high of an accuracy due to the other challenges in cross language translation.

6. Conclusions

In this paper, we presented the converter we built for the task of cross language translation among programming languages, more specifically, from the wide-use language Python to Java. Our converter takes in Python source code and outputs Java source code as close as the original code. Our converter is an algorithm handcrafted in Python, the philosophy it uses is enumeration, which means testing out every possible structure contained in a Python code. The translation process is done line by line, with the algorithm detecting certain structures in the Pythoncode and turning each piece into its corresponding Java version. The accuracy of our converter is about 90%, without considering the involvement of Python dependency in the input. This is supported by the two experiments we had conducted. The first one tested the converter on 5 USACO programming problems, the Java translation has about 9% error compared to the original Python solutions. The second experiment tested the converter with code inputs from 10 different users, as well as varying length. The accuracy fluctuates around 92% to 98%.

Accuracy: all though that full accuracy in the translation from Python to Javacannot be achieve with the enumeration method that we had implemented, the current accuracy at roughly 92% is definitely not the upper limitation

Accessibility to the public: the goal of building this converter is never about keeping it to ourselves, instead it should be a tool benefiting the public, in the current moment it is still a piece of code. We will work on how to make it accessible to everyone.

Error detection system: since the converter will encounter parts of the code which it cannot convert into Java, it will be reasonable for it to have a feature of labeling those parts in the code to notify the users.

Our team will continue to work on modifying our algorithms to improve the accuracies of the converter and add the error detection function to it. Thus, it will be ready as a convenient tool open to the public. We will publish it as a web application, anyone can go to the domain name and use our tool for their own purpose.

REFERENCES

[1] Pierce, Benjamin C., and C. Benjamin. Types and programming languages. MIT press, 2002.

[2] Dijkstra, E. W., and EdsgerWybe Dijkstra. "Programming languages." Co-operating Sequential Processes. Academic Press, 1968. 43-112.

[3] Qiu, Lili. Programming language translation. Cornell University, 1999.

[4] Cass, Stephen. "The 2015 top ten programming languages." IEEE Spectrum, July 20 (2015).

[5] Lo, Chieh-An, Yu-Tzu Lin, and Cheng-Chih Wu. "Which programming language should students learn first? A comparison of Java and Python." 2015 International Conference on Learning and Teaching in Computing and Engineering. IEEE, 2015.

[6] Floyd, Robert W. "The syntax of programming languages-a survey." IEEE Transactions on Electronic Computers 4 (1964): 346-353.

[7] Chaganti, Prabhakar. Google Web Toolkit. Packt Publishing, 2007.

[8] Zhao, Haiping, et al. "The HipHop compiler for PHP." ACM SIGPLAN Notices 47.10 (2012): 575-586.

[9] Juneau, Josh, et al. The definitive guide to Jython: Python for the Java platform. Apress, 2010.

[10] Beckwith, Mary, and Frank Restle. "Process of enumeration." Psychological Review 73.5 (1966): 437.

[11] Arnold, Ken, James Gosling, and David Holmes. The Java programming language. Addison Wesley Professional, 2005.

[12] Lutz, Mark. Programming Python. " O'Reilly Media, Inc.", 2001.

[13] USACO, www.usaco.org/.

[14] Lachaux, Marie-Anne, et al. "Unsupervised translation of programming languages." arXiv preprint arXiv:2006.03511 (2020).

[15] Abazyan, Suren, Narek Mamikonyan, and Vakhtang Janpoladov. "Interlanguage Translation Utility with Integrated Machine Learning Algorithms." Open Access Library Journal 7.5 (2020): 1-5.

[16] Aggarwal, Karan, Mohammad Salameh, and Abram Hindle. Using machine translation for converting Python 2 to Python 3 code. No. e1817. PeerJPrePrints, 2015.

BUSINESS INTELLIGENCE AND DATA WAREHOUSE TECHNOLOGIES FOR TRAFFIC ACCIDENT DATA ANALYSIS IN BOTSWANA

Monkgogi Mudongo, Edwin Thuma, Nkwebi Peace Motlogelwa,
Tebo Leburu-Dingalo and Pulafela Majoo

Department of Computer Science,
University of Botswana, Gaborone, Botswana

ABSTRACT

Road traffic accidents are a serious problem for the nation of Botswana. A large amount of money is used to compensate those who are affected by road accidents. Traffic accidents are one of the major causes of Deaths in Botswana. It is important for relevant organizations to have a reliable source of data for accurate evaluation of traffic accidents. Similarly, data on vehicle registration must be transformed and be readily available to assist managerial decision makers. In this article, we deploy a Business Intelligence (BI) and Data Warehouse (DW) solution in an attempt to assist the relevant departments in their road traffic accidents and vehicle registration evaluation. In Our evaluation of the traffic accidents our findings suggest that across accident severity, Damage Only accidents had the most interesting recent trend with a 11.93% decrease in the last 3 years on record. Count of Accident Severity for Damage Only accidents dropped from 13,491 to 11,881 between 2018 and 2020 whilst Minor accidents experienced the longest period of growth. Most accidents take place in rural locations and more accidents take place during the weekend. At 28,439, Sunday had the highest number of accidents and was 47.59% higher than Wednesday, which had the lowest count of accidents at 19,269. The results for vehicle registration reveal that the number of vehicle registration decreased for the last 3 years on record. The number of vehicles registered dropped from 65535 to 24457 during its steepest decline between 2019 and 2021.

KEYWORDS

Business Intelligence, Data Warehousing, ETL, Accident and Vehicle registration.

1. INTRODUCTION

Organizations generate and accumulate data on daily basis. These data become so huge with time and can be of no help if the right tools are not deployed to process it. It is imperative that data be converted to information then knowledge and ultimately to wisdom that decision makers can use to make quality decisions thus the need to manage information and knowledge in any organization. Due to constant change in business environment, organizations are forced or prompted to respond quickly to the changes in the environment [1]. It is imperative for organizations to have tools and technologies that will help them transform data into valuable information to assist in comprehensive managerial decision making. Traffic departments in Botswana are not exempt to the ever-increasing data. For example, the traffic department under the Botswana police captures data on accidents while the Department of Road Transport and Safety (DRTS) registers new vehicles and issue driver licenses daily. Road accidents are a serious

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO, NLDM - 2021
pp. 231-247, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111720

concern as a lot of lives have been lost to accidents. Furthermore, there has been a growing number of vehicles in the roads in the past years. It is against this background that decision makers must be provided with the right information, in the right format in order to help them make decisions that will assist in injury prevention and in vehicle control and management. This can be made possible by building a BI platform that will integrate data from different data sources and transform these data to quality information. This process will help identify the information needed to support decisions at different organizational levels more especially the strategic decision makers. Even though various departments deal with different aspects of traffic issues, the need for a consolidated data repository cannot be over emphasized. Currently data on traffic and accidents is stored in different and dispersed data sources. This poses a problem because there is no central data store for easy access and evaluation of accidents and vehicle registrations as well as other issues pertaining to traffic data. The capturing and recording of road traffic accidents is done by police department and they are the custodians of that data while vehicle registration is done by DRTS. Therefore, there is a need for a BI platform to integrate data from disparate system sources, which can then be prepared for analytical use. The other limitation is that the data is kept in transactional databases, which are limited in the sense that they store current transactions for ongoing business processes. On the other hand, a BI platform has a data warehouse which has the capability to store large quantities of historical and current data, which enables fast, complex queries across all the data, typically using Online Analytical Processing (OLAP) and Power BI technologies.

In this article, we develop a BI solution for traffic accidents and vehicle registration data analysis in order to avail information required for proper decision making. First, we develop a data warehouse for our BI solution. In developing this data warehouse, we carry out the following steps: we identify the data sources, then identify dimensions, design the DW star schema and we demonstrate the ETL process. This solution model is developed to address the following selected strategic areas:

- **Traffic Accidents Analysis**

This area seeks to investigate the traffic accidents in relation to location, causalities, driver details, vehicle details and road details. This investigation will unearth details pertaining to accident circumstances. The following are sample strategic questions that will be addressed later;

1. How many accidents have been registered across the years as classified by severity and year?
2. Who are involved in these accidents as classified by gender, class and age?
3. Where do most of the accidents happen?
4. When do these accidents happen as classified by time and day of the week?
5. What is the status of the roads where accidents happen?

- **Vehicle Registrations Analysis**

This strategic area provides an investigation into the vehicle registration to understand vehicle registration by model/make, year of registration and registration location. The following are some of the strategic questions which managerial decision makers may be interested in;

1. How many vehicles were registered in the past five years?
2. Where are most vehicles registered?
3. How many vehicles are currently active?
4. Which vehicle model is mostly registered?

The rest of this article is organized as follows: In Section II, we present Related Works, this is followed by a description of data and its sources in Section III. Section IV describes the data warehouse development methodology while section V is about identifying data warehouse dimensions. In section VI, we represent the data warehouse dimensions with a star schema, it is followed by business intelligence platform, the ETL process, data analysis and results and lastly conclusion and future work.

2. RELATED WORKS

To date, several studies have demonstrated that Business Intelligence solutions can be used in critical decision making such as in insurance as well as in traffic accident evaluation. Traffic accidents are a serious worldwide problem that requires robust action from decision makers. According to Hemalatha & Krishnaveni [2] a traffic collision occurs when a road vehicle collides with another vehicle, pedestrian, animal, or geographical or architectural obstacle resulting in injury, property damage, and death. Data warehouse is one of the important BI components that plays a vital role in integrating data sources. The concept of data warehousing arose in mid 1980s with the intention to support huge information analysis and management reporting [3]. Kimball [4] defines a Data Warehouse as a source of data in the enterprise that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making. Moreover, a Data Warehouse is a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions [5]. Data warehousing is critical in data consolidation where there exists dispersed and isolated data systems. For the remainder of this section, we present previous works that describe how other countries have consolidated traffic data from different departments through BI and data warehousing. Lastly, we summarize what has been found through literature in section C.

2.1. Business Intelligence

Companies accumulate and generate data through transactions as they interact with customers on daily basis. BI systems and tools play a vigorous role in bringing together data sources in order to extract the right information from data generated by companies. Without the right tools and technologies, it is difficult to make use of the data as it increases over the years. In modern business, increasing standards, automation and technologies have led to vast amounts of data becoming available [6].

BI is simply a combination of various components and tools that work together collectively to transform operational data into quality information useful for decision making. BI is the process of taking large amount of data, analysing that data and presenting a high-level set of reports of business action to enable management to make fundamental daily business decisions [7]. The objective of BI is to improve the timeliness and quality of information, and enable managers to be able to better understand the position of their firm as in comparison to competitors [8]. The authors further explain that BI explores several technological tools, producing reports and forecasts, in order to improve the efficiency of the decision making. BI components and tools include Data Warehouse (DW), Extract-Transform and Load (ETL), On-Line Analytical Processing (OLAP), Data Mining (DM), Text Mining, Web Mining, Data Visualization, Geographic Information Systems (GIS), and Web Portals. All these tools are combined to create BI platform to facilitate data analytics.

