

**Computer Science & Information Technology 210**

**Machine Learning and Soft Computing**



David C. Wyld  
Dhinaharan Nagamalai (Eds)

## **Computer Science & Information Technology**

- 4<sup>th</sup> International Conference on Machine Learning and Soft Computing (MLSC 2023)
- 9<sup>th</sup> International Conference on Information Technology Convergence and Services (ITCSS 2023)
- 9<sup>th</sup> International Conference on Advances in Computer Science and Information Technology (ACSTY 2023)
- 9<sup>th</sup> International Conference on Software Engineering (SOFE 2023)
- 9<sup>th</sup> International Conference on Natural Language Processing (NATP 2023)
- 4<sup>th</sup> International Conference on Big Data and Blockchain (BDAB 2023)
- 4<sup>th</sup> International Conference on Artificial Intelligence and Big Data (AIBD 2023)

**Published By**



**AIRCC Publishing Corporation**

## **Volume Editors**

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

Dhinaharan Nagamalai (Eds)  
Wireilla, Australia  
E-mail: dhinaharann@gmail.com

ISSN: 2231 - 5403  
ISBN: 978-1-925953-86-2  
DOI: 10.5121/csit.2023.130201 - 10.5121/csit.2023.130214

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

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

## Preface

4<sup>th</sup> International Conference on Machine Learning and Soft Computing (MLSC 2023), January 28 ~ 29, 2023, Copenhagen, Denmark, 9<sup>th</sup> International Conference on Information Technology Convergence and Services (ITCSS 2023), 9<sup>th</sup> International Conference on Advances in Computer Science and Information Technology (ACSTY 2023), 9<sup>th</sup> International Conference on Software Engineering (SOFE 2023), 9<sup>th</sup> International Conference on Natural Language Processing (NATP 2023), 4<sup>th</sup> International Conference on Artificial Intelligence and Big Data (AIBD 2023) was collocated with 4<sup>th</sup> International Conference on Big Data and Blockchain (BDAB 2023). 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 MLSC 2023, ITCSS 2023, ACSTY 2023, SOFE 2023, NATP 2023, AIBD 2023, BDAB 2023. 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, MLSC 2023, ITCSS 2023, ACSTY 2023, SOFE 2023, NATP 2023, AIBD 2023 BDAB 2023 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 MLSC 2023, ITCSS 2023, ACSTY 2023, SOFE 2023, NATP 2023, AIBD 2023, BDAB 2023.

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

David C. Wyld,  
Dhinaharan Nagamalai (Eds)

## Organization

Southeastern Louisiana University, USA  
Wireilla Net Solutions, Australia

## Program Committee Members

Abdel-Badeeh M. Salem,  
Abdelhadi Assir,  
Abdellatif I. Moustafa,  
Abderrahim Siam,  
Abdessamad Belangour,  
Abhishek Roy,  
Abhishek Shukla,  
Adnan Salman,  
Afaq Ahmad,  
Ahmad A. Saifan,  
Akhil Gupta,  
Alexander Gelbukh,  
Ali Abdrhman Mohammed Ukasha,  
Alia Karim abdulhassan,  
Alireza Valipour Baboli,  
Alla Anohina-Naumeca,  
Álvaro Rocha,  
Amal Azeroual,  
Amina El murabet,  
Amizah Malip,  
Ana Leal,  
Anita Yadav,  
Anouar Abtoy,  
António Abreu,  
Archit Yajnik,  
Ashkan Ebadi,  
Ashwin Viswanathan Kannan,  
Asif Irshad Khan,  
Asimi Ahmed,  
Assem Moussa,  
Atul Garg,  
B Nandini,  
Bhagyashree SR,  
Bhavtosh Rath,  
Bogdan Wiszniewski,  
Bouchra Marzak,  
Brahim Lejdel,  
Carlos Becker Westphall,  
Carlos Guardado da Silva,  
Chang Gao,  
Cheng Siong Chin,  
Christian Mancas,  
Dan Wan,

Ain Shams University, Egypt  
Hassan 1st University, Morocco  
Umm AL-Qura University, Saudi Arabia  
University of Khenchela, Algeria  
University Hassan II Casablanca, Morocco  
Adamas University, India  
R D Engineering College, India  
An-najah National University, Palestine  
Sultan Qaboos University, Oman  
Yarmouk university, Jordan  
Lovely Professional University, India  
Instituto Politécnico Nacional, Mexico  
Sebha University, Libya  
University of technology, Iraq  
University Technical and Vocational, Iran  
Riga Technical University, Latvia  
University of Coimbra, Portugal,  
Mohammed V University, Morocco  
Abdelmalek Essaadi University, Morocco  
University of Malaya, Malaysia  
University of Macau, China  
Harcourt Butler Technical University, India  
Abdelmalek Essaâdi University, Morocco  
Polytechnic Institute of Lisbon, Portugal  
Sikkim Manipal university, India  
Concordia University, Canada  
Oklahoma State University, USA  
King Abdulaziz University, Saudi Arabia  
University Ibn Zohr, Morocco  
GGA, Egypt  
Chitkara University, India  
Telangana University, Nizamabad  
ATME College of Engineering, India  
Target Corporation, USA  
Gdansk University of Technology, Poland  
Hassan II University, Morocco  
University of El-Oued, Algeria  
Federal University of Santa Catarina, Brazil  
Universidade de Lisboa, Portugal  
Waymo, United States  
Newcastle University, Singapore  
Ovidius University, Romania  
Hunan Normal University, China

Dário Ferreira,  
Dibya Mukhopadhyay,  
Didar Urynbassarova,  
Domenico Rotondi,  
Dumitru Dan Burdescu,  
Elmurabet Amina,  
Elzbieta Macioszek,  
Elzbuieta Macioszek,  
Felix J. Garcia Clemente,  
Francesco Zirilli,  
Grigorios N. Beligiannis,  
Grzegorz Sierpinski,  
H.Hamidi,  
Hala Abukhalaf,  
Hamed Taherdoost,  
Hamid Ali Abed AL-Asadi,  
Hamid Khemissa,  
Hamidreza Bolhasani,  
Hamidreza Rokhsati,  
Hamza Ouarnoughi,  
Hasan Kadhem,  
Heba Mahmoud Afify,  
Hedayat Omidvar,  
Hiba Zuhair,  
Hiromi Ban,  
Hlaing Htake Khaung Tin,  
Hon Hai,  
Ilango Velchamy,  
Ilango,  
Ilham Huseyinov,  
Ines Bayouth Saadi,  
Isa Maleki,  
Israa Shaker Tawfic,  
Iyad Alazzam,  
Jagadeesh HS,  
Janaki Raman Palaniappan,  
Jawad K. Ali,  
Jaymer Jayoma,  
Jesuk Ko,  
Jia Ying Ou,  
Joao Antonio Aparecido Cardoso,  
Jong-Ha Lee,  
Juntao Fei,  
Kamel Benachenhou,  
Kamel Jemai,  
Kanstantsin MIATLIUK,  
Keneilwe Zuva,  
Kevin Caramacion,  
Khalid M.O Nahar,  
Ki-II Kim,  
Kire Jakimoski,  
Kirtikumar Patel,  
University of Beira Interior, Portugal  
Principal Data Scientist at Zuora, Inc., USA  
Beijing Institute of Technology, China  
Fincons SpA, Italy  
University of Craiova, Romania  
Abdelmalek Essaadi University, Morocco  
Silesian University of Technology, Poland  
Silesian University of Technology, Poland  
University of Murcia, Spain  
Sapienza Universita Roma, Italy  
University of Patras, Greece  
Silesian University of Technology, Poland  
K.N. Toosi University of Technology, Iran  
Palestine Polytechnic University, Palestine  
Hamta Business Corporation, Canada  
Basra University, Iraq  
USTHB University Algiers, Algeria  
Islamic Azad University, Iran  
Sapienza university of Rome, Rome  
LAMIH, France  
American University of Bahrain, Bahrain  
Cairo university, Egypt  
National Iranian Gas Company, Iran  
Al-Nahrain University, Iraq  
Nagaoka University of Technology, Japan  
University of Computer Studies, Myanmar  
Research Institute, Taiwan  
CMR Institute of Technology, India  
CMR Institute of Technology, India  
Istanbul Aydin University, Turkey  
Tunis University, Tunisia  
Science and Research Branch, Iran  
Ministry of Science and Technology, Iraq  
Yarmouk University, Jordan  
APS College Of Engineering(VTU), India  
Brunswick Corporation, USA  
University of Technology, Iraq  
Caraga State University, Philippines  
Universidad Mayor de San Andres, Bolivia  
York University, Canada  
The Federal Institute of São Paulo, Brazil  
Keimyung University, South Korea  
Hohai University, P. R. China  
Blida University, Algeria  
University of Gabes, Tunisia  
Bialystok University of Technology, Poland  
University of Botswana, Botswana  
University at Albany, USA  
Yarmouk University, Jordan  
Chungnam Natiuonal University, Korea  
FON University, Republic of Macedonia  
Hargrove Engineers and Constructors, USA

Klenilmar Lopes Dias,  
Koh You Beng,  
Koh You Beng,  
Konstantinos Karampidis,  
Larry De Guzman,  
Litao Guo,  
Loc Nguyen,  
Luca De Cicco,  
Luisa Maria Arvide Cambra,  
MA. Jabbar,  
Mabroukah Amarif,  
Mario Versaci,  
Michail Kalogiannakis,  
Mirsaeid Hosseini Shirvani,  
Mohamed Fakir,  
Mohamed Ismail Roushdy,  
Mohammad A. Alodat,  
Mohammad Nasfikur Rahman Khan,  
Morteza Alinia Ahandani,  
Mourad Chabane Oussalah,  
Mu-Song Chen,  
Nadia Abd-Alsabour,  
Nahlah Shatnawi,  
Narinder Singh Gorla,  
Nikola Ivković,  
Nikolai Prokopyev,  
Oleksii K. Tyshchenko,  
Omid Mahdi Ebadati,  
Pascal Lorenz,  
Pellumb Killogjeri,  
Piotr Kulczycki,  
Piyush Behre,  
Ramadan Elaïess,  
Ramgopal Kashyap,  
Rao Li,  
Robert Ssali Balagadde,  
Rodrigo Pérez Fernández,  
Rupak Vignesh Swaminathan,  
Saad Aljanabi,  
Said Agoujil,  
Samir Kumar Bandyopadhyay,  
Santosh Kumar Bharti,  
Sebastian Fritsch,  
Shah Khalid Khan,  
Shahid Ali,  
Shahram Babaie,  
Shashikant Patil,  
Shing-Tai Pan,  
Siarry Patrick,  
Siddhartha Bhattacharyya,  
Sikandar Ali,  
Federal Institute of Amapa, Brazil  
Universiti Malaya, Malaysia  
University of Malaya, Malaysia  
Hellenic Mediterranean University, Greece  
Isabela State University, Philippines  
Xiamen University of Technology, China  
Loc Nguyen's Academic Network, Vietnam  
Politecnico di Bari, Italy  
University of Almeria, Spain  
Vardhaman College of Engineering, India  
Sebha University, Libya  
DICEAM - Univ. Mediterranea, Italy  
University of Crete, Greece  
Islamic Azad University, Iran  
Sultan Moulay Slimane university, Morocco  
Ain Shams University, Egypt  
Sur University College, Oman  
Independent University, Bangladesh  
University of Tabriz, Iran  
University of Nantes, France  
Da-Yeh University, Taiwan  
Cairo university, Egypt  
Yarmouk University, Jordan  
Punjabi University, India  
University of Zagreb, Croatia  
Kazan Federal University, Russia  
University of Ostrava, Czech Republic  
Kharazmi University, Tehran  
University of Haute Alsace France, France  
GeoGebra Institute of Albania, Albania  
Systems Research Institute, Poland  
Microsoft, USA  
University of Benghazi, Libya  
Amity University Chhattisgarh, India  
University of South Carolina Aiken, USA  
Kampala international University, Uganda  
Universidad Politécnica de Madrid, Spain  
Amazon.com, Inc. USA  
Alhikma College University, Iraq  
Moulay Ismail University, Morocco  
University of Calcutta, India  
Pandit Deendayal Energy University, India  
IT and CS enthusiast, Germany  
RMIT University, Australia  
AGI Education Ltd, New Zealand  
Islamic Azad University, Iran  
ViMEET Khalapur Raigad MS, India  
National University of Kaohsiung, Taiwan  
Universite Paris-Est Creteil, France  
Rajnagar Mahavidyalaya, India  
China University of Petroleum, China



Stefano Michieletto,  
Subhendu Kumar Pani,  
Suhad Faisal Behadili,  
sukhdeep kaur,  
Sun-yuan Hsieh,  
Taleb zouggar souad,  
Tanzila Saba,  
Titas De,  
Tzung-Pei Hong,  
Umesh Kumar Singh,  
Usman Naseem,  
V.Ilango,  
Vilem Novak,  
Wei-Chiang Hong,  
William R. Simpson,  
WU Yung Gi,  
Yang Cao,  
Yekini Nureni Asafe,  
Yilun Shang,  
Zayar Aung,  
Ze Tang,

University of Padova, Italy  
Krupajal Engineering College, India  
University of Baghdad, Iraq  
punjab technical university, India  
National Cheng Kung University, Taiwan  
Oran 2 University, Algeria  
Prince Sultan University, Saudi Arabia  
Data Scientist - Glance Inmobi, India  
National University of Kaohsiung, Taiwan  
Vikram University, India  
University of Sydney, Australia  
CMR Institute of Technology, India  
University of Ostrava, Czech Republic  
Yuan Ze University, Taiwan  
Institute for Defense Analyses, USA  
Chang Jung Christian University, Taiwan  
Southeast University, China  
Yaba College Of Technology, Nigeria  
Northumbria University, UK  
National Research University, Russia  
Jiangnan University, China

## Technically Sponsored by

**Computer Science & Information Technology Community (CSITC)**



**Artificial Intelligence Community (AIC)**



**Soft Computing Community (SCC)**



**Digital Signal & Image Processing Community (DSIPC)**



## TABLE OF CONTENTS

### **4<sup>th</sup> International Conference on Machine Learning and Soft Computing (MLSC 2023)**

**Micam: Visualizing Feature Extraction of Nonnatural Data.....01-19**  
Randy Klepetko and Ram Krishnan, University of Texas at San Antonio, USA

**Comparative Study of Anxiety Symptom's Predictions From Discord Chat  
Messages using Automl.....21-35**  
Anishka Duvvuri, Navya Kovvuri, Sneka Kumar, Rebecca Victor, Tanush Kaushik,  
Basis Independent Silicon Valley High school, USA

**A First-Person Shooter Game Designed to Educate and Aid the Player Movement  
Implementation.....37-48**  
Chunhei Zhu<sup>1</sup>, Yujia Zhang<sup>2</sup>, <sup>1</sup>Beckman High School, <sup>2</sup>University of California Irvine,  
USA

### **9<sup>th</sup> International Conference on Information Technology Convergence and Services (ITCSS 2023)**

**Augmented Reality Assisted Maintenance and Monitoring at Onpremise Data  
Center.....49-58**  
Muhamad Adib Bahari, Shah Runnizam Mohd Salleh and Muhammad Kamal Abdul  
Kiram, TNB Research Sdn. Bhd., Malaysia

### **9<sup>th</sup> International Conference on Advances in Computer Science and Information Technology (ACSTY 2023)**

**Review of Metrics to Measure the Stability, Robustness and Resilience of  
Reinforcement Learning.....59-78**  
Laura L. Pullum, Oak Ridge National Laboratory, USA

**A Cryptocurrency Analysis Tool based on Social Metrics.. .....79-91**  
Bill Xu<sup>1</sup>, Yu Sun<sup>2</sup>, <sup>1</sup>École Internationale de Montréal, USA <sup>2</sup>California State  
Polytechnic University, USA

**A Novel System for Regional Twitter Hate Speech Analysis and Detection using  
Deep Learning Models and Web Scraping.....93-103**  
Nicole Ma<sup>1</sup>, Yu Sun<sup>2</sup>, <sup>1</sup>Sage Hill School, USA, <sup>2</sup>California State Polytechnic  
University, USA

**A Smart Mobile Application Designed to Educate and Aid the Public in Combating Climate Change.....105-116**  
Kerry Zhang, University High School, USA

### **9<sup>th</sup> International Conference on Software Engineering (SOFE 2023)**

**A NLP-learning Powered Customizable Approach Towards Auto-blocking Distracting Websites.....117-125**  
Yulin Zhang<sup>1</sup>, Yu Sun<sup>2</sup>, <sup>1</sup>University High School, USA, <sup>2</sup>California State Polytechnic University, USA

### **9<sup>th</sup> International Conference on Natural Language Processing (NATP 2023)**

**Evaluating and Improving Context Attention Distribution on Multi-Turn response generation using Self-Contained Distractions.....127-143**  
Yujie Xing and Jon Atle Gulla, Norwegian University of Science and Technology, Norway

### **4<sup>th</sup> International Conference on Big Data and Blockchain (BDAB 2023)**

**Implementation of a New E-voting System based on Blockchain using ECDSA with Blind Signatures.....145-152**  
Lina Lumburovska, Vesna Dimitrova, Aleksandra Popovska-Mitrovikj, Ss. Cyril and Methodius University of Skopje, North Macedonia

### **9<sup>th</sup> International Conference on Information Technology Convergence and Services (ITCSS 2023)**

**Privacy-Preserving Online Sharing Charging Pile Scheme with Different Needs Matching.....153-166**  
Zhiyu Huang<sup>1,2</sup>, <sup>1</sup>Hunan University of Science and Technology, China, <sup>2</sup>Hunan Key Laboratory for Service Computing and Novel Software Technology, China

**9<sup>th</sup> International Conference on Advances in Computer Science and  
Information Technology (ACSTY 2023)**

**Research on Construction of RDF with HBase.....167-178**  
Hui Hu, Nanjing University of Aeronautics and Astronautics, China

**4<sup>th</sup> International Conference on Artificial Intelligence and Big Data  
(AIBD 2023)**

**Ethical Algorithms in Human-Robot-Interaction. A Proposal.....179-186**  
Joerg H. Hardy, Free University of Berlin, Germany

# MICAM: VISUALIZING FEATURE EXTRACTION OF NONNATURAL DATA

Randy Klepetko and Ram Krishnan

Department of Electrical and Computer Engineering  
University of Texas at San Antonio, San Antonio, Texas, USA

## ***ABSTRACT***

*Convolutional Neural Networks (CNN) continue to revolutionize image recognition technology and are being used in non-image related fields such as cybersecurity. They are known to work as feature extractors, identifying patterns within large data sets, but when dealing with nonnatural data, what these features represent is not understood. Several class activation map (CAM) visualization tools are available that assist with understanding the CNN decisions when used with images, but they are not intuitively comprehended when dealing with nonnatural security data. Understanding what the extracted features represent should enable the data analyst and model architect tailor a model to maximize the extracted features while minimizing the computational parameters. In this paper we offer a new tool Model integrated Class Activation Maps, (MiCAM) which allows the analyst the ability to visually compare extracted feature intensities at the individual layer detail. We explore using this new tool to analyse several datasets. First the MNIST handwriting data set to gain a baseline understanding. We then analyse two security data sets: computers process metrics from cloud based application servers that are infected with malware and the CIC-IDS-2017 IP data traffic set and identify how re-ordering nonnatural security related data affects feature extraction performance and identify how reordering the data affect feature extraction performance.*

## ***KEYWORDS***

*Convolutional Neural Networks, Security, Malware Detection, Visualizations, Deep Learning*

## **1. INTRODUCTION**

Improvements in CNN have achieved better than human performance in computer image recognition [1]. They also have applications in non-image related research. Other sources of data include text [2], sound [3], and in the medical diagnostics of DNA [4]. These are examples where the data is organized in a grid like fashion by nature. But what about cases where the data wasn't ordered by natural phenomena? Sensors on an automated vehicle [5] for example, does the order matter? In most “*nonnatural*” applications the grid order is defined by some man made structure, usually defined by an arbitrary specification. The term “*nonnatural*” is for those data ordering schemes not defined by nature, as opposed “unnatural” which infers the “supernatural”.

Using CNN in detecting cyber-security issues has shown significant interest. Raw IP traffic [6, 7] computer process metrics [8], and industrial sensors [9] are all data sets where researchers are evaluating CNN use in security. CNN are successful as they identify patterns from large data sets to extract features. CNN are often applied as a feature extractor source which is supplied as the input stage to decision network such as a densely connected, recurrent, or another machine learning procedure. To maximize the patterns detected, order of the grid supplied to a CNN should be of concern when it was arbitrarily defined.

In our previous research we showed that using the structural order in detecting malware using computer process metrics is not preferred when training a shallow or deep CNN model if high accuracy and precision are desired. We found that using statistical relationships as a basis for order does improve performance. We showed that grouping our data points created *artificial objects* that most CNN models could better identify as malware features. Do these findings hold true when analysing raw IP data traffic?

CNN models consist of various layers each performing a specific task. Some run convolutions via a series of filters, some pool data points together, while others perform mathematical operations over either one or a pair of grids. Comprehending what could be going on within these “black boxes” is improved with *visualization* techniques that let the user by eyesight understand what the network is doing.

By providing transparency and an explanation [10] as to the network parameter intensities they assist the researcher in all stages of the network development life cycle. Early in model construction visualizations provide failure details letting the engineer to see how performance is affected by model changes. Visualizing the hidden layers enhance confidence that the model is identifying a proper set of features during network maturity. As the network exceeds human performance, the visualization tools provide a computer instructor, teaching novel ways of examining the data to the researcher.

A number of visualization tools have been created to assist in the engineering and development of CNN. Some image generating tools create graphs to provide a higher level understanding of the data flow within the model. Other visualization tools provide histograms of the parameters as they adjust over the training period. One important class of visualization tools are classification response graphs which are designed to show the how responsive a pixel is to that particular classification made on a tested sample. These include Saliency and CAM graphs. Most of these latter tools apply well with image data, but are not as well suited for data that is not visual in nature like cyber-security. These novel cases is where this research is focused. To find the patterns that the CNN layers are extracting from non-natural security data as features, we built a better visualization tool.

The contributions of this paper are:

- Present a new visualization tool, Model integrated Class Activation Maps (MiCAM), a confluence of several visualization tools, and show how MiCAM assists in identifying feature extraction response.
- Test previous defined ordering algorithms with a new security data set, raw IP traffic from CIC-IDS-2017, showing again that statistical correlation provides a better than randomly ordered performance.

The remainder of the paper is organized as follows: Section 2 discusses related work using CNN with nonnatural data and a background on visualization tools. Section 3 outlines the methodology including a description of MiCAM and data organization. Section 4 describes the analysis procedure and evaluation results. Section 5 summarizes and concludes this paper.

## 2. RELATED WORK

### 2.1. Convolutional Neural Networks and NonnaturalData

CNN have matured to where they have many applications, beyond the recognition of images. Their ability is to identify patterns in large data sets when that data can be arranged in a grid. For instance, in the analysis of tire tread using the parameters measured during the manufacturing process. Lihao and Yanni [11] with eleven metrics sampled from four manufacturing levels, they arranged a 4x11 matrix and were able to identify faulty tires with a 94% accuracy.

Golinko et al. in [12] used a one dimensional CNN as a feature extractor front for other machine learning algorithms (k-Nearest Neighbour with k=1, Support Vector Machine, and Random Forest), examining if the ordering of nonnatural "Generic" source data for the CNN has a performance impact on the final classifying algorithm. They found that using statistical correlation as a method for identifying relationships of adjacent data performed well, but not pre-ordering the data for CNN feature extraction was detrimental. Using a correlation ordering scheme offered improvement in most cases, especially for kNN and SVN, improving accuracy from 76% with no feature extraction to 82% if the data points were ordered by correlation prior to CNN feature extraction.

In a collision detection system Park, et. al. [13] used information from robotic sensors and actuators creating 66 data points. Testing both a one-dimensional CNN and a Support Vector Machine Regression they were able to show that the CNN would perform better if it trained with enough data, but the SVMR performed better with less training.

With cross-related sensor data (local speed, GPS location, and accelerometer) from automated vehicles, Van Wyk, et. al. [5] used an analyser to identify whenever any of the sensors behaved anomalously. The different analysers tested included a Kalman Filter, CNN, and a CNN-KF hybrid. Each had its unique benefits.

### 2.2. Convolutional Neural Networks and Security

CNNs have found value in cyber-security applications. Their ability to find patterns instead of statically looking for distinct signatures provide feature extraction from large data sets and using the algorithm's nonlinear space enables the dynamic/online detection of zero-day attacks. These data sources are usually nonnatural.

From hypervisors in a cloud environment Abdelsalem et al. [8] places process metrics as they are reported into a grid as they look for malware as it is injected into virtual machines. This produced a set of 35 metrics that were captured per time segment for every running process. They were supplied to a Lenet-5 [14] CNN. Using the order as found in the logs and specifications, they achieved an 89% accuracy. McDoleet. al. [15] follow up with research analysing deeper CNN architectures using the same data set and ordering scheme. Kimmellet. al. [16] includes using recurrent neural networks (RNN), by testing the validity of using long short term memories (LSTM) and Bi-Direction LSTMs. They also explore if the order has an effect on training and discover that it does affect performance for all models.

Arranging raw IP traffic packets in a grid after the physical layer was stripped, Zhang et. al. [7] analysed them using CNN, LSTM, and a hybrid of the two. They tested for both binary classification (benign/maleficent) and multi-classification (benign + 10 maleficent types). They show all systems achieve quite remarkable, near-perfect results. For binary classification from



the best in precision was the hybrid which was better than CNN, followed by LSTM. With multi-classification, CNN had some minor advantage in precision over the hybrid, but LSTM was behind both.

### 2.3. Visualizing Convolutional Neural Networks

Visually revealing the hidden layers provides researchers comprehension behind neural network decisions. They are also evolving as the field matures. They are some form of flow and layer diagrams, class activation maps [17] (CAM), gradient visualization [18] sensitivity to perturbations [19], or a confluence of these.

Flow and model diagrams were introduced since the very first deep learning models were published. They provide a visual representation of the mathematical processing objects that are coded into the software. They represent these abstracts as spheres or cubes, and as multiple mathematical objects are aligned in a layer, the graphical constructs are placed next to each other in a row. A line between objects represent communication or parameter passing pathways. For convolutional layers, a plane of objects is used, and stacks of planes are a symbol which includes the third filter dimension. For brevity when the interpretation is understood, sometimes a higher dimension abstract is represented by a lower level visual construct.

CAMs were initially generated using a weighted sum and up-sampling the class activation maps from the penultimate layer to generate activation regions of the original image. CAMs have evolved using different parameters as the weight values for the ratio in summing the class activation maps. Detailed by Selvarajuet. al in 2016, GradCAM [10] uses gradients in a back propagation step with a *relu* function. LayerCAM[20] published by Jiang, et. al. collects the GradCAM maps from all of the individual layers and then sums them together in a normalized total that includes higher amount of detail from the shallower layers within the network.

GradCAM++ [21] by Chattopadhyay et. al. modified GradCAM by adjusting a normalizing factor used to determine the weights for the individual gradients from the feature activation maps. Devised by Wang et. al. in 2020 ScoreCAM [22], goes further by dropping the gradients altogether and include a contribution value to measure the importance of each activation map. EigenCAM submitted by Muhammad et. al. [23] replaces the gradients with an eigenvector that is derived from a combinations of the weights from all of the layers.

All of these CAM systems have several things in common. They attempt to produce a two dimensional region that shows how the features on the penultimate layer are related to the objects within the sample image, and they do so with only a single degree of the resulting image, grey scale. This works fine with shallower networks since the features within the penultimate layer are closely related to the pixels within the source image, but what about CNN models that are deep, and the final feature set have no direct relationship to the initial image, e.g. a source image of 75x75 pixels (75 x 75 x 3) and the resulting DenseNet-121 penultimate layer (2 x 2 x 1028). A 2x2 grid does not distinctly map to points on a 75x75 grid. A better visualization tool is needed to understand these deeper models.

### 2.4. CNN Models

Many models have been derived as CNN technology matures. Each new model uses a novel technique to accomplish higher degree of computer image object identification and classification precision. We examine three in this research. LeNet model [24], ResNet[25], and DenseNet[26]. They were chosen for their distinct architecture and their place as milestones in CNN evolution.

In 1989, LeCun et al. introduced the LeNet-5 model in [24]. The first to use back propagation in a practical application as it identifies and classifies black and white images of hand written numbers provided by the US postal system. The goal, a 1% error rate, was reached after 23 epoch of training. It was sequential in structure and consisted of three convolutional and two dense layers. The data set they used closely resembles one used in this paper, the MNIST [27] data set of handwritten numbers.

He et al. in late 2015 [25], introduced ResNet which added a new feature in network topology, the residual connection. This is a new link from the input of a convolution stage directly to the output, using addition, which feeds the next stage's input. This reintroduces the input data to the following stages, greatly reducing vanishing gradient, a major issue when training deep networks. They were able to win first in the 2015 ImageNet competition taking the prize in all categories: classification, localization, and detection. They also won the 2015 COCO competition in the categories of detection and segmentation. This research uses the smallest published version, ResNet-18.

Revised in 2018, Huang et al. [26] published DenseNet. Like residual links, they have connections around layers but instead of using addition as the function for combining the input source with the output, they used concatenation. Each stage increases in depth from the previous, creating a depthwise *denser* input cluster. This forwards all of the input information and details previously gathered from earlier stages to the latter stages. This reduces the data lost by the addition process used in residual links, maintaining input integrity, further mitigating the vanishing gradient. They use bottleneck stages to reduce parameter count in the latter layers. These include a depth separable convolution to reduce the depth and a pooling layer for a reduction in width and height. This study uses DenseNet-121.

Our previous research expands on the techniques discussed by Abdelsalem et al. [8] by exploring the relationship between ordering of the rows, columns, and various CNN models' performance analysing cyber-security computer process metric data. We identified several structural relationships on which to base our ordering scheme, we included the use of a statistical relationship as an option for ordering the metric columns, and we compared those against a background of random orderings. We showed that using structural relationships as an ordering appeared to have no more advantage than a random order and statistical relationships as a foundation for order offered some performance improvement. We also shared that although the visualization tools available showed some response, the plots were difficult to interpret.

In this research we test these statistical ordering techniques using a different cyber-security data set, raw IP traffic from CIC-IDS-2017 following the work done by Zhang et al. [7], and compare it to the structural order used in Zhang's research. We share a new tool, the Model integrated Class Activation Map (MiCAM), a confluence of model diagrams with activation maps displayed per layer. We use with the MNIST data set to establish a baseline so we can understand the visual representations as they are constructed for features extracted from black and white images. We then use this tool to analyse the features generated for two cyber-security data sets, computer process metrics and raw IP traffic, and show how it better displays feature extraction.

### 3. METHODOLOGY

#### 3.1. Model integrated Class Activation Maps MiCAM

To fully visualize feature extraction we built a tool that is a combination of a model diagram with class activation maps. A model diagram is a flow plot that has the network layers displayed with

the data pathways identified so the engineer can visually see the related connections between layers. This flow diagram is rather trivial when working with sequential models, but can be quite complex when dealing with network like Inception Net, that have multiple interconnections between layers. A class activation map (CAM) is a combination via a weighted sum of all of the activation maps for the filters a single layer. The weights for this sum define the type of CAM. This tool takes the model diagram and instead of displaying an object (i.e. layer) as a graphical construct (sphere or rectangle) it displays the CAM for that layer. After the MiCAM diagram is complete the result is a map clearly showing the various features that each layer defines as important in identifying the class of a tested sample. A diagram of the process steps used to generate MiCAM plots is found in Figure 1.

The multiple steps to the process are identified in alphabetical order. In the beginning the researcher has the chosen model and the data seen in (A). The model is trained in step (B) while at the same time, the model layout is extracted from the model definition. From the result, the trained model in (C) and the layout the layers are pulled out and the activation model is defined (D). This model has the pre-trained layers from the trained model laid out with the filters' outputs exposed for sampling later.

With the activation model, we take a sample (E) and test it determine how it is classified in (F). Using the activation model post-test and the model layout, we now extract the outputs or activations (G) for all of the filters and the associated filters' weights in (H). In step (I), using an inverse Fourier transform, we take the inverted convolution between a filters' activation and its kernels' weights. We then take the result for each filter and use the weight for the particular filter to sum a single CAM plot for each layer. This CAM plot is then up-sampled to match the original input grids dimensions.

To enhance the details within the CAM plots, we use the full RGBA pallette, by associating different variations of the CAM data within the plotted pixels. We note that every plot has a maximum and minimum range that is scaled to 256 discrete intensities. These pixel values can be positive or negative, so we use a set of *relu* functions to display these variations in intensities by matching one of the 4 degrees to a specific range of values. For blue we use the full range of minimum to maximum for this plot, scaled to the 256 colour levels. For red we display the positive peaks using the *relu* of the values, scaling from zero to the maximum of this plot. For green we display the negative peaks using *relu* of the negative value or zero if the values are positive, scaling from zero to the minimum. For alpha and size, we use the full range for the plot, but scale the results to the minimum and maximum values for all of the CAM plots within the model. The results are very dynamic images that display a full range of the extracted features.

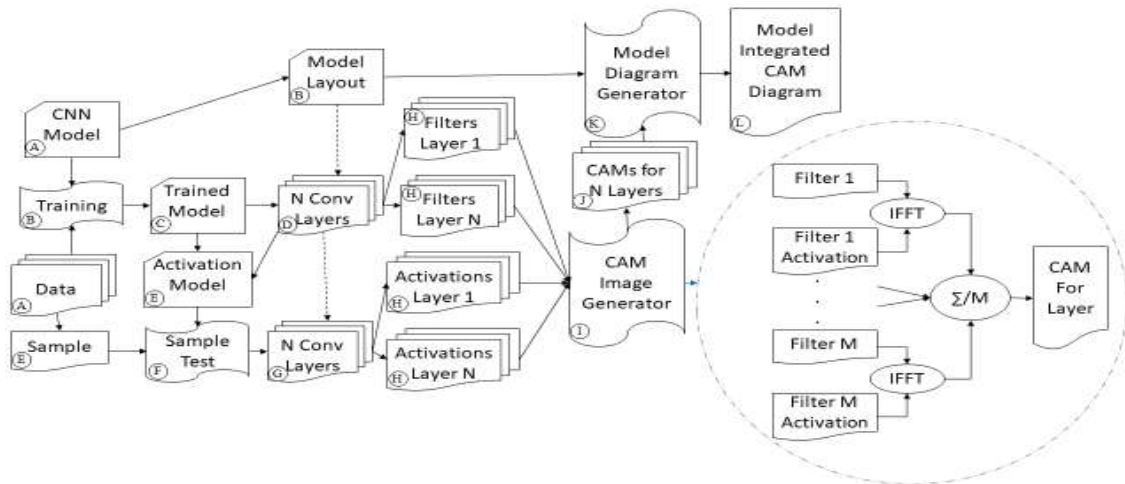


Figure 1. MiCAM Generation Process

After generating the images, we have a stack of CAM plots for all of the layers within the model (J). For layers that are not convolutional, we simply use a weighted sum of the outputs across the filter dimension, and then up-sample them to provide a graphic for each layer. For layers that are one-dimensional (flatten and dense) MiCAM fits the linear data within the input grid, scaling elements up if there are fewer data points within the layer than the width and height of the source data. The CAM plots are then integrated with the Model Layout in the Model Diagram Generator (K) which produces the final MiCAM diagram.

The code uses the "pydot/graphviz" graphical diagram module which has an interface for integrating images in place of objects. We added some slight modification for passing two list of parameters. One the list of layers than had CAM plot images, and the second was the list of the image files for the CAM plots. Both lists must be the same length, and for proper diagram generation the layer names in the first list should align with the filenames in the second list. The code is under open source license and found at <https://github.com/rklepetko/MiCAM.git> for easy access.

### 3.2. Dataset-1: MNIST Handwritten Numbers

The MNIST data set, compiled and released by Deng [27], consist of a library of images of hand written numerical text. The 10 image classes are from "0" to "9" and consist 60,000 samples from 250 census takers and 250 high school students. Another set of testing data was compiled from a separate group of 250 census and high school students, but comprised of only 10,000 samples. We join the two, shuffle them and use 20% of the data for testing, 20% in validation, or 14,000 of the samples per set, with the remaining used for training. Each sample was fitted in to a 20x20 grid, normalized for shading, and centered on a 28x28 image. For our analysis on deeper models, we further up-sampled the image to 75x75 pixels in size. Visual examples of our MNIST data are seen in Figure 2. We use several MNIST samples with the MiCAM diagrams to give us a base line on evaluating feature extraction.



Figure 2. MNIST Data Samples

### 3.3. Dataset-2: Malware Infected Computer Metric by Process Grids

The second data source is process metric samples taken from virtual machines in a cloud IaaS environment. They were application servers arrayed in a LAMP stack hosted web-site. The machines were injected with malware halfway through the experiment. There were 114 infections each from different malware packages. During the experiment, the server was polled for process log samples. Each sample is for a unique process running on the VM kernel and contains a set of  $M$  number of metrics per process during a section of time. Stacking  $P$  processes that are captured during a single time slice results in the matrix:

$$\mathbf{X}_t = \begin{bmatrix} p_1 & x_{m1p1} & x_{m2p1} & \dots & x_{mMp1} \\ p_2 & x_{m1p2} & x_{m2p2} & \dots & x_{mMp2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_P & x_{m1pP} & x_{m2pP} & \dots & x_{mMpP} \end{bmatrix}$$

Our initial research was identifying how order of nonnatural data within the grid affects performance. We generated ten random rows and ten random columns for 100 options. We also identified several structural ordering methods and after examining the mathematical relationships within images derived several statistical relationships to see if they provide any improved performance. Since objects in images have pixels that are statistically correlated, we use the statistical functions used are detailed in Table 3 of Appendix A, at the end of this paper. The metric columns calculation were independent per sample, so we used correlation between two metrics (Eq. 1), absolute value of correlation (Eq. 3), and one minus the absolute value of the correlation, or what we called anticorrelation (Eq. 4) to test a counter hypothesis.

Unlike the independent metric columns, process rows calculations were dependent between samples, so the correlation function (Eq. 2) was derived per metric for a pair of processes. A sum of the correlation between two processes (Eq. 5) was used as the base process relationship function, from which we also derived an absolute correlation (Eq. 6) and anticorrelation (Eq. 7) relationship functions.

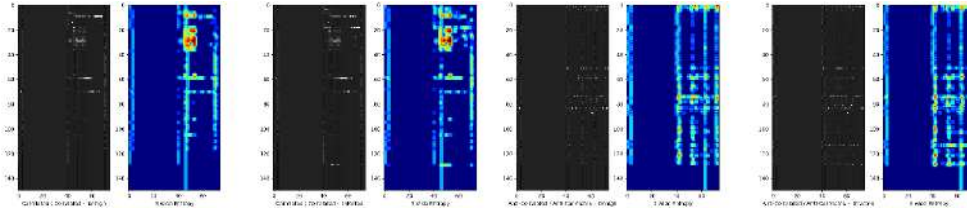


Figure 3. Correlated Rows & Columns (left) And Anticorrelated Rows & Columns(right) Benign & Infected Samples

These functions are then processed through the ordering algorithm, shown in Algorithm 1 found within the Appendix A which generates the ordering for each row or column along an axis. It can be seen in the samples ordered with correlation, Figure 3 left, and anticorrelation, Figure 3 right, our correlation functions generate artificial objects while the anticorrelation disperses them. The 35 metrics were expanded through one hot encoding to  $M = 75$  metric columns and we made available room in the matrix for as many as  $P \leq 150$  process rows. The 29+ million process samples from 114 experiments (malware infections), and consisted of 31,064 grids, about half of which are considered infected. The experiments were split between 60% training, 20% validation, and 20% testing. The entire sample set for each experiment was included in the group

it was assigned, so no experiment was split between training, validation, and testing. Every training and test set was reorganized among the 252 different ordering schemes we generated. We test all of our samples on several models, identified the best and worst ordering schemes (Table 6 in Appendix A) for each CNN model we trained, and then analysed the results of the best and worst ordering schemes with MiCAM.

### **3.4. Dataset-3: CIC-IDS-2017 Raw IP Data with Attack Vectors**

The CIC-IDS-2017 data set has captured live, raw IP traffic that is intentionally subjected to various forms of attack vectors. There were 12 attack classes, ten of which were of a sizable sample. The sample count and break down by class is included with the results in Table 1 found in the next section. This traffic is compiled by session, with the sessions labeled benign or by attack class. Each packet in the session has the physical layer of the IP packet stripped, the first fourteen bytes, and only the following 160 bytes kept. If the original packet wasn't 174 bytes long, the remaining portion of the 160 bytes are supplied with zeros. The first ten packets of the session are then compiled in order of transmission, and if there aren't ten packets, the remaining are filled with zeros. The result is a 10x160 byte grid.

This is the basic single sample from the data set before it is reorganized into a 40x40 square. The current order of this grid is IP specification for the columns and transmission time for rows. Transmission time is a natural order, an instance in a sequence, but IP specification, human defined, is a nonnatural order. Is IP specification the best order? Will statistical correlation on the data be a high performing order? These are secondary questions this study is trying to resolve.

To test these hypothesis we first generated 100 random column ordering schemes to process and compare. Since the calculations between bytes are independent per sample we used the function Eq. 1 and the ordering algorithm shared in Algorithm 1, both found within the Appendix A. To diversify the number of ordering options available to analyse we used correlation relationships within different data subsets. The first data set was total of all samples. Next, we separate between the benign and maleficent and use the correlation of each of these data subsets. We then extract each of the attack types as subsets and generate correlated orderings from each of these. The idea is to see if it is possible to focus on a specific artificial objects by re-arranging the order to match the correlation generated from that subset sample type. We also generate an absolute value of the correlation (Eq. 3) and anticorrelation (Eq. 4) orderings for each of the datasets.

This resulted in 146 ordering schemes to analyse. After reordering, the samples were then translated into a 40x40 grid by splitting the 160 bytes into four sections and stacking them on top of each other in order. We randomly reordered the samples and split them into 60% training, 20% validation, and 20% testing sets. We cover the evaluation in the next section.

## **4. EVALUATION**

### **4.1. MiCAM and MNIST**

The resulting MiCAM plots are large when compared to other CAM plots. They are usually vertically aligned following the model layout as the CNN is constructed. Since not only the convolutional layers, but the pooling, adding and concatenation layers, along with the final flatten and dense layers at the end of the convolutional stages are all plotted, the combined plot contains a visual representation of each layer. For example, DenseNet-121, with 121 convolutional layers has a total of 429 individual layers within the model. For brevity the diagrams are not all

included but can be found on GitHub at: <https://github.com/rklepetko/MiCAM.git>. We do share snapshots of elements that illuminate the value of this visualization tool.

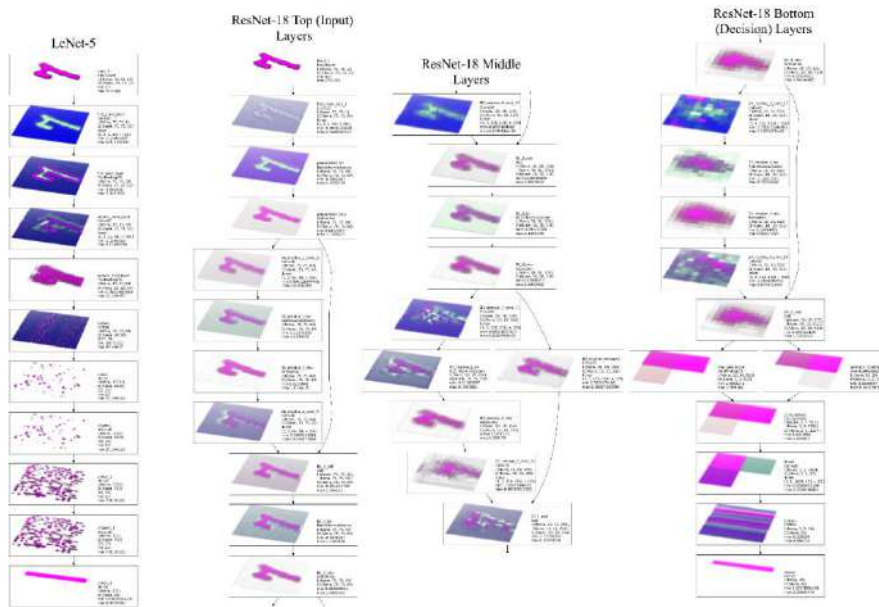


Figure 4. MiCAM Plots of LeNet-5 (left one) and MiCAM Plot Clips of ResNet-18(right three) analysing an MNIST sample”7”

To start we examine the LeNet-5 MiCAM plot (Left side of Figure 4) which clearly shows how the convolution layers build the identifying features. Examining the dense layers closely it can be seen the variation in the colour pixelintensities relate to specific features the network has identified.

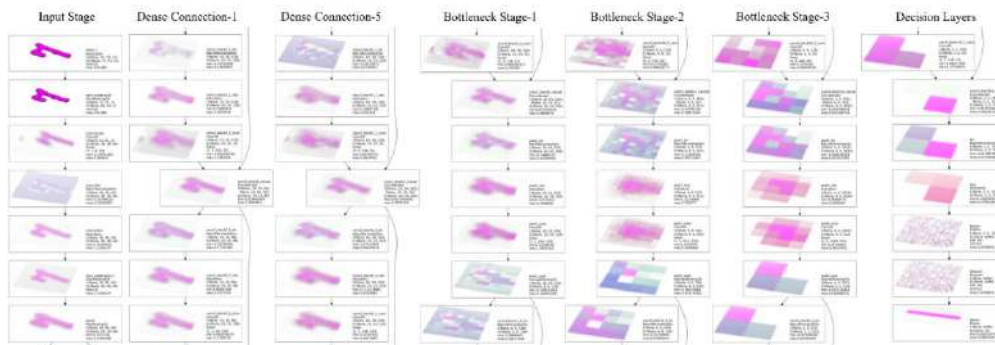


Figure 5. Clips of MiCAM Plot from a DenseNet-121 analysing an MNIST sample”7”

It is even clearer when examining ResNet-18 MiCAM plot (the right three plots of Figure-4) as we display the top, or input stages, the middle of the model, and the final bottom or decision layers. It's seen in these graphs how the residual links re-introduce features extracted from earlier layers. It can also be viewed within the final layers how the ResNet-18 network collapses the number of extracted features to relatively few, 40, as compared to LeNet-5 which was 20736.



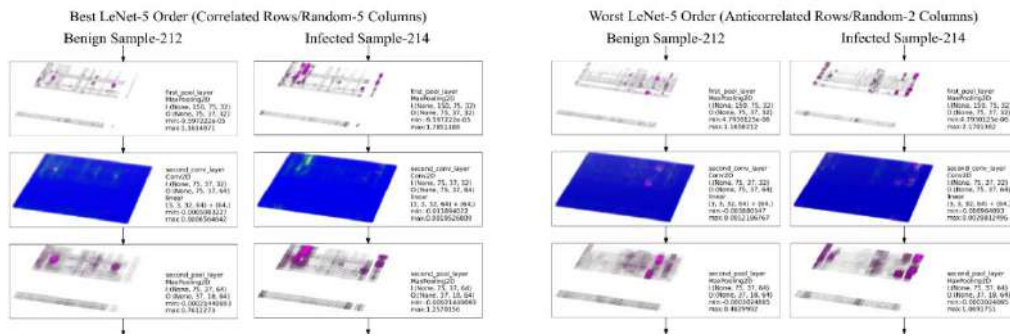


Figure 6. MiCAM Plots of the Lower Quarter for the LeNet-5 Best (left) and Worst (right) Ordering of Samples Benign #212 and Infected #214

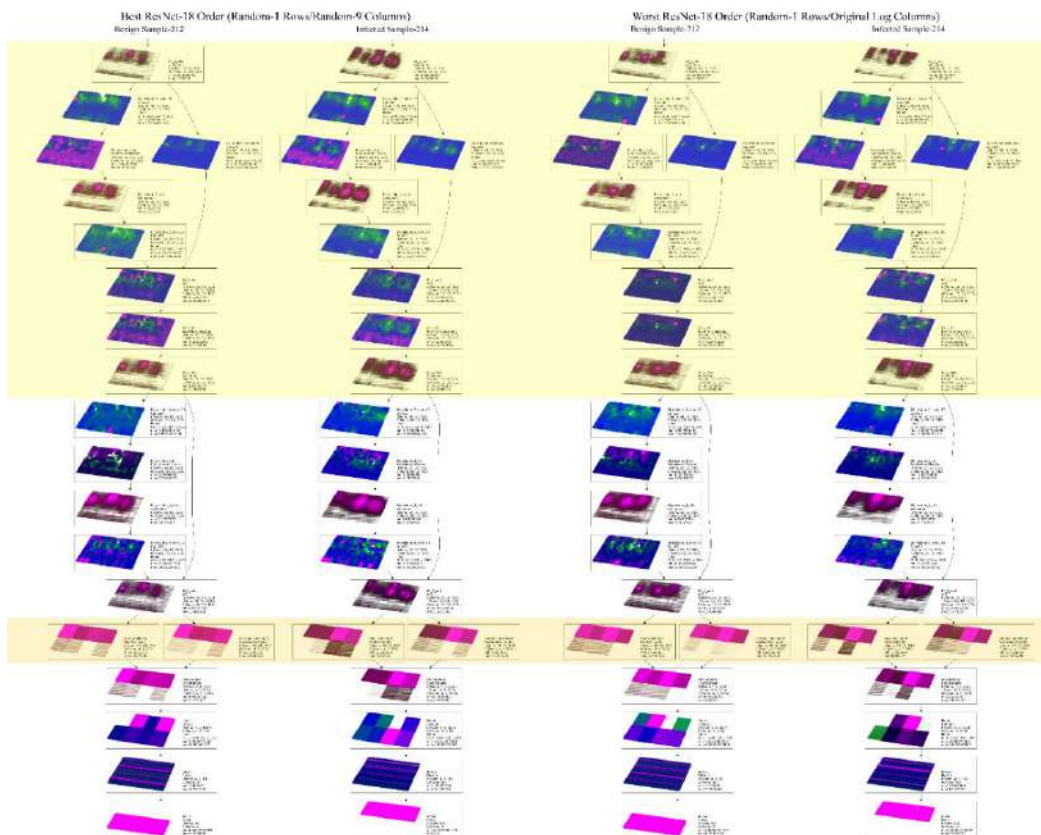


Figure 7. MiCAM Plots of Pooling and Last Convolution Layers for the ResNet-5 Best (left) and Worst (right) Ordering of Samples Benign #212 and Infected #214

Examining the DenseNet-121 MiCAM plot of the same sample (Figure 5), we choose to share 49 of the 429 layers. From left to right we include the details of the input layers, the first and last dense connection before the first bottleneck, the three bottle neck stages, and the final decision layers. In the dense connection plots, the reintroduction of the input stages initial features (outline of a "7") is visible as the data cascades through all the way to the first bottleneck stage, maintaining a higher level of details for feature extraction precision. We can also see that it is these bottle neck layers that are compiling the features for discrimination later.



Table 1. Sample Counts by Class Set and Analysis Results by Order.

Sample			Prec/Recal mAP			Improve/Degrade			
Random Average			99.545%			0%			
Internet Protocol Specification			<b>99.703%</b>			<b>34.675%</b>			
Column Order	Sample Set	Count	%	Corr	ABS	Anti	Corr	ABS	Anti
	Bot	1228	0.151%	99.54%	99.58%	99.57%	-0.36%	6.98%	6.58%
	DDoS	44918	5.539%	99.61%	99.65%	99.55%	14.19%	22.21%	1.70%
	DoS Hulk	5952	0.734%	99.58%	99.57%	99.53%	7.86%	5.70%	-3.04%
	DoS								
	Slowhttptest	4216	0.520%	99.54%	99.59%	99.56%	-0.27%	9.00%	3.60%
	DoS sloworis	3872	0.477%	99.46%	99.55%	99.66%	-18.97%	1.98%	25.29%
	FTP - Patator	3974	0.490%	99.59%	99.55%	99.52%	8.93%	0.34%	-5.35%
	Infiltration	6	0.001%	99.59%	99.61%	99.64%	9.57%	13.52%	21.20%
	PortScan	158410	19.534%	99.57%	99.55%	99.57%	4.51%	0.13%	4.48%
	SSH-Patator	2978	0.367%	99.59%	99.56%	99.54%	9.82%	3.71%	-0.19%
	Web Attack - Brute Force	1363	0.168%	99.61%	99.57%	99.48%	14.67%	5.37%	-15.16%
	Web Attack - Sql Injection	12	0.001%	99.61%	99.59%	99.55%	15.33%	10.09%	0.84%
	Web Attack - XSS	625	0.077%	99.58%	99.48%	99.52%	8.80%	-13.74%	-6.56%
	Malfeicent	227554	28.060%	99.56%	99.57%	99.53%	2.98%	6.30%	-4.36%
	Benign	583411	71.940%	99.52%	99.54%	99.51%	-6.38%	-0.55%	-6.81%
	Total	810965	100%	99.59%	99.59%	99.57%	10.95%	9.63%	6.47%
	Average Improvement	-	-	-	-	-	5.44%	5.38%	1.91%

## 4.2. MiCAM and Malware Infections

As mentioned in the previous section, we use MiCAM to analyse the difference between the best and worst ordering schemes (Table 6 in Appendix A) when searching for malware. Between the LeNet-5 MiCAM plots we found the pooling layers to have the most distinguishing characteristics. It is visible in Figure 6 which is divided by the best and worst ordering schemes. We can see how the features are better defined in the pooling layers with the stronger intensities, and the range on the infected sample of the best order is noticeably larger in the second pooling layer than the worst order.

Within the ResNet-18 plots we see a number of items to take notice of in Figure 7. Several of the CAM plots are identifying clusters of data points they have some significance on the decision. In particular the B4 residue convolution layers and associated additions and activation layers, highlighted in yellow, perhaps point to particular data points the CNN identifies as maleficent or benign. Also noticed is that the features from the best ordering are distinct in the final pooling layers for the benign and infected samples, highlighted orange, but the worst order displays those layers as having similar features by comparison.

To keep this report within the space limit, we are not displaying the DenseNet-121 graphs, but they are available at the Git site mentioned earlier. Things to note, the CAM plots most relatable to the source data are the last convolution stage before the first bottle neck stage. We see a number of highlighted pixels of interest for the different classifications. In particular we notice a highlighted row within the best ordering scheme for an infected sample, perhaps informing us that we have an infected process on that row.

### 4.3. MiCAMand IP Attacks

One of the unique details this study considers relevant is analysing the affect that order has on nonnatural data, and one data set, the CIC-IDS-2017 raw IP-traffic data, poses a scenario to tests our hypothesis. As described previously, we devised 146 different columns related ordering schemes, and compare them with the results when using the order devised using the IP-specification as a scheme. We trained a shallow LeNet-3 CNN model (2 convolution and one dense layer), matching previously published research and the results are found in Table 1. They include the PR curve mAP for every non-random ordering scheme we devised including a percentage of improvement over the average mAP for all of the randomly generated schemes. We include a breakdown of the results in our conclusion section.

Table 2. Best and Worst Ordering Schemes for Maleficent IP-Traffic.

CNN Architecture	Best Column Order	mAP Score	Worst Column Row Order	mAP Score
Lenet-3 (10 Epoch)	IP Specification	99.70%	Random-40	99.45%

To analyse the differences between the best and worst ordering schemes with the MiCAM diagrams, we identified them and include their details in Table2.

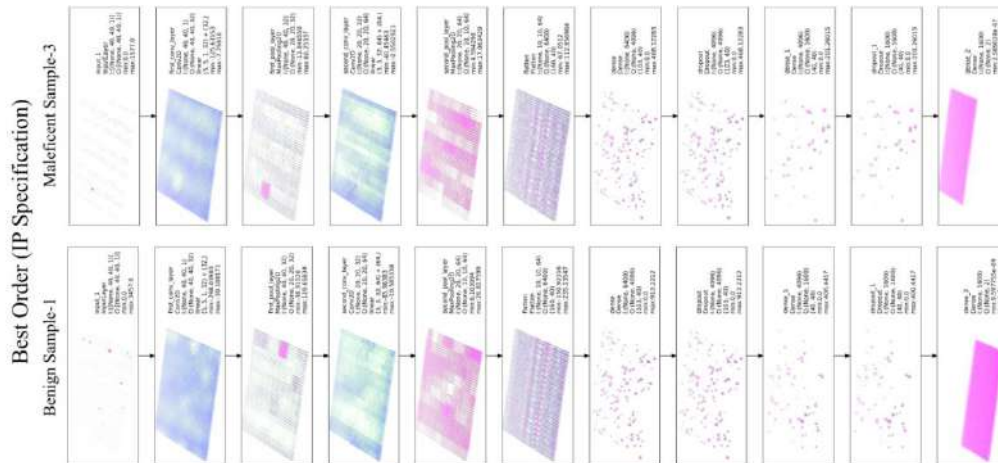


Figure 8. MiCAM Plots of LeNet-3 Analysing Best Order (IP Spec) IP Packets with Benign and Maleficent Packages

Examining the MiCAM plots, in Figures 8 and 9, we can see how the best order has a wider range, with the peak negative values showing very distinct regions within the convolutional layers. Also in both orders, in several layers it shows the first quarter of the sample is significant in finding the maleficent sample's attack vector, while several areas within the packet are identified significant in the benign.

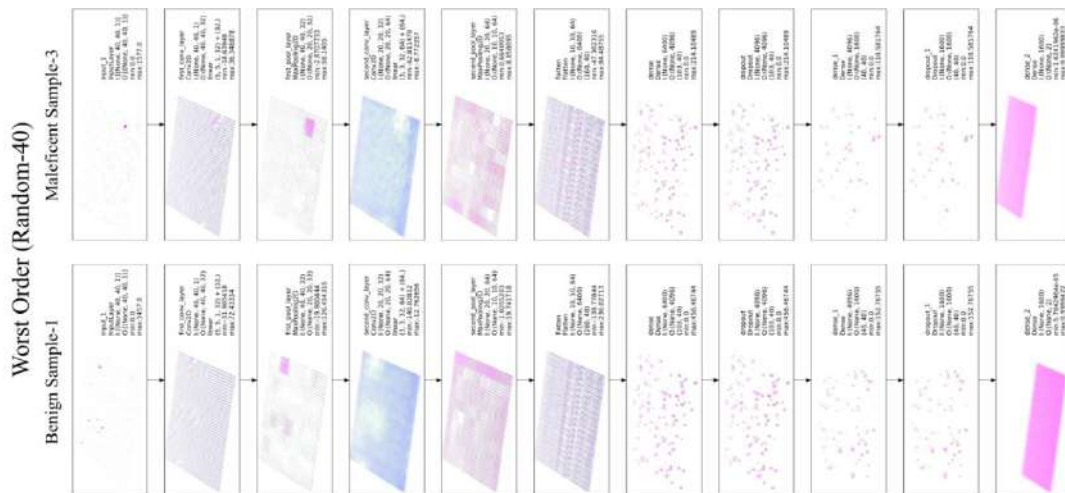


Figure 9. MiCAM Plots of LeNet-3 Analysing Worst Order (Random-40) IP Packets with Benign and Maleficent Packages

## 5. CONCLUSIONS

The MiCAM diagrams offer more detail regarding feature extraction within the CNN models. They visually expose the layers allowing the user to further understand the intensities of the features extracted within the CNN structure. We've seen and identified several capabilities that allow us to further compare how minor variation in a model or process can affect feature extraction. This offers an additional tool for engineers as they tailor CNN models to non natural cybersecurity applications. We used CAM plots that normalized the sum of activation maps with the filters weights for the individual map, but one could enhance this tool to include other CAM variations, and better methods for displaying the one dimensional (flatten and dense) layers.

There is some processing cost related to generating the MiCAM plots. We went to some length to take advantage of the graphics engine by plotting all of the pixels within a single layer at one time which greatly improved the rendering speed.

When comparing the CIC-IDS-217 dataset ordering schemes, counter our hypothesis, the ordering scheme derived when following the IP specification exceeded expectations out performing all other ordering options. This shows the care to which IEEE specification was laid to logically organize the data packets as they relate to each other.

It is also interesting to note that the majority of the ordering schemes devised around a statistical relationship between data bytes within subsets of the data also performed better than average. The surprise regarding the subsets was the correlation of the benign samples. Only two other correlation subsets showed a major degradation in performance compared to the random average, and those sample sizes were less than one percent of the total samples. The benign correlation had 70% of the samples, but resulted in more than a 6% degradation. Focusing on benign samples to find maleficent actors proved detrimental. These findings support our hypothesis that statistical correlation does produce a better than average precision, as long as the data subset that the correlation is taken from has enough maleficent samples.

It's also notable that although anticorrelation ordering did have some significant improvement for some subsets, the majority of the subsets showed a poorer performance. Absolute value of correlation produced only one significantly detrimental ordering using a subset, which comprised

of less than 1/10th of 1% of the total samples, so appears to be a relatively safe when using with a shallow network.

To further our understanding on how order affects CNN performance when analysing non-natural data, we plan on continuing our research by:

- Using MiCAM to further analyse the differences in CNN model response when comparing ordering schemes.
- Identifying other security and nonsecurity datasets on which to test ordering hypothesis and techniques.
- Integrating the CNN feature extraction with other models to see if order can improve performance of ML hybrids.

## ACKNOWLEDGEMENTS

This work is partially supported by NSF grants HRD-1736209 and CNS-1553696.

## REFERENCES

- [1] He, K., Zhang, X., Ren, S., Sun, J. (December 2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: The IEEE International Conference on Computer Vision (ICCV).
- [2] Lee, J.Y., Derroncourt, F. (2016) Sequential short-text classification with recurrent and convolutional neural networks. CoRR abs/1603.03827, <http://arxiv.org/abs/1603.03827>.
- [3] Deng, L., Hinton, G., Kingsbury, B. (May 2013) New types of deep neural network learning for speech recognition and related applications: an overview. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8599-8603 <https://doi.org/10.1109/ICASSP.2013.6639344>.
- [4] Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velazquez Vega, J.E., Brat, D.J., Cooper, L.A.D. (2018) Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* 115(13), E2970-E2979. <https://doi.org/10.1073/pnas.1717139115>, <https://www.pnas.org/content/115/13/E2970>.
- [5] van Wyk, F., Wang, Y., Khojandi, A., Masoud, N. (2020) Real-time sensor anomaly detection and identification in automated vehicles. *IEEE Transactions on Intelligent Transportation Systems* 21(3), 1264-1276. <https://doi.org/10.1109/TITS.2019.2906038>.
- [6] Liu, C., Dai, L., Cui, W., Lin, T. (2019) A byte-level cnn method to detect dns tunnels. In: 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). pp. 1-8. <https://doi.org/10.1109/IPCCC47392.2019.8958714>.
- [7] Zhang, Y., Chen, X., Jin, L., Wang, X., Guo, D. (2019) Network intrusion detection: Based on deep hierarchical network and original flow data. *IEEE Access* 7, 37004-37016. <https://doi.org/10.1109/ACCESS.2019.2905041>.
- [8] Abdelsalem, M., Krishnan, R., Huang, Y., Sandu, R. (2018) Malware detection in cloud infrastructure using convolutional neural networks. *IEEE 11th International Conference on Cloud Computing*.
- [9] Hu, Y., Zhang, D., Cao, G., Pan, Q. (2019) Network data analysis and anomaly detection using CNN technique for industrial control systems security. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). pp. 593-597. <https://doi.org/10.1109/SMC.2019.8913895>.
- [10] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (Oct 2019) Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128(2), 336-359. <https://doi.org/10.1007/s11263-019-01228-7>, <http://dx.doi.org/10.1007/s11263-019-01228-7>.
- [11] Lihao, W., Yanni, D. (Nov 2018) A fault diagnosis method of tread production line based on convolutional neural network. In: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). pp. 987-990. <https://doi.org/10.1109/ICSESS.2018.8663824>.

- [12] Golinko, E., Sonderman, T., Zhu, X. (Dec 2018) Learning convolutional neural networks from ordered features of generic data. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 897-900 <https://doi.org/10.1109/ICMLA.2018.00145>.
- [13] Park, K.M., Kim, J., Park, J., Park, F.C. (2021) Learning-based realtime detection of robot collisions without joint torque sensors. *IEEE Robotics and Automation Letters* 6(1), 103–110. <https://doi.org/10.1109/LRA.2020.3033269>.
- [14] Liu, G., Zhao, F. (2007) An efficient compression algorithm for hyperspectral images based on correlation coefficients adaptive three dimensional wavelet zerotree coding. In: 2007 IEEE International Conference on Image Processing. vol. 2, pp. II-341 -- II-344. <https://doi.org/10.1109/ICIP.2007.4379162>.
- [15] McDole, A., Abdelsalam, M., Gupta, M., Mittal, S. (2020): Analyzing CNN based behavioural malware detection techniques on cloud IAAS. In: Zhang, Q., Wang, Y., Zhang, L.J. (eds.) *Cloud Computing - CLOUD 2020*. pp. 64-79. Springer International Publishing, Cham.
- [16] Kimmel, J.C., Mcdole, A.D., Abdelsalam, M., Gupta, M., Sandhu, R. (2021) Recurrent neural networks based online behavioural malware detection techniques for cloud infrastructure. *IEEE Access* 9, 68066-68080. <https://doi.org/10.1109/ACCESS.2021.3077498>.
- [17] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. (2015) Learning deep features for discriminative localization.
- [18] Erhan, D., Bengio, Y., Courville, A., Vincent, P. (2009) Visualizing higher layer features of a deep network. *University of Montreal* 1341(3), 1.
- [19] Ribeiro, M.T., Singh, S., Guestrin, C. (2016) " why should i trust you?" explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1135-1144
- [20] Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y. (June 2021) Layercam: Exploring hierarchical class activation maps. *IEEE Transactions on Image Processing* pp. 1-1. <https://doi.org/10.1109/TIP.2021.3089943>.
- [21] Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N. (Mar 2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). <https://doi.org/10.1109/wacv.2018.00097>, <http://dx.doi.org/10.1109/WACV.2018.00097>.
- [22] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X. (2020) Score-cam: Score-weighted visual explanations for convolutional neural networks
- [23] Muhammad, M.B., Yeasin, M.(Jul 2020) Eigen-CAM: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). *IEEE*. <https://doi.org/10.1109/ijcnn48605.2020.9206626>.
- [24] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D. (1989) Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4), 541-551. <https://doi.org/10.1162/neco.1989.1.4.541>.
- [25] He, K., Zhang, X., Ren, S., Sun, J. (2015) Deep residual learning for image recognition. *CoRR* abs/1512.03385, <http://arxiv.org/abs/1512.03385>.
- [26] Huang, G., Liu, Z., Weinberger, K.Q. (2016) Densely connected convolutional networks. *CoRR* abs/1608.06993, <http://arxiv.org/abs/1608.06993>.
- [27] Deng, L. (2012) Themnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29(6),141-142.