2.2. Application of BI and data warehousing in traffic accident evaluation

The data warehouse is an important component of BI as it consolidates data from various operational sources and make it readily available for data analysis. BI solutions have been used and applied in many areas. A BI solution was proposed in a study in Serbia on traffic accidents analysis and evaluation. In the proposed BI solution, the accidents data is extracted from operational databases to establish the possibility of using those databases as the sources of accident data required for quality analysis of road traffic safety [9]. Quality traffic accident data analysis are enabled in the proposed model by application of the OLAP (Online Analytical Processing) data warehouse concept, which represents the possibility of creating a multidimensional database with a large number of options. The data sources are traffic accident databases by Ministry of Interior (MoI), health authorities, insurance companies, road directorate, statistical office and results of researches. This Serbian model demonstrates how data is extracted from various sources and loaded into the data warehouse by integrating data from multiple sources to create a central storage of data.

Another BI solution was demonstrated in a study conducted in Victory on traffic accident analysis. This BI approach was applied within a SAS Enterprise-based Data Warehouse called Statistical Application for Population Health and Intelligence (SAPHaRI) [10]. The data collections held within SAPHaRI were available for analysis through a secure internet connection that requires password and security token access. This data warehouse includes some indicators that relate to road traffic fatal and non-fatal injuries. From this data warehouse, annual trend graphs, the annual numbers and incidence rate can all be easily accessed. The data sources for this solution were from road safety agencies and hospital's emergency departments [10].

The European Union (EU) also has a system for accident evaluation known as Community database on Accidents on the Roads in Europe (CARE). This system's main objective is to identify and quantify road safety problems, evaluate the efficiency of road safety measures, determine the relevance of community actions and facilitate the exchange of experience in this field. CARE is the European centralized data storage on road accidents which result in death or injury across the EU [11]. This data warehouse provides member states access to this central data store which is hosted by the European Commission at the Luxembourg data center. Through EU data warehouse decision makers are able to identify and quantify of road safety problems, evaluate the efficiency of road safety measures and also determine the relevance of community actions. The CARE DW pulls together non-confidential data from across the EU member states into one central database [11]. Each country is responsible for producing road safety statistics every year, which it then submits in the form of a report to the European Commission. The categories of information used to build statistics on road traffic accidents include: Person Class, Gender, Age group, Vehicle group, Area type, Motorways, Junctions, Collision type, Lighting conditions, Weather conditions and Day of the week.

Furthermore, The Accident Compensation Corporation (ACC) in New Zealand also applied the BI techniques in traffic accidents through their accident injury compensation scheme, which provides injury insurance for all citizens, residents and temporary visitors to New Zealand [13]. The primary source of crash data is the crash analysis system (CAS) database, which contains and summarizes police reported crashes. The database categorizes accident into fatal, injury and non-injury crash types [13]. Information about the nature of the injury, like what, when, where, how, by what, is also captured and stored in the database. The data is extracted from these sources in order to assist decision makers.

2.3. Summary

All the examples above acutely demonstrate that BI tools can be and have been successfully applied in the traffic, accident analysis. The literature indicates that indeed data warehouse plays a pivotal role in BI solutions required for traffic data analysis and evaluation. In Botswana, we are proposing a BI solution with integration of the accident data and vehicle registration data into a central repository for easy access and comprehensive data analysis. The data sources are traffic database system and vehicle registration system. The solution provides an integrated and total view of the enterprise data to facilitate easy decision making. This will enable data analysis through a set of powerful tools for calculations, analysis and report generation. The results of the analysis and evaluation will result in development of measures and accident prevention strategies.

3. DATA SOURCES

Traffic departments in Botswana have accumulated huge amount of data in the past years therefore drowning in data and yet starving in strategic information needed for good decision making. Most of data are kept in transactional information systems for different departments. The data sources for our platform are the files from the traffic accident database from Botswana police and vehicle registration system from DRTS. The traffic accident system has three important tables; Attendant circumstances, Causality details and the Vehicle details. All these tables have different attributes that describe circumstances of the accidents. The vehicle registration file was provided as an excel file with attributes that are captured when a new vehicle is registered. The accident data comprise data from 2012-2020 whereas data on vehicle registration is from 2015-2020.

4. DATA WAREHOUSE DEVELOPMENT METHODOLOGY

In this work, we develop a BI and Data Warehouse solution using the Kimball approach which uses dimensional modelling and bottom-up methodology [4]. Dimensional modelling is a technique for making databases simple to ensure that users can easily understand the data, as well as allow software to navigate and deliver results quickly and efficiently [14]. In building the dimensional model, the following phases are critical; identifying business process requirements, identifying the grain, identifying the dimensions, identifying the facts, verifying the model, physical design considerations and metadata management [15]. Dimensions are developed around identified strategic questions. These dimensions help in the interrogation of different aspects of the business in order to answer the desired strategic questions.

The Data Warehouse was implemented using SQL server platform which supports SQL Server Integration Services (SSIS), SQL Server Analysis Services (SSAS) and SQL Server Reporting Services (SSRS). Power BI was also deployed for interactive data visualization and reports. Microsoft visual studio was used to develop and build the required packages.

5. DATA WAREHOUSE DIMENSIONS

Based the data provided a number of dimensions were identified as follows;

5.1. Accident Dimension

This dimension was created to interrogate the accident data in order to answer the important queries pertaining accident circumstances.

5.2. Causality Dimension

This dimension was created to interrogate the details of those who die because of accidents. Such details include; gender, age, class, injury and other factors pertaining to victims.

5.3. Vehicle Dimension

This dimension is loaded with data about the details of the vehicles involved in the accident. Details such as vehicle make, year and vehicle ownership are capture in this dimension.

5.4. Road Dimension

The road dimension will provide the details of the road when the accident occurred. This dimension has attributes such as road curvature, road slope as well as slippery.

5.5. Driver Dimension

The driver dimension will help answer questions about the driver. These will include details such as driver age and driver gender.

5.6. Vehicle Registration Dimension

This dimension is created to respond to questions regarding all registered vehicles in Botswana. Details such as vehicle model, year of registration and others are captured in this dimension.

5.7. Time Dimension

The time dimension consists of all the dates in the data warehouse. It also consists of the levels of year, quarter, and month. The dimension hierarchies are the paths for drilling down or rolling up in our analysis [17].

6. DATA WAREHOUSE SCHEMA

The dimensional model can be represented with a star schema. Kimball and Ross [14] define a star schema as the generic representation of a dimensional model in a relational database in which a fact table with a composite key is joined to a number of dimensional tables, each with a single primary key.

Figure. shows the star schema for the identified dimensions. The schema also shows the attributes that are essential in answering the strategic questions.

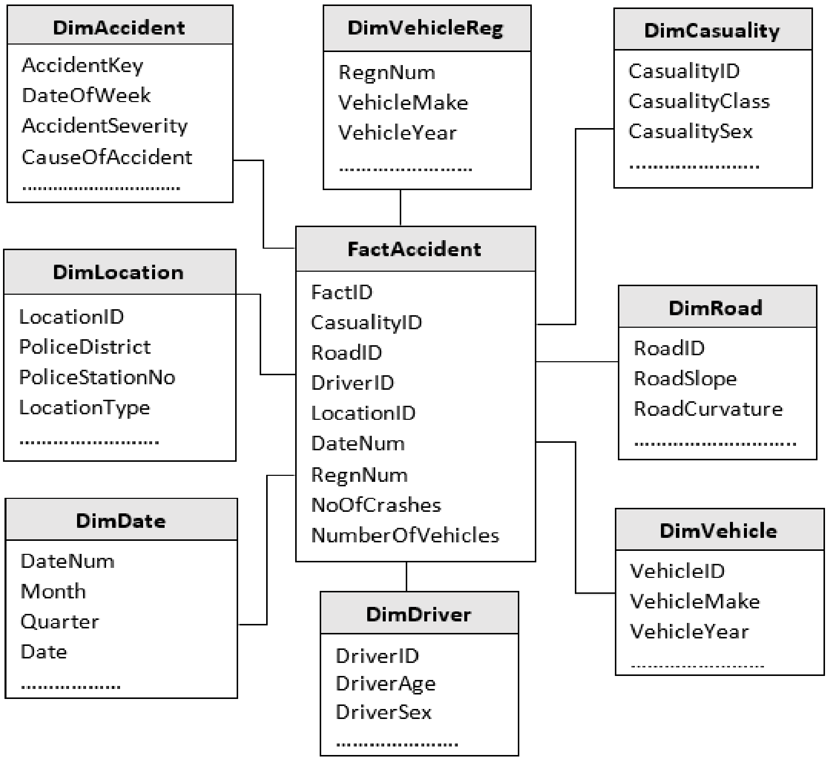


Figure. 1. DW star schema

7. BUSINESS INTELLIGENCE PLATFORM FOR ROAD TRAFFIC

Business intelligence plays a pivotal role in business analytics. Sharda et al. [1] describe BI as an umbrella term that combines architectures, tools, databases, analytics tools, applications and methodologies with the aim of extracting valuable information and knowledge from data. For our proposed solution, data is extracted from the operational systems from Botswana police database and DRTS system. The data goes through the Extraction, Transformation and Loading (ETL) process. ETL is a data integration function that involves extracting data from operational systems sources, transforming it to fit business needs, and ultimately loading it into a Data Warehouse [17]. Relevant data is extracted to address the strategic questions, followed by the transformation which involves the cleaning and correcting of errors and inconsistencies to ensure that only quality data is loaded in the Data Warehouse. Data loading refers to loading the data into the end target [18]. A comprehensive description of this ETL process is provided in Section VIII. Figure.2 represents the architectural blueprint of the proposed solution.

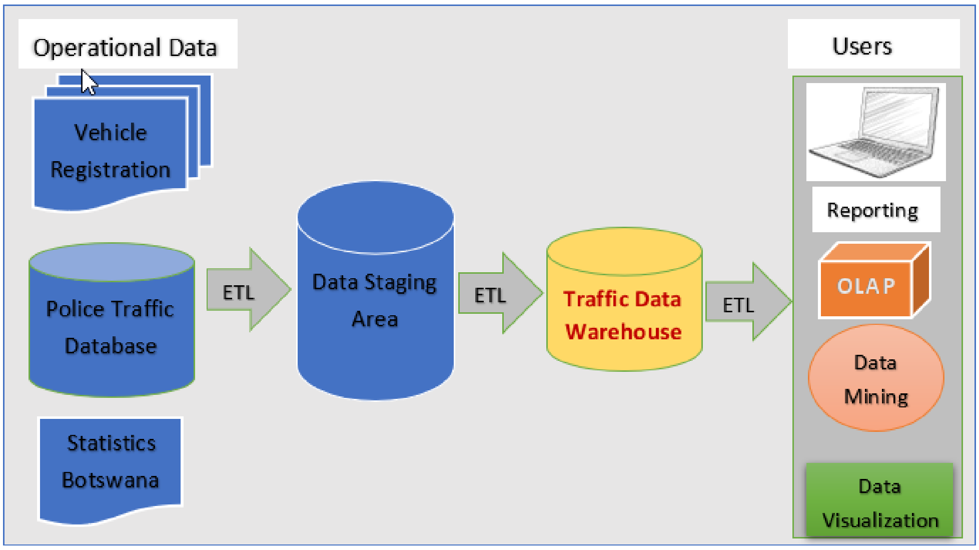


Figure. 2. Business Intelligence platform

8. THE ETL PROCESS

In this section, we demonstrate the implementation of the ETL process for the traffic data warehouse we are proposing. ETL is the process of extracting data from source systems and carrying it into the data warehouse [19]. This involves extraction, transformation and loading of relevant data into the target dimensions to provide strategic information. The ETL process was done through the SQL Server platform which supports data integration, analysis and reporting. SQL Server Integration Services (SSIS) contain a data-flow engine to transfer and transform data to and from various data sources through its operations such as Aggregate, Sort, Lookup, Merge, Merge Join, Union All, Data Conversion and Audit. It also has graphical tools and wizards for creating an extraction, transform, and loading. SQL Server Analysis Services (SSAS) enable users to accommodate multiple analytic needs within one solution and also creation of the cubes. SQL Server Reporting Services (SSRS) provide a platform that supports the authoring, management and delivery of interactive reports to the entire organization. Microsoft visual studio was used to develop and build the integration packages. Power BI was then used to generate reports to address strategic question for traffic accident managers. After the development and loading, the data was used to generate reports for data analysis. Data was extracted from different source files and stored in a central location through the use of SSIS for the purpose of answering strategic questions.