## APPENDIX-A

In Table 3 on the next page are the equations we presented as ordering foundations in our previous research and test MiCAM analysis with in this study. They are used as parameters for ordering Algorithm 1 found on the following page.

Table 3. Detailed Set of Statistical Relationship Functions.

Equation 1: Metric/Byte Column Statistical Correlation Function
$\rho_{mi mj} = \frac{E(x_{mi}x_{mj}) - E(x_{mi})E(x_{mj})}{\sqrt{E(x_{mi}^2) - E(x_{mi})^2} - \sqrt{E(x_{mj}^2) - E(x_{mj})^2}}$
Equation 2: Process Row Statistical Correlation Function
$\rho_{mk pi pj} = \frac{E(x_{mk pi} x_{mk pj}) - E(x_{mk pi})E(x_{mk pj})}{\sqrt{E(x_{mk pi}^2) - E(x_{mk pi})^2} - \sqrt{E(x_{mk pj}^2) - E(x_{mk pj})^2}}$
Equation 3: Metric/Byte Column ABS-Correlation Function
$\rho_{ABS mi mj} =  \rho_{mi mj} $
Equation 4: Metric/Byte Column Anticorrelation Function
$\rho_{ANTI mi mj} = (1 -  \rho_{mi mj} )$
Equation 5: Process Row Correlation (Sum) for All Metrics Function
$\rho_{SUM pi pj} = \sum_{k=1}^M \rho_{mk pi pj}$
Equation 6: Process Row ABS-Correlation for All Metrics Function
$\rho_{ABS pi pj} = \sum_{k=1}^M  \rho_{mk pi pj} $
Equation 7: Process Row Anticorrelation for All Metrics Function
$\rho_{ANTI pi pj} = \sum_{k=1}^M (1 -  \rho_{mk pi pj} )$
Equation 8: Metric/Byte Column Total Relationship Function
$\rho_{TOT mi} = \sum_{j=1}^M (\rho_{mi mj})$
Equation 9: Process Row Total Relationship Function
$\rho_{TOT pi} = \sum_{j=1}^P (\rho_{SUM pi pj})$

Algorithm 1: Derive Statistical Relationship Order.

<p>For features along an axis, <math>f_i</math>, define a function, <math>\rho_{f_i f_j} \forall i, j</math>;</p> <p>From <math>\rho_{f_i f_j}</math> define <math>\rho_{TOT f_i} \forall i</math>;</p> <p>Create a selection pool of features <math>P \ni f_i</math>;</p> <p><b>While</b> <math>P \neq \emptyset</math> do:</p> <p>    Create an empty bidirectional queue <math>Q</math> for features <math>f_i</math>;</p> <p>    Find <math>\max(\rho_{TOT f_i}) \forall f_i \in P</math>;</p> <p>    Place corresponding feature <math>f_{\max(\rho)}</math> onto <math>Q</math>;</p> <p>    Remove feature <math>f_{\max(\rho)}</math> from <math>P</math>;</p> <p>    Create two pointers left, <math>L</math>, and right, <math>R</math>; <math>L, R \in Q</math>;</p> <p>    Point <math>L</math> and <math>R</math> towards <math>f_{\max(\rho)}</math> in <math>Q</math>;</p> <p>    <b>While</b> <math>P \neq \emptyset</math> and not(<b>STOP</b>) do:</p> <p>        <b>If</b> <math>\exists \rho_{f_L f_i} \forall f_i \in P</math> or <math>\exists \rho_{f_R f_i} \forall f_i \in P</math> then:</p> <p>            Find <math>\max(\rho_{f_L f_i}, \rho_{f_R f_i}) \forall f_i \in P</math>;</p> <p>            Place new feature <math>f_{\max(\rho)}</math> next to the appropriate <math>f_L</math> or <math>f_R</math> on <math>Q</math>;</p> <p>            Remove new feature <math>f_{\max(\rho)}</math> from <math>P</math>;</p> <p>            Move the appropriate pointer <math>L</math> and <math>R</math> towards the new <math>f_{\max(\rho)}</math> in <math>Q</math>;</p> <p>        <b>Else:</b></p> <p>            Stack current queue <math>Q</math> into final ordered axis <math>V</math>;</p> <p>            <b>STOP</b>;</p> <p>    <b>End if else</b>;</p> <p>    <b>End while</b>;</p> <p><b>End while</b>;</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Following that are two tables which are the results from our previously published VM Malware analysis but are now using with MiCAM. The first (Table 4) is the PR curve mAP results from the various ordering schemes. The second is (Table 5) the percentage of improvement (or degradation) over the observed average. The last (Table 6) shows the best and worst performing ordering schemes that we use to analyse with MiCAM.

Table 4. Mean AUC for Precision Recall Curves for Malware Analysis.

CNN Architecture	All Options	Correlated Rows	ABS-Corr Rows	Anti-Corr Rows	Correlated Columns	ABS-Corr Columns	Anti-Corr Columns
LENET-5 (20 epoch)	99.550%	99.680%	99.580%	99.090%	99.590%	99.600%	99.440%
ResNet-18	89.850%	87.020%	86.560%	94.530%	91.240%	89.230%	95.130%
DenseNet-121	99.530%	99.700%	99.430%	99.200%	99.600%	99.520%	99.560%

Table 5. Percentage Improvement over Average (Mean) Performance for Malware Analysis.

CNN Architecture	100% minus All Mean	Correlated Columns	ABS-Corr Columns	Anti-Corr Columns	Correlated Rows	ABS-Corr Rows	Anti-Corr Rows
LENET-5 (20 epoch)	0.450%	8.889%	11.111%	-24.444%	28.889%	6.667%	-102.222%
ResNet-18	10.150%	13.695%	-6.108%	52.020%	-27.882%	-32.414%	46.108%
DenseNet-121	0.470%	14.894%	-2.128%	6.383%	36.170%	-21.277%	-70.213%

Table 6. Best and Worst Ordering Schemes for Malware Analysis by CNN Model.

CNN Architecture	Best Combined		mAP	Worst Combined		mAP
	Row Order	Column Order	Score	Row Order	Column Order	Score
LENET-5 (20 epoch)	Correlated	Random-5	99.82%	Anticorrelated	Random-2	98.64%
ResNet-18	Random-1	Random-9	99.99%	Random-1	Original	50.31%
DenseNet-121	VMPID	Random-1	99.87%	ABS-Correlated	Random-5	96.36%

## AUTHORS

Randy Klepetko graduated from Texas A&M University in May of 1990 with a Bachelor's in Computer Science. Since, he has enjoyed a 30 year career in a broad range of engineering and digital fields. He continued his education with courses in electrical engineering and electronics at Texas A&M University, San Antonio College, and University at San Antonio during the 1990's and 2000's, saturating his knowledge in audio and video engineering protocols, methods and techniques. In 2017 he returned to academia at the University of Texas at San Antonio to enhance his competency regarding digital security where he was encouraged to join the University's Center for Security and Privacy Enhanced Cloud Computing (C-SPECC), receiving his Masters in May 2022 and scheduled PhD completion in December



Ram Krishnan is a Professor of Electrical and Computer Engineering at the University of Texas at San Antonio, where he holds Microsoft President's Endowed Professorship. His research focuses on (a) applying machine learning to strengthen cybersecurity of complex systems and (b) developing novel techniques to address security/privacy concerns in machine learning. He actively works on topics such as using deep learning techniques for runtime malware detection in cloud systems and automating identity and access control administration, security and privacy enhanced machine learning and defending against adversarial attacks in deep neural networks. He is a recipient of NSF CAREER award (2016), the University of Texas System Regents' Outstanding Teaching Award (2015) and the UTSA President's Distinguished Award for Research Achievement (2016). He received his PhD from George Mason University in 2010.





# COMPARATIVE STUDY OF ANXIETY SYMPTOM'S PREDICTIONS FROM DISCORD CHAT MESSAGES USING AUTOML

Anishka Duvvuri, Navya Kovvuri, Sneka Kumar, Rebecca Victor, Tanush Kaushik

Basis Independent Silicon Valley High school, USA

## ABSTRACT

*Anxiety is a chronic illness especially during the Covid and post-pandemic era. It's important to diagnose anxiety in its early stages. Traditional Machine learning (ML) methods have been developmental intense procedures to detect mental health issues, but Automated machine learning (AutoML) is a method whereby the novice user can build a model to detect a phenomenon such as Generalized Anxiety Disorder (GAD) fairly easily. In this study we evaluate a popular AutoML technique with recent chat engine (Discord) conversation dataset using anxiety hashtags. This multi-symptom AutoML Random Forest predictive model is at least 75+% accurate with the most prevalent symptom, namely restlessness. This could be a very useful first step in diagnosing GAD by medical professionals and their less skilled hospital's IT area using pre diagnostic textual conversations. But it lacks high quality in predicting GAD in most symptoms as found by a low 50% precision on most symptoms (except 5). The AutoML technology is quicker for IT professionals and gives a decent performance, but it can be improved upon by more sophisticated ANN methods like Convolution neural networks that plug AutoML's symptom's deficiencies with at least 80+% precision and 0.4+% in F1 score, namely in detecting poorly predicted symptoms of concentration and irritability.*

## KEYWORDS

*General Anxiety Disorder, machine learning, Discord chat, AutoML, Convolutional neural network*

## 1. INTRODUCTION

Feelings of worry, nervousness, or unease about uncertain outcomes are usually associated with an emotion called anxiety. Anxiety is a feeling many people experience throughout their lives in varying degrees of intensity. However, there are individuals with anxiety so severe it impedes their day-to-day functioning. Those with general anxiety disorder (GAD) experience this worry and unease chronically, and they are unable to control the behaviors that come along with it [1]. The prevalence of GAD in the general population ranges from 1.9% to 5.4% [2]. The prevalence in the United States was 3.1% in 2014 and 5.7% over the course of a patient's lifetime, according to epidemiological surveys [1]. While the age of onset is highly variable, there is a higher rate of females having GAD compared to males at a rate of 2:1 [1]. Risk factors of GAD include low socioeconomic status, exposure to childhood adversity, and being female; twin studies have also shown anxiety has a moderate chance (about 15-20%) of being inherited through genetics [1]. General anxiety disorder also commonly co-occurs with major depressive disorder, with many of their symptoms overlapping, thus making it difficult to distinguish the two diagnoses [1]. Although the inability to experience pleasure does not overlap with GAD, individuals with this condition feel hopelessness like patients with major depressive disorder. According to the NCS, 65% of patients with GAD also had at least one other disorder at the time of their assessment [2].

Additionally, patients with GAD have a higher risk for other mental health disorders and physical symptoms such as chronic pain and asthma [1]. About 35% of those with GAD turn to alcohol and medications to reduce their symptoms of anxiety, which can increase the risk for substance and drug-related problems [1].

This research focuses on the following questions:

- How can one get a complete picture of GAD using symptoms?
- Can we predict the severity of GAD using symptoms as the guiding post?
- How can social media conversations help with diagnosing GAD?
- Can Auto ML be a quick way to model GAD?
- What intelligent techniques are there beyond Auto ML to better predict GAD?

The main contributions of this research are to:

- Analyze GAD symptoms with the help of popular chat engine conversations.
- Predict severity GAD based on symptoms.
- Evaluate off the shelf Auto ML techniques to predict GAD.
- Build a dataset for Anxiety to be used for further research purposes.

## 2. RELATED WORK

Machine learning algorithms are present in the process of diagnosing and predicting future outcomes related to mental health. The science field uses machine learning algorithms in a variety of areas because they can apply solutions without human input [4]. They have also been used to detect the prognosis of various mental health disorders such as bipolar disorders and panic disorder [3]. In the context of anxiety disorders, these machines have been used to predict and detect anxiety. More specifically, they can potentially be used to diagnose anxiety, predict future risk of anxiety, and predict responses to medical treatment. [4] Prior studies have made use of a Bayesian network, a probable graphic model that represents certain attributes amongst others; Artificial Neural Networks (also known as ANNs), adaptive processing units made for discovering new knowledge; Support Vector Machines (also known as SVM), supervised learning models meant for analyzing and classifying data based on training data it has been provided with; decision trees, they predict responses and branches based on specific features present in data; linear regression (also known as LR), explains the relationship between the outcome and another variable, and Neuro-Fuzzy Systems (also known as NFSs), combines neural networking and fuzzy logic to develop new fuzzy rules or functions of the inputs and outputs in the system [3].

Individually, these algorithms and machines can do little. This is why the scientific community has turned to hybrid models, models that combine two or more existing methods to create a more efficient product [3,11]. The community has made use of logistic regression, Naive Bayes, and a Bayesian Network while feeding the machines input data based on inferred heart-rate measurements. [12] used a Bayesian joint model paired with a linear mixed effects model and a generalized linear model to analyze input data from self-esteem data and anxiety diagnosis in regards to examining the development of self-esteem on adult onset anxiety disorder. [13] used ANN to analyze a dataset of patients. [14] used ANN, RF, NFS, and SVM to predict affective states of an individual based on five defined classes without any input data. [15] used a SVM nestled within a leave-one-out-cross-validation framework to separate GAD diagnoses from healthy subjects and major depressive disorder with input data from questionnaires, cortisol release and white and grey matter volumes.

[11] found the Bayesian Network model was the most accurate machine learning algorithm they had tested with an accuracy of 73.33%. [12] found the joint-model to be more effective with a 75% accuracy rate. [13] found an overall 82.35% accuracy rate using ANN. [14] found the NFS to be the most accurate than the other models with a 84.3% being the highest accuracy level. Taking all the data into account, 15 found an improved accuracy in detecting GAD from healthy individuals and differentiating it from major depressive disorder at 90.10% and 67.46% respectively. According these results, ANN was concluded to be the most accurate in predicting GAD [3].

Recent studies [16, 17] have given rise to more accurate machine learning algorithms in detecting anxiety with an accuracy of 90% – 96%. But these are cumbersome for hospital IT staff to implement. AutoML [8] is a recent technique to enable novice users to build ML intelligence. This has several advantages to evaluate the traditional ML models fairly quickly and focus on the optimizations and business question at hand. However, they have a severe limitation: “But although automation and efficiency are among AutoML’s main selling points, the process still requires human involvement at a number of vital steps, including understanding the attributes of domain-specific data, defining prediction problems, creating a suitable training dataset, and selecting a promising machine learning technique” [8]. This study evaluates the efficacy of AutoML using recent technique (Navigator by Pyxeda [9]) in building a high performance machine learning model for GAD diagnosis. Secondly, the study digs deep into symptoms with the anticipation that ML techniques using AutoML would not be as performant as per observation in [18]. Finally, there is no such in-depth study in recent times leveraging AutoML. The study aims to validate the hypothesis that AutoML techniques can be a low hanging fruit for hospital IT staff to use to diagnose GAD and similar mental health problems.

### 3. THORETICAL FRAMEWORK

The DSM-IV [7] highlights six symptoms of tension/negative affect which are associated with symptoms of anxiety and for consideration when making a diagnosis for GAD. Three of the six symptoms (only one for children) must occur along with excessive anxiety being present for more days than not, for at least six months to qualify. This iteration of the DSM also indicates these symptoms and worry connected to them must be perceived by the individual to be difficult to control. According to [2], this revision from the previous edition of the DSM was made due to evidence showcasing the difference between the anxiety of the general population and those with GAD; while they both had worries about similar content, the controllability of the worry was reported to be vastly diminished in individuals with GAD [2]. More criteria explain that the anxiety, in order to qualify for a diagnosis, should cause distress and/or impairment in important functioning and should not be the result of substances or their side effects.

In the DSM-V, the criteria have not changed significantly, but there is a short addition: “the disturbance should not be better explained by another mental disorder” as per American Psychiatric Association. From DSM [7], the six criteria of anxiety are specified as:

1. restlessness or feeling on edge
2. being easily fatigued
3. difficulty concentrating or the mind going blank
4. irritability
5. muscle tension
6. sleep disturbance

One of the ways to diagnose GAD is via non-elicited speech (e.g. chats). Text chat can be used as a source for both the second and third ways, both elicited and non-elicited speech. Our plan is to

create a corpus of sentences from ‘Discord’ anxiety health conversations, tag them for each of the six text-based criteria, and create a prediction model for each.

#### 4. DATA COLLECTION

The study uses Discord [8] to collect the data related to anxiety. Discord is a free audio, video, and text chat service used by tens of millions of individuals aged 13 and above to communicate and socialize with their communities and friends. People use Discord on a regular basis to discuss a wide range of topics, from art projects and family vacations to homework and mental health. The vast majority of servers are private, invite-only locations where friends and communities may communicate and spend time together. Because all discussions are opt-in, users have complete control over who they connect with and how they interact on Discord. It's a place where they can be themselves while still spending time with others who share their interests and hobbies.

The public chats are communities of "servers." Servers are groups of persistent chat rooms. The channel #R/SocialAnxiety and #Kai Havanon Discord as in specified in Table 1 is a popular place to express and discuss anxiety. Almost 42,000 posts and comments have been downloaded from various individuals. The data has been preprocessed as per Table 2 to correlate a user's sentences to anxiety.

Table 1: Discord Original Uncondensed Dataset

Channel	Totals	Collection date
#Kai Havan	10,482	02/09/2022 – 04/10/2022
#R/Social Anxiety	31,959	05/20/2021- 04/18/2022

#### Pre-processing

After Data Collection, the study involved looking at distribution of each variable (all the six symptoms) as well as the Anxiety Intensity variable to understand the spread of the data. Once the data gaps were identified, the data was balanced using SMOTE techniques. Then feature engineering was performed to create new variables using the existing data provided. Using Natural Language Processing, techniques were performed, such as the removing stop words, analysing the errors using spell checker, and cleaning the suspected errors by correcting their spellings. During explorations, visualizations were generated for the content variable to understand distribution of text length and build a word cloud to understand the top or the most common words used by the people during anxiety.

### 5. EXPERIMENTATION

#### 5.1. Tagging & Labeling

The first involved tagging the individual data pieces and required a clear foundation for what qualified for each of the six symptoms. This would assist the study and future researchers as well. The symptoms were described as the following: “edginess or restlessness,” “tiring easily; more fatigued than usual,” “impaired concentration or feeling as though the mind goes blank,” “irritability (which may or may not be observable to others),” “increased muscle ache or soreness,” and “difficulty sleeping (due to trouble falling asleep or staying asleep, restlessness at night, or unsatisfying sleep).” The process involved looking out for behaviors that fit the symptoms such as descriptions of not knowing what to say which implies feelings of the mind

going blank, tones of voice to attribute to edginess or irritability, expressions of exhaustion or hints of fatigue to ascribe to the easily fatigued and/or difficulty sleeping symptoms, and more. During tagging the data, it was found noticeable edginess and the mind going blank were frequent occurrences, as they were both strongly related to social situations in which the individual did not know how to converse with other people, this topic being a common topic of discussion amongst the messages.

However, there were messages that described symptoms not covered clearly within the six categories that had been defined, such as experiencing shaky hands, inability to maintain eye contact, general shyness, nausea and dizziness, panic attacks, and extensive worry about other anxiety-related thoughts and behaviors. While there were some that could fit into the defined symptom categories, some could not qualify for any, despite the display of behavior/content of the message clearly showcasing anxiety or being a result of anxiety. Additionally, many messages giving advice to others expressing concerns over their anxiety could not properly be accounted for, as they did not provide a clear understanding of how the individual experienced their symptoms of anxiety, even if they did reference their own personal experiences; furthermore, these messages were very general and mainly provided emotional support.

A number of messages spoke about medication use and feelings surrounding different brands of prescription medications, as well as feelings of depression that occurred alongside symptoms of anxiety. Many of these messages could not properly be categorized within our defined categories due to the symptoms being more aligned with depression rather than anxiety in the provided context. Medication use also appeared to not be related to our categories as the messages only vaguely mentioned increasing or decreasing dosage or generally if the medication did or did not help with managing symptoms. While a majority of messages talked about general anxiety, a vast portion of them concerned social anxiety, and our categories of symptoms were not equipped to properly assess symptoms related specifically to social anxiety.

Finally, an intensity scale 1-5 (\*0 none) was assigned to curate the level of anxiety based on the severity of the symptoms as in specified in Table 2.

Table 2: Severity of GAD based on symptoms

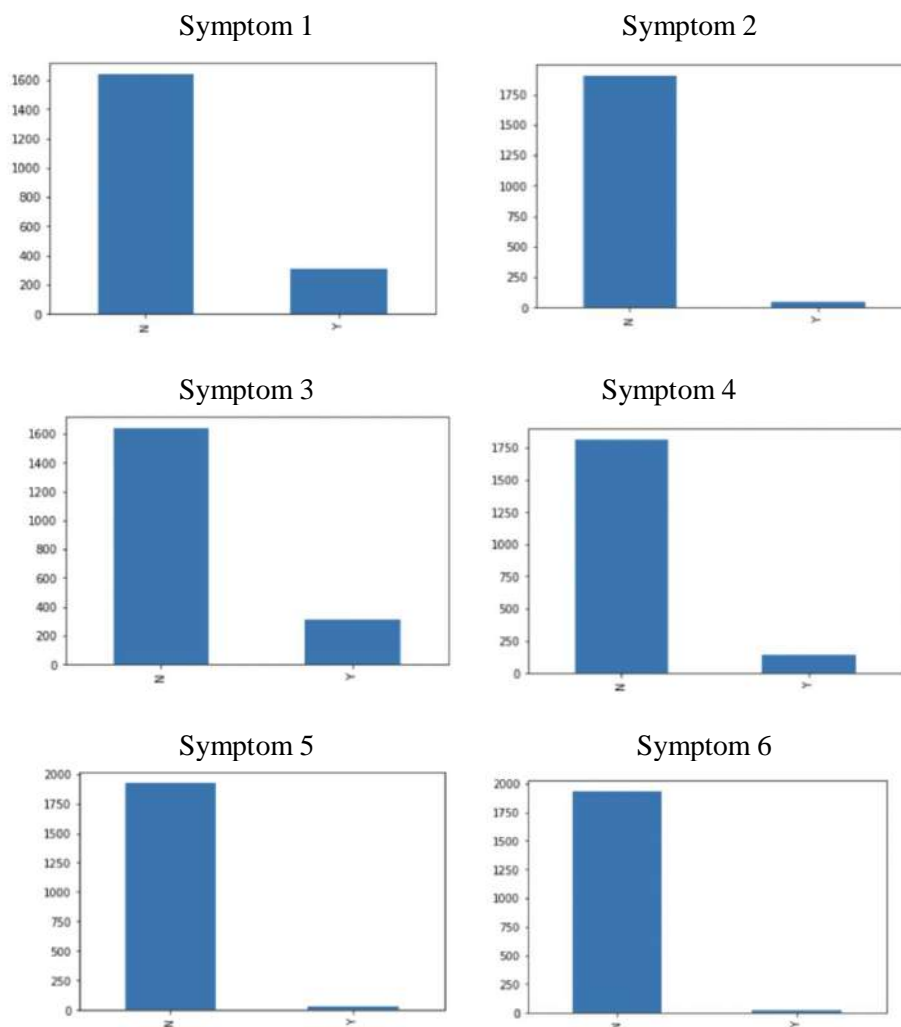
Intensity Level	%
1	11.6
2	3.85
3	1.54
4	0.62
5	0.15
0*	82.2

## 5.2. Data Analysis

The Discord dataset has a majority of the sentences which do not have any symptoms as shown in Table 2. But symptom #1 (restlessness) is more prominent amongst the comments where there is some symptom. The distribution of the records when the symptoms are exhibited is shown in Figure 1.

Figure 1: Symptoms distribution

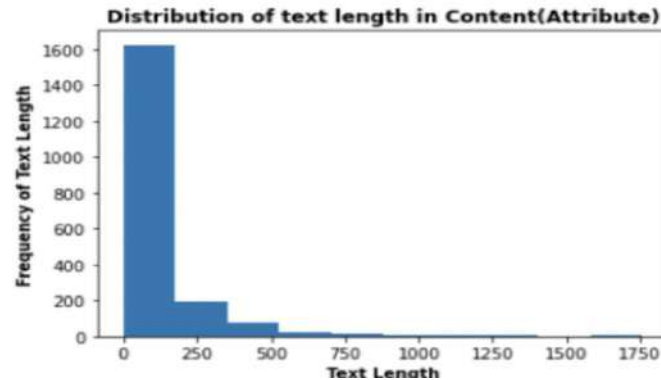
- 1: restlessness or feeling on edge
- 2: being easily fatigued
- 3: difficulty concentrating or the mind going blank
- 4: irritability
- 5: muscle tension
- 6: sleep disturbance
- \*7: None



Thus, restlessness is the most prominent symptom followed by difficulty concentrating. Sleep disturbance is the least prominent symptom in this dataset. This seems consistent with the general observations in patients suffering GAD and leads us to believe our dataset from discord has merits in diagnosing GAD.

Figure 2 shows that lengths of text in each labelled user comment is roughly 0-500 (ignoring long tail). So, the comments are fairly long which gives us confidence in the expressivity of the users in their deliberation on the depression topic in Discord.

Figure 2: Distribution of text length in Content



The plot below (in Figure 3) shows top N words after removing the stop words vs number of occurrences of each word. As seen in the plot, the top words now are 'anxiety, feel, talk, not' with high frequency. The word 'anxiety' has a frequency close to 276 and negativity expressed by 'not' is amongst the top. This is again indicative that this is a good dataset to develop GAD models.

Figure 3. Distribution of Top N words (after removing stop words)

```
from collections import Counter
Counter(" ".join(Anxiety["Content"]).split()).most_common(30)

[('not', 391),
 ('like', 326),
 ('anxiety', 276),
 ('people', 261),
 ('feel', 234),
 ('know', 212),
 ('social', 194),
 ('get', 175),
 ('anyone', 151),
 ('talk', 144),
 ('really', 144),
```

Figure 4 shows the word cloud of top frequent words after removing stop words from users' comments. We see that 'people' occurred more followed by 'feel,' 'anxiety,' 'help,' and 'anyone'. These indicate that the users want to discuss depression and are looking for help.

Figure 4: Word Cloud showing top words after lemmatizing and stopword processing





## 6. MODELING

### 6.1. Auto ML

Pyxeda Navigator [9], as in Figure 5, is an Auto Machine Learning (AutoML) tool that uses the Amazon Web Services cloud. Using this tool, a basic model was created. First, “Create an AI service” was chosen.

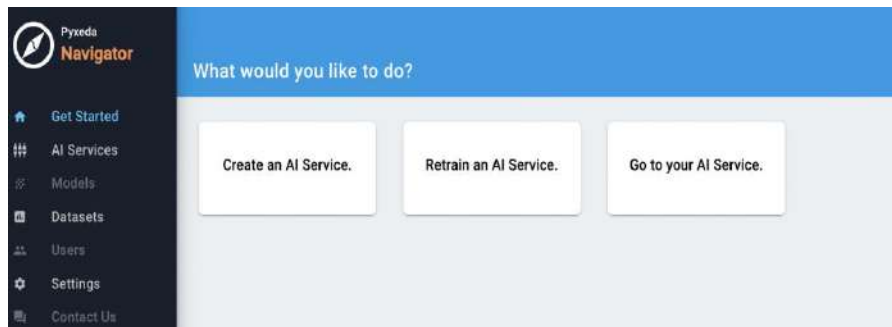
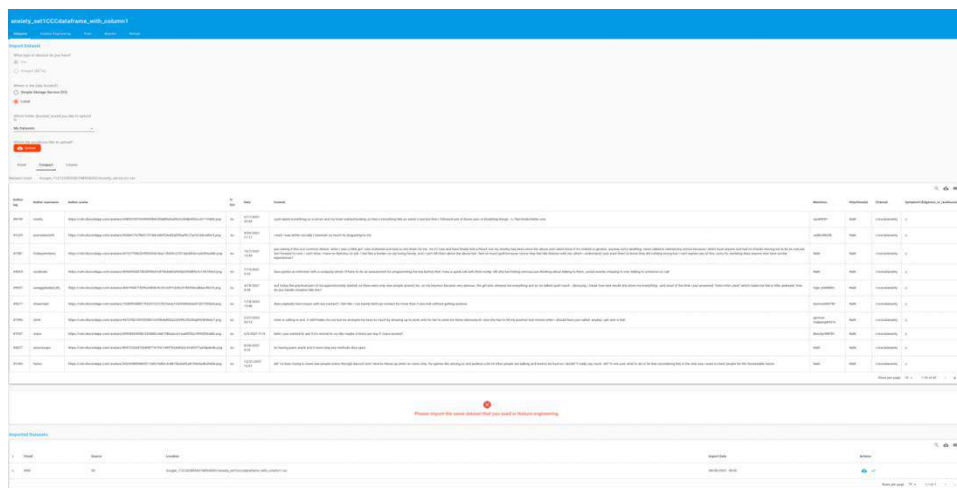


Figure 5: Home screen of Pxyeda Navigator

A CSV file “Anxiety set” was then preprocessed as in Figure 6. Using pandas coding, data frames were created for each of the symptoms, excluding the “Anxiety Intensity Scale.” Then one of the resulting CSV files with the feature “Corrected\_Content” and label “Symptom1(Edginess\_or\_restlessness)” was uploaded and stored on the navigator.

Figure 6: CSV file “Anxiety\_set” uploaded onto the AutoML tool



Feature Engineering was then done automatically as in Figure 7. Data was formatted and cleaned, and the relevant label, “Symptom1(Edginess\_or\_restlessness)” was selected.

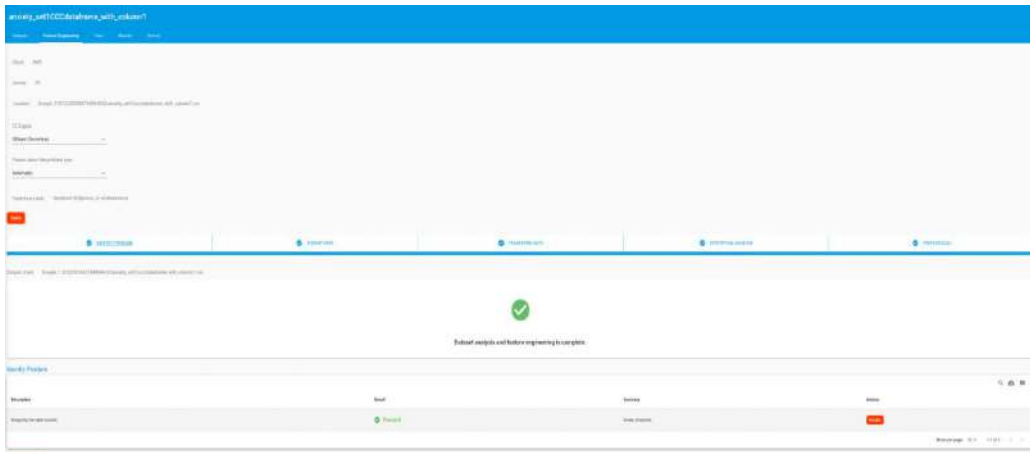


Figure 7: Feature Engineering in Pyxeda (Navigator)

Next, training occurred using Random Forest Classifier, MLP Classifier, and Logistic Regression as in Figure 8.

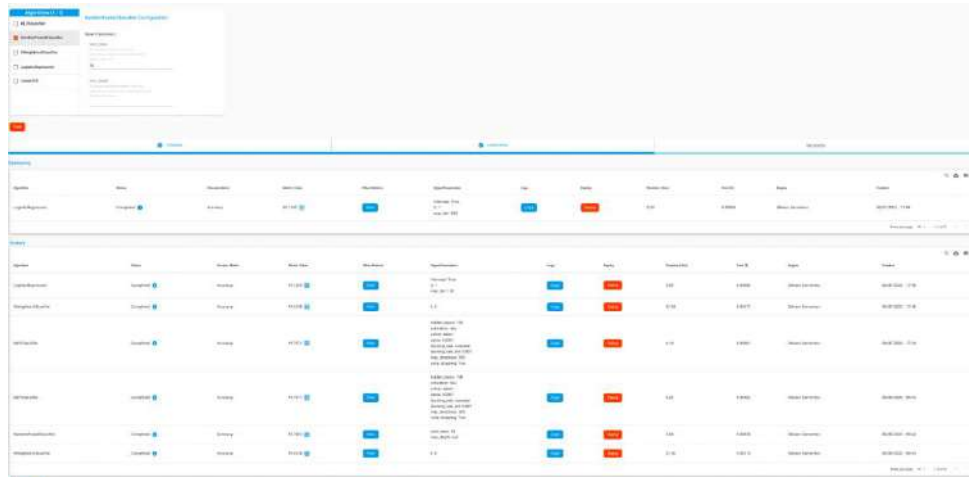


Figure 8: Training in Pyxeda (Navigator)

Source Code was then generated as Figure 10. Then, the methodology included creating a separate validation data set of 300 records of unseen data for testing the models. Finally, the source code was modified to test on unseen data to get final results.



Figure 9: Generate Code button

## 7. RESULTS

### 7.1. Auto ML Results

The study evaluates Auto ML as in Table 4 (<https://github.com/sneka/anxiety>) based on accuracy precision, recall and F1 score. We focus on these metrics as our dataset consists of potential GAD patients. So, it's more important for us to use precision as this model can serve as a guiding post. Tables 3, 4, and 5 shows the resulting metrics for the AutoML generated models using the algorithms Random Forest Classifier, MLP Classifier, and Logistic Regression. The algorithms were tuned with max\_text\_features set to 20, 200, 2000, and 5000.

Symptom with max_text_features determined for best output	Accuracy	Precision	Recall	F1 Score
Symptom 1 (5000)	75.0	0.789	0.174	0.286
Symptom 2 (5000)	95.333	1.000	0.125	0.222
Symptom 3 (2000)	84.333	1.000	0.041	0.078
Symptom 4 (2000)	85.33	0.500	0.023	0.043
Symptom 5 (5000)	98.33	0.000	0.000	0.000
Symptom 6 (5000)	98.0	1.000	0.250	0.400
Symptom 7 (200)	75.67	0.784	0.874	.827

Table 3: Metrics for the Random Forest Classifier model

Symptom with max_text_features determined for best output	Accuracy	Precision	Recall	F1 Score
Symptom 1 (200)	76.33	0.632	0.419	0.503
Symptom 2(200)	94.66	0.500	0.250	0.333
Symptom 3(2000)	84.67	0.600	0.184	0.281
Symptom 4(200)	78.33	0.216	0.182	0.198

Symptom 5(5000)	98.33	0.000	0.000	0.000
Symptom 6(5000)	98.0	1.000	0.250	0.400
Symptom 7(2000)	72.0	0.731	0.915	0.812

Table 4: Metrics for the MLP Classifier model

Table 5: Metrics for the Logistic Regression Model

Symptom with max_text_features determined for best output	Accuracy	Precision	Recall	F1 Score
Symptom 1 (200)	76.33	0.632	0.419	0.503
Symptom 2(5000)	95.33	1.000	0.125	0.222
Symptom 3(5000)	84.67	0.667	0.122	0.207
Symptom 4(200)	79.0	0.268	0.250	0.259
Symptom 5(5000)	98.0	0.000	0.000	0.000
Symptom 6(5000)	98.0	1.000	0.250	0.400
Symptom 7(200)	75.33	0.808	0.824	0.816

\*Symptom 5's low positive samples affected its metrics as shown by the Table 3, 4, and 5.

## 7.2. Deep Learning Model Results

As pointed by [8], AutoML has its challenges. This study looks beyond into Artificial Neural Networks to evaluate them as compared to Auto ML models. Notably, the study evaluates Feed forward networks and convolution neural networks.

### Feed forward ANN

Keras is an open-source Python framework used for creating and analyzing deep learning models. It is part of the TensorFlow library and allows us to define and train neural network models. After loading the dataset, we split the data into input (X) and output (y) variables and then create a Sequential model and add layers to our network architecture. Fully connected layers are defined using the Dense class. One can specify the number of neurons or nodes in the layer as the first argument and the activation function using the activation argument. Also, one can use the rectified linear unit activation function referred to as 'relu' on the first two layers and the Sigmoid function in the output layer. By using a sigmoid on the output layer, one can easily transfer our network output to a probability of class 1, or, with a default threshold of 0.5, snap to a hard classification of either class. After adding the layers, one can compile the model because it has been specified. For training and producing predictions on our hardware, such as CPU, GPU, or even distributed, the backend automatically determines the appropriate method to represent the network. There are a few more characteristics that must be specified during compilation in order

to train the network. Keeping this in mind, we can determine the optimal set of weights to translate our dataset's inputs to outputs while training a network.

The study used cross entropy as the loss justification. This loss, known in Keras as "binary\_crossentropy". The study uses the effective stochastic gradient descent method "adam" to define the optimizer. This variant of gradient descent is well-liked since it automatically fine-tunes itself and produces effective solutions to a variety of issues. The classification accuracy described by the metrics argument will be collected and reported because it is a classification problem. Now we build the model with 85% training data. By using the fit() method on the model, one can train or fit our model using the loaded data. The training process runs for a fixed number of epochs (iterations) through the dataset that will be specified using the epochs argument. The study used 15% validation dataset to assess its performance. The evaluate() function returns loss, accuracy, f1, precision and recall for the validation dataset.

Symptom	Accuracy	Precision	Recall	F1 Score
Symptom 1	76	0.8	0.20	0.31
Symptom 2	93	0.10	0.10	0.10
Symptom 3	85	0.5	0.18	0.25
Symptom 4	85	0.43	0.18	0.24
Symptom 5*	95	0.000	0.00	0.00
Symptom 6*	94	0.0	0.00	0.0

Table 6: Feed forward metrics for GAD symptoms

\*Symptom 5's & 6's low positive samples affected its metrics as in Table 6.

### Convolutional neural network - CNN

In Keras, one may simply add the necessary layer one at a time to build up layers. The Sequential object's add method is then called to add layers. The layers themselves are examples of classes like Dense, which denotes a layer that is fully linked and uses a certain number of neurons with a certain activation function. The study adds a first convolutional layer using Conv1D (). The rectified linear unit activation function, often known as relu, was then to be used on the first layer. Next, were added the max-pooling layer with MaxPooling1D() and so on. The last layer is a dense layer that signifies sigmoid activation. After the model is created, it was compiled using Adam optimizer, one of the most popular optimization algorithms, and subsequently used cross entropy as the loss justification. This loss, known in Keras as "binary\_crossentropy". Similar methods like previous sections on 15% validation dataset were used to generate accuracy, f1, precision and recall.

Symptom	Accuracy	Precision	Recall	F1 Score
Symptom 1	76	0.8	0.20	0.32
Symptom 2	96	0.4	0.25	0.29
Symptom 3	<b>87</b>	<b>0.8</b>	0.29	<b>0.41</b>
Symptom 4	<b>92</b>	<b>0.9</b>	0.51	<b>0.62</b>
Symptom 5	98	0.1	0.10	0.10
Symptom 6	99	0.5	0.45	0.46

Table 7: CNN metrics for GAD symptoms

As in Table 6 basic Feed forward Neural networks performance is not way superior to Auto ML, but as in Table 7 when more sophisticated Convolutional Neural networks were adopted, symptoms 3's 4's & 6's detection was well above the Auto ML capabilities with F1 score north of 0.4+. Thus additional sophistication gets to detect more difficult to detect symptoms.

For the final GAD prediction, we used the regressor model stacked on top of previous symptom model using Random Forest regressor to get the predicted GAD intensity. The ground truth symptom labels were regressed as well using the same model to compare its efficacy over the predicted symptom label's ones. A 19% discrepancy was observed in MAE as in Table 8 which indicates that we can predict GAD intensity using symptom models within 80% error rates.

Table 8: Anxiety intensity model

Model	MAE diff w. Ground Truth
<b>RF Regressor</b>	0.19

## 8. CONCLUSION

This research recommends an Auto ML approach to predict General Anxiety symptoms to get a complete picture of depression. Such methods are simpler and easy to use by less technical IT folks in Hospitals. Social media conversations, especially in Discord chat engine, can help fuel and seed an accurate analysis of GAD symptoms. The multi-symptom's AutoML random forest model predicts GAD with 50+% precision (except 5) and is at least 75+% accurate including the most prevalent symptom of restlessness. This could be very useful in diagnosing GAD by medical professionals using pre diagnosis chats as recommended by DSM. The AutoML technology gives quick and gives decent performance, but it can be improved upon by more sophisticated ANN methods like Convolution neural networks that plug rest of AutoML's Symptom's deficiencies with at least 80+% precision and 0.4+% in F1 score, namely in detecting concentration and irritability lapses symptoms.

## REFERENCES

- [1] Stein, Murray B.&Sareen,Jitender, (2015) "Generalized anxiety disorder",*New England Journal of Medicine*, Vol. 373, No. 21, pp2059-2068.
- [2] Gale, Christopher K & Mark Oakley-Browne (2003) "Generalized anxiety disorder." *American family physician*, Vol. 67, No. 1, pp135-8.

- [3] Pintelas, Emmanuel &Kotsilieris, Theodore &Livieris, Ioannis&Pintelas, P, (2018) “A review of machine learning prediction methods for anxiety disorders”,doi: 10.1145/3218585.3218587.
- [4] Arif, M., *et al.* (2020) “Classification of anxiety disorders using machine learning methods: a literature review”,*Insights of Biomedical Research*, Vol. 4, No. 1, pp95-110.
- [5] Zulfiker, Md Sabab, *et al.* (2021)“An in-depth analysis of machine learning approaches to predict depression”,*Current research in behavioral sciences*,Vol.2, No.100044.
- [6] Sau, Arkaprabha&Ishita Bhakta (2017) "Predicting anxiety and depression in elderly patients using machine learning technology",*Healthcare Technology Letters*,Vol. 4, No. 6, pp238-243.
- [7] Bell CC,(1994) “DSM-IV: Diagnostic and Statistical Manual of Mental Disorders”, *JAMA*, Vol. 272, No. 10, pp828–829, doi:10.1001/jama.1994.03520100096046.
- [8] Verma, Anirudh&Tyagi, Shashikant &Mathur, Gauri, (2021) "A Comprehensive Review on Bot-Discord Bot", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, Vol. 7, No. 2, pp532-536.
- [9] Talagala,Nisha (2022)*Pydeda*, Retrieved from <https://www.pyxeda.ai>.
- [10] Daly, Michael &Robinson,Eric, (2022) "Depression and anxiety during COVID-19",*The Lancet*, Vol. 399, No. 10324, p518.
- [11] M., Chatterjee *et al.*, (2014) "Context-based signal descriptors of heart-rate variability for anxiety assessment",*2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3631-3635, doi: 10.1109/ICASSP.2014.6854278.
- [12] Chen, Yilin, *et al.*, (2022) "Machine learning classification model using Weibo users' social appearance anxiety",*Personality and Individual Differences*,Vol. 188, No. 111449.
- [13] Dabek, Filip &Caban, Jesus J, (2015) "A neural network based model for predicting psychological conditions",*International conference on brain informatics and health. SpringerLink*, Cham.
- [14] Katsis, Christos D., Katertsidis,Nikolaos S.&Fotiadis,Dimitrios I, (2011) "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders",*Biomedical Signal Processing and Control*,Vol. 6 No. 3, pp261-268.
- [15] Hilbert, Kevin, *et al.*,(2017) "Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: A multimodal machine learning study",*Brain and behaviour*,Vol. 7, No. 3e00633.
- [16] Sribala, M., (2015) "An approach of artificial neural networks for prediction of generalized anxiety disorder",*International Journal of Research in Computer Applications and Robotics*,Vol. 3, No. 3, pp118-124.
- [17] Hussain-Shamsy, Neesha, *et al.*,(2020)"Mobile health for perinatal depression and anxiety: Scoping review",*Journal of medical Internet research*,Vol. 22, No. 4e17011.

## AUTHORS

**Anishka Duvvuri** is a high schooler who interned for this project at Jeeva Health. She is familiar with using AutoML tools by the AiClub.world organization. She strives to continue engaging with AI, especially for healthcare. She goes to BASIS Silicon Valley High school. Her passion is playing table tennis and teaching AI in her school.



**Naavya Kovuri** is currently working as a Senior Data Engineer at Amtrak. She received her Master’s degree in Data Analytics from Northeastern University this May. She also worked as a Data Analyst at Bluestar for 3 years and before joining Bluestar she graduated in 2016 with a Bachelor’s degree in Computer Science. She has a great passion for Data Science and for developing Machine Learning Models.



**Sneha Kumar** is an experienced analytics Professional, currently working as Operations Analyst at IronRidge and recently graduated analytics degree from Northeastern University. She loves working with data & statistics and that is her passion towards career. She is focused and committed to deliver the best possible solution to any business problems.



**Rebecca Victor** is currently pursuing a Master's degree in psychology at Pepperdine University, California. She recently graduated from University of California, Merced with a Bachelor's degree in Psychology earning Graduation Honors (GPA 3.79). While at University of California, Rebecca aided the research of experimental protocols and methodology in studies such as "Eye Movements and Abstract Thinking", "Validating EEG Source Localization with fNIRS", and "Adults' Understanding of Social Categories". An excellent communicator and innovator, Rebecca is passionate about influencing health care policy by exploring and measuring mechanisms behind health behaviors to promote good health and prevent illness.



**Tanush Kaushik** is a rising senior at a suburban high school near Boston, with an active interest in Python as a Data Science & Analysis tool for diving into his passions of technology and digital health. He is also certified in Machine Learning at Stanford utilizing Unsupervised Learning use cases. With COVID-19's advent, he has taken a particular interest in use of technology and tools like data science, NLP, ML to develop an understanding of the intricacies of causes, symptoms, and prevention-techniques for various mental health issues.







# A FIRST-PERSON SHOOTER GAME DESIGNED TO EDUCATE AND AID THE PLAYER MOVEMENT IMPLEMENTATION

Chunhei Zhu<sup>1</sup>, Yujia Zhang<sup>2</sup>

<sup>1</sup>Beckman High School, 3588 Bryan Ave, Irvine, CA 92602

<sup>2</sup>University of California Irvine, Irvine, CA 92697

## **ABSTRACT**

*The issue of finding a clean and simple player movement implementation that the general public will find intuitive and easy to use has been tackled over the years in various ways. With FPS (first-person shooter) games, the need for a simple and fun style of movement is monumentally crucial, as that will be a core aspect of the gameplay [4]. To address this issue, an FPS game was created with the ability to maintain momentum while crouching with intention of providing a smoother and more intuitive gaming experience for players. This movement implementation was tested by having participants play the game for a sufficient amount of time, then asking the participants to rate the experience of movement in the game and the overall enjoyment of playing the game. The results indicate that the implemented movement would be well-received by the general public, as the vast majority of the participants viewed the new form of movement as a welcome feature based on the optional feedback and the quantitative ratings. However, the other aspects of the gameplay were not as polished and therefore lowered the overall enjoyment of the game for the participants, particularly the shooting in the game that does not yet have proper audio or visual cues to let the player know that the weapon has been fired.*

## **KEYWORDS**

*FPS Game, Player Movement, Unity, Artificial Intelligence*

## **1. INTRODUCTION**

First-person shooter games are a genre of shooter game generally based around guns or ranged weapons, in which users would control the in-game character in a three-dimensional space (3-D) [5]. Many FPS games' goals are similar; multiplayer games usually either involve different teams fighting against each other in a player versus player setting (PvP) or players working together to defeat a common enemy in a player versus environment setting (PvE), whereas single player games tend to involve elements of exploration and adventure [6] [7]. The history of first-person shooter games started in 1973 when the first FPS titled "Maze War" was developed. However, it is believed that the concept of FPS games was solidified in 1992 with the release of the game "Wolfenstein 3D" [10]. First-person shooters rely on the first-person point of view, which is through the eyes of the character in the game. A benefit to having a game be an FPS game is that it can be more realistic and/or immersive since the game is in a 3-D environment. People can enjoy the first-person perspective and feel like they are in the game by controlling the character in the game from this first-person point of view, which would not be possible if they were playing a game from a third-person perspective. Most first-person shooter games have some realistic factors, and an example is walking slower after getting damaged by an enemy. First-person shooter games are significant for their massive popularity and their large impact on the gaming

industry as a whole. FPS games are relatively simple to get into since almost all FPS games have the same keys to control the character and move around. Furthermore, many multiplayer games have a competitive aspect, which may draw some players in. The aforementioned immersion is also what some players like to experience when playing FPS games [8].

Many FPS games currently exist, and they have been a popular genre of game. An example of an FPS game is Valorant, which is an online multiplayer game that pits a team of five against another team of five. However, some potential issues exist within these games, specifically with the movement of the player. In Valorant, the player slows down while crouching, which is quite common. However, while this game development decision makes sense from a logical standpoint and keeps a tradition that was passed on from previous FPS games, such an implementation may come off as jarring and limiting to some players. Some players prefer to maintain their momentum in the game whether they run normally or crouch, which can be frustrating and disappointing when a game is not designed to do that. Another issue that Valorant has is that as an online multiplayer game, a poor connection could result in an unenjoyable experience. Movements and shots may not register, and what the player sees may not be up to date with what is actually occurring in the game.

Therefore, only those with stable and fast internet connections will be able to enjoy the smoothest gaming experience. does not suffer from the issue. The final drawback of this game also has to do with the nature of multiplayer games, which is the toxic environment that players can potentially be placed in when interacting with both voice chat and typed chat. Disgruntled players may insult the opposing team or even their own teammates in frustration due to losing the game. Those who play the game with this type of player will generally not be able to enjoy the game as much, as encountering such players creates a stressful and uncomfortable situation that may leave others in a bad mood long after finishing the game.

An FPS game created in Unity was the tool used to address the issue of player movement in other video games. In the game, the player can move around a map that resembles a city with houses and buildings and shoot enemies with a gun. The player can jump on top of the buildings, and the player can destroy the enemies once enough the enemy takes enough shots from the player. The enemies will navigate toward the player and attempt to walk into the player. If the player stays in contact with an enemy for long enough, the player loses and has the option to either restart or quit the game. One important implementation in the game that is intended to enhance the player's movement is the maintaining of player momentum, whether the player is running or crouching [9]. As many other games slow down the player while the player crouches, this game takes a different approach in hope that the player has a more pleasant experience operating the character. The player also has a high movement speed, which can allow for more control. As this is a single-player game that does not require any internet whatsoever to play, a poor internet connection will not hinder the players' enjoyment of the game.

The experiment that was chosen to test the Unity game was a survey. Several participants were selected to play the game for at least five minutes. Then, they were asked to fill out a Google Forms survey that asked two questions. The first question was "How do you rate the player movement of the game?", and the second question was "How do you rate the overall enjoyment of the game?". Each of the questions would provide the participant with a scale from one to ten to choose their rating. After answering these two questions, the participants would also have the opportunity to fill out an optional free-response section that asked the users if they had any other feedback to provide. This would offer participants the ability to express their thoughts in a way that would not be possible by only using the previous two questions.

With this experiment, a high overall score regarding the player movement indicates that

controlling the character was a smooth and intuitive experience. A high overall score regarding the general enjoyment of the game could also indicate that members of the general public would be willing to play more games like this in the future. By taking these two scores, we can also identify any possible issues with the game. For example, if the player movement was rated highly and the overall enjoyment was rated much lower, this could imply that other areas in the game are lacking. On the other hand, if the player movement was rated much lower than the overall enjoyment, the general public might prefer a different player movement implementation.

The remainder of the paper is split into five sections, which will be labeled 2 through 6. Section 2 details the difficulties faced when trying to design and implement the game as well as come up with an experiment to test the game. Section 3 explains how the general implementation of the Unity game was done, and more details are provided on particular aspects of the game such as the player movement. Section 4 brings up how effective the Unity game is at providing its players with an enjoyable gaming experience by surveying participants on the intuitiveness of the player movement and the overall enjoyment of the game, and Section 5 introduces a few related works and states how they compare to this work. Section 6 provides a conclusion that includes a summary of the application, some current limitations to the application, and what can be done in the future to resolve these limitations.

## **2. CHALLENGES**

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

### **2.1. Thinking ideas of the setting and the Design for the Game**

There were many challenges while building the game, and one major challenge was thinking of ideas for the game, specifically with the setting of the game. There are lots of elements in the game such as the movement, the map, and the game's overall experience that should be taken into consideration. First, the character in the game was created using 3D objects in Unity. Then, after brainstorming how the map would look and what the scene should be based on, a city was selected to be the environment in which the player would be placed. To build the city environment, city buildings were selected from the Unity Assets Store to be imported into the project. Creating the map in the game was a lengthy process, as the imported building objects had to be placed in a specific manner to resemble a modern city in the real world. Lastly, the buildings in the city were colored with different textures. Besides the settings, there are many different smaller aspects in the game that had to be considered in the design of the project.

### **2.2. Implementing Health into the FPS game**

The second challenge that was faced when making the FPS game was implementing health into the shooter game. First, both the player and the enemies needed to have health so that they would die after they are attacked enough times. A script was written that determined how much health a player and each of the enemies would have. However, with the implementation of the game, the enemies would not have any weapons and would not be able to deal damage with far-range attacks. Therefore, the only way that the enemies could deal damage was by walking into them. The enemies would have a certain amount of health and would die when the character used the gun to shoot the enemy in the head once. The user's screen would have a dotted crosshair and when left clicked, the gun would shoot and subtract the health from the enemy. An end screen was later designed in which if the character dies, a screen would appear showing the words "Game Over" and a quit button to close the game.

### 2.3. Implementing Shooting in the FPS Game

The last challenge that was faced was implementing shooting in the shooter game. A model of the gun first had to be chosen from the Unity Assets store, then the gun was imported into the game. Then, a crosshair was created in the character's first-person point of view, and a good angle was carefully chosen to place the gun in the character's hand. A script was written to control what will happen when the gun shoots. The gun was designed so that when left clicked, the gun would fire and do damage to enemies if the crosshairs are placed directly on the enemy. The gun will take away one health from the enemy if it hits the enemy, while the character will also lose health if the enemy touches him. The shooting was implemented into the game in the code by returning a Boolean value after a target has been successfully hit.

## 3. SOLUTION

The game is created using the Unity engine, and the game has three main screens. The first one is the main menu screen, which provides the player with two buttons; one button says "Play" and brings the player into the game screen, while the other button says "Quit" and closes the game when pressed. Once brought into the game, the player will spawn into a city scene, and the user is equipped with a gun. The player can find enemies that spawn around the map and pursue the player. The player can shoot the enemies in the head to defeat them and make them disappear from the map. However, if the player lets the enemies come into contact with their character for too long, the player will lose and the game over screen will appear. From here, the player can choose to either play again using the respawn button or quit the game using the quit button.

Objects were a major part of the Unity game and are responsible for much of the game implementation. The city buildings, the enemies and the player's character, and even the buttons on the main menu and game over screen would not exist without using Game Objects in Unity. To provide the functionality to the game, C# scripts were written and attached to certain objects in the game. [12] These scripts would be able to control various aspects of the game, from the movement of the character to the implementation of shooting and player/enemy health.

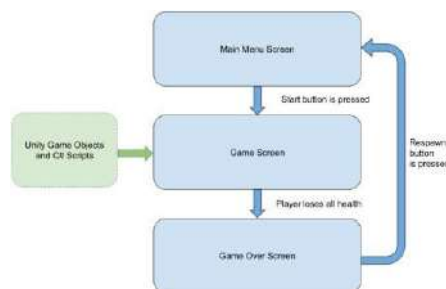


Figure 1. Overview of the solution

The shooting in the game was implemented by using Unity Raycast, which is the Raycast system where data is returned after a target is successfully hit by a ray. With the shooting implementation, a point was added to the screen as a crosshair, in which the crosshair is a ray in the system. In a C# script, Unity would first check whether the button to fire the weapon (left mouse click) was being pressed and whether enough time has elapsed between now and the last time that the player fired the weapon using if statements. The way that this was implemented was by creating variables called nextFire and fireRate. The fireRate was a set number, while nextFire would add the current time and the fireRate to determine when the next time that the player could fire was. By doing so, a hard limit could be set on how often the player could fire the weapon. An

implementation like this could prevent people from having an advantage from clicking faster and incentivize them to click as rapidly as possible. Without such a limit, some players may use auto-clickers to cheat; an auto-clicker is a form of computer software that can automate the process of clicking a mouse, and many of them have settings for the users to freely adjust how quickly the clicks occur. Alternatively, other people who play legitimately may suffer carpal tunnel syndrome or other injuries or strains related to the fingers and hands by clicking too much and too quickly. Therefore, this fixed firing rate system can allow people to play on equal grounds. After the object has been hit, the data would be stored in a Raycast variable and could later be used to determine how much health they have remaining.

The Raycast shoot system, which is done in the Update method, involves first getting the data back to see if an enemy had been hit by checking whether the collision involved an element with the tag of "Target". [11] If an enemy was shot in the head, the enemy would disappear by setting the game object to inactive, as the enemies in the game only have enough health to be defeated in one shot. On the other hand, the player has 100 health and loses 10 health with every collision with an enemy, enough to endure 10 collisions. If the character reaches 0 health, the game will end and the game over screen would be displayed. The value returned can later be used in the health system to determine the health of the enemies. Enemies were coded to spawn randomly around the map so the player wouldn't face lots of enemies in one direction at once.

The C# script responsible for enemy spawning was implemented to have one enemy spawn every ten seconds. A pause menu was later created in the game where the user can click "ESC" to pause the game. The get key function was used to check if the user had clicked the pause button, then the time scale function would play a role in pausing the game, because the time scale would change to 0 if the game is paused and to 1 when the user exits out of the pause menu. The pause menu has a quit button, and if the user clicks the button, the game would be exited. Player movement was added to the game by using a MonoBehaviour class. First, a collision script was made where players wouldn't fall through the ground when they are walking on it. Then the sensitivity was set for the player's walk speed, player run speed, player jump height, and crouch height. The movement keys are set similarly to other games, "WASD" to move around and space to jump. In many current FPS games, the movement speed of the player is brought to only a fraction of the original movement speed when crouching (which can be done by multiplying the player's speed by a certain amount), as this mimics how quickly a person would move when crouching in real life and allows for more precise movement. However, as this game does not require precision in character movement, the implementation of crouching involved the movement speed staying consistent and uninterrupted, which may create smoother player movement.

```
if (Input.GetButtonDown("Fire1") && Time.time > nextFire)
{
    nextFire = Time.time + fireRate;

    Vector3 rayOrigin = fpsCam.ViewportToWorldPoint(new Vector3(0.5f, 0.5f, 0.0f));
    RaycastHit hit;

    if (Physics.Raycast(rayOrigin, fpsCam.transform.forward, out hit, weaponRange))
    {
        if(hit.collider.CompareTag("Target"))
        {
            hit.collider.gameObject.SetActive(false);
        }
    }
}
```

```
public class EnemySpawn : MonoBehaviour
{
    public GameObject enemyPrefab;
    private float spawnTime = 10.0f;

    void Start()
    {
        StartCoroutine(Spawn());
    }

    public IEnumerator Spawn()
    {
        Instantiate(enemyPrefab);
        //enemy.transform.position = this.transform.position;
        yield return new WaitForSeconds(spawnTime);
        StartCoroutine(Spawn());
    }
}
```

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

[RequireComponent(typeof(Collider), typeof(Rigidbody))]
public class PlayerMovement : MonoBehaviour
{
    //Look
    public float sensitivity = 1f;
    public float smoothing = 2f;

    private Transform charCamera;
    private Vector2 mouseLook;
    private Vector2 mouseDelta;

    //Move
    public float walkSpeed = 5f;
    public float runSpeed = 10f;
    public KeyCode runKey = KeyCode.LeftShift;

    private Rigidbody rb;

    //Jump
    public float jump = 5f;
    public KeyCode jumpKey = KeyCode.Space;

    private bool isGrounded = true;

    //Crouch
    public float crouch = 0.25f;
    public KeyCode crouchKey = KeyCode.LeftControl;

    private float normalYposition = 1f;

    void Start()
    {
        //Look
        Cursor.lockState = CursorLockMode.Locked;
        Cursor.visible = false;
        charCamera = Camera.main.transform;

        //Move
        rb = GetComponent<Rigidbody>();