8.1. Implementation of ETL Process

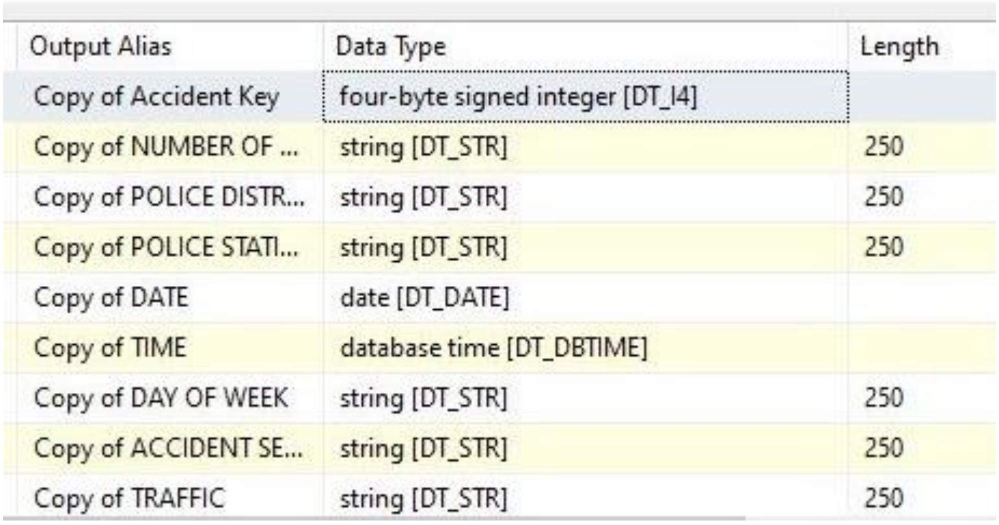
The ETL was implemented using SQL Server Integration Services and the Visual studio which develops the packages required to build the solution. The data from the traffic accident database had three tables. Dimensions were identified around these three files. The Accident details files had data about the accident and road details. This file was targeting the accident dimension and the road dimension. The causality details table captured data on the gender, age and class of the accident victims. This file was used to populate the causality dimension. The vehicle data file had data about vehicles and driver details. This file was targeted at the vehicle and driver dimension. Vehicle Registration file contained details on the registration of new vehicles. The data were extracted from this file and loaded in the vehicle registration dimension.

8.2. Data Staging Area

The staging area which can also be referred to as landing area sit between the data sources and the data warehouse. It is the immediate data storage before data can be loaded in the data warehouse. The ETL process is used to extract data sets from the various sources, bring them to a data staging area, apply a sequence of processes to prepare the data for migration into the data warehouse, and actually load them [20]. Lans [20] further asserts that the data that is loaded in the staging areas undergo a lot of processing before it's in a form suitable for storage in a data warehouse. Incorrect values have to be transformed, missing data values have to been replaced and so on.

8.3. Data Transformation

During the extraction process data types were converted through the transformation process. Data types were changed to the ones that conforms to SSIS. The varchar data types are converted to String [DT_STR] and integer values are converted to four-byte signed integer [DT_I4]. The conversion is shown in figure 3.



Output Alias	Data Type	Length
Copy of Accident Key	four-byte signed integer [DT_I4]	
Copy of NUMBER OF ...	string [DT_STR]	250
Copy of POLICE DISTR...	string [DT_STR]	250
Copy of POLICE STATI...	string [DT_STR]	250
Copy of DATE	date [DT_DATE]	
Copy of TIME	database time [DT_DBTIME]	
Copy of DAY OF WEEK	string [DT_STR]	250
Copy of ACCIDENT SE...	string [DT_STR]	250
Copy of TRAFFIC	string [DT_STR]	250

Figure. 3. Data type conversion

For our solution data was extracted from the source files into the staging area as demonstrated in figure 4.

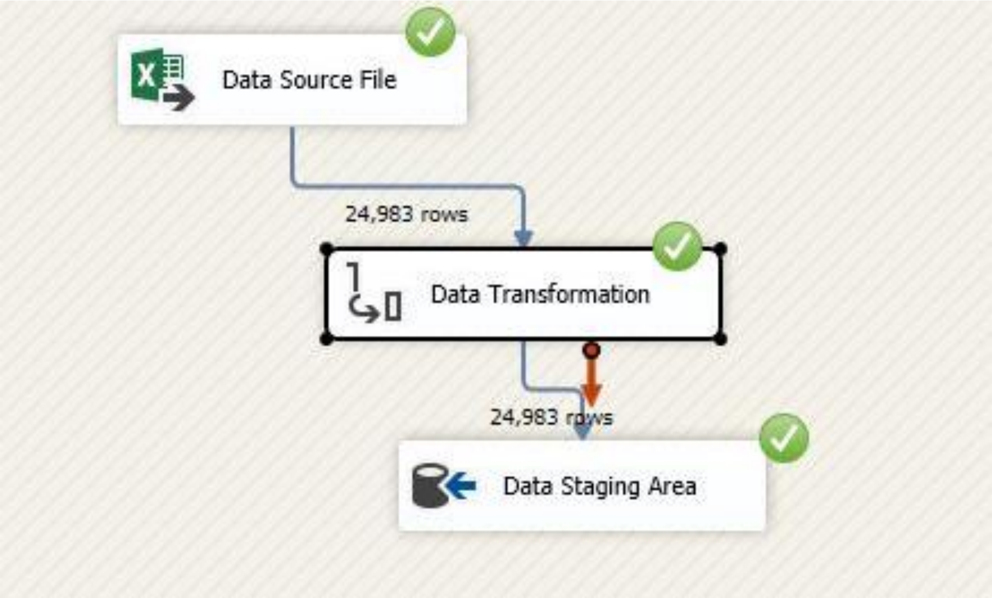


Figure. 4. ETL Process

When the data in the staging area has been transformed it can be copied to the dimensions in a warehouse. Data are copied from the staging area into the identified dimensions. Figure 5 demonstrates the loading of the Road dimension and the Location dimension.

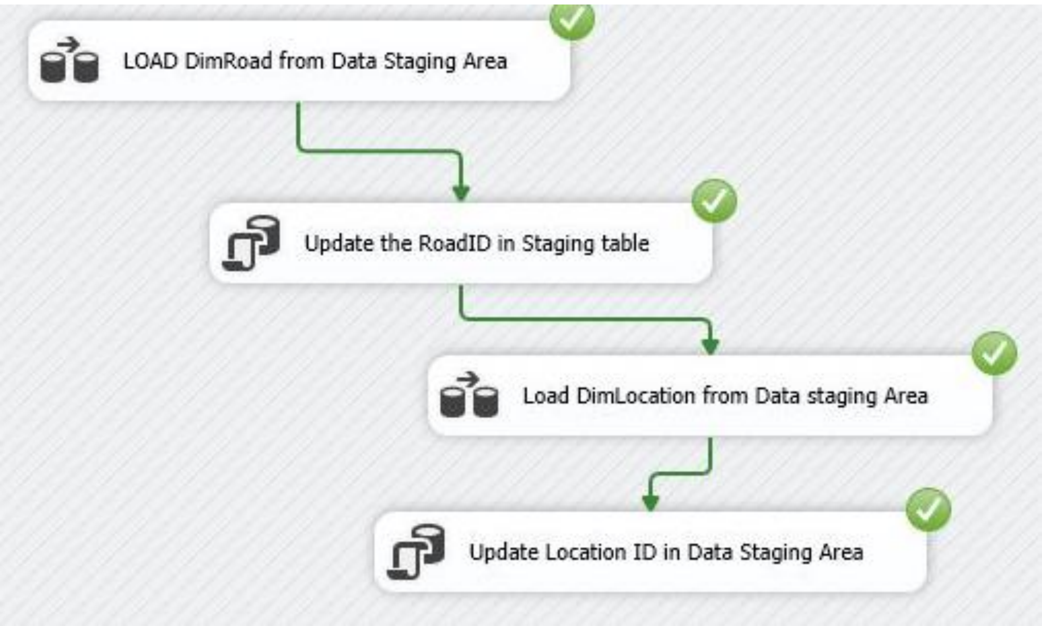
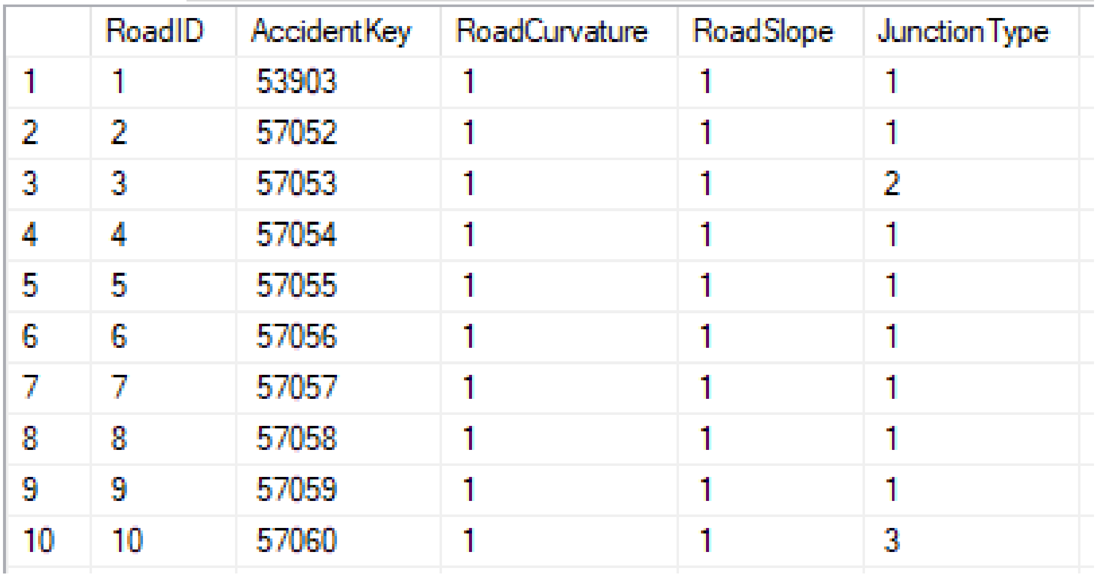

LOAD DimRoad from Data Staging Area
Update the RoadID in Staging table
Load DimLocation from Data staging Area
Update Location ID in Data Staging Area

Figure. 5. Data Loading

After the data from these files are loaded into the data warehouse, decisions makers should be able to query the data warehouse in order to answer all questions pertaining to accidents and vehicle registration. For this research, the dimensions should provide answers to the aforementioned research questions as will be demonstrated in results section. Figure 6 shows the data in the road dimension after successful loading.