        //Crouch
        normalYposition = rb.transform.localScale.y;
    }
}
```

Figure 2. Screenshots of the FPS game's code







Figure 3. Screenshots of the FPS game

## 4. Experiment

### 4.1. Experiment 1

The experiment that was selected to test the effectiveness of the implementation of player movement in the game as well as the effectiveness of the game as a whole at providing players with enjoyment and entertainment is a Google Forms survey [14]. First, fifteen participants were individually given a ZIP folder containing the first-person shooter game. After unzipping the folder and opening the game, they would play the game for at least five minutes. Before playing the game, each participant was given the message “Try to test the movement of the game as much as you would like.” Because the participants were given the same version of the game and the same message, this reduces any confounding variables in the experiment. The first question that the survey asks is “How do you rate the player movement of the game?” and a scale from one to ten is provided for the participants to answer.

Participant Number	Rating of Player Movement
1	9
2	10
3	7
4	6
5	8
6	8
7	9
8	7
9	9
10	10
11	10
12	7
13	9
14	8
15	7

Figure 4. Table of participants

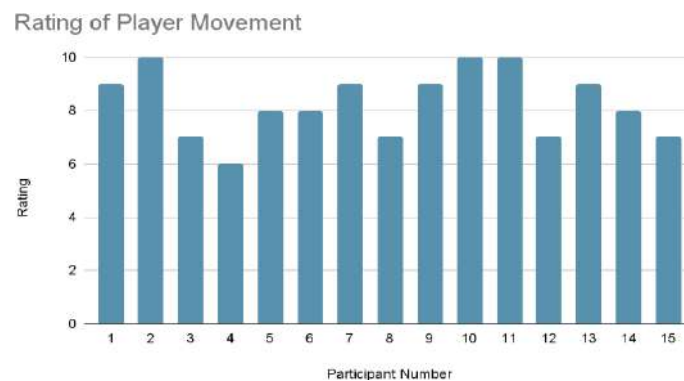


Figure 5. Rating of Player Movement

According to the results, it appears that the participants rated the player movement in the FPS game relatively high overall. The lowest rating of 6 was only given by one participant, while the highest rating of 10 was given by three participants. These ratings could indicate that participants found the implementation of the game's player movement to be done well. This idea is backed up further by the optional feedback that some of the participants provided. Participants stated that the player movement seemed smooth, intuitive, and uninterrupted when they played the game. In particular, a few participants pointed out that crouching while moving would not reduce the momentum, and this was a feature that they gladly welcomed. However, one of the participants felt the opposite and believed that since the player movement was implemented so much differently from other traditional modern FPS games, not slowing down when crouching was jarring.

## 4.2. Experiment 2

The second question in the survey is "How do you rate the overall enjoyment of the game?". To keep the survey consistent, a scale from 1 to 10 was also provided for the user to answer. At the end of the Google Forms survey, an optional free-response section was provided to the participants. In case they have any feedback that would not be possible to provide from the two previous questions or would like to expand upon something from the previous questions, they would be able to do so. Since there are fifteen participants, sample size is large enough to account for variability.

Participant Number	Rating of Player Movement
1	9
2	7
3	8
4	5
5	6
6	7
7	8
8	7
9	9
10	10
11	8
12	6
13	7
14	8
15	7

Figure 6. Table of participants

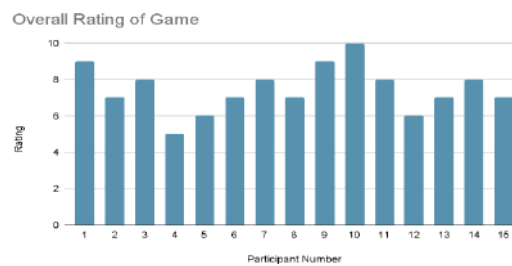


Figure 7. Overall Rating of Game

The responses to the second question were relatively positive. One participant gave the highest score of 10, while another participant provided the lowest score of 5. The results indicate that the participants were moderately pleased with the entertainment value of the game overall. A better explanation of the results can be seen in the feedback provided in the free-response section. While most participants stated that playing the game was fun, there were several issues that still had to be addressed. The largest issue among them was the shooting mechanic. While shooting in the game worked as normal and players could defeat enemies by shooting them enough times, there was no visual or audio cue to indicate that the player was shooting. Some participants stated that they thought the shooting mechanic did not work at all as a result. While the player movement in the game may be polished, the other features of the game did not match that same level of polish. In the first question that measured how well the player movement was implemented in the game, participants generally seemed to agree that the movement of the player was fun, intuitive, and somewhat unique to traditional FPS games. The second question that gauged the enjoyment of a game as a whole did not receive responses as positive as the previous one, since the participants found issues in quality in aspects of the game besides player movement that reduced the entertainment value of the game.

Taking both results into account, the game appears to be much more well-received with its player movement than as a whole. This was to be expected, as the player movement was one of the primary aspects of the game that was emphasized during its development. However, this could also indicate that the other areas in the game are lacking, such as the shooting mechanic or the enemy spawning mechanic.

## 5. RELATED WORK

A first-person shooter game was developed to study the effects of implicit and explicit biofeedback, where explicit biofeedback is consciously displayed and/or felt while implicit biofeedback is felt at the subconscious level. The study concluded that explicit biofeedback had a much more significant impact on the players than implicit biofeedback did and indicated that explicit biofeedback interaction may have a bright future in video games [1]. Both the study on biofeedback and this work focus on first-person shooter games as a primary topic. While the study on biofeedback had a stronger focus on the biosensors' data for results, this research places an emphasis mainly on the players' perceived experience regarding character movement in a newly developed FPS game.

Another study evaluates how high players of a first-person shooter game believe the quality of the game is and creates a model using ping, jitter, and packet loss values of players to predict the perceived quality of the game from other players. The model from this research indicates that the perceived quality of a game has a high correlation with reduced amounts of ping and jitter values [2]. While the study involving a perceived quality model focuses on how poor connection issues can impact the players' experience of a game, this work emphasizes adjustments in the implementation of player movement to enhance players' experience in FPS games.

A related work presents a study regarding how player performance and enjoyment in first-person shooter games are affected by frame rates. According to the results of the study, the number of frames per second that a player experiences greatly affects their performance in all aspects of the game (including both movement and shooting), but the improvement of the player becomes smaller the higher the number of frames per second becomes [3]. This related work is similar to our work in that the experience of players in FPS games is tested, but the related work emphasizes frame rate to do so while this work utilizes development choices regarding the player movement to do so.

## 6. CONCLUSIONS

Our method of finding a way to improve player movement in first-person shooter video games is creating a game that implemented player movement in a different manner. The most notable difference from traditional FPS games is that when crouching, the momentum of the player remains unchanged. In the game, the player is free to jump on top of buildings in a city and shoot large enemies to defeat them. The game is currently endless, meaning that the player's goal is to survive as long as possible and stop the enemies from walking into the player and dealing contact damage [13].

To test the effectiveness of the player movement in the game at maintaining player interest and providing an enjoyable gaming experience, an experiment was conducted in which fifteen participants were gathered to play the game and test the controls, then fill out a Google Forms survey that asked the participants how much they enjoyed the player movement and the game as a whole. From the results, the player movement was rated higher than the enjoyment of the overall game. While this indicates that the player movement implementation was done well, the other areas of the game do not seem to match that level of quality. The feedback also reinforced this notion, as the shooting mechanic appeared to be a feature that reduced the level of enjoyment from most participants who provided feedback. Nevertheless, the maintaining of momentum while crouching was widely well-received, which can indicate that such an implementation in future games may prove successful in the right circumstances, such as games that do not require precision in movement.

One of the biggest limitations is that the game currently does not provide an indicator of when the gun has been fired. The game can still register if the player shoots and hits enemies, and the enemies disappear once they have taken enough shots. However, a new player may not know this, and they may assume the game is bugged instead. If there is no visual or audio cue that indicates when the gun has been fired, the player may not have the information they need to make the correct decisions in the game. The lack of effects may also ruin the immersion of the game for some players.

An effective way to solve the issue is adding a sound effect whenever the gun has been fired. This will let the player know that the gun has been fired. Eventually, a visual effect like a flash can be used as well [15]. With more effects and polish to the game, the players can become more immersed.

## REFERENCES

- [1] Kuikkaniemi, Kai, et al. "The influence of implicit and explicit biofeedback in first-person shooter games." *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010.
- [2] Wattimena, A. F., et al. "Predicting the perceived quality of a first person shooter: the Quake IV G-model." *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*. 2006.
- [3] Claypool, Kaja T., and Mark Claypool. "On frame rate and player performance in first person shooter games." *Multimedia systems* 13.1 (2007): 3-17.
- [4] Voorhees, Gerald A., Joshua Call, and Katie Whitlock, eds. *Guns, grenades, and grunts: First-person shooter games*. Bloomsbury Publishing USA, 2012.
- [5] Hew, KheFoon, and Wing Sum Cheung. "Use of three-dimensional (3-D) immersive virtual worlds in K-12 and higher education settings: A review of the research." *British journal of educational technology* 41.1 (2010): 33-55.
- [6] Shafer, Daniel M. "Causes of state hostility and enjoyment in player versus player and player versus environment video games." *Journal of Communication* 62.4 (2012): 719-737.
- [7] Pirker, Johanna, et al. "Analyzing player networks in *Destiny*." *Entertainment Computing* 25 (2018): 71-83.
- [8] Jennett, Charlene, et al. "Measuring and defining the experience of immersion in games." *International journal of human-computer studies* 66.9 (2008): 641-661.
- [9] Strömberg, Hanna, Antti Vääänen, and Veli-Pekka Rätty. "A group game played in interactive virtual space: design and evaluation." *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*. 2002.
- [10] Breuer, Johannes, et al. *Violent video games and physical aggression: Evidence for a selection effect among adolescents*. Vol. 4. No. 4. Educational Publishing Foundation, 2015.
- [11] Niu, Hanlin, et al. "Accelerated sim-to-real deep reinforcement learning: Learning collision avoidance from human player." *2021 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2021.
- [12] Konzack, Lars. "Computer game criticism: A method for computer game analysis." *CGDC Conf.*. 2002.
- [13] Adiwikarta, Rendy, and Harya Bima Dirgantara. "Pengembangan permainan video endless running berbasis android menggunakan framework game development life cycle." *Indonesia: KALBIScientia, ISSN* (2017): 2356-4393.
- [14] Argyris, Chris, and Donald A. Schon. *Theory in practice: Increasing professional effectiveness*. Jossey-Bass, 1974.
- [15] Lv, Zhihan, et al. "Game on, science-how video game technology may help biologists tackle visualization challenges." *PloS one* 8.3 (2013): e57990.

# AUGMENTED REALITY ASSISTED MAINTENANCE AND MONITORING AT ON- PREMISE DATA CENTER

Muhamad Adib Bahari, Shah Runnizam Mohd Salleh and  
Muhammad Kamal Abdul Kiram

Information Technology & Analytics, TNB Research Sdn. Bhd.,  
Selangor, Malaysia

## **ABSTRACT**

*Augmented Reality is a combination of a real-world and a computer-generated environment that can enhance the operation and maintenance experience. This paper aims to describe the implementation process of Augmented Reality in monitoring the health and condition of selected assets in Tenaga Nasional Berhad Research (TNBR) Data Center. This research focuses on asset maintenance and monitoring that requires frequent maintenance and close monitoring that rigorously requires human interventions. Integration with the off-the-shelf solution (Zabbix) and custom-made web-based applications, mobile app and AR viewer is established via Application Programming Interface (API) and respective Software Development Kit (SDK). Artificial Intelligence (AI) modules are embedded with the existing knowledge base to rank the best possible solution for each alert. As a result, the solution shortens decision-making time and the troubleshooting process, especially with limited expertise.*

## **KEYWORDS**

*Augmented Reality, Knowledge Base, Artificial Intelligence, Asset Maintenance, Monitoring System*

## **1. INTRODUCTION**

Augmented Reality (AR) is the augmentation of digital images on a real-world object using various AR apps [1]. AR includes graphic images then added to the real world which significantly enhance learn ability and the user experience. Image augmentation is done interactively in realtime when the AR system is combined with the real world. AR is an enabler for disruptive technologies listed under the fourth industrial revolution and slowly assimilates into industries. For instance, the utility industry takes advantage of AR technology by enhancing the employee experience, keeping employees safer, and also closing the knowledge gap [3].

Augmented Reality allows users to enhance their field of view with real-time superimposed digital information [4]. This allows users to gain information on an asset or step-by-step instructions on how to repair or maintain an asset. The end in mind of an AR system is to enhance the user's perception by supplementing the real-world environment with 3D virtual objects that appear similar to the same space or direction as the real world [6]. The 3D objects need to be developed using a 3D game engine such as Unity or Unreal that allows customizations in terms of sizing, camera distance, and 3D object placement.

From a system development perspective, AR functionality is not a standalone module and utilized to visualize interactive data while allowing human-computer interaction. The foundation of AR requires a strong digitalization process to provide data, either from transactional processing, sensor outputs, databases and calculations. Although the interest in AR increases over time, publications regarding the implementation of the technology in the industry are still scarce. This has proven to be a strong motivation for researchers to further contribute in this area despite the mixed opinion on the practicality and safety aspects.

The objective of this research is to experiment with AR technologies in on-premise data center maintenance and operation workflow. In this research, a knowledge base with simple AI features is integrated into the system to assist in decision-making. The main component is a database that stores the historical events, possible root causes, possible solutions, and verification techniques. A simple calculation is applied to rank the best solution based on the frequency of similar cases that were previously solved. Finally, the solution will be displayed interactively on the AR viewer and the user will benefit in terms of efficiency of maintenance, incident data reliability, and data analytics for decision-making.

## **2. METHODOLOGY**

To ensure the project objectives are achieved, project methodologies are planned carefully before the development process begins. A feasibility study has been conducted to select the critical data center equipment that requires frequent maintenance and close monitoring. The selection of equipment covers multiple types of servers, network equipment, and uninterruptible power supplies (UPS) to enhance variation in the knowledge base. Each equipment type consists of different attributes, for instance, servers store the attributes of CPU utilization, memory utilization and CPU temperature.

The technical design follows which consists of the overall system architecture, data structure, network interconnectivity, desired mock-up user interface and user experience (UI/UX), and process flow for each module. The platform heavily relies on system development, hence, a proper design was produced by brainstorming, storyboarding and participatory design. Next, the system development phase started where each project deliverable was programmed using a suitable framework such as Unity, Vuforia, Laravel and Java. In parallel, the development of a database was initiated by extracting expert knowledge into key attributes such as historical events, possible root causes, possible solutions, and verification techniques.

After data has been collected, system integration takes place where network connections were established to enable data exchange between databases and backend monitoring API (Zabbix) that is crucial for AR viewer and mobile app. Once the development is readily deployed, an acceptance test was conducted to verify business logic, system stability and functionalities according to system requirements.

## **3. SCOPE OF WORK**

The scope of work that has been carried out through the project is site and equipment selection that match the project criteria. Firstly, we need to identify the critical or working operational environment that requires frequent maintenance and monitoring. The inspection and maintenance of the environment must be carried out on-site.

Secondly, we need to identify the most suitable hardware and software for AR implementation. Online data collection mechanisms such as built-in sensors, add-on Internet of Things (IoT)

sensors, and data acquisition systems need to be integrated to ensure data interconnectivity. Multiple and different devices need to communicate with each other using different types of data exchange protocols. Therefore, the method and protocol used to transport data from online sensors to the central database are crucial to produce accurate result.

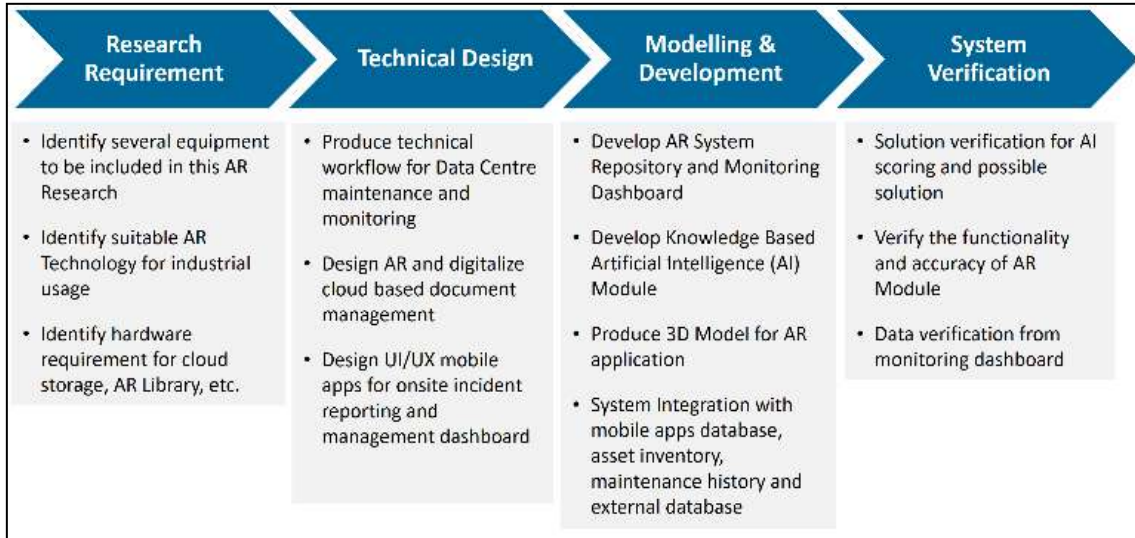


Figure 1. Overview of the Project Methodology

Thirdly, the development phase of modules is required for AR-assisted maintenance. A web-based application with API integration is developed to exchange data from the central database to the mobile application and the AR viewer. The mobile application complements the web-based functionality and acts as a companion tool to capture the data for inspection and maintenance on-site. A solution for each issue can be captured, thus developing the knowledge-based AI features for the root cause analysis.

Lastly, 3D models for markerless AR tracking are developed and embedded with live data from the database of the web-based application. Markerless tracking requires no marker e.g. QR, logo, or image, and solely responds to object recognition. This demonstrates both marker-based and markerless AR-assisted are capable to enhance inspection work on-site more conveniently and efficiently.

#### 4. RESULTS AND DISCUSSIONS

The AR-assisted module has given a rounded solution that supports maintenance activities as well as asset monitoring. Information data on the asset could be displayed on both AR Web App (Figure 2) and AR Viewer (Figure 4).

AR Web App is a web-based application developed using Laravel PHP framework. This application was hosted at TNBR Data Center and accessible via the public network. The AR Web Server runs using Apache web server on Ubuntu 20.04 Long Term Support (LTS) operating system. API is utilized to simplify software development and innovation that enables applications to exchange data and functionality easily and securely through web services [2]. The API is hosted on the same server thus, increase data confidentiality. To gather real-time data from selected equipment, another API is configured using Zabbix to simulate data exchange to the AR Viewer.





Figure 2. Web App Process Flow

#### 4.1. Asset Monitoring Dashboard

Asset Monitoring Dashboard allows the system administrator to monitor real-time data visualized through gauges and graphs. The dashboard summarizes critical information on existing assets such as server, network, and UPS health status. The dashboard displays feature for all assets as follows:

##### 4.1.1. To display server health and condition status



Figure 3. Server Health Status

#### 4.1.2. To display network health and condition status



Figure 4. Network Health Status

#### 4.1.3. To display UPS health and condition status



Figure 5. UPS Health Status

### 4.2. User Access Role

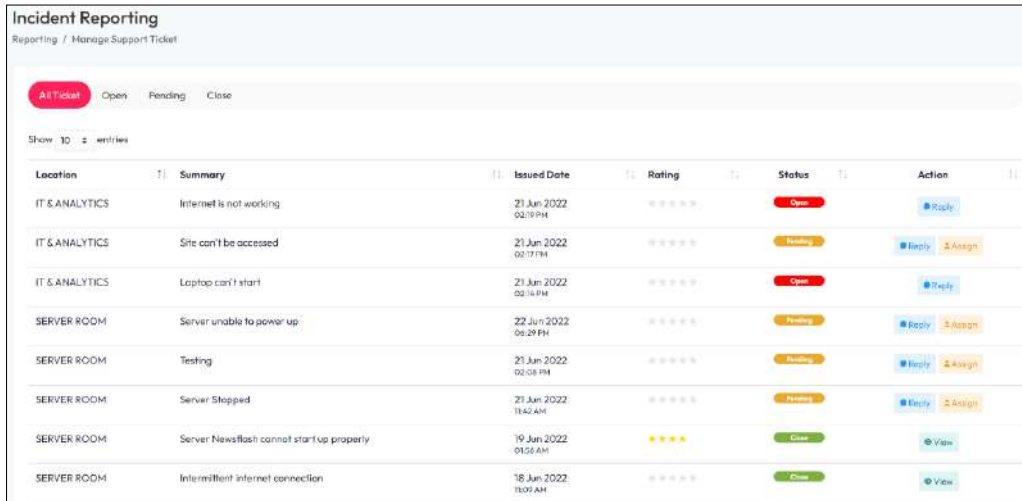
Multiple types of roles can be assigned to a designated user. Different types of roles will have different types of access limits such as personalized dashboards and permission control. The authentication is linked with Tenaga Nasional Berhad (TNB) Lightweight Directory Access Protocol (LDAP) to restrict access to only TNB users using their email authentication password. To enhance security measures, this Web App utilizes Secure Sockets Layer (SSL) which only enables the web browser to access the site via HTTPS. It is the standard technology for keeping an internet connection secure, thus safeguarding sensitive data.

### 4.3. Application Programming Interface (API)

API is a set of rules that computers or applications communicate with one another. APIs act as an intermediary layer that process data transfer between system. Data that is requested by a user or other application will be authorized and granted by API. Similar to this AR Web App, API integrates with both Mobile App and AR Viewer App to display selected data information securely. The AR system interprets the data and presented it on AR Dashboard in a structured manner.

#### 4.4. Support Ticket

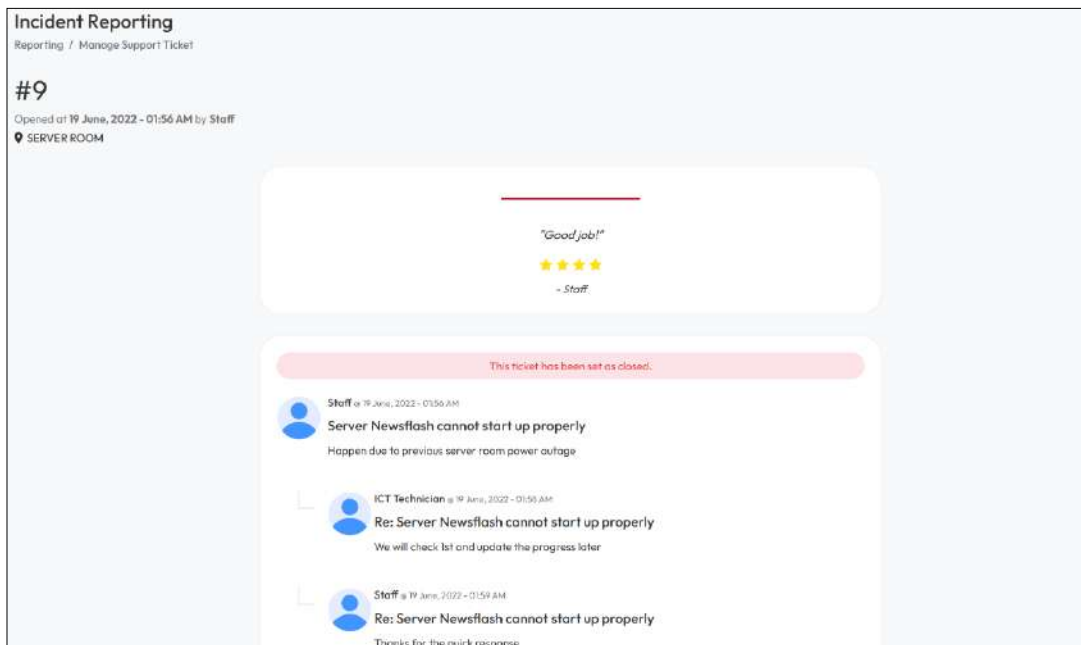
Businesses or organizations are increasingly dependent on ICT infrastructure and services to support business processes that lead to positive revenue growth. Therefore, the support ticket module platform is important for IT Department to monitor and optimize customer/user experience during the resolution process.



Location	Summary	Issued Date	Rating	Status	Action
IT & ANALYTICS	Internet is not working	21 Jun 2022 02:19 PM	★★★★★	Open	Reply
IT & ANALYTICS	Site can't be accessed	21 Jun 2022 02:17 PM	★★★★★	Pending	Reply Assign
IT & ANALYTICS	Laptop can't start	21 Jun 2022 02:14 PM	★★★★★	Open	Reply
SERVER ROOM	Server unable to power up	22 Jun 2022 09:29 PM	★★★★★	Pending	Reply Assign
SERVER ROOM	Testing	21 Jun 2022 02:08 PM	★★★★★	Pending	Reply Assign
SERVER ROOM	Server Stopped	21 Jun 2022 11:42 AM	★★★★★	Pending	Reply Assign
SERVER ROOM	Server Newsflash cannot start up properly	19 Jun 2022 01:56 AM	★★★★	Close	View
SERVER ROOM	Intermittent internet connection	18 Jun 2022 11:01 AM	★★★★★	Close	View

Figure 6. Incident Reporting

This is crucial for IT Department to comply with Service Level Agreement (SLA) that has been agreed upon. Chat box feature was included in AR Web as an additional module to capture the conversation between technician and staff.



**Incident Reporting**  
Reporting / Manage Support Ticket

#9  
Opened at 19 June, 2022 - 01:56 AM by Staff  
SERVER ROOM

"Good job!"  
★★★★★  
- Staff

This ticket has been set as closed.

Staff @ 19 June, 2022 - 01:56 AM  
Server Newsflash cannot start up properly  
Happen due to previous server room power outage

ICT Technician @ 19 June, 2022 - 01:50 AM  
Re: Server Newsflash cannot start up properly  
We will check 1st and update the progress later

Staff @ 19 June, 2022 - 01:59 AM  
Re: Server Newsflash cannot start up properly  
Thanks for the quick response

Figure 7. Star Rating Feedback

Star ratings are included inside the module allowing the staff to rate the technician service or share their feedback. This is important to evaluate their services and competencies, as well as to monitor Key Performance Index (KPI). This provides valuable information and room for improvement by the SLAs or customer charters.

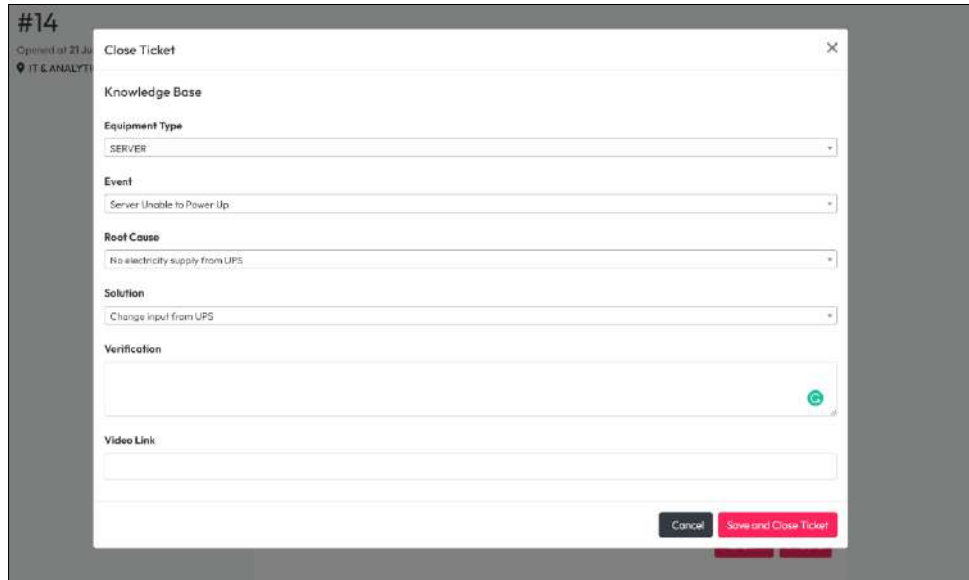
The image shows a 'Close Ticket' form with a white background and a grey border. At the top left, it says '#14' and 'Opened at 21 Jul'. Below that, there's a 'Knowledge Base' section. The form contains several dropdown menus: 'Equipment Type' with 'SERVER' selected, 'Event' with 'Server Unable to Power Up', and 'Root Cause' with 'No electricity supply from UPS'. Below these is a 'Solution' dropdown with 'Change input from UPS'. There is a 'Verification' section with a text area and a green checkmark icon. At the bottom, there is a 'Video Link' text area and two buttons: 'Cancel' and 'Save and Close Ticket'.

Figure 8. Knowledge Base Data Entry

Upon closing the ticket, the technician is required to key in the knowledge base. All information such as events, root causes, and solutions is already predefined. With this, the data can be collected and stored inside the database accurately and in a standardized manner.

To lodge an incident report, support tracking and asset monitoring can be accessed through AR Mobile App (Fig. 2). The mobile app complements the functionality of the web application in terms of flexibility and practicality. The functionality is trimmed from the web version and personalized to the different types of roles such as system administrator, technician, and user. The app consists of a personalized dashboard, asset management, and incident reporting. Intuitively designed, the app interface and user experience are engaging thus reducing the learning curve. The app is ready on both the Android and iOS platforms.

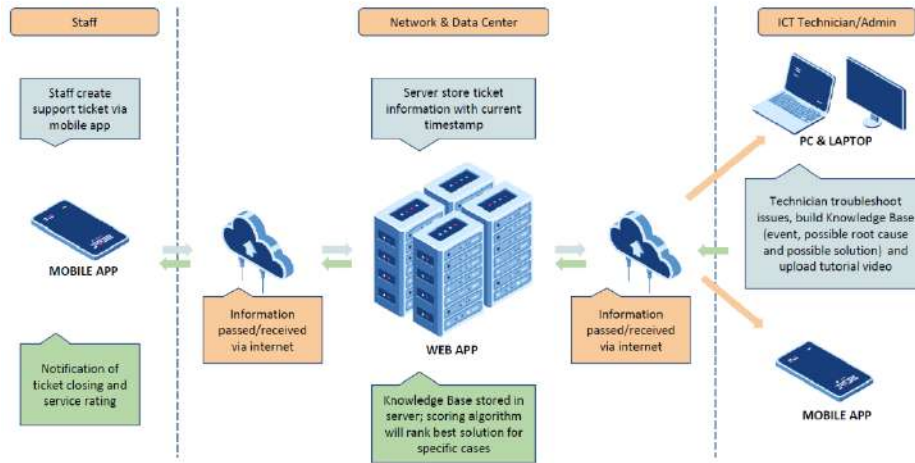


Figure 9. Mobile App Process Flow

AR Viewer has been designed to support on-site inspection, regular maintenance, and troubleshooting. With AR, machine maintenance can be performed faster and with fewer errors. That reduces mean time to repair, improves equipment availability, and reduces unplanned downtime.

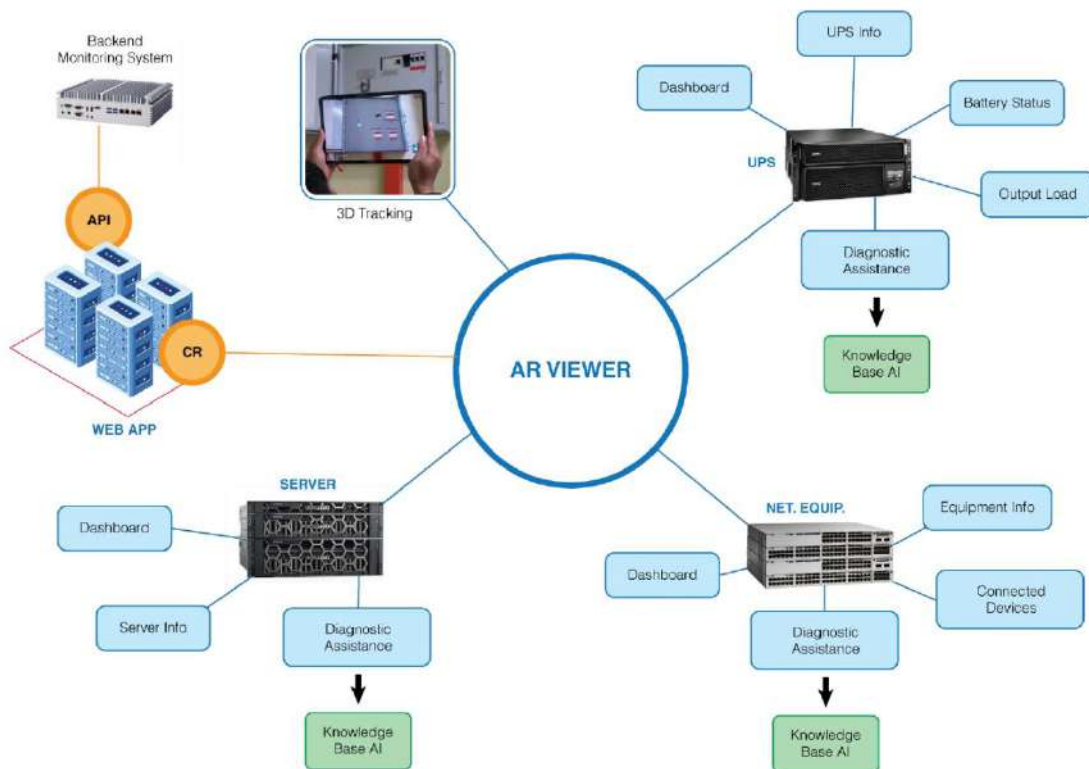


Figure 10. AR Viewer Process Flow

With all modules ready, the technical person in charge can easily identify the problem that arises and respond to the issue with the solution given by AR Viewer. This will shorten the time of troubleshooting and increase work productivity.

## 5. LIMITATIONS

The limitation of the research can be divided into three aspects, which are knowledge base maturity, practicality and safety, and organizational change management. To produce the best user experience, the knowledge base should be mature in terms of historical data and contain an adequate amount of use cases. This requires a lot of effort to extract expertise from domain experts and experienced users, especially in a niche area. The collected data must be cleansed for further analysis.

In terms of practicality and safety, the AR viewer can be designed as glasses, head mount units, wearables and mobile apps. Each of the design must be suited to the working environment to ensure usability and complies with safety policy. Lastly, a new approach introduced to organizations may result in lack of user support and engagement. Change management plans should be taken into consideration in the early development stage to ensure a successful outcome.

## 6. FUTURE WORK

This paper suggests further studies with larger knowledge base datasets to enhance the accuracy of the diagnostic analysis. Machine learning can be incorporated into the existing approach to mimic the human decision-making process. Wearables and head mount devices in the market should be studied in practicality and safety aspects to fully appreciate what AR technology has to offer from time to time.

## 7. CONCLUSIONS

In summary, the project has successfully executed and produced a functional system that enhances user and customer experience, as theoretically, AR increases user engagement and interaction. This proves that AR technology is capable to retain maintenance expertise by digitalizing knowledge [5]. This AR-assisted project can be applied not only to TNBR Data Center but also to other equipment such as system at Hydro Power Plant or other industry that requires frequent maintenance.

## REFERENCES

- [1] A.Riya and S.Abhishek, "Augmented Reality and its effect on our life" – IEEE, 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)
- [2] What is an Application Programming Interface (API) - <https://www.ibm.com/cloud/learn/api>
- [3] How AR fits into Industry 4.0 - <https://electronics360.globalspec.com/article/18215/how-ar-fits-into-industry-4-0>
- [4] What Is Augmented Reality (AR)? A Practical Overview - <https://www.threekit.com/blog/what-is-augmented-reality>
- [5] Augmented reality tools: A digital bridge for the skill gap - <https://www.controleng.com/articles/augmented-reality-tools-a-digital-bridge-for-the-skill-gap>
- [6] R. Azuma, Y. Bailot, R. Behringer, S. Feiner, S. Julier and B. MacIntyre, "Recent Advance in Augmented Reality" – Computers & Graphics, November 2001

**AUTHOR**

**Muhamad Adib** is an IT Executive in one of the subsidiaries company for Tenaga Nasional Berhad (TNB) in Malaysia. He was born in Kuala Lumpur. He holds a degree in Business Computing and further his Master in Economics. Passionate about implementing technology, especially in IT that could help people make life easier.



© 2023 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# REVIEW OF METRICS TO MEASURE THE STABILITY, ROBUSTNESS AND RESILIENCE OF REINFORCEMENT LEARNING

Laura L. Pullum

Mathematics and Computer Science Division, Oak Ridge National Laboratory,  
Bethel Valley Road, Oak Ridge, USA

## **ABSTRACT**

*Reinforcement learning (RL) has received significant interest in recent years, primarily because of the success of deep RL in solving many challenging tasks, such as playing chess, Go, and online computer games. However, with the increasing focus on RL, applications outside gaming and simulated environments require an understanding of the robustness, stability, and resilience of RL methods. To this end, we conducted a comprehensive literature review to characterize the available literature on these three behaviors as they pertain to RL. We classified the quantitative and theoretical approaches used to indicate or measure robustness, stability, and resilience behaviors. In addition, we identified the actions or events to which the quantitative approaches attempted to be stable, robust, or resilient. Finally, we provide a decision tree that is useful for selecting metrics to quantify behavior. We believe that this is the first comprehensive review of stability, robustness, and resilience, specifically geared toward RL.*

## **KEYWORDS**

*Reinforcement Learning, Resilience, Robustness, Stability*

## **1. INTRODUCTION**

Recent literature on the robustness of machine-learning models has focused almost entirely on the robustness of deep neural networks for imaging applications. However, at the time of this study, there were no published surveys on the robustness of reinforcement learning (RL). We pursued this review because of the increasing use of RL, particularly in control systems. Along with robustness, stability and resilience are included. Stability was included because the term has been used interchangeably with robustness, and resilience was included because the term has been used as a state beyond robustness.

RL involves agents that act in an environment and experience a reward for their actions. The agent learns the policy that maximizes the cumulative reward. Formally, consider an agent operating at time  $t \in \{1, \dots, T\}$ . At time  $t$ , the agent is in environment state  $s_t$  and produces an action  $a_t \in A$ . The agent then observes a new state  $s_{t+1}$  and receives reward  $r_t \in R$ . A set of possible actions  $A$  can be discrete or continuous. The goal of reinforcement learning is to find a policy  $\pi(a_t|s_t)$  for choosing an action in state  $s_t$  to maximize the utility function or (expected return). [252]



$$J(\pi) = \mathbf{E}_{s_0, a_0, \dots} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \quad (1)$$

Where  $0 \leq \gamma \leq 1$  is a discount factor,  $a_t \sim \pi(a_t | s_t)$  is drawn from the policy, and  $s_{t+1} \sim P(s_{t+1} | s_t, a_t)$  is generated by environmental dynamics. The state value function

$$V^\pi(s_t) = \mathbf{E}_{a_t, s_{t+1}, \dots} [\sum_{i=0}^{\infty} \gamma^i r(s_{t+1}, a_{t+1})] \quad (2)$$

is the expected return by policy  $\pi$  from state  $s_t$ . The state action function

$$Q^\pi(s_t, a_t) = \mathbf{E}_{s_{t+1}, a_{t+1}, \dots} [\sum_{i=0}^{\infty} \gamma^i r(s_{t+1}, a_{t+1})] \quad (3)$$

is the expected return by policy  $\pi$  after taking action  $a_t$  at state  $s_t$ . [252].

The objective of this study is to present a systematic review of RL literature to identify metrics for measuring the stability, robustness, and resilience of RL. We limit RL to general reinforcement learning and not to specialized RL, such as inverse RL. We reviewed studies that attempted to measure or otherwise characterize stability and robustness, and resilience of RL, seeking metrics for these behaviors.

We searched computer science and technical literature databases for eligible papers, combining RL, behavior terms, and terms related to measuring, metrics, and quantification. The result comprised 16,015 items, and after removal of duplications and extraneous material, a collection of 546 items was established. Through the process of elimination described in full in this paper, we reduced the set to 248 papers. We systematically reviewed 248 papers and presented the results in this analysis. We classified the papers by behavior (i.e., stability ( $n=76$ ), robustness ( $n=169$ ), and resilience ( $n=3$ )), and identified the primary domains of application as robotics, network systems, power system control, and vehicle/traffic control and navigation. We identified approaches to determine or measure each behavior individually and across behaviors. The approaches were categorized as quantitative or theoretical, and the quantitative approaches were further classified as being applied internally (e.g., in training) or externally (e.g., performance measures on outputs) to the model. The metrics, approaches, and objectives were identified for each paper reviewed. The objective indicates the metric or approach intended to be stable, robust, or resilient. We close by indicating the need to define stability, robustness, and resilience behaviors for RL and identify quantitative and theoretical approaches to achieve measurement and determination of these behaviors.

There is a rich set of domains (i.e., 53 identified in this survey) in which the measurement of RL stability, robustness, and resilience has been conducted. The domains ranged from robotics and network systems to sheep herding and fish behavior. The most frequently mentioned domains include robotics, general control, and network systems, with numerous studies not specifying a domain. Many studies used Gym [254] and other environments for demonstration purposes. Though the search focused on the quantitative measurement of stability, robustness, and resilience, theoretical approaches were identified as well. The quantitative approaches were categorized as internal or external depending on where the evaluation was conducted in the model. Internal measures quantified the performance of the training and external measures quantified the ultimate performance of the model.

The goal of this systematic review is to identify metrics for measuring the stability, robustness, and resilience of RL. To initiate the search for this review, we identified keywords and phrases related to reinforcement learning, the *behaviors* of interest (stability, robustness, and resilience), and measurement. The *key phrase* is reinforcement learning. The *measurement* keywords are metric, measure, index, score, quantifier and indicator.

We believe that this is the first comprehensive review of stability, robustness, and resilience specifically geared toward RL. The remainder of this paper is organized as follows. Section 2 describes the methods used in the systematic review. Section 3 presents the results of the review. Section 4 discusses the results of the review and introduces a decision tree for metric selection based on the review.