	RoadID	AccidentKey	RoadCurvature	RoadSlope	JunctionType
1	1	53903	1	1	1
2	2	57052	1	1	1
3	3	57053	1	1	2
4	4	57054	1	1	1
5	5	57055	1	1	1
6	6	57056	1	1	1
7	7	57057	1	1	1
8	8	57058	1	1	1
9	9	57059	1	1	1
10	10	57060	1	1	3

Figure. 6. Road data dimension

9. Data Analysis and Results

In this Section, we deploy power BI and SSRS to build reports based on the strategic questions across the dimensions first introduced in Section I. The results demonstrate how information can become available for strategic decisions by addressing just a few questions from each dimension.

9.1. Traffic Accidents Results

9.1.1. Results from strategic questions

The results in this section are based on queries from different dimensions. A few sample questions that a manager may need answers to are addressed. One of the powerful aspects of a data warehouse is that users can change how they ask their questions from time to time. The following are results based on the stated strategic questions.

- **Strategic question 1:** How many accidents have been registered across the years as classified by severity and year?

The distribution of accidents by type reveals four types of accidents being fatal, serious injuries, minor injuries and damage only. Across accident severity, damage Only accidents had the most interesting recent trend with a 11.93% decrease in the last 3 years on record. The number for damage only dropped from 13,491 to 11,881 during its steepest decline between 2018 and 2020. Minor accidents experienced the longest period of growth in between 2012 and 2020, and serious accidents experienced the longest decline (-283) during the same period as demonstrated in table 1.

Table 1. distribution of accidents by type

Year	Damage Only	Fatal	Minor	Serious	Total
2012	13931	328	2467	795	**17521**
2013	13473	321	2486	778	**17058**
2014	13077	288	2536	732	**16633**
2015	13950	329	2607	768	**17654**
2016	14425	348	2861	739	**18373**
2017	13984	366	2726	710	**17786**
2018	13491	379	2835	636	**17341**
2019	14652	361	2909	701	**18623**
2020	11881	265	2417	512	**15075**
Total	**122864**	**2985**	**23844**	**6371**	**156064**

- **Strategic question 2**: How many accidents are recorded in each police district as classified by severity and location?

  The distribution of accidents by location indicates that many accidents occur in the Broadhurst area where the district accounted for highest number of 27778 followed by G/West and Serowe. The least number of accidents was recorded in Tsabong which recorded 659 accidents. This information can be used to inform managers to deploy more resources there, such as road blocks, patrols and speed traps. Table 2 shows accidents by police district.

Table 2. Accidents by police district

Police District	2015	2016	2017	2018	2019	2020	Total
Broadhurst	4715	5000	4685	4612	4861	3905	**27778**
Francistown	792	827	756	657	685	564	**4281**
G/West	4746	4759	4627	4618	4941	4025	**27716**
Gantsi	217	263	235	223	207	168	**1313**
Kasane	119	195	237	217	200	172	**1140**
Kutlwano	1011	1007	926	879	1015	763	**5601**
Letlhakane	236	232	237	196	274	191	**1366**
Lobatse	268	263	218	184	260	215	**1408**
Mahalapye	598	735	771	796	823	560	**4283**
Maun	663	679	659	585	709	645	**3940**
Mochudi	435	483	485	543	624	425	**2995**
Molepolole	612	600	680	544	758	611	**3805**
Sejelo	609	838	804	764	867	669	**4551**
Selibe Phikwe	460	359	301	249	283	225	**1877**
Serowe	988	1026	1005	1088	1006	735	**5848**
Tsabong	123	91	123	100	123	99	**659**
Grand Total	**16592**	**17357**	**16749**	**16255**	**17636**	**13972**	**98561**

- **Strategic Question 3:** When do these accidents happen as classified by time and day of the week?

Most accidents were recorded over the weekend. At 28,439, Sunday had the highest number of recorded accidents and was 47.59% higher than Wednesday, which had the lowest count of accidents at 19,269. Sunday accounted for 18.22% of the accidents. Across all 7 Days of the week, count of number of accidents ranged from 19,269 to 28,439. Accident prevention measures must be intensified over the weekend to curb the ever-increasing number of accidents. Figure.7 shows the distribution of these accidents by day of the week from list to highest.

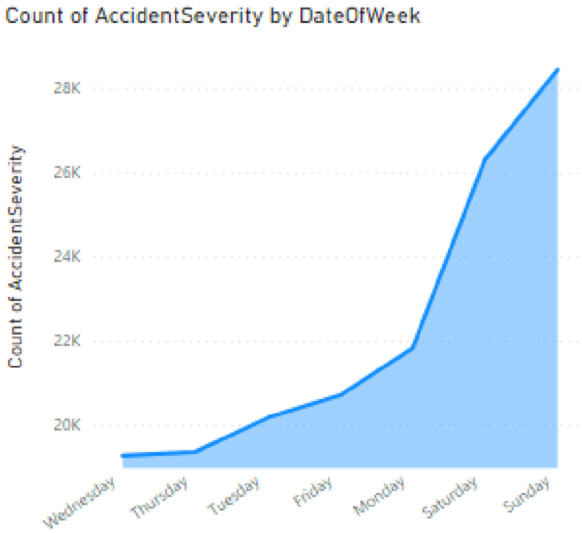


Figure.7. Accident by day of the week

- **Strategic question 4: Who are involved in these accidents as classified by age and accident severity?**

The distribution of causalities by age reveals that the 20-40 age group recorded the highest number at 33002 whilst there are less accidents for age group 81-100. Table 1 shows the distribution of accident by severity and age.

Table 3. accident by severity and age

Age	**Damage Only**	**Minor**	**Serious**	**Grand Total**
1-20	151329	6383	2273	**10000**
21-40	685058	20209	7667	**33002**
41-60	311685	6057	2485	**10258**
61-80	4421	1031	428	**1884**
81-100	50	85	38	**173**
Total	**1188543**	**33765**	**12891**	**55317**

9.1.2. Data Visualization through interactive Dashboards

A dashboard is a data visualisation tool that facilitates information delivery to the business and data warehouse users. Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at single glance and easily drilled in and further explored [2]. Power BI allows for a creation of interactive dashboards where user can see many other factors or details just by way of hovering the mouse over the required fields. Dashboards can include graphs from different dimensions of the data warehouse as well as maps to show locations. Figure 8 shows a dashboard for our traffic data warehouse solution.

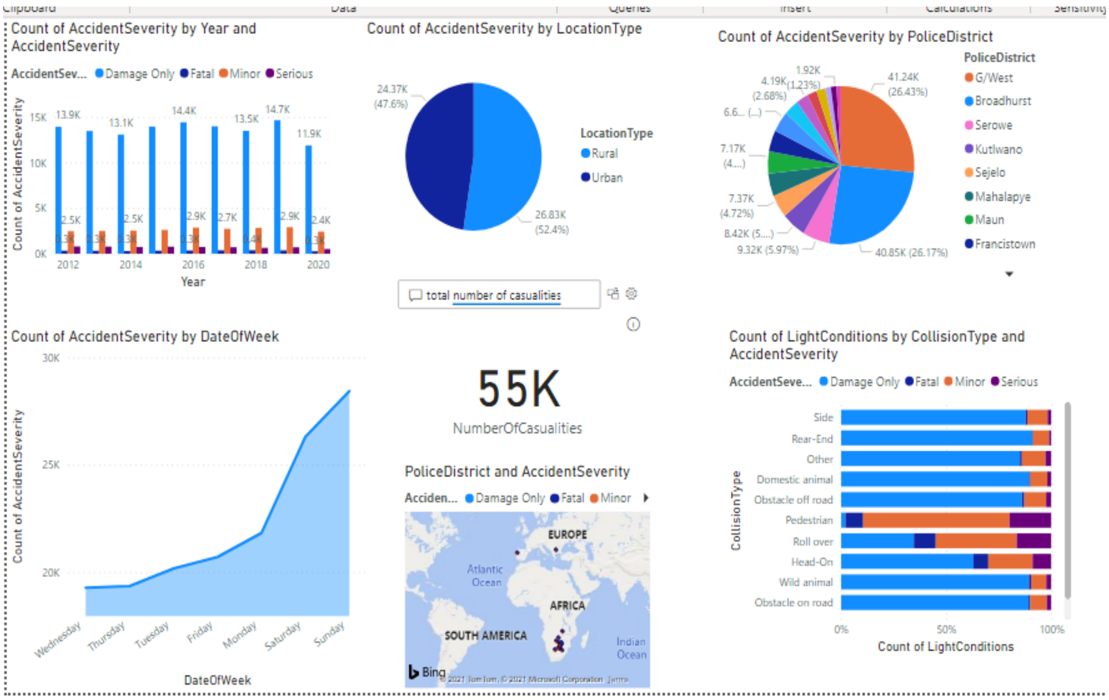


Figure. 8. Data visualization dashboard

9.2. Vehicle Registration Results

The vehicle registration dimension was used to answer sample questions pertaining to vehicle registration. Queries were based on the strategic questions below;

- **Strategic question 1:** How many vehicles were registered in the past five years?

The number of vehicle registration decreased for the last 3 years on record. Number of vehicles dropped from 65535 to 24457 during its steepest decline between 2019 and 2021. Vehicle registration experienced the longest period of growth (+9243) between 2015 and 2018. At 65,535, 2019 had the highest number of registered vehicles and was 57.00% higher than 2020, which had the lowest number at 41,741. The year 2019 accounted for 21% of the registered vehicles. Across all 6 Year, the numbers of registration ranged from 41,741 to 65,535.

- **Strategic question 2:** Where are most vehicles registered?

At 75,438, ROAD TRANSPORT & SAFETY (HQ) had the highest count of registered cars. Decision makers can deploy more resources at this center to prevent slow service delivery. Sowa registered the lowest number of 70 vehicles across a period of 2015 to 2021. ROAD TRANSPORT & SAFETY (HQ) accounted for 22.40% of registered vehicles. Across all 29 transport offices, the number of registered vehicles ranged from 70 to 75,438.

- **Strategic question 3**: How many vehicles are currently active?

An evaluation of the vehicle status indicates that of all the registered vehicles, Run (327,026) was higher than EXPORTED (9,666). Run refers to active vehicles. RUN accounted for 97.13%. The results also reveal that 142 vehicles have been scrapped. Figure 9 shows the vehicle status.

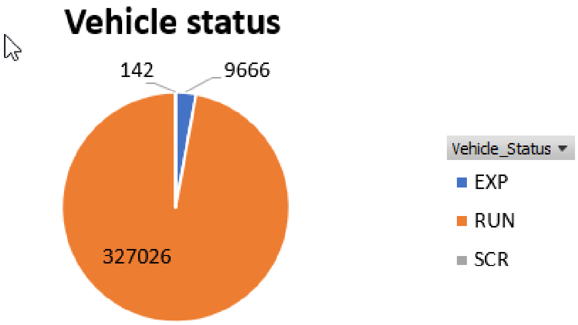


Figure. 9. Vehicle Status

- **Strategic question 4**: Which is the mostly registered vehicle model?

Of all the vehicle registered, a total count was highest for TOYOTA at 136,988, followed by HONDA, and MAZDA. FIT made up 9.14% as the least. We can safely conclude that Toyota is the mostly registered car model in Botswana.

10. Conclusion And Future Work

The development of the BI platform for traffic accidents has proven to have an immense impact on information delivery for accident data analysis and vehicle registration as demonstrated by our results. Quality decision making is achieved when decision makers have the information they need at the right time and in the right quality. Through the data visualisation tools demonstrated above, the injury prevention departments can view the statistics on different categories and can query the DW to get answers to any questions they have. Easy accessibility to accident data will help decision makers to formulate accident prevention strategies and measures for traffic safety improvement.

The vehicle registration data mart also has shown the possibility of providing answers to questions pertaining to new vehicles registration. This information will inform managerial decision makers about the number of vehicles that are active in our roads and other finer details about vehicles.

In conclusion BI is a viable computerized support tool for decision making. The framework developed in this research is a solution for accidents and vehicle registration data evaluation that other organization in the same domain must consider. Data integration is critical in ensuring easy access to strategic information required to make decisions at strategic level.

In future other analysis techniques can be employed such as OLAP, drilling and data mining to further unearth interesting revelations from the data.

ACKNOWLEDGMENTS

Our sincere gratitude goes to the Botswana police and DRTS for providing us with relevant data to model our solution.

REFERENCES

[1] R. Sharda, D. Delen and E. Turban, Business Inteligence and Data Analytics: Systems for Decision Making, Pearson, 2014.

[2] M. Hemalatha and S. Krishnaveni, "A Perspective Analysis of Traffic Accident using Data Mining Techniques," International Journal of Computer Applications, 2011.

[3] O. E. Sheta and A. N. Eldeen, "The technology of using a data warehouse to support decision making in Health care," International journal of database management systems (IJDMS), pp. 75-86, 2013.

[4] R. Kimball, Data warehouse Life cycle toolkit, Wiley, 2013.

[5] H. W. Inmon, Building the Data Warehouse, New York: John Wiley & Sons, Inc, 2002.

[6] J. Ranjan, "Business Intelligence: Concepts, Components, Techniques and Benefits," Journal of theoretical and Applied Information Technology, 2009.

[7] R. Stackowiak, J. Rayman and R. Greenwald, Oracle Data Warehousing and Business Intelligence Solutions, Wiley Publishing, 2006.

[8] A. R. Khan and. K. S. Quadri, "Business Intelligence: An Integrated Approach," Business Intelligence Journal, 2012.

[9] O. Pantelic, D. Pesic, M. Vujanic and D. B. Vujaklija, "Towards an analytical information system of traffic accidents in the function of traffic safety monitoring," Scientific Research and Essays, pp. 398-409, 2012.

[10] R. Mitchell, M. Babanch, A. Williamson and R. Grzebieta, "Transport And Road Safety (TARS) Research submission to the Victoria serious injury Inquiry of the road safety committee," University of New South Wale, victoria, 2013.

[11] IDABC, "Idabc," 2004. [Online]. Available: http://ec.europa.eu/idabc/en/document/2281/5926.html. [Accessed 5 April 2020].

[12] European Commission, 2017. [Online]. Available: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/care_flowchart_a0.jpg. [Accessed 26 January 2018].

[13] Statistics New Zealand, "Injury Statistics Project Pilot: Output Report," Statistics New Zealand, Wellington, 2004.

[14] R. Kimball and M. Ross, The Data Warehouse toolkit: The Complete Guide to Dimensional Modelling, Second ed., New York: John Wiley and Sons, Inc., 2002.

[15] C. Ballard, D. M. Farrell, A. Gupta, C. Mazuela and S. Vohnik, Dimensional Modelling: In a Business Intelligence Enviroment, First ed., IBM Corp, 2006.

[16] P. Ponniah, Data warehouse fundamentals, New york: John Wiley and Sons, INC, 2001.

[17] V. Gour, S. Sarangdevot, G. S. Tanwar and A. Sharma, "Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse," International Journal on Computer Science and Engineering, pp. 786-789, 2010.

[18] F.S. Esmail Ali, "A Survey of Real-Time Data Warehouse and ETL," International Scientific Journal of Management Information Systems, vol. 9, no. 3, 2014.

[19] P. Dhadha and N. Sharma, "Extract Transform Load Data with ETL Tools," Internationl Journal of Advanced Research in Computer Science, vol. 7, pp. 152-160, 2016.

[20] R. v. d. Lans, Data Virtualization for Business Intelligence Systems, Morgan Kaufmann, 2012.

AUTHORS

Mr Monkgogi Mudongo is a lecturer in the department of Computer Science in the University of Botswana. He obtained his MSc in Computer Information Systems from the university Of Botswana. His research interests are in the areas of Data Warehousing, Knowledge Management, Business Information Systems, Big data analytics, and Information Retrieval.

Dr Edwin Thuma has a broad background in Computing Science with specific expertise in Information Retrieval (the science of search engines) and Big Data Systems. In particular, his research has been focused primarily on the development of search engines tailored to support health professionals and laypeople when searching for health content on the web. Recently he has started working on search engines that are tailored to support legal professionals when searching for precedent cases or statutes that support the current case.

Mr Nkwebi P. Motlogelwa works as a Lecturer in the Department of Computer Science, University of Botswana. He is currently actively involved in two research areas: Information retrieval in the medical domain research, and Natural Language Processing specifically focusing on the Setswana Language. In the past he was engaged in a Microsoft funded research that explored how wireless and mobile technologies could improve public health in under-served communities

Mrs Tebo K. Leburu-Dingalo is a Lecturer in the Department of Computer Science, University of Botswana. Her areas of interest are Health Informatics, Knowledge Based Systems and High-Performance Computing. She is also interested in the use of Information and Communications Technology towards social and economic development. Tebo holds a Bachelor of Science (BS) in Computer Engineering from Florida Institute of Technology, and Master of Science in Information Systems from the University of Botswana.

Ms Pulafela A. Majoo is a Graduate Teaching Assistant in the Disability Support Services Unit in the University of Botswana. She is currently awaiting to be awarded with Msc Computer Information Systems after submission of her dissertation from University of Botswana. Her areas of interest are in information retrieval in the medical domain, data warehousing, software development frameworks and data analytics.

CASCADED SEGMENTATION NETWORK BASED ON DOUBLE BRANCH BOUNDARY ENHANCEMENT

Li Zeng[1], Hongqiu Wang[1], Xin Wang[3], Miao Tian[1*] and Shaozhi Wu[1, 2*]

[1]University of Electronic Science and Technology of China,
Chengdu, 611731, China
[2]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang 324000, China
[3]Department of Abdminal Oncology, Cancer Center,
West China Hospital, Sichuan University, China

ABSTRACT

Cervical cancer is one of the most common causes of cancer death in women. During the treatment of cervical cancer, it is necessary to make a radiation plan based on the clinical target volume (CTV) on the CT image. At present, CTV is manually sketched by physicists, which is time-consuming and laborious. With the help of deep learning model, computer can accurately draw the outline of CTV in Colleges and universities. The CDBNet proposed in this paper is a cascaded segmentation network based on double-branch boundary enhancement. First, classification network determines whether a single image contains a region of interest (ROI), and then the segmentation network uses DBNet to segment more accurately at the ROI contour. In this paper, we propose CDBNet, a cascaded segmentation network based on double-branch boundary enhancement. First, classification network determines whether a single image contains a region of interest (ROI), and then the segmentation network uses DBNet to segment more accurately at the ROI contour. The CDBNet proposed in this paper was verified on the cervical cancer dataset provided by the Department of Radiation Oncology, West China Hospital, Sichuan Province. The average dice and 95HD of the delineation results are 86.12% and 2.51mm. At the same time, the classification accuracy rate of whether the image contains ROI can reach 93.19%, and the average Dice of the image containing ROI can reach 70%.

KEYWORDS

CTV delineation, cascade, segmentation, boundary enhancement.

1. INTRODUCTION

The International Agency for Research on Cancer (IARC) estimates that by 2020, more than 600,000 women worldwide have been diagnosed with cervical cancer, and about 340,000 women have died from the disease. Cervical cancer is one of the most common causes of cancer death in women. Early detection and treatment of the diseased can greatly improve the survival rate of patients. Cervical cancer can be prevented by the HPV vaccine, but less than 25% of girls in the world can get HPV vaccine. In the diagnosis and treatment of cervical cancer, Computed Tomography (CT) is one of the most widely used imaging methods and plays an important role in assisting diagnosis [1]. To facilitate the assessment of cancer and the development of treatment plans, it is necessary to accurately locate the lesion area in the CT image [2]. Currently, medical physicists mainly rely on the manual delineation of CT images to determine the lesion area. The

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO, NLDM - 2021
pp. 249-261, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111721

lesion area is also called a region of interest (ROI) in the segmentation task. The closer to the real ROI delineation, the more accurately the lesion can be treated and the better the surrounding healthy tissue can be protected. However, this manual delineation method is highly subjective and labor-intensive, which affects the efficiency of diagnosis. Therefore, a standardized automatic segmentation method is very necessary.

In recent decades, with the development of Computer-Aided Diagnosis (CAD) [3], some automated segmentation methods based on machine learning have been used in ROI segmentation tasks, including traditional machine learning methods and deep learning methods. Hong [4] used fuzzy C-means clustering, and Bilello et al. [5] used intensity-based histograms and lesion contour refinement to segment. Due to the simple structure of traditional machine learning algorithms, it can only extract some simple features, such as texture, contour and other features, resulting in limited segmentation effects, and there is still a lot of room for improvement.