## 2. METHODS

Keywords salient to RL, system behavior, and measurement were identified for the research topic. The typical search was of the form:

<Key Phrase> + <Behavior> + <Measurement>

with <Key Phrase>, <Behavior> and <Measurement> defined above. A specific example is

“reinforcement learning” AND robust\* AND (“metric” OR “measure” OR “index” OR “score” OR “quantifier” OR “indicator”)

Multiple searches were conducted using bibliographic databases covering broad areas of computer science, physical and biological sciences, and engineering. The information sources used in this study are the open-access arXiv covering 1991-present and the subscription services Scopus (1823-present) and Web of Science (1900-present). No restrictions were placed on the publication date or language. Journal articles, books, books in a series, book sections or chapters, edited books, theses and dissertations, conference papers, and technical reports containing keywords and phrases were included in the search. The publication date of the returned search results is bound by the dates of coverage of each database and the date on which the search was performed; however, all searches were completed by October 31, 2020. The range of dates for the documents ultimately included in the review was from 2002 to 2020.

The queried databases yielded 16,015 citations. Irrelevant citations were also retrieved. We excluded extraneous studies, resulting in a collection of 699 publications. Furthermore, the removal of duplicate papers resulted in 580 publications. Citations for “full conference proceedings were removed if the relevant paper(s) within the associated conference were otherwise collected, resulting in 546 publications. Further refinement excluded publications that were not on RL, which were not on the searched behavior, or those that had no metrics or theoretical content, resulting in 248 documents. We systematically reviewed 248 papers, and the results are presented in this analysis.

The 248 papers that made it through the screening process were grouped by search behavior: stability, robustness, and resilience. We also identified papers on one behavior that mentioned one or both other behaviors. Some studies that mentioned other behaviors did so interchangeably. For instance, stability and robustness have been used interchangeably in several studies, which can lead to some confusion in the definitions of these behaviors. The primary domains of application were identified and categorized as robotics, network systems, general control systems, Gym [254], and other environments. We also identified publications that mentioned the RL policy.

The primary focus of this study was to identify approaches to determine or measure each behavior. Of course, most publications reviewed focused on quantitative approaches because of the search terms used. Those that use a theoretical approach provide additional insight into the behavior-determination problem. The quantitative approaches were further classified as being applied internally (e.g., in training) or externally (e.g., performance measures on outputs) to the model.

Metrics, approaches, and objectives were identified for each study (see Figure 1). The objective indicates the metric or approach intended to be stable and robust, or resilient.

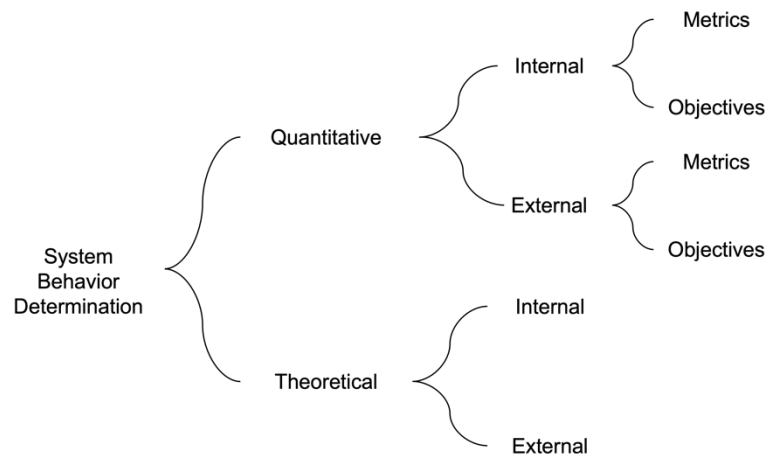


Figure 1. Categorization and resulting metrics, approaches, and objectives

There is little agreement in the literature on the definitions of stability, robustness, and resilience. In fact, there are few distinct definitions of these behaviors. In this review, we used the following definitions:

*Stability* is a property of the learning algorithm (i.e., a small change in the training set results in a similar model) and refers to the ranking of the variance of a model [253]. For example, if we use the variance of the loss function over all datasets as a performance measure, we test a set of models. The smallest loss indicated a more stable model. Given this definition, stability analysis is an application of sensitivity analysis to machine learning.

*Robustness*, when used with respect to computer software, refers to an operating system or other program that performs well not only under ordinary conditions but also under unusual conditions that stress its designers' assumptions (<http://www.linfo.org/robust.html>). Robustness is a property of the model and is measured by, for example, loss over all datasets (as opposed to the variance of the loss).

Throughout the literature, *resilience* has been used interchangeably with robustness; however, it is used most often with production machine learning systems to indicate robustness to different datasets and different data added to the dataset.

### 3. RESULTS AND ANALYSIS

Publications were categorized by behavior as follows: stability ( $n=76$ ) [4-80], robustness ( $n=169$ ) [81-169], and resilience ( $n=3$ ) [1-3]. Studies on one behavior often mention other behaviors, especially stability and robustness. Resilience was mentioned in five stability papers and 11 robustness papers. Robustness was mentioned in 50 stability papers and in one resilience paper. Stability was mentioned in 104 Robustness papers and in all (3) Resilience papers.

Given the recent explosion of literature on the robustness of neural networks to adversarial attacks, one might expect it to be a cornerstone of the robustness papers reviewed herein. The term "adversarial" was mentioned in a quarter ( $n=61$ ,  $N=248$ ) of the papers reviewed. That is, 1 resilience paper, 56 robustness papers, and 4 stability papers mention "adversarial". Some papers on

one behavior used one of the other behaviors interchangeably, notably stability and robustness, specifically [91, 93, 105, 145, 146, 179, 194, 225, and 237] and generally in several other articles.

### 3.1. Application Domains

The publication application domains are provided in the supplementary information and summarized in Figure 2. The primary domains were robotics, with 16.4% ( $n=44$ ) of the total citations ( $N=268$ ), followed by network systems and general control ( $n=7.8%$ ,  $n=21$ ), with 9.3% ( $n=25$ ) using Gym or other environments as their experimental domain. Just as many ( $n=25$ , 9.3%) papers did not specify a domain. These top 5 ( $n=53$ ) domains comprised over 50% (52.9%,  $n=136$ ) of citations. Most (52.8%,  $n=28$ ) domains ( $n=53$ ) had a single citation.

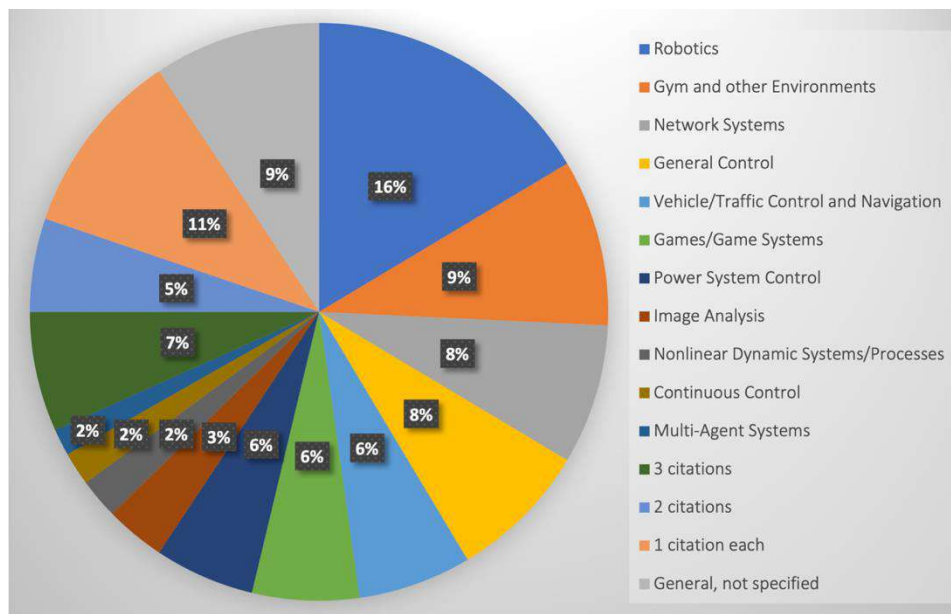


Figure 2. Application Domain Categories

### 3.2. Reinforcement Learning Policies

Twenty-one (21) RL policies were mentioned in the articles. Most documents did not identify the policies used. Of the 21 types of policies mentioned, the top 4 – Actor-Critic ( $n=18$ ), Q-learning ( $n=16$ ), Proximal Policy Optimization (PPO) ( $n=8$ ), and Adaptive Critic Design ( $n=5$ ) comprised 72.3% of the total citations that included policy ( $N=65$ ).

### 3.3. Approach to Determining or Measuring Behavior

The publications' approaches to determining or measuring each behavior are categorized as either quantitative or theoretical. Most of the publications focused on quantitative approaches ( $n=205$ , 82.0%), which is understandable given that the search focused on quantifying behaviors. For publications on stability behavior, there was an almost even split between the quantitative ( $n=42$ ) and theoretical ( $n=43$ ) approaches. However, publications on robustness behavior have primarily focused on quantitative approaches ( $n=160$ ) vice theoretical ( $n=35$ ). All (3) publications on resilience applied quantitative approaches.

### 3.3.1. Types of Quantitative Approaches

Next, we further categorized the quantitative approaches according to whether they were focused internal or external to the model. Internal quantitative approaches measure aspects within the model, such as its training and associated measures, including the value of rewards over time or the number of episodes until convergence. External quantitative approaches measure performance-related aspects of a model, such as variations in accuracy or throughput. Most ( $n=142$ , 63.1%) quantitative approaches were categorized as performance-related or external measures. Of these, most ( $n=103$ ) were for robustness, followed by those for stability ( $n=36$ ). The 3 papers on resilience focused on performance-related quantitative measures. Robustness also led to internal approaches ( $n=69$ ) with stability ( $n=14$ ). This is primarily due to the large number of robustness papers ( $n=170$ ) and paucity of resilience papers ( $n=3$ ). Of the robustness papers, 40.0% ( $n=69$ ) contained internal quantitative measures, and 60.6% contained external quantitative measures. The stability values were 18.2% and 46.8%, respectively.

### 3.3.2. Types of Internal Quantitative Approaches

Looking at the types of internal quantitative approaches, we see a narrow set of aspects considered in the papers. These metrics are specifically designed to measure stability rather than the variance of the output. They measured the variation in training performance. The vast majority ( $n=75$ , 88.2%) of the internal quantitative approaches calculated the reward- or score-based metrics. Other types of internal quantitative approaches include two each of policy entropy, variations in control strategy approximation weights, and convergence rate, and one each of policy weight, calculation of the Lyapunov stability criteria, and calculation of the Wasserstein function lower bound. In RL context, convergence refers to the stability of the learning process (and the underlying model) over time [11].

### 3.3.3. Types of External Quantitative Approaches

External or performance-based quantitative approaches for measuring behaviors primarily ( $n=39$ ) used deviations or variations in performance-related metrics other than precision, accuracy, or recall (Figure 3). The next highest category ( $n=28$ ) of quantitative metrics used error, failure, and success rates. Statistics on the performance of the tracking or estimation error follow, with  $n=23$  papers. Papers in the network domain used network-related metrics ( $n=15$ ) to measure behavior. Statistics on precision, accuracy, and recall ( $n=12$ ) were also used. Five papers used variance in loss or regret estimation, three papers used game-related performance measures to quantify behavior, and two papers each used bounds on or the size of the stability region and terminal wealth and inventory. Eighteen (18) additional different types of external quantitative metric categories were represented by a single paper each.

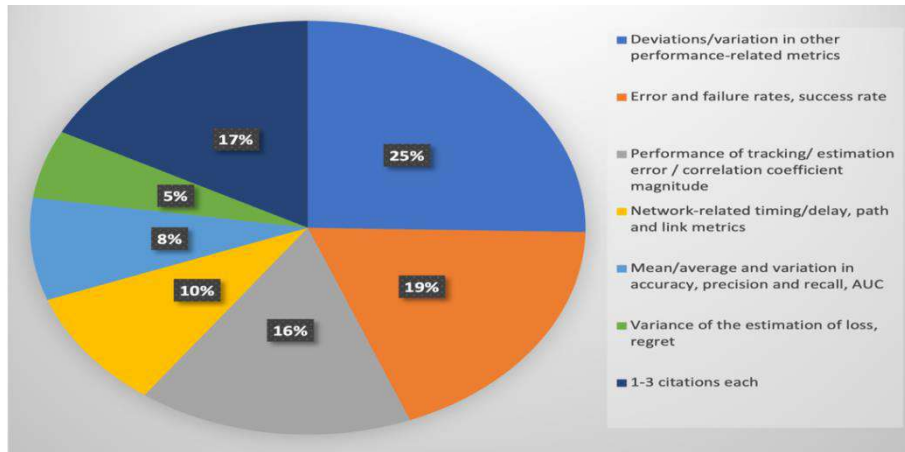


Figure 3. External Quantitative Metrics

### 3.3.4. Quantitative Approach Objectives

An additional aspect reviewed was to determine to what actions or events were the quantitative approaches attempting to be stable, robust, or resilient. We call this the *<behavior> objective*. The *<behavior>* objective category (see Figure 4), with the highest number of citations, was geared toward handling changes in the operational environment, dynamic environment, or network ( $n=41$ ). Papers that did not specifically state their objectives comprised the next most populous category ( $n=35$ ). The objective of handling uncertainties and disturbances in the environment also contained  $n=35$  papers. The remaining objectives included input variation/perturbations ( $n=20$ ), differences between training and test or operational environments ( $n=19$ ), differences or uncertainties in model parameters ( $n=16$ ), adversarial attack ( $n=14$ ), different domains, environments, or settings ( $n=8$ ), errors or failures in the operational environment ( $n=5$ ), differences in training datasets or initializations ( $n=5$ ), high variability ( $n=2$ ), and one paper each in systematic pressure, spamming, incomplete data, and unknown control coefficients.

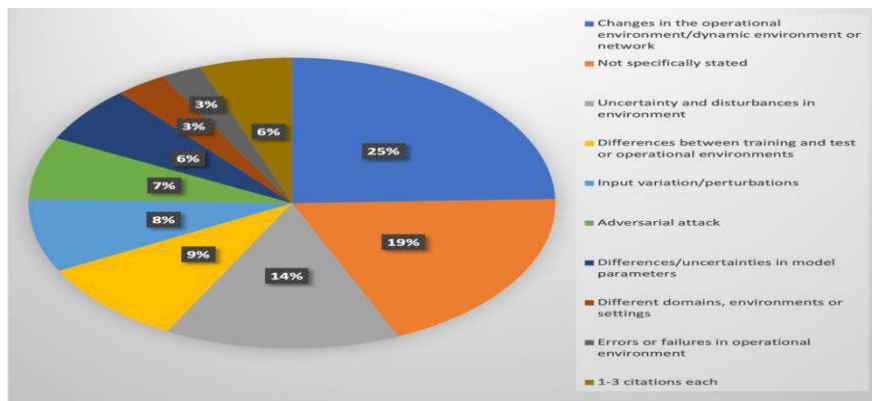


Figure 4. Quantitative *<behavior>* Objectives

### 3.3.5. Types of Theoretical Approaches

Most of the theoretical approaches in the papers reviewed were based on the Lyapunov theory ( $n=50$ , 61.0%) (Figure 5). The next highest types of theoretical approaches used are convergence to Nash equilibrium ( $n=10$ ) and value-based guarantees, such as error and output deviation

bounds ( $n=8$ ). Of the remainder, three papers used the Wasserstein distance to explore stability, three studies proved that the methods were doubly robust, two papers proved that the methods exhibited Lipschitz continuity, and stochastic stability theory to prove stability, stability guarantees, policy-based guarantees, regret bounds, minimization of the Jacobian on input, and per-episode Bellman-error regret guarantees/bounds were used by a single paper each to establish the stability of the RL methods discussed.

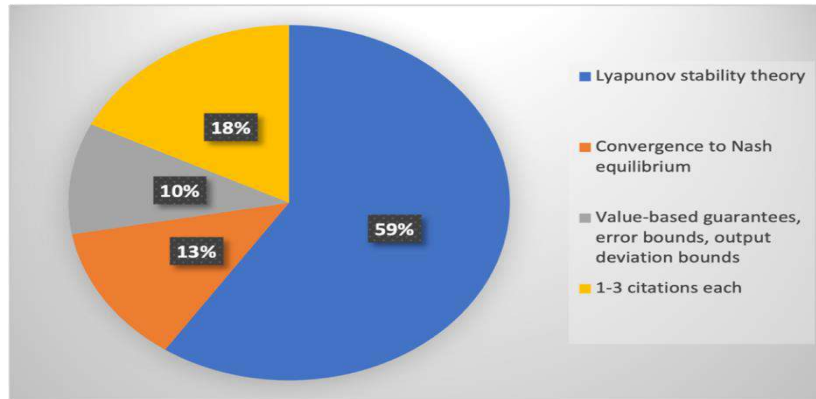


Figure 5. Theoretical approaches

### 3.3.6. Theoretical Approach Objectives

We also reviewed the *<behavior>* objective for theoretical papers (Figure 6). Most papers ( $n=42$ , 54.5%) on theoretical approaches did not state their objectives. Of the few that did, changes or dynamics in the operational environment were the most frequent objective ( $n=10$ ), followed by differences or uncertainties in model parameters ( $n=7$ ), adversarial attack ( $n=6$ ), error or failure ( $n=5$ ), differences between training and test or operational environments ( $n=2$ ), input variation ( $n=2$ ), and one each for domain shifts, different function approximation architectures, and differences in quantization levels.

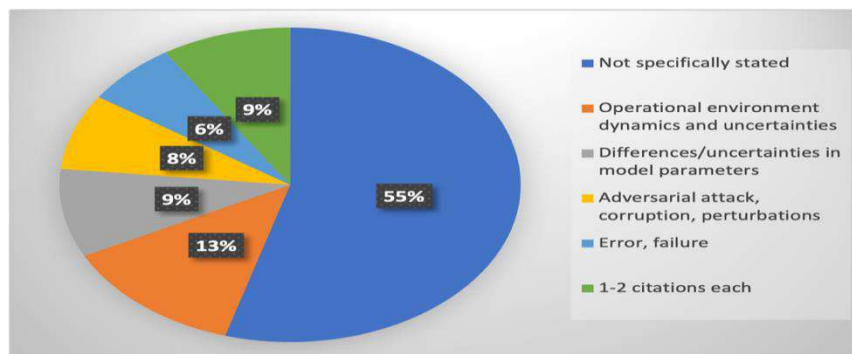


Figure 6. Theoretical *<behavior>* Objectives

## 4. DISCUSSION

Our study was conducted to characterize the published methods of measuring or determining the stability, robustness, or resilience of RL. Of an initial collection of 16,015 items, 248 papers met the inclusion criteria and were systematically reviewed. Approaches to measuring or determining behavior are classified as either quantitative or theoretical. Quantitative approaches were further classified as internal or external depending on whether they evaluated the training, test, or

operational phases. For both categories of quantitative approaches, we categorized the metrics used, with internal approaches primarily using the reward or score (and statistics on the same) and external approaches primarily using variations in performance-related metrics (although not precision, accuracy, or recall). The theoretical approaches were dominated by Lyapunov stability theory. We further characterized the objectives of stability, robustness, and resilience. Quantitative approaches to measuring behavior focused on the ability to handle differences in the operational environment, whereas most theoretical approaches to determining behavior did not specifically state an objective. However, the objective of the theoretical approaches can be implied using Lyapunov stability theory, that is, to prove the stability of the system. Lyapunov was used, regardless of whether the article was on stability or robustness.

To determine the metric to use, we developed a decision tree based on the information obtained in this literature review. It is a collapsible tree, so that branches are not exposed unless selected, and open branches can be closed or collapsed. There are several levels in the decision tree, starting with i) behavior (stability, robustness, or resilience); ii) the domain; iii) a list of quantitative and theoretical objectives; iv) the next level divides the metrics into external, internal, and theoretical metrics; and v) the last level, that is, the leaves, is the set of metrics for that branch of the decision tree. For example, suppose we want to find a suitable metric to measure the robustness of a control system expected to face changes in the operational environment. From the metric decision tree shown in Figure 7, we can see that the first selection is for a robustness metric. This selection displays the domains in which the robustness metrics are described. Selecting the General Control domain reveals 9 objectives, including the objective “Dynamic Environment.” An external metric found in the literature for this case is “blood glucose response” which is not applicable for this control system. The more appropriate metrics and approaches are the size of the stability region, value-based guarantees, error bounds, and Lyapunov stability theory and calculation. Any or all of these can be used to measure the robustness of a general control system in a dynamic operational environment.

Supplementary information for this review is provided at <https://arxiv.org/pdf/2203.12048.pdf>, including a) PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [251] diagrams for Stability, Robustness and Resilience, respectively; b) the data reduction methodology for Stability, Robustness and Resilience, respectively; and the PRISMA checklist. In addition, the site provides detailed tables of the results described in Section 3.

## ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
DOE	Department of Energy
ORNL	Oak Ridge National Laboratory
PPO	Proximal Policy Optimization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RL	Reinforcement Learning
US	United States

## ACKNOWLEDGMENTS

The author would like to acknowledge Rama Vasudevan, PhD of the Oak Ridge National Laboratory (ORNL) for intellectual discussions and collaborative research on reinforcement learning. The author would also like to thank Nathan Martindale (ORNL) for assistance in improving the functionality and usability of the decision tree.



This work was funded initially by the AI Initiative at the Oak Ridge National Laboratory and subsequently funded by the US Department of Energy, National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research and Development (NA-22). This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). The publisher, by accepting the article for publication, acknowledges that the US government retains a non-exclusive, paid up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for U.S. Government purposes. The DOE will provide public access to these results in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

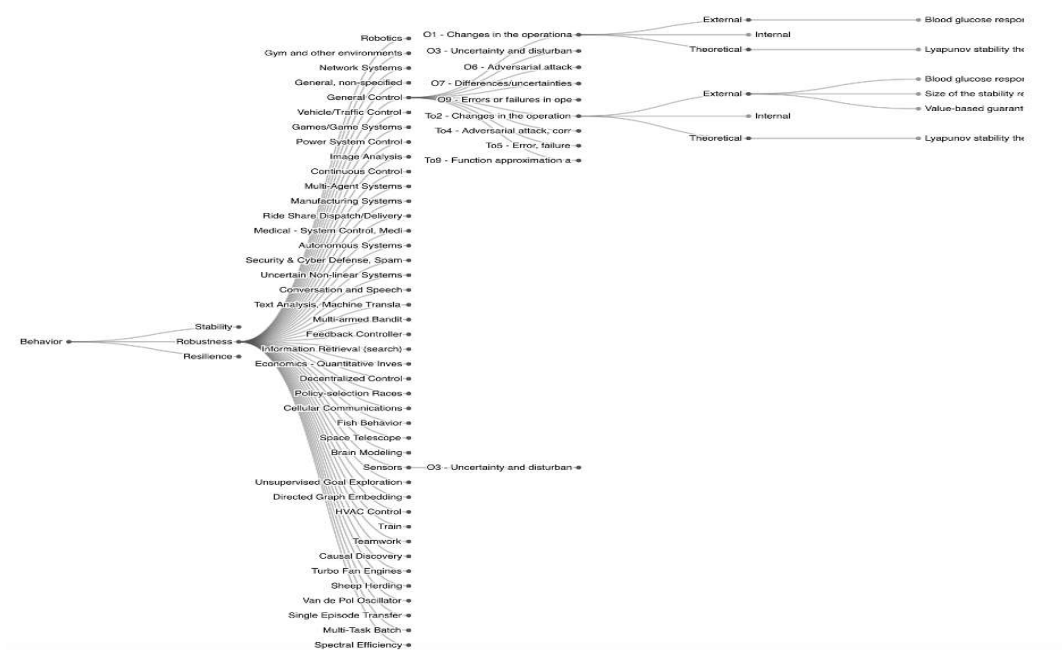


Figure 7. Metric Selection Decision Tree Section

## REFERENCES

- [1] V. Behzadan and A. Munir, "Adversarial exploitation of emergent dynamics in smart cities," Proc 2018 IEEE Intl Smart Cities Conf, doi 10.1109/ISC2.2018. 8656789.
- [2] S. Enjalber and F. Vanderhaegen, "A hybrid reinforced learning system to estimate resilience indicators," Eng Appl of AI, vol. 64, pp. 295-301, 2017.
- [3] M. Bunyakitanon, et al, "End-to-end performance-based autonomous vnf placement with adopted reinforcement learning," IEEE Trans on Cognitive Comms and Networks, vol. 6, no. 2, pp. 534-547, 2020, doi 10.1109/TCCN.2020.2988486.
- [4] Z. Dong, X. Huang, Y. Dong, and Z. Zhang, "Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system," J. Appl Energy, vol. 259, 2020, doi 10.1016/j.apenergy.2019.114193.
- [5] G. Wen, et al, "Optimized adaptive nonlinear tracking control using actor-critic reinforcement learning strategy," IEEE Trans on Industrial Informatics, vol. 15, no. 9, pp. 4969-4977, 2019.
- [6] B. Muneeswari and M.S.K. Manikandan. "Energy efficient clustering and secure routing using reinforcement learning for three-dimensional mobile ad hoc networks," IET Commun, vol. 13, no. 12, pp. 1828-1839, 2019.
- [7] B. A. G. de Oliveira, C. A. P. da S. Martins, F. Magalhaes, L. Fabricio, and W. Goes, "Difference based metrics for deep reinforcement learning algorithms," IEEE Access, vol. 7, pp. 159141-159149, 2019.

- [8] K. Zhang, et al, "Policy search in infinite-horizon discounted reinforcement learning: advances through connections to non-convex optimization," in Proc: 53rd CISS, Baltimore, MD, USA, 2019.
- [9] Z. Du, W. Wang, Z. Yan, W. Dong, and W. Wang, "Variable admittance control based on fuzzy reinforcement learning for minimally invasive surgery manipulator," *Sensors*, vol. 17, no. 4, 2017.
- [10] H. Jiang, et al, "Optimal tracking control for completely unknown nonlinear discrete-time Markov jump systems using data-based reinforcement learning method," *Neurocomputing*, vol. 194, pp. 176-182, 2016.
- [11] S. Abdallah, "Why global performance is a poor metric for verifying convergence of multi-agent learning," arXiv:0904.2320v1 [cs.MA] 15 April 2009.
- [12] N. Talele and K. Byl, "Mesh-based tools to analyze deep reinforcement learning policies for underactuated biped locomotion," arXiv:1903.12311v2 [cs.RO] 1 November 2019.
- [13] Y.-L. Tuan, J. Zhang, Y. Li, and H.-y. Lee, "Proximal policy optimization and its dynamic version for sequence generation," arXiv:1808.07982v1 [cs.CL] 24 August 2018.
- [14] A. Serhani, et al, "AQ-Routing: mobility-, stability-aware adaptive routing protocol for data routing in MANET-IoT systems," *Cluster Comp*, vol. 23, pp. 13-27, 2020, doi 10.1007/s10586-019-02937-x.
- [15] Z. Dong, et al, "Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system," *Applied Energy*, vol. 259, 2020, doi 10.1016/j.apenergy.2019.114193.
- [16] H. Zhang, K. Zhang, Y. Cai, and J. Han, "Adaptive fuzzy fault-tolerant tracking control for partially unknown systems with actuator faults via integral reinforcement learning method," *IEEE Trans on Fuzzy Systems*, vol. 27, no. 10, 2019, doi 10.1109/TFUZZ.2019.2893211.
- [17] D. Cohen, S. M. Jordan, and W. B. Croft, "Learning a better negative sampling policy with deep neural networks for search," in *ICTIR '19*, Santa Clara, CA, USA, 2019, doi 10.1145/3341981.3344220.
- [18] C. Mu, et al, "Q-learning solution for optimal consensus control of discrete-time multiagent systems using reinforcing learning," *J Frank Inst*, vol. 356, pp. 6946-6967, 2019, 10.1016/j.jfranklin.2019.06.0070016-0032.
- [19] X. Tang, et al, "A deep value-network based approach for multi-driver order dispatching," in *KDD 19*, Anchorage, AK, USA, 2019, doi 10.1145/3292500.3330724.
- [20] M. Abouheaf, and W. Gueaieb, "Model-free adaptive control approach using integral reinforcement learning," in *Proc. IEEE Intl Symp on Robotic and Sensors Environ*, 2019.
- [21] D. Seo, H. Kim, and D. Kim, "Push recovery control for humanoid robot using reinforcement learning," *Third IEEE IRC*, 2019, doi 10.1109/IRC.2019.00102.
- [22] Y. Lv, X. Ren, and J. Na, "Online Nash-optimization tracking control of multi-motor driven load system with simplified RL scheme," *ISA Trans*, 2019, doi 10.1016/j.isatra.2019.08.025.
- [23] L. Tang, Y.-J. Liu, and C. L. P. Chen, "Adaptive critic design for pure-feedback discrete-time MIMO systems preceded by unknown backlashlike hysteresis," *IEEE Trans on Neural Networks and Learning Syst.*, vol. 29, no. 11, 2018, doi 10.1109/TNNLS.2018.2805689.
- [24] P. Mertikopoulos, and W. H. Sandholm, "Riemannian game dynamics," *J of Econ Theory*, vol. 177, pp. 315-364, 2018, doi 10.1016/j.jet.2018.06.002.
- [25] D. Liu, and G.-H. Yang, "Model-free adaptive control design for nonlinear discrete-time processes with reinforcement learning techniques," *Intl J of Systems Science*, vol. 49, no. 11, pp. 2298-2308, 2018, doi 10.1080/00207721.2018.1498557.
- [26] A. Bentaleb, et al, "ORL-SDN: Online reinforcement learning for SDN-enabled HTTP adaptive streaming," *ACM Trans. Multimedia Comput. Commun. Appl*, vol. 14, no. 3, 2018, Art. no. 71, doi 10.1145/3219752.
- [27] Y. Hu and B. Si, "A reinforcement learning neural network for robotic manipulator control," *Neural Computation*, vol. 30, no. 7, pp. 1983-2004, 2018, doi 10.1162/neco\_a\_01079.
- [28] Y. Mei, et al, "Chaotic time series prediction based on brain emotional learning model and self-adaptive genetic algorithm," *Acta Physica Sinica*, vol. 67, no. 8, 2018, doi 10.7498/aps.67.20172104.
- [29] Z.-W. Hong, S.-Y. Su, T.-Y. Shann, Y.-H. Chang, and C.-Y. Lee, "A deep policy inference Q-network for multi-agent systems," in *Proc. AAMAS 2018*, Stockholm, Sweden, 2018.
- [30] Y. Xiong, H. Chen, M. Zhao, and B. An, "HogRider: Champion agent of Microsoft Malmo collaborative AI challenge," in *AAAI-18*, pp. 4767-4774, 2018.
- [31] W. Wu and L. Gao, "Posture self-stabilizer of a biped robot based on training platform and reinforcement learning," *Robotics and Autonomous Systems*, vol. 98, pp. 42-55, 2017, doi 10.1016/j.robot.2017.09.001.

- [32] M. Boushaba, A. Hafid, and M. Gendreau, "Node stability-based routing in wireless mesh networks," *J of Network and Computer Appl*, vol. 93, pp. 1-12, 2017, doi 10.1016/j.jnca.2017.02.010.
- [33] G. C. Chasparis, "Stochastic stability analysis of perturbed learning Automata with constant step-size in strategic-form games," in *Proc. ACC*, Seattle, WA, USA, 2017, pp. 4607-4612.
- [34] N. W. Prins, J. C. Sanchez and A. Prasad, "Feedback for reinforcement learning based brain-machine interfaces using confidence metrics," *J of Neural Eng*, 2017, doi 10.1088/1741-2552/aa6317.
- [35] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Trans on Neural Networks and Learning Systems*, vol. 28, no. 3, 2017, doi 10.1109/TNNLS.2016.2582849.
- [36] R. Yousefian, et al, "Hybrid transient energy function-based real-time optimal wide-area damping controller," *IEEE Trans on Industry Appls*, vol. 53, no. 2, 2017, doi 10.1109/TIA.2016.2624264.
- [37] F. Tatari, M.-B. Naghibi-Sistani, and K. G. Vamvoudakis, "Distributed learning algorithm for non-linear differential graphical games," *Trans of the Inst of Measurement and Control*, pp. 1-10, 2015.
- [38] C. Lu, J. Huang, and J. Gong, "Reinforcement learning for ramp control: an analysis of learning parameters," *Promet – Traffic & Transportation*, vol. 28, no. 4, pp. 371-381, 2016.
- [39] K. G. Vamvoudakis, "Optimal trajectory output tracking control with a Q-learning algorithm," in *Proc of the American Control Conf*, pp. 5752-5757, 2016, doi 10.1109/ACC.2016.7526571.
- [40] P. H. M. Rêgo, et al, "Convergence of the standard RLS method and UDUT factorisation of covariance matrix for solving the algebraic Riccati equation of the DLQR via heuristic approximate dynamic programming," *Intl J of Systems Science*, 2013, doi 10.1080/00207721.2013.844283.
- [41] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans on Neural Networks and Learning Systems*, vol. 25, no. 3, 2014.
- [42] A. Alharbi, A. Al-Dhalaan, and M. Al-Rodhaan, "Q-routing in cognitive packet network routing protocol for MANETs," in *Proc NCTA-2014*, pp. 234-243, 2014, doi 10.5220/0005082902340243.
- [43] R. Yousefian and S. Kamalasadani, "An approach for real-time tuning of cost functions in optimal system-centric wide area controller based on adaptive critic design," in *IEEE PESGM*, 2014, doi 10.1109/PESGM.2014.6939224.
- [44] B. Dong and Y. Li, "Decentralized reinforcement learning robust optimal tracking control for time varying constrained reconfigurable modular robot based on ACI and Q-function," *Mathematical Problems in Eng*, 2013, Art. no. 387817, doi 10.1155/2013/387817.
- [45] C. Teixeira, et al, "Biped locomotion - improvement and adaptation," in *Proc. ICARSC*, Espinho, Portugal, 2014.
- [46] N. T. Luy, et al, "Reinforcement learning-based intelligent tracking control for wheeled mobile robot," *Trans of the Inst of Measurement and Control*, vol. 36, no. 7, pp. 868–877, 2014, doi 10.1177/0142331213509828.
- [47] L. vS. Hager, et al, "Series-parallel approach to on-line observer based neural control of a helicopter system," in *Proc 19th World Congress the Intl Fedn of Autom Cntrl*, Cape Town, South Africa, 2014.
- [48] Q. Wei and D. Liu, "Data-driven neuro-optimal temperature control of water-gas shift reaction using stable iterative adaptive dynamic programming," *IEEE Trans on Industrial Electr*, vol. 61, no. 11, 2014.
- [49] D. Zhao, B. Wang, and D. Liu, "A supervised Actor-Critic approach for adaptive cruise control," *Soft Comput*, vol. 17, pp. 2089–2099, 2013, doi 10.1007/s00500-013-1110-y.
- [50] M. Kashki, et al, "Power system dynamic stability enhancement using optimum design of PSS and static phase shifter based stabilizer," *Arab J Sci Eng*, vol. 38, pp. 637–650, 2013, doi 10.1007/s13369-012-0325-z.
- [51] C. Li, R. Lowe, and T. Ziemke, "Humanoids learning to walk: a natural CPG-actor-critic architecture," *Frontiers in Neurorobotics*, vol. 7, 2013, Art. no. 5, doi 10.3389/fnbot.2013.00005.
- [52] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598-1611, 2012, doi 10.1016/j.automatica.2012.05.074.
- [53] P. Moradi, et al, "Automatic skill acquisition in reinforcement learning using graph centrality measures," *Intelligent Data Analysis*, vol. 16, no. 1, pp. 113-135, 2012, doi 10.3233/IDA-2011-0513.
- [54] S. Bhasin, et al, "Asymptotic tracking by a reinforcement learning-based adaptive critic controller," *J Control Theory Appl*, vol. 9, no. 3, pp. 400-409, 2011, doi 10.1007/s11768-011-0170-8.
- [55] R. Hafner and M. Riedmiller, "Reinforcement learning in feedback control: Challenges and benchmarks from technical process control," *Machine Learning*, vol. 84, pp. 137-169, 2011, doi 10.1007/s10994-011-5235-x.

- [56] N. T. Luy, "Reinforcement learning-based tracking control for wheeled mobile robot," in *IEEE Intl Conf on Systems, Man, and Cybernetics*, Seoul, Korea, 2012.
- [57] P. Shih, B. C. Kaul, S. Jagannathan, and J. A. Drallmeier, "Reinforcement-learning-based output-feedback control of nonstrict nonlinear discrete-time systems with application to engine emission control," *IEEE Trans on Systems, Man, and Cybernetics—Part B: Cybernetics*, vol. 39, no. 5, 2009.
- [58] M. J. L. Boada, et al, "Active roll control using reinforcement learning for a single unit heavy vehicle," *Intl J of Heavy Vehicle Systems*, vol. 16, no. 4, pp. 412-430, 2009, doi 10.1504/IJHVS.2009.027413.
- [59] L. Guo, Y. Zhang, and J.-L. Hu, "Adaptive HVDC supplementary (lamping controller based on reinforcement learning," *Electric Power Automation Equip*, vol. 27, no. 10, pp. 87-91, 2007.
- [60] C.-K. Lin, "A reinforcement learning adaptive fuzzy controller for robots," *Fuzzy Sets and Systems*, vol. 137, no. 3, pp. 339-352, 2003, doi 10.1016/S0165-0114(02)00299-3.
- [61] S. Jagannathan, "Adaptive critic neural network-based controller for nonlinear systems," in *Proc 2002 IEEE Intl Symp on Intelligent Control*, Vancouver, Canada, 2002.
- [62] B. H. Kaygisiz, A. M. Erkmén, and I. Erkmén, "Smoothing stability roughness of fractal boundaries using reinforcement learning," in *IFAC Procs Vols*, vol. 15, no. 1, pp. 481-485, 2002.
- [63] J. N. Li, et al, "Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes," *IEEE Trans on Cybernetics*, vol. 50, no. 9, pp. 4132-4145, 2020, doi 10.1109/TCYB.2019.2950262.
- [64] K. Zhang, H. G. Zhang, Y. L. Cai, and R. Su, "Parallel optimal tracking control schemes for mode-dependent control of coupled Markov jump systems via integral RL method," *IEEE Trans on Automation Science and Eng*, vol. 17, no. 3, pp. 1332-1342, 2020, doi 10.1109/TASE.2019.2948431.
- [65] Q. Zhang, et al, "Route planning and power management for phev's with reinforcement learning," *IEEE Trans on Vehicular Tech*, vol. 69, no. 5, pp. 4751-4762, 2020, doi 10.1109/TVT.2020.2979623.
- [66] A. Serhani, et al, "AQ-Routing: mobility-, stability-aware adaptive routing protocol for data routing in MANET-IoT systems," *Cluster Comp*, v 23, no. 1, pp. 13-27, 2020, doi 10.1007/s10586-019-02937-x.
- [67] Z. Dong, et al, "Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system," *Appl Energy*, vol. 259, 2020, doi 10.1016/j.apenergy.2019.114193.
- [68] Q. Wang, "Integral reinforcement learning control for a class of high-order multivariable nonlinear dynamics with unknown control coefficients," *IEEE Access*, vol. 8, pp. 86223-86229, 2020, doi 10.1109/ACCESS.2020.2993265.
- [69] J. Zhang, Z. Peng, J. Hu, Y. Zhao, R. Luo, B. K. Ghosh, "Internal reinforcement adaptive dynamic programming for optimal containment control of unknown continuous-time multi-agent systems," *Neurocomputing*, vol. 413, pp. 85-95, 2020, doi 10.1016/j.neucom.2020.06.106.
- [70] A. Mitriakov, et al, "Staircase traversal via reinforcement learning for active reconfiguration of assistive robots," in *Proc IEEE Intl Conf on Fuzzy Systems*, 2020, doi 10.1109/FUZZ48607.2020.9177581.
- [71] Y. Lv, et al, "Online Nash-optimization tracking control of multi-motor driven load system with simplified RL scheme," *ISA Trans*, vol. 98, pp. 251-262, 2020, doi 10.1016/j.isatra.2019.08.025.
- [72] C. E. Thornton, et al, "Deep reinforcement learning control for radar detection and tracking in congested spectral environments," *IEEE Trans on Cognitive Commun and Networking*, 2020, doi 10.1109/TCCN.2020.3019605.
- [73] J. D. Prasanna, et al, "Reinforcement learning based virtual backbone construction in manet using connected dominating sets," *J of Critical Reviews*, vol. 7, no. 9, pp. 146-152, 2020, doi 10.31838/jcr.07.09.28.
- [74] J. Pongfai, X. Su, H. Zhang, and W. Assawinchaichote, "PID controller autotuning design by a deterministic Q-SLP algorithm," *IEEE Access*, vol. 8, pp. 50010-50021, 2020, doi 10.1109/ACCESS.2020.2979810.
- [75] S. Hoppe and M. Toussaint, "Q graph-bounded Q-learning: stabilizing model-free O-policy deep reinforcement learning," *arXiv:2007.07582v1 [cs.LG]* 15 Jul 2020.
- [76] P. Osinenko, L. Beckenbach, T. Göhrt, and S. Streif, "A reinforcement learning method with closed-loop stability guarantee," *arXiv:2006.14034v1 [math.OC]* 24 Jun 2020.
- [77] M. Han, L. Zhang, J. Wang, and W. Pan, "Actor-critic reinforcement learning for control with stability guarantee," *arXiv:2004.14288v3 [cs.RO]* 15 Jul 2020.

- [78] S. A. Khader, H. Yin, P. Falco and D. Kragic, "Stability-guaranteed reinforcement learning for contact-rich manipulation," arXiv:2004.10886v2 [cs.RO] 27 Sep 2020.
- [79] M. Han, Y. Tian, L. Zhang, J. Wang, and W. Pan, "H infinity model-free reinforcement learning with robust stability guarantee," arXiv:1911.02875v3 [cs.LG], 2019.
- [80] C. Tessler, N. Merlis, and S. Mannor, "Stabilizing deep reinforcement learning with conservative updates," arXiv:1910.01062v2 [cs.LG], 2019.
- [81] N. Abuzainab, et al, "QoS and jamming-aware wireless networking using deep reinforcement learning," arXiv:1910.05766v1 [cs.NI] 13 October 2019.
- [82] M. Ahn, "ROBEL: Robotics benchmarks for learning with low-cost robots," arXiv:1909.11639v3 [cs.RO] 16 Dec 2019.
- [83] V. Dhiman, S. Banerjee, B. Griffin, J. M. Siskind, and J. J. Corso, "A critical investigation of deep reinforcement learning for navigation," arXiv:1802.02274v2 [cs.RO], 2018.
- [84] N. Naderializadeh, et al, "When multiple agents learn to schedule: a distributed radio resource management framework," arXiv:1906.08792v1 [cs.LG] 20 Jun 2019.
- [85] K. Nguyen, H. Daumé III, and J. Boyd-Graber, "Reinforcement learning for bandit neural machine translation with simulated human feedback," arXiv:1707.07402v4 [cs.CL] 11 Nov 2017.
- [86] N. Talele and K. Byl, "Mesh-based tools to analyze deep reinforcement learning policies for underactuated biped locomotion," arXiv:1903.12311v2 [cs.RO] 1 Nov 2019.
- [87] M. Turchetta, A. Krause, and S. Trimpe, "Robust model-free reinforcement learning with multi-objective Bayesian optimization," arXiv:1910.13399v1 [cs.RO] 29 Oct 2019.
- [88] Y. Yuan and K. Kitani, "Ego-pose estimation and forecasting as real-time PD control," arXiv:1906.03173v2 [cs.CV] 4 Aug 2019.
- [89] B. Muneeswari and M. S. K. Manikandan, "Energy efficient clustering and secure routing using reinforcement learning for three-dimensional mobile ad hoc networks," IET Commun, vol. 13, no. 12, pp. 1828-1839, 2019, doi 10.1049/iet-com.2018.6150.
- [90] B. Zhao, et al, "Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration," IEEE Trans on Sys Man Cybernetics-Syst, vol. 48, no. 10, 2018.
- [91] Y. Zhang, et al, "Optimal design of residual-driven dynamic compensator using iterative algorithms with guaranteed convergence," IEEE Trans on Systems, Man, and Cybernetics: Systems, vol. 46, no. 4, 2016, doi 10.1109/TSMC.2015.2450203.
- [92] M. Tokic, "Adaptive epsilon-greedy exploration in reinforcement learning based on value differences," in Lecture Notes in Artificial Intelligence, 33rd Annual German Conf on AI, Karlsruhe, Germany, 2010.
- [93] Y. Xiong, L. Guo, Y. Huang, and L. Chen, "Intelligent thermal control strategy based on reinforcement learning for space telescope," J of Thermophysics and Heat Transfer, vol. 34, no. 1, pp. 37-44, 2020.
- [94] R. F. Isa-Jara, G. J. Meschino, and V. L. Ballarin, "A comparative study of reinforcement learning algorithms applied to medical image registration," in IFMBE Procs, pp. 281-289, 2020.
- [95] F. Guo, et al, "A reinforcement learning decision model for online process parameters optimization from offline data in injection molding," Applied Soft Computing J, vol. 85, 2019, doi 10.1016/j.asoc.2019.105828.
- [96] S. Li, et al, "Design and implementation of aerial communication using directional antennas: Learning control in unknown communication environments," IET Control Theory and Apps, vol. 13, no. 17, pp. 2906-2916, 2019, doi 10.1049/iet-cta.2018.6252.