With the development of deep learning, many ROI segmentation methods based on Convolutional Neural Networks (CNN) have been proposed. Jonathan Long et al. [6] proposed a Full Convolutional Network (FCN) to supervise the training of the model at the pixel level. In the same year as FCN, Olaf Ronneberger et al. proposed UNet [7], which uses skip connection to combine feature maps in the encoding and decoding paths. Semantic information such as positions and contours that are ignored in the encoding path are compensated in the decoding path. Unet performs very well in the field of medical image segmentation, but many researchers believe that Unet and FCN can be improved. Dolz, Jose et al. combined UNet and DenseNet architecture [8] to get DenseUNet [9]. The Dense block in DenseNet can retain more semantic information in the encoding path. Zongwei Zhou et al. proposed the UNet++ [10] network, which uses more nodes to replace jump connections in Unet, so that more semantic information of each feature map can be retained in the network. These methods of retaining semantic information also retain a lot of redundant or erroneous noise information. Therefore, there are many ways to improve segmentation accuracy by supplementing semantic information. Meng et al. [11] used the local path and the global path to complement each other to obtain sufficient 3D spatial information, but the training of 3D network requires higher computer hardware. KiUnet proposed by Jose [12] et al. adds a complete convolution branch to supplement the contour information. The complete convolution part uses a feature map with a larger size, which will cause a larger computational cost. In addition, the model mixes the two branches without deep supervision of the complete convolution branch, which will cause a lot of noise in the boundary features extracted by the branch. In addition, KiUnet mixes the two branches without deep supervision of the complete convolution branch, which will cause a lot of noise in the boundary features extracted by the branch.

The segmentation accuracy of the ROI area is not only reflected in the accuracy of the ROI outline, but also in whether all the slices containing the lesion can be found. In clinical data, the cervical cancer dataset is case-based. A case contains multiple CT slices. In addition to accurately delineating the outline of the lesion area on the slices with lesions, it is also necessary to determine which slices have lesions. Many researchers believe that data balancing can solve this problem. Tran et al. [13] resample the data, that is, exclude 2/3 of the no-ROI slices to produce a more balanced dataset. But it is not sure that the ratio is the best one, and more experiments are needed to find out the ratio of ROI slices to no-ROI slices. Wardhana et al. [14] apply the class weight that adjusting the cost of the class error. And similarly [15][16], these methods also require the manual setting of weights. Some people use feature selection to reduce the number of no-ROI slices. Chlebus [17] trained a conventional random forest classifier (RF) with 256 trees using 36 hand-crafted features to filter false positive. However, hand-crafted features may not be available for other datasets.

So far, there are few studies on deep learning models for cervical cancer target segmentation. In 2019, Rhee, DJ et al. [18] integrated the CNN network into an auto-contouring tool. In 2020, Peking Union Medical College Hospital [19] was inspired by Unet and double-path network DPN on the basis of CNN and proposed a network DpnUNet designed to perform advanced semantic feature extraction and high-quality CTV delineation.

We propose a novel cascade segmentation network (CDBNet, Cascade Double-Branch Net) based on double-branch boundary enhancement to perform cervical cancer clinical target volume (CTV) segmentation. We use a cascade structure to complete the segmentation of the ROI. The cascade structure contains a classification module and a segmentation module. The classification module can predict whether each CT slice contains ROI, so that the segmentation module can focus on the ROI segmentation. In addition, we propose a novel double-branch boundary enhancement module, which focuses on segmenting the boundary part of the ROI to improve the segmentation accuracy. The main contributions of this paper can be summarized into the following two points:

1) We designed a cascade structure to determine whether the CT slice contains ROI. Among them, we use the segmentation network + post-processing method to perform the classification task, which improves the classification accuracy rate and reduces the classification false positive rate.
2) We propose a novel double-branch boundary enhancement segmentation network (DBNet). The boundary branch conducts independent training for ROI contour segmentation and effectively integrates with the ROI segmentation network of the main baseline branch. This improves the accuracy of ROI segmentation in CT slices. And we put DBNet into the cascade structure to form a stronger CDBNet.

The rest of the paper is arranged as follows: Section 2 describes the proposed method, Section 3 is related to experimental results, and Section 4 summarizes and discusses.

2. METHOD

In this section, we introduce the details of the proposed cascaded double-branch segmentation network (CDB-Net). The network architecture is shown in Figure 1. After the CT slice is input, it is divided into two types with ROI and no-ROI by LCM. LCM is the key module to reduce false positives in classification. Then use DBSM to perform accurate target segmentation on the CT slices of the ROI.

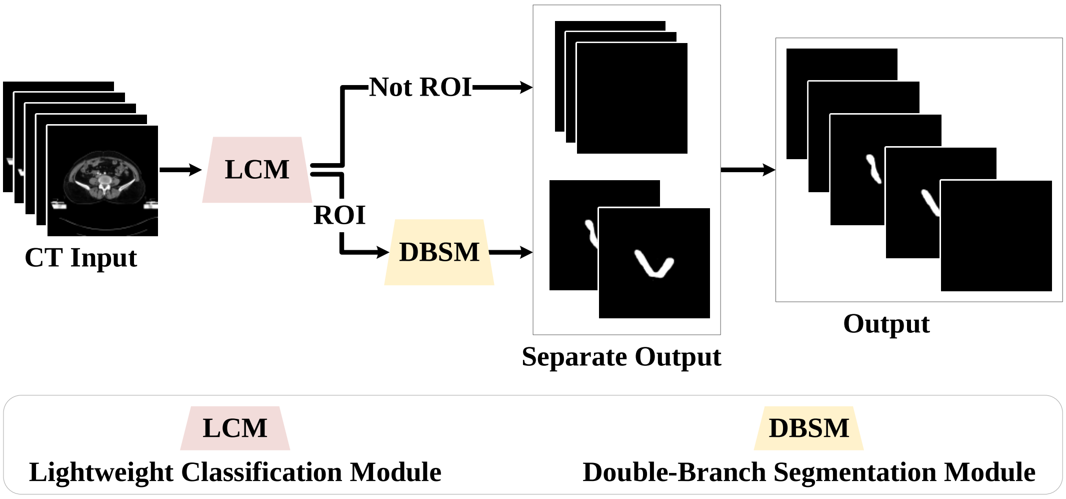

CT Input
LCM
Not ROI
ROI
DBSM
Separate Output
Output
LCM
Lightweight Classification Module
DBSM
Double-Branch Segmentation Module

Figure 1 Illustration of Cascade Double-Branch Net (CDBNet). Our proposed framework is cascaded by Lightweight Classification Module (LCM) and Double-Branch Segmentation Module (DBSM). The inputs are Two-dimensional CT slices. Then LCM predicts each CT slice contains ROI or not. The slices containing ROI will be sent to the DBSM to obtain more accurate ROI segmentation.

2.1. Lightweight Classification Module

The LCM module is used to determine whether the CT slice contains an ROI so that the DBSM only needs to pay attention to the slice with ROI. Due to the blurred and changeable ROI boundary, the classification method at the CT slice level is not sufficient for accurate classification. Therefore, it is necessary to perform classification at the pixel level and then judge at the slice level, which can improve the classification accuracy. In the LCM module, because it does not require precise segmentation results, the Unet [7] structure with fewer parameters is used to segment CT images. The Unet model consists of a down-sampling encoder and an up-sampling decoder, as shown in Figure 2. The Unet down-sampling path extracts and filters features, while the up-sampling path restores size and supplementary features. The InConv operation in Figure 2 converts the input data to a size of 256×256, and then expands the channel of the feature map through convolution. The DownConv operation convolves and downsamples the feature map twice. The Skip + UpConv operation includes skipping connection and two convolution operations. For example, in the Skip + UpConv process of Dn5, the Dn5 feature map is first up-sampled to obtain the feature map Up4, then it is merged with the feature map Dn4, and finally, convolution is performed. The OutConv operation makes the model output a channel feature map through convolution.

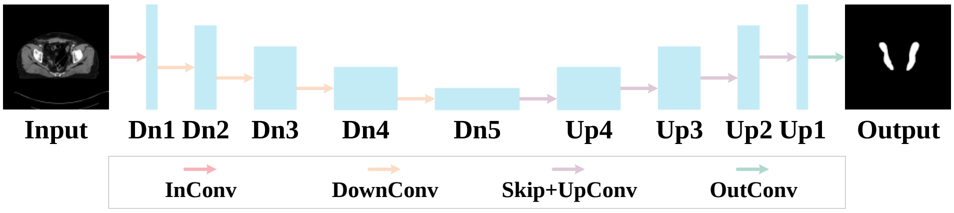


Figure 2. U-net architecture **Error! Reference source not found.** uses the Encoder-Decoder structure. And it uses skip connections in the decoding path to retain more information.

Finally, it is judged whether each CT slice contains ROI according to the result of segmentation. Because the area of ROI in a small number of slices is very small, a threshold n is set in this paper to assist classification. The value of n is the optimal threshold determined based on the performance of different thresholds on the validation set. When the sum of ROI in each pixel in a single segmentation prediction image is greater than n, it is considered that the CT slice contains ROI, otherwise, it does not.

2.2. Double-Branch segmentation Module

Although the structure of the encoder and decoder has been well done in the field of medical image segmentation, some information will be ignored in the encoding path. The Unet model uses Skip connection to solve this problem. Skip connection supplements the large-size feature map as complete information to the up-sampling decoder path, but we may need to pay more attention to the boundary segmentation in the large-size feature map. Figure 4 vividly shows the correctness of this idea. When the Unet feature map has a smaller size (such as Up3), the target area is not clear; in the Up2 and Up1 stages with a larger feature map, the target area is gradually clear. At this time, the fusion of clearer boundary information will be more beneficial to the training of the network.

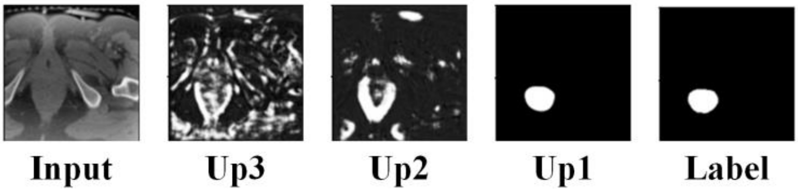


Figure 3. The feature maps of each layer and label. Up3, Up2, Up1 correspond to the feature map in Figure 2

KiUnet mentioned that using a complete convolution branch to provide boundary information can improve segmentation accuracy. But the boundary mentioned in that paper is a feature map with more noise. Therefore, we are inspired by the double-branch structure of KiUnet. Our DBSM includes the baseline branch and border branch. The Baseline branch uses the Unet structure to provide the main ROI segmentation. The Border branch retains the larger layer of the feature map in Unet, and uses the ground truth border image as supervision for training. And merge the baseline branch model with the border branch at each layer to achieve the purpose of effectively enhancing the border. The specific structure diagram is shown in Figure 4. The border branch takes the Input feature map as input, and then performs two down-sampling DownConv operations and two up-sampling Skip+UpConv operations. During down-sampling, the contour information is supplemented in the baseline by ConcatConv-Down operation, and the up-sampling is supplemented by ConcatConv-Up. Finally, after the ConcatConv operation, the single-channel feature map is output as the ROI prediction map, and the border branch outputs a contour prediction map.