- [97] X. Tang, et al, "A deep value-network based approach for multi-driver order dispatching," in Proc of the ACM SIGKDD Intl Conf on Knowl Discovery and Data Mining, pp. 1780-1790, 2019, doi 10.1145/3292500.3330724.
- [98] A. Chowdhury, et al, "DA-DRLS: Drift adaptive deep reinforcement learning based scheduling for IoT resource management," J of Network and Computer Appls, vol. 138, pp. 51-65, 2019, doi 10.1016/j.jnca.2019.04.010.
- [99] X. Wang, et al, "UAV first view landmark localization with active reinforcement learning," Pattern Recognition Letters, vol. 125, pp. 549-555, 2019, doi 10.1016/j.patrec.2019.03.011.
- [100] B. Lütjens, et al, "Safe reinforcement learning with model uncertainty estimates," in Procs IEEE Intl Conf on Robotics and Automation, pp. 8662-8668, 2019, doi 10.1109/ICRA.2019.8793611.

- [101] A. Balakrishnan and J. V. Deshmukh, "Structured reward functions using STL," in Proc of the 2019 22nd ACM Intl Conf on Hybrid Systems: Comput and Control, pp. 270-271, 2019, doi 10.1145/3302504.3313355.
- [102] C. Tang, W. Zhu, and X. Yu, "Deep hierarchical strategy model for multi-source driven quantitative investment," IEEE Access, vol. 7, pp. 79331-79336, 2019, doi 10.1109/ACCESS.2019.2923267.
- [103] Q. Cheng, X. Wang, Y. Niu, and L. Shen, "Reusing source task knowledge via transfer approximator in reinforcement transfer learning," Symmetry, vol. 11, no. 1, 2019, doi 10.3390/sym11010025.
- [104] Y.-S. Jeon, H. Lee, and N. Lee, "Robust MLSD for wideband SIMO systems with one-bit ADCs: reinforcement-learning approach," in Proc ICC Workshops, pp. 1-6, 2018, doi 10.1109/ICCW.2018.8403665, 2018.
- [105] X. Yang and H. He, "Self-learning robust optimal control for continuous-time nonlinear systems with mismatched disturbances," Neural Networks, vol. 99, pp. 19-30, 2018, doi 10.1016/j.neunet.2017.11.022.
- [106] H. Jiang, H. Zhang, Y. Cui, and G. Xiao, "Robust control scheme for a class of uncertain nonlinear systems with completely unknown dynamics using data-driven reinforcement learning method," Neurocomputing, vol. 273, pp. 68-77, 2018, doi 10.1016/j.neucom.2017.07.058.
- [107] H. Shayeghi and A. Younesi, "An online Q-learning based multi-agent LFC for a multi-area multi-source power system including distributed energy resources," Iranian J of Electrical and Electronic Engineering, vol. 13, no. 4, pp. 385-398, 2017, doi 10.22068/IJEEE.13.4.385.
- [108] D. Zhao, Y. Ma, Z. Jiang, and Z. Shi, "Multiresolution airport detection via hierarchical reinforcement learning saliency model," IEEE J of Selected Topics in Applied Earth Observ and Remote Sensing, vol. 10, no. 6, pp. 2855-2866, 2017, doi 10.1109/JSTARS.2017.2669335.
- [109] A. W. Tow, S. Shirazi, J. Leitner., N. Sünderhauf, M. Milford, and B. Upcroft, "A robustness analysis of deep Q networks," in Australasian Conf on Robotics and Automation, pp. 116-125, 2016.
- [110] E. Hatami, and H. Salarieh, "Adaptive critic-based neuro-fuzzy controller for dynamic position of ships," Scientia Iranica, vol. 22, no. 1, pp. 272-280, 2015.
- [111] J. Xiang and Z. Chen, "Adaptive traffic signal control of bottleneck subzone based on grey qualitative reinforcement learning algorithm," in Proc ICPRAM, vol. 2, pp. 295-301, 2015.
- [112] R. Padmanabhan, et al, "Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning," Biomed Signal Process and Control, vol. 22, pp. 54-64, 2015, doi 10.1016/j.bspc.2015.05.013.
- [113] R. Bruno, et al, "Robust adaptive modulation and coding (AMC) selection in LTE systems using reinforcement learning," in IEEE Vehicular Technology Conf, 2014, doi 10.1109/VTCFall.2014.6966162.
- [114] N. Jamali, P. Kormushev, S.R. Ahmadzadeh, and D. G. Caldwell, "Covariance analysis as a measure of policy robustness," in OCEANS, Taipei, Taiwan, 2014, doi 10.1109/OCEANS-TAIPEI.2014.6964339.
- [115] S. Tati, S. Silvestri, T. He, and T. L. Porta, "Robust network tomography in the presence of failures," in Proc Intl Conf on Distributed Computing Systems, pp. 481-492, 2014, doi 10.1109/ICDCS.2014.56.
- [116] N. T. Luy, N. T. Thanh, and H. M. Tri, "Reinforcement learning-based robust adaptive tracking control for multi-wheeled mobile robots synchronization with optimality," in Proc 2013 IEEE Workshop on Robotic Intelligence in Info Structured Space, pp. 74-81, 2013, doi 10.1109/RiiSS.2013.6607932.
- [117] M. Kashki, M. A. Abido, and Y. L. Abdel-Magid, "Power system dynamic stability enhancement using optimum design of PSS and static phase shifter based stabilizer," Arabian J for Science and Engineering, vol. 38, no. 3, pp. 637-650, 2013, doi 10.1007/s13369-012-0325-z.
- [118] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer, "Exploration in model-based reinforcement learning by empirically estimating learning progress," Advances in Neural Info Process Systems, vol. 1, pp. 206-214, 2012.
- [119] M. S. Llorente and S. E. Guerrero, "Increasing retrieval quality in conversational recommenders," IEEE Trans on Knowl and Data Eng, vol. 24, no. 10, pp. 1876-1888, 2012, doi 10.1109/TKDE.2011.116.
- [120] F. Maes, et al, "Learning to play K-armed bandit problems," in 4th Intl Conf on Agents and AI, pp. 74-81, 2012.
- [121] S. Bhasin, et al, "Asymptotic tracking by a reinforcement learning-based adaptive critic controller," J of Control Theory and Apps, vol. 9, no. 3, pp. 400-409, 2011, doi 10.1007/s11768-011-0170-8.

- [122] A. Tjahjadi, et al, "Robustness analysis of genetic network programming with reinforcement learning," in Proc Jt 5th Intl Conf on Soft Comp and Intell Sys and 11th Intl Symp on Advanced Intelligent Systems, pp. 594-601, 2010.
- [123] S. A. Kulkarni and G. R. Rao, "Vehicular ad hoc network mobility models applied for reinforcement learning routing algorithm," Comms in Computer and Info Science, pp. 230-240, 2010, doi 10.1007/978-3-642-14825-5\_20.
- [124] C. Molina, et al, "Maximum entropy-based reinforcement learning using a confidence measure in speech recognition for telephone speech," IEEE Trans on Audio, Speech and Language Processing, vol. 18, no. 5, pp. 1041-1052, 2010, doi 10.1109/TASL.2009.2032618.
- [125] N. T. Luy, et al, "Robust reinforcement learning-based tracking control for wheeled mobile robot," in 2nd Intl Conf on Computer and Automation Eng, pp. 171-176, 2010, doi 10.1109/ICCAE.2010.5451973.
- [126] V. Heidrich-Meisner and C. Igel, "Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search," in Proc of the 26th Intl Conf on Machine Learning, pp. 401-408, 2009.
- [127] H. Satoh, "A nonlinear approach to robust routing based on reinforcement learning with state space compression and adaptive basis construction," IEICE Trans on Fundamentals of Electronics, Comms and Comp Sci, vol. 7, pp. 1733-1740, 2008, doi 10.1093/ietfec/e91-a.7.1733.
- [128] K. Conn, and R. A. Peters, "Reinforcement learning with a supervisor for a mobile robot in a real-world environment," in Proc of the 2007 IEEE Intl Symp on Computl Intell in Robotics and Automation, pp. 73-78, 2007, doi 10.1109/CIRA.2007.382878.
- [129] X.-S. Wang, et al, "A proposal of adaptive PID controller based on reinforcement learning," J of China Univ of Mining and Tech, vol. 17, no. 1, pp. 40-44, 2007, doi 10.1016/S1006-1266(07)60009-1.
- [130] J.B. Leem, and H. Y. Kim, "Action-specialized expert ensemble trading system with extended discrete action space using deep reinforcement learning," PLOS One, vol. 15, no. 7, 2020, doi 10.1371/journal.pone.0236178.
- [131] Y. Xiong, et al, "Intelligent thermal control strategy based on reinforcement learning for space telescope," J of Thermophysics and Heat Transfer, vol. 34, no. 1, pp. 37-44, 2020, doi 10.2514/1.T5774.
- [132] A. Balakrishnan and J. V. Deshmukh, "Structured reward functions using STL," Proc HSCC '19, pp. 270-271, 2019. doi 10.1145/3302504.3313355.
- [133] G. Chen, et al, "Distributed non-communicating multi-robot collision avoidance via map-based deep reinforcement learning," Sensors, vol. 20, no. 17, 2020, doi 10.3390/s20174836.
- [134] C. Sun, X. Li, and C. Belta, "Automata guided semi-decentralized multi-agent reinforcement learning," in Proc of the American Control Conf, pp. 3900-3905, 2020, doi 10.23919/ACC45564.2020.9147704.
- [135] X. Wang and X. Ye, "Optimal robust control of nonlinear uncertain system via off-policy integral reinforcement learning," in Proc of the Chinese Control Conf, pp. 1928-1933, 2020, doi 10.23919/CCC50068.2020.9189626.
- [136] Z. Yan, J. Ge, Y. Wu, L. Li, and T. Li, "Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks," IEEE J on Selected Areas in Commun, vol. 38, no. 6, pp. 1040-1057, 2020. doi 10.1109/JSAC.2020.2986662.
- [137] K. Alhazmi and S. M. Sarathy, "Continuous control of complex chemical reaction network with reinforcement learning," in Proc ECC, pp. 1066-1068, 2020.
- [138] A. Ghasemkhani, et al, "DeepGrid: robust deep reinforcement learning-based contingency management," in Proc IGST, 2020, doi 10.1109/ISGT45199.2020.9087633.
- [139] A. Pitti, M. Quoy, C. Lavandier, and S. Boucenna, "Gated spiking neural network using Iterative Free-Energy Optimization and rank-order coding for structure learning in memory sequences (INFERNO GATE)," Neural Networks, vol. 121, pp. 242-258, 2020, doi 10.1016/j.neunet.2019.09.023.
- [140] M. Vecerik, et al, "S3K: self-supervised semantic keypoints for robotic manipulation via multi-view consistency," arXiv:2009.14711, 2020.
- [141] Y. Jiao, et al, "Learning to swim in potential flow," arXiv:2009.14280, 2020.
- [142] P. Almási, R. Moni, and B. Gyires-Tóth, "Robust reinforcement learning-based autonomous driving agent for simulation and real world," arXiv:2009.11212, 2020.

- [143]W. Ding, B. Chen, B. Li, K. J. Eun, and D. Zhao, “Multimodal safety-critical scenarios generation for decision-making algorithms evaluation,” arXiv:2009.08311, 2020.
- [144]G. Schamberg, M. Badgeley, and E. N. Brown, “Controlling level of unconsciousness by titrating propofol with deep reinforcement learning,” arXiv:2008.12333, 2020.
- [145]B. Pang and Z.-P. Jiang, “Robust reinforcement learning: a case study in linear quadratic regulation,” arXiv:2008.11592, 2020.
- [146]T. Kobayashi and W. E. L. Ilboudo, “t-Soft update of target network for deep reinforcement learning,” arXiv:2008.10861, 2020.
- [147]A. Zavoli and L. Federici, “Reinforcement learning for low-thrust trajectory design of interplanetary missions,” arXiv:2008.08501, 2020.
- [148]O. Limoyo, et al, “Heteroscedastic uncertainty for robust generative latent dynamics,” arXiv:2008.08157, 2020.
- [149]W. Zhao, J. P. Queralt, L. Qingqing, and T. Westerlund, “Towards closing the sim-to-real gap in collaborative multi-robot deep reinforcement learning,” arXiv:2008.07875, 2020.
- [150]X. Qu, Y.-S. Ong, A. Gupta, and Z. Sun, “Defending adversarial attacks without adversarial attacks in deep reinforcement learning,” arXiv:2008.06199, 2020.
- [151]P. Swazinna, S. Udluft, and T. Runkler, “Overcoming model bias for robust offline deep reinforcement learning,” arXiv:2008.05533, 2020.
- [152]I. Ahmed, H. Khorasgani, and G. Biswas, “Comparison of model predictive and reinforcement learning methods for fault tolerant control,” arXiv:2008.04403, 2020.
- [153]G. Kovač, A. Laversanne-Finot, and P.-Y. Oudeyer, “GRIMGEP: learning progress for robust goal sampling in visual deep reinforcement learning,” arXiv:2008.04388, 2020.
- [154]J. L. Zhu, et al, “Adversarial directed graph embedding,” arXiv:2008.03667, 2020.
- [155]X. Ma, S. Chen, D. Hsu, and W. S. Lee, “Contrastive variational model-based reinforcement learning for complex observations,” arXiv:2008.02430, 2020.
- [156]T. Oikarinen, et al, “Robust deep reinforcement learning through adversarial loss,” arXiv:2008.01976, 2020.
- [157]E. Vinitsky, et al, “Robust reinforcement learning using adversarial populations,” arXiv:2008.01825, 2020.
- [158]H. Park, et al, “Understanding the stability of deep control policies for biped locomotion,” arXiv:2007.15242, 2020.
- [159]K. Steverson, J. Mullin, and M. Ahiskali, “Adversarial robustness for machine learning cyber defenses using log data,” arXiv:2007.14983, 2020.
- [160]X. Chen, et al, “Same-day delivery with fairness,” arXiv:2007.09541, 2020.
- [161]X. Chen, Y. Duan, Z. Chen, H. Xu, Z. Chen, X. Liang, T. Zhang, and Z. Li, “CATCH: context-based meta reinforcement learning for transferrable architecture search,” arXiv:2007.09380, 2020.
- [162]L. Zhang, H. Xiong, O. Ma, and Z. Wang, “Multi-robot cooperative object transportation using decentralized deep reinforcement learning,” arXiv:2007.09243, 2020.
- [163]K. L. Tan, Y. Esfandiari, X. Y. Lee, Aakanksha, and S. Sarkar, “Robustifying reinforcement learning agents via action space adversarial training,” arXiv:2007.07176, 2020.
- [164]A. Stooke, et al, “Responsive safety in RL by PID Lagrangian methods,” arXiv:2007.03964, 2020.
- [165]K. Abe and Y. Kaneko, “Off-policy exploitability-evaluation and equilibrium-learning in two-player zero-sum Markov games,” arXiv:2007.02141, 2020.
- [166]X. Wang, et al, “Falsification-based robust adversarial reinforcement learning,” arXiv:2007.00691, 2020.
- [167]H. Lee, M. Girnyk, and J. Jeong, “Deep reinforcement learning approach to MIMO precoding problem: optimality and robustness,” arXiv:2006.16646, 2020.
- [168]D. Xu, M. Agarwal, E. Gupta, F. Fekri, and R. Sivakumar, “Accelerating reinforcement learning agent with eeg-based implicit human feedback,” arXiv:2006.16498, 2020.
- [169]L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, “Multi-agent deep reinforcement learning for hvac control in commercial buildings,” arXiv:2006.14156, 2020.
- [170]A. Gleave, et al, “Quantifying differences in reward functions,” arXiv:2006.13900, 2020.
- [171]R. Raileanu, M. Goldstein, D. Yarats, I. Kostrikov, and R. Fergus, “Automatic data augmentation for generalization in deep reinforcement learning,” arXiv:2006.12862, 2020.
- [172]H. Liu and W. Wu, “Online multi-agent reinforcement learning for decentralized inverter-based voltage control,” arXiv:2006.12841, 2020.
- [173]Y. Zou and X. Lu, “Gradient-EM Bayesian meta-learning,” arXiv:2006.11764, 2020.



- [174]K. Panaganti and D. Kalathil, "Model-free robust reinforcement learning with linear function approximation," arXiv:2006.11608, 2020.
- [175]A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," arXiv:2006.10742, 2020.
- [176]A. Rahman, et al, "Open ad hoc teamwork using graph-based policy learning," arXiv:2006.10412, 2020.
- [177]H. Jeong, et al, "Learning to track dynamic targets in partially known environments," arXiv:2006.10190, 2020.
- [178]K.-P. Ning and S.-J. Huang, "Reinforcement learning with supervision from noisy demonstrations," arXiv:2006.07808, 2020.
- [179]Y. Dou, et al, "Robust spammer detection by nash reinforcement learning," arXiv:2006.06069, 2020.
- [180]X. Huang, F. Zhu, L. Holloway, and A. Haidar, "Causal discovery from incomplete data using an encoder and reinforcement learning," arXiv:2006.05554, 2020.
- [181]Y. Chow, et al, "Variational model-based policy optimization," arXiv:2006.05443, 2020.
- [182]T. Jafferjee, E. Imani, E. Talvitie, M. White, and M. Bowling, "Hallucinating value: a pitfall of dynamic style planning with imperfect environment models," arXiv:2006.04363, 2020.
- [183]Y. Tian, et al, "Real-time model calibration with deep reinforcement learning," arXiv:2006.04001, 2020.
- [184]N. Kallus and M. Uehara, "Efficient evaluation of natural stochastic policies in offline reinforcement learning," arXiv:2006.03886, 2020.
- [185]L. Hou, et al, "Robust reinforcement learning with Wasserstein constraint," arXiv:2006.00945, 2020.
- [186]J. Zhi and J.-M. Lien, "Learning to herd agents amongst obstacles: training robust shepherding behaviors using deep reinforcement learning," arXiv:2005.09476, 2020.
- [187]Y. Chandak, et al, "Optimizing for the future in non-stationary MDPs," arXiv:2005.08158, 2020.
- [188]Y. Ding, et al, "Mutual information maximization for robust plannable representations," arXiv:2005.08114, 2020.
- [189]S. Totaro, I. Boukas, A. Jonsson, and B. Cornélusse, "Lifelong control of off-grid microgrid with model based reinforcement learning," arXiv:2005.08006, 2020.
- [190]Z. Xie, et al, "ALLSTEPS: curriculum-driven learning of stepping stone skills," arXiv:2005.04323, 2020.
- [191]R. Singh, Q. Zhang, and Y. Chen, "Improving robustness via risk averse distributional reinforcement learning," arXiv:2005.00585, 2020.
- [192]I. Kostrikov, D. Yarats, and R. Fergus, "Image augmentation is all you need: regularizing deep reinforcement learning from pixels," arXiv:2004.13649, 2020.
- [193]J. Z. Chen, "Reinforcement learning generalization with surprise minimization," arXiv:2004.12399, 2020.
- [194]P. D. Ngo and F. Godtlielsen, "Data-driven robust control using reinforcement learning," arXiv:2004.07690, 2020.
- [195]M. Everett, et al, "Certified adversarial robustness for deep reinforcement learning," arXiv:2004.06496, 2020.
- [196]M. Koren and M. J. Kochenderfer, "Adaptive stress testing without domain heuristics using go-explore," arXiv:2004.04292, 2020.
- [197]B. Anahtarci, et al, "Q-Learning in regularized mean-field games," arXiv:2003.12151, 2020.
- [198]B. Lindenberg, et al, "Distributional reinforcement learning with ensembles," arXiv:2003.10903, 2020.
- [199]Q. Shen, et al, "Deep reinforcement learning with robust and smooth policy," arXiv:2003.09534, 2020.
- [200]H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh, "Robust deep reinforcement learning against adversarial perturbations on state observations," arXiv:2003.08938, 2020.
- [201]X. Guo, et al, "A general framework for learning mean-field games," arXiv:2003.06069, 2020.
- [202]A. Touati, A. A. Taiga, and M. G. Bellemare, "Zooming for efficient model-free reinforcement learning in metric spaces," arXiv:2003.04069, 2020.
- [203]S. Gao, P. Dong, Z. Pan, and G Y. Li, "Reinforcement learning based cooperative coded caching under dynamic popularities in ultra-dense networks," arXiv:2003.03758, 2020.
- [204]J. Lin, K. Dzeparoska, S. Q. Zhang, A. Leon-Garcia, and N Papernot, "On the robustness of cooperative multi-agent reinforcement learning," arXiv:2003.03722, 2020.

- [205]E. Derman, and S. Mannor, “Distributional robustness and regularization in reinforcement learning,” arXiv:2003.02894, 2020.
- [206]T. Spooner, R. Savani, “Robust market making via adversarial reinforcement learning,” arXiv:2003.01820, 2020.
- [207]M. Chancán and M. Milford, “MVP: unified motion and visual self-supervised learning for large-scale robotic navigation,” arXiv:2003.00667, 2020.
- [208]W. E. L. Ilboudo, et al, “TAdam: a robust stochastic gradient optimizer,” arXiv:2003.00179, 2020.
- [209]A. Tschantz, et al, “Reinforcement learning through active inference,” arXiv:2002.12636, 2020.
- [210]S. Kuutti, et al, “Training adversarial agents to exploit weaknesses in deep cntrl policies,” arXiv:2002.12078, 2020.
- [211]N. D. Nguyen, T. T. Nguyen, and S. Nahavandi, “A visual communication map for multi-agent deep reinforcement learning,” arXiv:2002.11882, 2020.
- [212]C.-H. H. Yang, J. Qi, P.-Y. Chen, Y. Ouyang, I-T. D. Hung, C.-H. Lee, and X. Ma, “Enhanced adversarial strategically-timed attacks against deep reinforcement learning,” arXiv:2002.09027, 2020.
- [213]T. Sun, et al “Adaptive temporal difference learning with linear function approximation,” arXiv:2002.08537, 2020.
- [214]N. Naderializadeh, J. Sydir, M. Simsek, and H. Nikopour, “Resource management in wireless networks via multi-agent deep reinforcement learning,” arXiv:2002.06215, 2020.
- [215]P. Kamalaruban, Y.-T. Huang, Y.-P. Hsieh, P. Rolland, C. Shi, and V. Cevher, “Robust reinforcement learning via adversarial training with langevin dynamics,” arXiv:2002.06063, 2020.
- [216]N. Kallu, and M. Uehara, “Statistically efficient off-policy policy gradients,” arXiv:2002.04014, 2020.
- [217]G. Lee, B. Hou, S. Choudhury, and S. S. Srinivasa, “Bayesian residual policy optimization: scalable Bayesian reinforcement learning with clairvoyant experts,” arXiv:2002.03042, 2020.
- [218]V. Pacelli and A. Majumdar, “Learning task-driven control policies via information bottlenecks,” arXiv:2002.01428, 2020.
- [219]J. Yao, et al, “Policy gradient based quantum approx optimization algorithm,” arXiv:2002.01068, 2020.
- [220]D. Nishio, et al, “Discriminator soft actor critic without extrinsic rewards,” arXiv:2001.06808, 2020.
- [221]T. Dai, K. Arulkumaran, T. Gerbert, S. Tukra, F. Behbahani, and A. A. Bharath, “Analysing deep reinforcement learning agents trained with domain randomisation,” arXiv:1912.08324, 2019.
- [222]X. Zhang, J. Liu, X. Xu, S. Yu, and H. Chen, “Learning-based predictive control for nonlinear systems with unknown dynamics subject to safety constraints,” arXiv:1911.09827, 2019.
- [223]T. Lykouris, et al, “Corruption robust exploration in episodic reinforcement learning,” arXiv:1911.08689, 2019.
- [224]S. Salter, et al, “Attention-privileged reinforcement learning,” arXiv:1911.08363, 2019.
- [225]M. Han, et al, “ $H_\infty$  model-free reinforcement learning with robust stability guarantee,” arXiv:1911.02875, 2019.
- [226]B. Lütjens, et al, “Certified adversarial robustness for deep reinforcement learning,” arXiv:1910.12908, 2019.
- [227]M. Uehara, et al, “Minimax Weight and Q-Function Learning for Off-Policy Evaluation,” arXiv:1910.12809, 2019.
- [228]S. Li, and O. Bastani, “Robust model predictive shielding for safe reinforcement learning with stochastic dynamics,” arXiv:1910.10885, 2019.
- [229]R. B. Slaoui, et al, “Robust visual domain randomization for reinforcement learning,” arXiv:1910.10537, 2019.
- [230]K. Zhang, B. Hu, and T. Başar, “Policy optimization for  $H_2$  linear control with  $H_\infty$  robustness guarantee: implicit regularization and global convergence,” arXiv:1910.09496, 2019.
- [231]Z. Liu, et al, “Regularization matters in policy optimization,” arXiv:1910.09191, 2019.
- [232]J. Yang, et al, “Single episode policy transfer in reinforcement learning,” arXiv:1910.07719, 2019.
- [233]S. Chen, et al, “Zap q-learning with nonlinear function approximation,” arXiv:1910.05405, 2019.
- [234]E. Schwartz, G. Tennenholtz, C. Tessler, and S. Mannor, “Language is power: representing states using natural language in reinforcement learning,” arXiv:1910.02789, 2019.
- [235]D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau, and R. Fergus, “Improving sample efficiency in model-free reinforcement learning from images,” arXiv:1910.01741, 2019.
- [236]G. Kalweit, M. Huegle, and J. Boedecker, “Composite q-learning: multi-scale q-function decomposition and separable optimization,” arXiv:1909.13518, 2019.

- [237]M. Ryu, et al, “CAQL: continuous action q-learning,” arXiv:1909.12397, 2019.
- [238]J. Li, et al, “Multi-task batch reinforcement learning with metric learning,” arXiv:1909.11373, 2019.
- [239]M. Shen, and J. P. How, “Robust opponent modeling via adversarial ensemble reinforcement learning in asymmetric imperfect-information games,” arXiv:1909.08735, 2019.
- [240]N. Kallus and M. Uehara. “Double reinforcement learning for efficient off-policy evaluation in Markov decision processes,” arXiv:1908.08526, 2019.
- [241]J. Roy, P. Barde, F. G. Harvey, D. Nowrouzezahrai, and C. Pal, “Promoting coordination through policy regularization in multi-agent deep reinforcement learning,” arXiv:1908.02269, 2019.
- [242]Y. Urakami, A. Hodgkinson, C. Carlin, R. Leu, L. Rigazio, and P. Abbeel, “DoorGym: A scalable door opening environment and baseline agent,” arXiv:1908.01887, 2019.
- [243]Q. Wang, K. Feng, X. Li, and S. Jin, “PrecoderNet: hybrid beamforming for millimeter wave systems with deep reinforcement learning,” arXiv:1907.13266, 2019.
- [244]M. Bogdanovic, et al, “Learning variable impedance control for contact sensitive tasks,” arXiv:1907.07500, 2019.
- [245]D. J. Mankowitz, et al, “Robust reinforcement learning for continuous control with model misspecification,” arXiv:1906.07516, 2019.
- [246]A. C. Li, et al, “Sub-policy adaptation for hierarchical reinforcement learning,” arXiv:1906.05862, 2019.
- [247]M. Assran, J. Romoff, N. Ballas, J. Pineau, and M. Rabbat, “Gossip-based actor-learner architectures for deep reinforcement learning,” arXiv:1906.04585, 2019.
- [248]B. Gravell, P. M. Esfahani, and T. Summers, “Learning robust control for LQR systems with multiplicative noise via policy gradient,” arXiv:1905.13547, 2019.
- [249]A. Francis, A. Faust, H.-T. L. Chiang, J. Hsu, J. C. Kew, M. Fiser, and T.-W. E. Lee, “Long-range indoor navigation with PRM-RL,” arXiv:1902.09458, 2019.
- [250]J. Wang, Y. Liu, and B. Li. “Reinforcement learning with perturbed rewards,” arXiv:1810.01032, 2018.
- [251]D. Moher, et al, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” PLoS Med, vol. 6, no. 7, e1000097, 2009, DOI 10.1371/journal.pmed1000097.
- [252]Y. Liu, Y., et al, “Stein Variational Policy Gradient,” arXiv:1704.02399v1 [cs.LG], April 2017.

## AUTHORS

**LAURA L. PULLUM** (IEEE M’86–SM’03) received the B.S. degree in mathematics from the University of Alabama in Huntsville (UAH) in 1983, the M.S. degree in operations research from UAH in 1987, the MBA from the Southeastern Institute of Technology (SIT), Huntsville, AL in 1990, the D.Sc. in systems engineering and operations research from SIT in 1992, and the M.S. in geology from the University of Tennessee (Knoxville) in 2015. Since 1982, she worked in industry (large and small businesses), non-profit research institutes, and academia. She was most recently a Senior Research Scientist at Oak Ridge National Laboratory, Oak Ridge, TN, USA. She is the author of 2 books (Software Fault Tolerance Techniques and Implementation, Artech House, 2001 and Guidance for the Verification and Validation of Neural Networks, Wiley, 2007), numerous book chapters and hundreds of articles/papers. Throughout her career, she has conducted research and development to enhance and ensure the dependability of intelligent and complex systems, including those incorporating machine learning, artificial intelligence and autonomy. Her current research is in the robustness and confidence of classifiers of non-traditional images and signals.



Dr. Pullum has received several best paper awards and certificates of appreciation from the software and artificial intelligence standards development organizations on which she has served.

# A CRYPTOCURRENCY ANALYSIS TOOL BASED ON SOCIAL METRICS

Bill Xu<sup>1</sup>, Yu Sun<sup>2</sup>

<sup>1</sup>École Internationale de Montréal, 11 Cm. de la Côte-Saint-Antoine,  
Westmount, QC H3Y 2H

<sup>2</sup>California State Polytechnic University, Pomona, CA, 91768,  
Irvine, CA 92620

## **ABSTRACT**

*Recent years have witnessed the dramatic popularity of cryptocurrencies, in which millions invest to join the cryptocurrency community or make financial gains [1]. Investors employ many ways to analyze a cryptocurrency, from a purely technical approach to a more utility-centred approach [2]. However, few technologies exist to help investors find cryptocurrencies with bright prospects through social metrics, an equally if not more important viewpoint to consider due to the importance of communities in the space. This paper proposes an application to evaluate cryptocurrencies based on social metrics by establishing scores and models with machine learning and other tools [3]. We verified the need for our application through surveys, applied it to test investment strategies, and conducted a qualitative evaluation of the approach. The results show that our tool benefits investors by providing them with a different lens to view cryptocurrencies and helps them make more thorough decisions.*

## **KEYWORDS**

*Cryptocurrencies, Machine learning, Analysis, Application*

## **1. INTRODUCTION**

The rise in popularity of cryptocurrencies, like Bitcoin, Ethereum and Dogecoin, has attracted much attention from investors worldwide who want a piece of that pie [4]. However, the lack of accessible resources and analytical tools to help investors identify cryptocurrencies with bright prospects left many clueless. There are numerous cryptocurrencies in the market, and without a proper analytical tool, it is challenging for an investor to detect opportunities and make proper commitments [5]. To better understand our target audience and consumer opinion, we conducted a customer outreach survey on online forums and within our communities. In total, it received 20 responses. The majority of the respondents (70%) felt the need for a cryptocurrency analysis tool and stated that they would like to know the future direction of each cryptocurrency (80%). Furthermore, half of the respondents have trouble finding the right cryptocurrency to invest in, and 70% think a cryptocurrency's community is essential to its success. For these reasons, we decided to develop Retrospect, a free and easy-to-use app designed to provide users with the latest cryptocurrency data and analysis based on social metrics, such as Twitter activity or GitHub commits, to determine a crypto's quality and prospects [6]. This solution will allow investors to analyze each cryptocurrency through community activity and an overview of the cryptocurrency's predicted direction. In our survey, 50% of the respondents would use Retrospect as their cryptocurrency analysis tool, with the other half being undecided.

The most important tools for cryptocurrencies analysis in the current market can be easily separated into three categories: charting tools, such as Tradingview, market data tools, such as Coinmarketcap, research reporting platforms, like cryptoresearch.report, network statistics tools, like BitcoinVisuals, and news aggregators, like CryptoPanic. Despite their efficiency and quality, these tools share one major issue: they only provide information to investors and expect them to come to conclusions themselves. However, the market lacks tools that come up with analysis results for the investor, like Retrospect. The current tools assume that investors want to analyze everything and do the job alone, which is often not the case. The rare analytical tools currently on the market offering similar services to Retrospect either fail to implement a consistent model (due to them employing different analysts) or do not directly compete with Retrospect by putting little to no emphasis on social metrics to give out prospects. Furthermore, tools with a plethora of features using highly sophisticated algorithms fail to address fundamental investor concerns, and their complicated interface makes them even less attractive to regular people. These critical issues make the current cryptocurrency analysis apps unattractive to investors.

Our tool, as stated, is a free-to-use application providing cryptocurrency data and analysis based on social metrics. In this paper, our goal is to explain the functionality of our app and our process of determining the perfect model to fit our data and give us prediction results. There are some excellent features of Retrospect. First, the user interface is straightforward, making navigation the slightest concern for our users. Second, our app provides a RETRO-SCORE<sup>©</sup> for each cryptocurrency, guiding users to understand each cryptocurrency's social state better and helping them make better investment decisions. Third, we provide our users with a market view score, allowing them to determine whether the community is currently optimistic or pessimistic. Fourth, we help our users determine the predicted price movement of the cryptocurrency in the next twenty-four hours based on our model. Compared to most tools available, we help investors do their job. Compared to similar apps, our analytical focus and approach set us apart, and compared to sophisticated tools, our easy UI and features will attract more users [7]. These features are our strengths and will allow us to provide users with the best experience possible. Therefore, we believe that Retrospect has its place in the cryptocurrency world.

In three application scenarios, we demonstrate how the above combination of techniques increases investors' ability to make better investment decisions and confidence. First, we show the usefulness of our approach by surveying our app testers to determine whether our application helps them solve some of their pains as investors: 1) Lack of analytical tools, 2) Hard to use the application, 3) Hard to understand analytical results. Second, we determined the model with the highest accuracy that will fit our data and give us predictions through a series of tests with different quantities of data. Third, we tested our model's "real" accuracy by back testing it and comparing the model's output with what happened.

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 conclusions remarks, as well as points out this project's future work.

## **2. CHALLENGES**

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

### **2.1. Establishing an Analytical Lens**

It is challenging to distinguish ourselves from others in finance and cryptocurrency if I do not have a different approach to the problem. That is why one of this project's significant challenges was finding a unique and adequate approach to cryptocurrency analysis [8]. After considering the factors that make a cryptocurrency successful, like usefulness, adoptability, and scalability, we realized that another underutilized factor also plays an essential role in a cryptocurrency's price movement: the community. After sufficient research, I concluded that it was indeed possible to develop a thorough model of a cryptocurrency through social metrics by measuring different factors: Tweet count, Tweets polarity, and GitHub commits.

Furthermore, I remarked that only a few tools considered social metrics, with little emphasis on them. Therefore, I concluded that having a community-focused analytical view is the best choice for a complete and unique model.

### **2.2. Finding a suitable model for the Data**

Another challenge was to find the right way to model the data. Indeed, with so many factors to consider in our final model (tweet count change, commits change, market cap rank, price, and price change 24h). During our experimentation phase, I struggled even to have a model with accuracy or  $R^2$  higher than 0. After countless tests and research, I finally found a suitable model yielding satisfactory results. As shown in section 4, I tested a wide variety of models from the library scikit-learn, including linear and random forest regression. This was the biggest challenge I faced while creating this project because of its difficulty and essential role in Retrospect.

### **2.3. Building the Application**

In addition, I also faced problems while building the application on Android Studio. Firstly, I struggled to make our application connect to our application backend server (AWS) due to problems with JSON reading. After fixing this, I also encountered issues with app formatting and functionalities like sort. Finally, there was an issue with the light theme not working on many devices. These issues kept me up at night multiple times to resolve them. Because of the time commitment put into fixing these simple yet complicated issues, the application-building process was also an uncomfortable ride.

### 3. SOLUTION

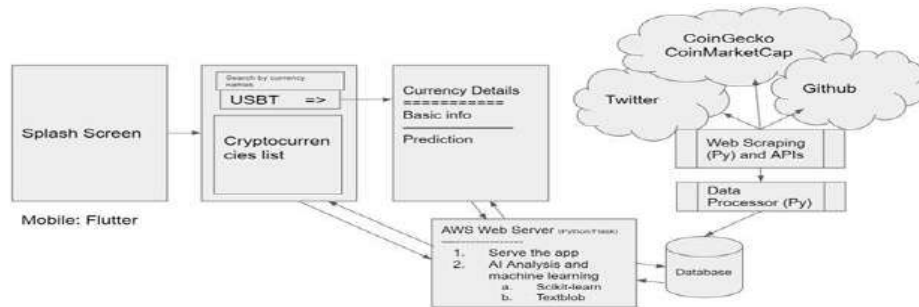


Figure 1. Overview of the solution

Retrospect is a free and easy-to-use app designed to provide users with the latest cryptocurrency data and analysis based on social metrics, such as Twitter activity or GitHub commits, to determine a crypto's quality and future prospects. Retrospect's home page consists of a list of the top 500 cryptocurrencies by market cap. For each cryptocurrency on that page, the user can view basic information, including price, 24 hours change, market cap, and volume. On the home page, users can also search for a specific cryptocurrency and sort the list by alphabetical order, market cap, or 24 hours change. Upon selecting a currency from the list, the user will be redirected to the cryptocurrency's details page. The user can access more information about the coin on that page, including 24 hours high and low, all-time high, symbol, total supply, and market cap rank. The user can then find cryptocurrency analysis in the bottom half of the page. This is where our AI analysis and machine learning come to play. Firstly, we have the RETRO-SCORE<sup>®</sup>, which determines the quality of a cryptocurrency (the higher, the better). Secondly, we have the market view score, which determines whether the crypto community is currently optimistic or pessimistic about the coin. Thirdly, we see the Tweets count change in the last seven days and GitHub activity change in the last seven days. Fourthly, we have the predicted change in the next 24h, telling the user the predicted price movement of the cryptocurrency in the next 24h. The predicted change ranges from very bearish to very bullish and is determined by our model.

Retrospect's data comes from many different sources and is obtained through 1) web scraping with beautifulsoup4 and 2) APIs [9]. The data is then processed and stored in our Database, where we analyze the cryptocurrency data using scikit-learn and Textblob. Our app then directly obtains the data from our Firebase database from Amazon Web Services.

### 3.1. Flutter APP