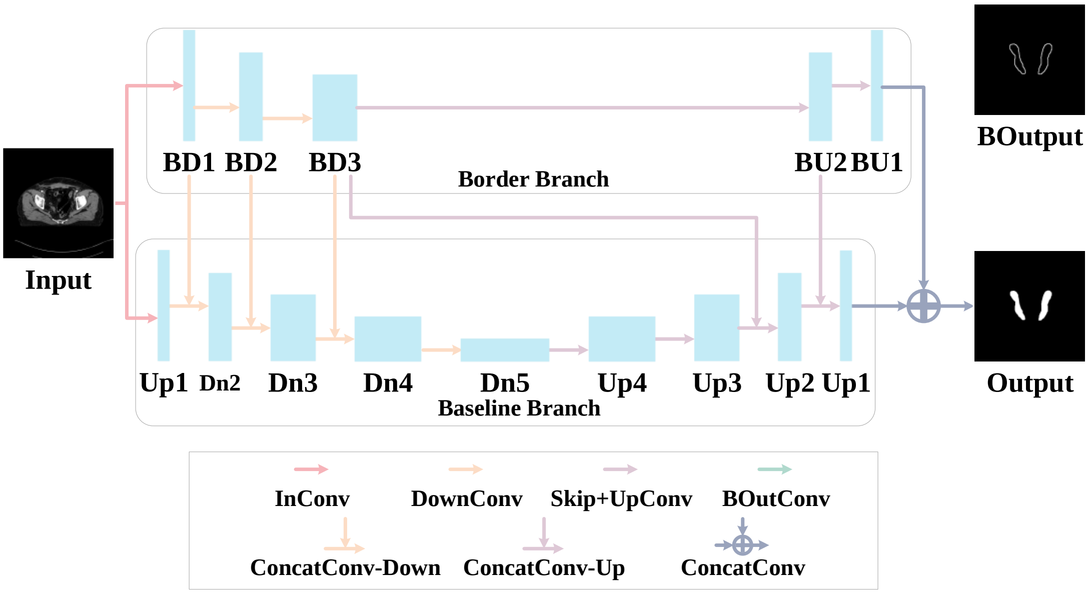


Figure 4. Double-Branch Segmentation Module (DBSM). The DBSM consists of the Baseline branch and the Border branch. The Baseline Branch performs segmentation prediction on the CT slice, and the Border Branch predicts the contour of the ROI. The two branches are merged through ConcatConv-Down, ConcatConv-Up, and ConcatConv. Finally, a more accurate segmentation prediction is output.

2.3. Loss Function

The loss function of DBNet training consists of two parts, one is the dice of the predicted contour and the real contour, defined as l_border, and the other is the dice of the predicted ROI and the real ROI, defined as l_end. The loss function used in the training process is as in equation (1). Among them, m is the length of the input feature map, n is the width of the input feature map, c is the number of channels, A is the predicted contour map, B is the real contour map marked by the physicist, and l_border is calculated as equation (2). The calculation of l_end is similar to equation (2).

$$L = l_{border} + l_{end} \qquad (1)$$

$$l_{border} = 1 - \frac{\sum_{k=1}^{c}\sum_{i=1}^{m}\sum_{j=1}^{n} A_{k,i,j} * B_{k,i,j}}{\left|\sum_{k=1}^{c}\sum_{i=1}^{m}\sum_{j=1}^{n} A_{k,i,j}\right| + \left|\sum_{k=1}^{c}\sum_{i=1}^{m}\sum_{j=1}^{n} B_{k,i,j}\right|} \qquad (2)$$

3. EXPERIMENTS

The dataset used in this paper comes from the Department of Radiation Oncology, West China Hospital, Sichuan Province. The experimental task is to segment the cervical cancer CTV in this dataset. Section 3.1 describes the dataset and related preprocessing operations. Section 3.2 explains the evaluation indicators used in this article. 3.3 shows the experimental results of cervical cancer clinical target volume segmentation. Section 3.5 shows the results of the ablation experiment.

3.1. Dataset and preprocessing

The dataset contains 276 patients, and the data comes from the contouring results of multiple physicists, and each case is contoured by only one physicist. In order to make the model universal, the data will only be used when a physicist outlines more than 30 cases, and the final dataset contains 196 patients. After random scrambling, the data of 19 patients were used as the test set, the data of 177 patients were used as the training set and the validation set, and the five-fold cross-validation was used for training. The data is read in from the Dicom medical format. The original size of the CT slice image is 512×512. Cut the HU range of the CT slice to [-128,256], set the HU value less than -128 to -128, and set it to 256 if the HU value is more than 256. And then normalized to [0, 1]. This paper uses GeForce RTX 2080Ti for network training, the epoch is set to 25, and batch size is set to the maximum value available for each model.

3.2. Evaluation

The values involved in the experiment are all averages in the same test set.

3.2.1. Dice

The Dice coefficient is used to measure the similarity between the predicted value and the ground truth value. The measurement value varies between 0-1, and 1 means that the two samples completely overlap. The specific formula is as shown in equation (3):$|A \cap B|$ represents the number of elements in common between the two samples, $|A|$ represents the number of elements in the A sample, and B is the same. The dice global used in the experiment is the average of all test slices, and the dice per case is the average of all test cases.

$$\mathrm{Dice} = \frac{2|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}|} \qquad (3)$$

3.2.2. Hausdorff

Use Hausdorff95 to measure the distance between two point sets, defined as the formula (4), H(A, B) is called the two-way Hausdorff distance, h(A, B) is the one-way distance from point set A to point set B, H(B, A) is the one-way distance from point set B to point set A.

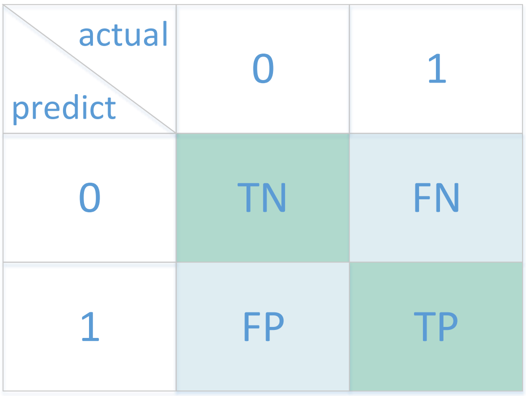
$$\mathrm{H}(\mathrm{A},\mathrm{B}) = \max[\mathrm{h}(\mathrm{A},\mathrm{B}),\mathrm{h}(\mathrm{B},\mathrm{A})] \qquad (4)$$

$$\mathrm{h}(\mathrm{A},\mathrm{B}) = \max_{\mathrm{a}\in\mathrm{A}}\min_{\mathrm{b}\in\mathrm{B}}\|\mathrm{a}-\mathrm{b}\| \qquad (5)$$

$$\mathrm{h}(\mathrm{B},\mathrm{A}) = \max_{\mathrm{b}\in\mathrm{B}}\min_{\mathrm{a}\in\mathrm{A}}\|\mathrm{b}-\mathrm{a}\| \qquad (6)$$

3.2.3. Accuracy, Recall, Precision

The confusion matrix in Figure 5 shows the classification of whether a single CT slice contains ROI. The three indicators of accuracy, recall, and precision can be a good assessment of the correctness of the network's classification of whether a single slice contains ROI. Since this paper judges whether each slice contains ROI according to the segmentation result, a threshold n is set. When the sum of pixels of a single predicted ROI image is greater than n, it is regarded as including ROI, otherwise, it does not.



actual predict	0	1
0	TN	FN
1	FP	TP

Figure 5. Confusion matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (9)$$

3.3. Experimental results and analysis

The experiment will have two parts to prove the effectiveness of the proposed Cascaded Double-Branch segmentation network (CDBNet). The first is the effectiveness of DBSM (Section 2.2), and the second is the effectiveness of the cascade structure.

3.3.1. Predicted result map

Figure 6 shows the segmentation results of the Unet network and the CDBNet network in the same test set. It can be seen that in the slice segmentation that includes ROI, CDBNet can perform better although the characteristics of ROI are different.

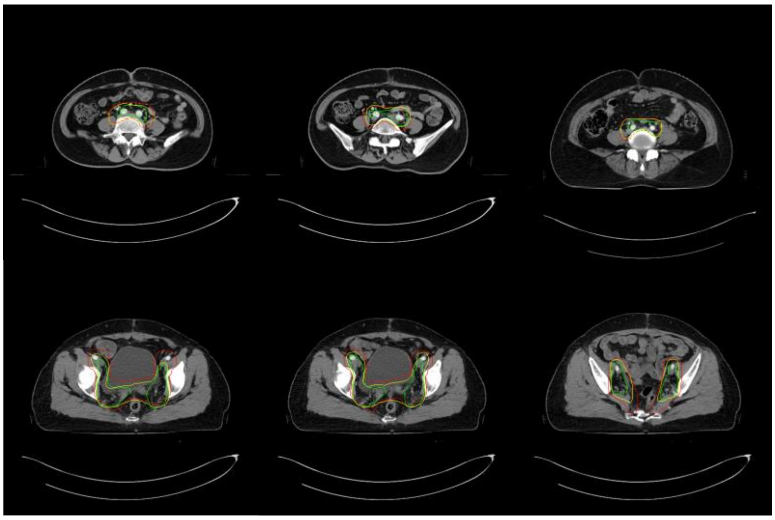


Figure 6. CTV contour result map (red is the standard contour from the physicist, yellow is the prediction result of CDBNet, and green is the prediction result of UNet)

3.3.1. DBSM

DBSM is shown in Figure 4, using five-fold cross-validation in the experiment. Table 1 shows the results of the five-fold model in the same test set, where Average means the average of the five-fold results of the same network. The DBSM used in this experiment is convolved twice on the boundary branch and contains feature maps of 3 kinds of sizes. This depth is obtained by the deep ablation experiment.

Table 1 Comparison of the segmentation effect of DBSM and Unet on the cervical cancer test dataset

Model	**validation set**	**dice global**	**dice per case**	**HD**	**acc**	**recall**	**dice(slices with ROI)**
Unet	1	83.26	83.30	2.67	92.97	92.94	61.15
	2	83.72	83.78	2.54	93.91	93.77	61.71
	3	84.60	84.63	2.58	93.13	91.61	66.35
	4	84.97	84.91	2.56	92.35	88.36	71.20
	5	84.64	84.70	2.54	93.60	92.07	66.05
	Average	**84.24**	**84.26**	**2.58**	**93.19**	**91.75**	65.29
DBSM	1	81.70	81.40	2.80	88.96	80.07	73.48
	2	82.76	82.70	2.71	90.01	83.31	71.41
	3	84.90	84.94	2.56	92.97	92.81	65.76
	4	82.77	82.69	2.73	90.58	83.46	71.58
	5	84.11	84.1	2.62	91.67	85.85	72.11
	Average	83.25	83.17	2.68	90.84	85.10	**70.87**

From the average point of view, the effect of Unet network in Dice global, Dice per case, HD, acc, and recall is better than DBSM. But on the ROI slice, the Dice value of DBSM is about 5.58% higher than that of Unet. This shows that on slices with ROI, DBSM segmentation is more accurate. However, due to the poor ability of DBSM to determine whether there is an ROI in a slice, the overall Dice value is relatively low.

3.3.2. Cascade structure

Experiment 3.3.1 shows the excellent segmentation ability of DBSM in ROI slices, but it cannot distinguish whether the slices have ROI. So this experiment will prove that the cascade structure can integrate the segmentation ability of DBSM and the classification ability of Unet.

Table 2. Comparison of segmentation effect of cascade and non-cascade structure on cervical cancer test dataset

Model	**validation set**	**dice global**	**dice per case**	**HD**	**acc**	**recall**	**dice(slices with ROI)**
Unet	1	83.26	83.30	2.67	92.97	92.94	61.15
	2	83.72	83.78	2.54	93.91	93.77	61.71
	3	84.60	84.63	2.58	93.13	91.61	66.35
	4	84.97	84.91	2.56	92.35	88.36	71.20
	5	84.64	84.70	2.54	93.60	92.07	66.05
	Average	84.24	84.26	2.58	93.19	91.75	65.29
CDBNet	1	85.86	85.82	2.51	92.97	92.94	68.22
	2	87.17	87.21	2.43	93.91	93.77	71.09
	3	86.05	86.08	2.55	93.13	91.61	70.17
	4	85.24	85.22	2.59	92.35	88.36	70.65
	5	86.38	86.40	2.49	93.60	92.07	70.77
	Average	**86.14**	**86.15**	**2.51**	93.19	91.75	**70.18**

Unet is a non-cascaded structure, and CDBNet is a cascaded structure. It can be seen that CDBNet with a cascade structure retains both the classification ability of Unet and the segmentation ability of DBSM, so the overall Dice coefficient value can reach 86.14%, which is about 1.9% higher than Unet.

3.4. Ablation experiment

This part conducts ablation experiments on the depth of the boundary segmentation branch module, and finds that the network performance is the best when downsampling twice. In addition, an ablation experiment was performed on the threshold n mentioned in 3.2.3.

3.4.1. Experiment on the depth of the boundary branch

This experiment randomly selected the training and verification data of the second fold in the five-fold cross-validation for network training, and showed the results of the same test set. The network is CDBNet, where CDBNet-2 means that the boundary branch is downsampled once, and there are a total of 2 types of feature maps.

Table 3. Depth comparison of boundary branches

Model	**dice global**	**dice per case**	**HD**	**Acc**	**Recall**	**dice(slices with ROI)**
CDBNet-2	86.90	86.93	2.45	93.91	93.77	70.34
CDBNet-3	**87.17**	**87.21**	**2.43**	93.91	93.77	**71.09**
CDBNet-4	86.93	86.98	2.45	93.91	93.77	70.42

It can be seen from Table 3 that the structure performs best when the boundary branch is sampled twice, and Dice global can reach 87.17%.

3.4.2. Experiment on the threshold n

This experiment explores the threshold n mentioned in 3.2.3. Use Unet to perform five-fold cross-validation training, set the value of n from 0 to 115, and increase by 5 each time. Figure 7 is the classification accuracy rate corresponding to different n, and the classification accuracy rate on the vertical axis is the average of the accuracy rates of the five-fold cross-validation in the respective validation sets. It can be seen that when the threshold is set to 65, the overall classification effect is the best.

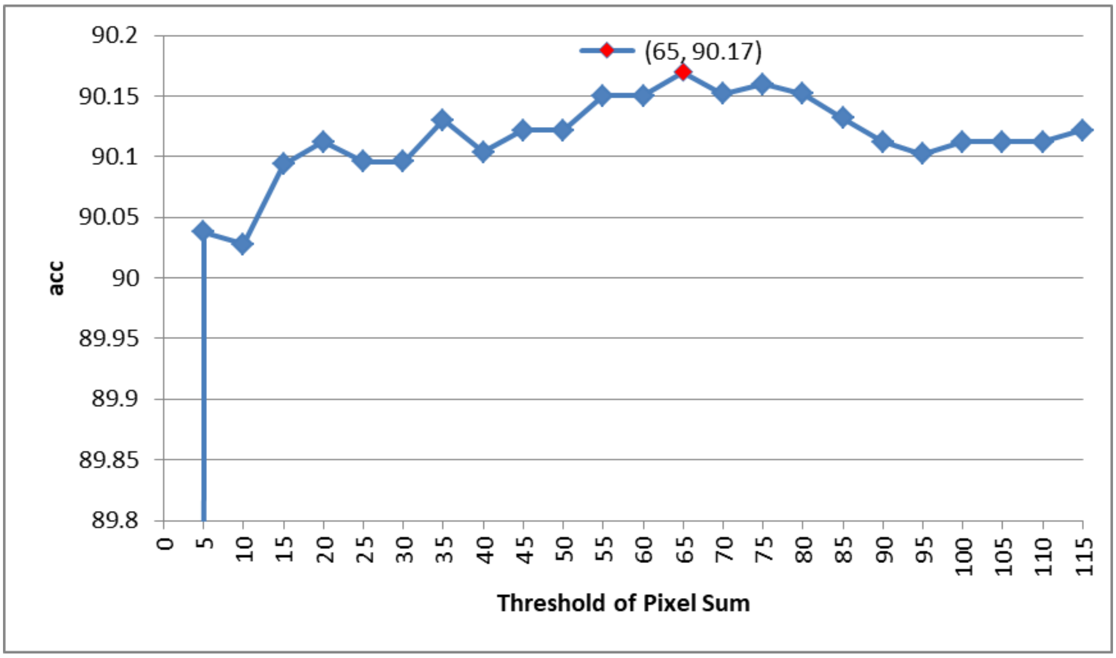


Figure 7. The relationship between the classification accuracy rate and the threshold n

4. CONCLUSION

The outline of ROI in medical images plays a very important role in the formulation of radiotherapy treatment plans. Manual delineation is time-consuming and laborious. It usually takes an experienced physicist two or three hours to delineate all the slices of one case. This paper proposes a model CDBNet that realizes automatic ROI contour delineation on CT images, which can describe the cervical cancer risk parts in CT images. This paper mainly focuses on CTV (Clinical Target Volume). Input the CT image, the network can predict the contour and position of the CTV. The main results of this paper are as follows:

First, we use a cascade structure to solve the problem that a single model cannot take into account classification tasks and segmentation tasks at the same time. This paper designs a cascade structure formed by the classification module and the segmentation module. The classification module can predict whether each slice contains ROI. Then the classification module puts the predicted ROI slice into the segmentation module for more accurate segmentation prediction. In this way, the overall segmentation performance of a case is better, and the false-positive classification of the ROI predicted by the slice without ROI is reduced.

Secondly, we designed the CDBNet network based on the cascade structure. The DBSM module in CDBNet adds a boundary branch on the basis of Unet to supplement the boundary information so that the network has a better segmentation performance at the boundary of the ROI.

In the whole process, the research mainly explored how to improve the accuracy of CT image delineation of cases, and designed CDBNet, which has an excellent performance in both the overall classification of cases and the segmentation of ROI slices. So that the contour information during segmentation is fully preserved, the segmentation Dice value of the case can reach about 86.14%, and the HD (95%) can reach 2.51mm.

ACKNOWLEDGEMENTS

This work has been supported by the scientific research projects of Quzhou Science and Technology Bureau, Zhejiang Province No.2020D012, No.2021D15,, and by the major projects of the National Bureau of Statistics No.2020LD07.

REFERENCES

[1] Crane C H , Koay E J . Solutions that enable ablative radiotherapy for large liver tumors: Fractionated dose painting, simultaneous integrated protection, motion management, and computed tomography image guidance[J]. Cancer, 2016, 122(13):1974-1986.

[2] Wei D, Ahmad S, J Huo, et al. Synthesis and Inpainting-Based MR-CT Registration for Image-Guided Thermal Ablation of Liver Tumors[J]. Springer, Cham, 2019.

[3] Fujita H, Uchiyama Y, Nakagawa T, et al. Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs[J]. Computer Methods & Programs in Biomedicine, 2008, 92(3):238-248.

[4] HONG, Jae-Sung, KANEKO, et al. Automatic Liver Tumor Detection from CT[J]. IEICE transactions on information and systems, 2001, 84(6):741-748.

[5] Bilello M, Gokturk S B, Desser T, et al. Automatic detection and classification of hypodense hepatic lesions on contrast - enhanced venous - phase CT[J]. Medical Physics, 2004, 31(9).

[6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.

[7] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[8] Iandola F , Moskewicz M , Karayev S , et al. DenseNet: Implementing Efficient ConvNet Descriptor Pyramids[J]. Eprint Arxiv, 2014.

[9] Huang G , Liu Z , Laurens V , et al. Densely Connected Convolutional Networks[J]. IEEE Computer Society, 2016.

[10] [Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation[J]. 2018.

[11] Meng L , Y Tian, Bu S . Liver tumor segmentation based on 3D convolutional neural network with double scale[J]. Journal of Applied Clinical Medical Physics, 2020, 21(1).

[12] Jose J M , Sindagi V , Hacihaliloglu I , et al. KiU-Net: Towards Accurate Segmentation of Biomedical Images using Over-complete Representations[C]// 2020.

[13] Tran S T , Cheng C H , Liu D G . A Multiple Layer U-Net, Un-Net, for Liver and Liver Tumor Segmentation in CT[J]. IEEE Access, 2020, PP(99):1-1.

[14] Wardhana G , Naghibi H , Sirmacek B , et al. Toward reliable automatic liver and tumor segmentation using convolutional neural network based on 2.5D models[J]. International Journal of Computer Assisted Radiology and Surgery, 2020, 16(12).

[15] Huang Q , Sun J , Hui D , et al. Robust liver vessel extraction using 3D U-Net with variant dice loss function[J]. Computers in Biology and Medicine, 2018, 101:S0010482518302385-.

[16] Tang Y , Tang Y , Zhu Y , et al. E^{2}Net: An Edge Enhanced Network for Accurate Liver and Tumor Segmentation on CT Scans[J]. 2020.

[17] Chlebus G , Schenk A , Moltz J H , et al. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing[J]. entific Reports, 2018, 8(1).

[18] Automatic detection of contouring errors using convolutional neural networks[J]. Medical Physics, 2019, 46(11).

[19] Zhikai L , Guan H . Development And Validation Of A Deep Learning Algorithm For Auto-Delineation Of Clinical Target Volume And Organs At Risk In Cervical Cancer Radiotherapy[J]. International Journal of Radiation OncologyBiologyPhysics, 2020, 108(3):e766.

AUTHORS

Shaozhi Wu received the Ph.D. degree in Computer software and theory from the University of Electronic Science and Technology of China (UESTC), Chengdu, China,. He is currently an associate researcher at UESTC. His academic interests include artificial intelligence, deep learning applications, and pattern recognition, big data, computer network, Intelligent Sense.

Zeng Li is a master's student at the University of Electronic Science and Technology of China, where he studies medical image processing and region-of-interest segmentation algorithms.

Miao Tian received the B.E. degree in electrical engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2003 and the M.S. degree in electrical engineering from the University of Tulsa, OK, USA, in 2005 and the Ph.D. degree in electrical engineering from the University of Colorado, Boulder, USA, in 2012. From 2013 to 2014, he worked as research associate in the Earth System Science Interdisciplinary Center (ESSIC) at University of Maryland at College Park, USA. From Dec. 2014 to 2016, he worked as senior engineer at the Earth Resources Technology (ERT), Inc., Maryland. He is currently an associate researcher at UESTC. His academic interests include deep learning applications, image processing and pattern recognition, passive remote sensing, radiative transfer, instrument calibration, and antenna design.

AUTHOR INDEX