```

1. return Scaffold(
2.   appBar: AppBar(
3.     title: const Text('Top 5 cryptocurrencies'),
4.     centerTitle: true,
5.     toolbarHeight: 40,
6.     leadingWidth: 80,
7.     leading: DropdownButton<String>(
8.       value: sortBy,
9.       isExpanded: true,
10.      items: const List<String>(['1h-2', '1d-2', '1Mkts', '1Mkts', '124h', '124h']),
11.      .map<DropdownMenuItem<String>>((String value) {
12.        return DropdownMenuItem<String>(
13.          value: value,
14.          child: Text(value),
15.        );
16.      }).toList(),
17.      onChanged: (String? newValue) {
18.        setState(() {
19.          sortBy = newValue!;
20.        });
21.      },
22.      style: const TextStyle(
23.        fontSize: 15,
24.      ),
25.    ),
26.    actions: [
27.      IconButton(
28.        icon: const Icon(Icons.search),
29.        onPressed: () {
30.          showSearch(
31.            context: context,
32.            delegate: CryptosSearchDelegate(CryptosList));
33.        },
34.      ),
35.    ],
36.    body: ListView.builder(
37.      //Implement cryptocurrencies list
38.    ),
39.    //more code below

```

Figure 2. Home page code snippet in main.dart

The flutter app was implemented in Android Studio using Dart. The flutter app is a summary of different files:

- main.dart contains the implementation for the home page with the cryptocurrencies list and the settings page. The application also fetches data from AWS in this file. Most of the app was implemented by myself, but I used GetX's for dark theme and a Settings\_ui's settings template.
- detailspage.dart contains the code for the app's details page, which includes basic information and the application's analysis results.
- cryptosearchdelegate.dart implements the search bar for the application.
- information.dart and updatelog.dart contain the code for the app's information and update log respectively.
- cryptosinfoclass.dart declares the class used to convert passed data from AWS from JSON to usable class.



```

1 class CryptoInfo {
2   final String market_cap_rank;
3   final String symbol;
4   final String market_cap;
5   final String low_24h;
6   final String high_24h;
7   //More final declarations
8
9
10  //constructor
11  //require every variable
12  ...
13  CryptoInfo.fromJson(Map<String, dynamic> json) {
14    String marketCap = Numbers(json['market_cap'].toDouble()).format().toString();
15    String volume = Numbers(json['volume_24h'].toDouble()).format().toString();
16    String price = json['current_price'].toString();
17    String marketCap = json['market_cap'].toString();
18    String low24h = json['low_24h'].toString();
19    String high24h = json['high_24h'].toString();
20    String id = json['id'].toString();
21    String prediction = json['prediction'].toString();
22    String marketCap = json['market_cap'].toString();
23    String tweets = json['tweets'].toString();
24    String commits = json['commits'].toString();
25    String score = json['score'].toString();
26
27    //convert each variable
28
29    return CryptoInfo(
30      //create a CryptoInfo class
31    );
32  }
33 }

```

Figure 3. CryptoInfo class implementation in cryptoinfoclass.dart

Figure 3. CryptoInfo class implementation in cryptoinfoclass.dart

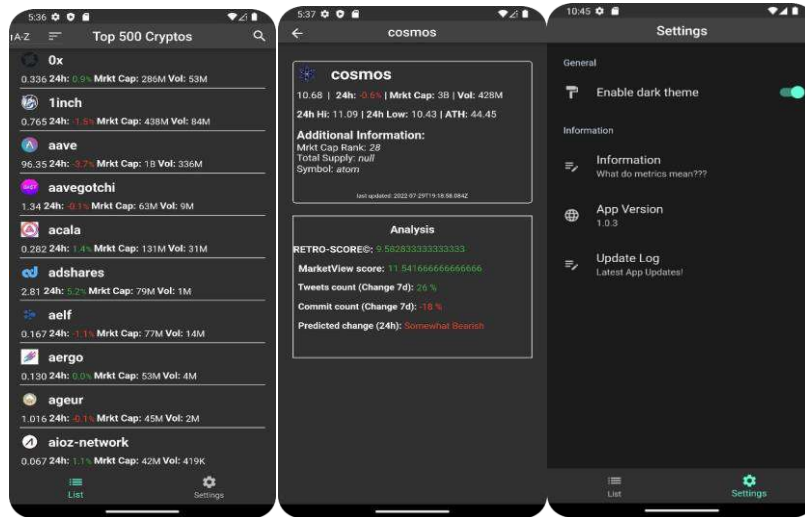


Figure 4. Flutter app overview

### 3.2. Data Fetching

```

1 for i in range(len(headersOfCrypto) / 100):
2     response = requests.get(
3         f'https://api.coinmarketcap.com/v3/all/price/{headersOfCrypto}
4         vs_currency=usd/order=desc/page_size=100/page={i+1}')
5     for res in response.json():
6         cryptoName = res['id'] - 1
7         crypto = res['data']
8         if len(not_in_walletCoins):
9             # filter data and only keep what we need
10            cryptoName = res['id']
11            # get source from GitHub/GitLab
12
13            # if API is down -- 0
14            counter = 0
15
16            for crypto in TopCrypto:
17                if doOnlyFetch == 1:
18                    # counter -- 50
19                    break
20                    URL = f'https://coinmarketcap.com/currencies/{crypto}
21                    ["id"]
22                    = requests.get(URL)
23                    soup = BeautifulSoup(soup.text, "html")
24                    check = soup.find('div', attrs={
25                        'class': 'AppHeader StyledAppHeader flex flex-wrap
26                        wrap justify-content-between align-items-center'})
27                    if check is None:
28                        table = soup.find('div', attrs={'class': 'no-funds no-
29                        data'})
30                    if table is None:
31                        cryptoName = TopCrypto[crypto]['id'] - ""
32                        continue
33                        link = ""
34
35                        for i in table.findAll('div', attrs={'class':
36                            'coin-item'}):
37                            if not link:
38                                link = soup.find('a')
39                                cryptoName = TopCrypto[crypto]['id'] - link
40                                # print(link)
41                                # link = ""
42                                cryptoName = TopCrypto[crypto]['id'] - ""
43                                continue
44
45                        counter += 1

```

Figure 5. Python code snipped to get data from CoinGecko API and webscrape CoinMarketCap for source code

The data fetching code is written in python and uses many libraries to extract data. The process is as follows:

- 1) Use CoinGecko's API to get individual cryptocurrency data
- 2) Webscrape CoinMarketCap to obtain the source code of each coin by forcing the URL (adding the crypto name at the end of <https://www.coinmarketcap.com/currencies/{cryptoname here}> )
- 3) Using the newly acquired source code to obtain GitHub commit count for each cryptocurrency using their API
- 4) Using Twitter's API to obtain the tweet count of each cryptocurrency
- 5) Storing all obtained data on Google Firebase Firestore's Database [10]

### 3.3. Data analysis

```

1 for crypto in TopCrypto:
2     if doOnlyFetch == 1:
3         # counter = 50
4         break
5         url = f'https://
6         www.coinmarketcap.com/currencies/{crypto}
7         ["id"]
8
9         if check is not None:
10            table = soup.find('div', attrs={'class': 'no-funds no-
11            data'})
12            if table is None:
13                cryptoName = TopCrypto[crypto]['id'] - ""
14                continue
15                link = ""
16
17                for i in table.findAll('div', attrs={'class': 'coin-item'}):
18                    if not link:
19                        link = soup.find('a')
20                        cryptoName = TopCrypto[crypto]['id'] - link
21                        # print(link)
22                        # link = ""
23                        cryptoName = TopCrypto[crypto]['id'] - ""
24                        continue
25
26                counter += 1

```

Figure 6. Python code

The data analysis code is also written in python. Let us view each part individually:

- 1) Get Tweet count and Commit count data from Database and save it under each cryptocurrency as a percentage change compared to the last seven days.
- 2) Obtain 10 tweets from each cryptocurrency using Twitter's API
  - a. clean them (using Word Net Lemmatizer from scikit-learn)
  - b. get the polarity of these 10 tweets and calculate the score from -100 to 100 (market view score)
- 3) Calculate the RETRO-SCORE© for each cryptocurrency:

$$R=0.5 \times MV+0.24 \times T+0.24 \times C+0.2 \times (501-MC)$$

Where:

R: RETRO-SCORE©  
 MV: market view score  
 T: Tweets change % 7d  
 C: Commits change % 7d  
 MC: market cap rank

- 4) Append each cryptocurrency's market cap rank, price, market view score, tweets count % change last seven days, and commits count % change last seven days as a list to the Random Forest Regressor model's X-axis.
- 5) Append each cryptocurrency's price change % 24h to the model's Y-axis
- 6) Obtain train data for the model
- 7) Fit newly obtained data and past data into Random Forest Regressor
- 8) Save new train data to the Database
- 9) Plug current data into the model to obtain predictions
- 10) Save model results to Database for each cryptocurrency

### 3.4. AWS and Flask Server

The Flask Server is the simplest part. Every time a user would make a request through AWS, the server would simply return each cryptocurrency's data and analysis from the Database as a JSON to the user through port 5000. If the Database is currently updating, the server will return{"refreshing\_data":"please wait"}, and the app will wait for the Database to complete to execute. Furthermore, data fetch, and analysis is implemented in the AWS server script and run every twohours to put the user to date.

## 4. EXPERIMENT

### 4.1. Experiment 1

Retrospect solves the major pains cryptocurrency investors experience: 1) Lack of analytical tools, 2) Hard to use the application, and 3) Hard to understand analytical results. Our solution is a free-to-use cryptocurrency analysis tool that provides investors with a cryptocurrency quality score and market view score while predicting future price movements for the currency. Retrospect's user interface is smooth and easy to use, thanks to its design and easy navigation. Furthermore, our analytical results are shown in a fashion that even people with no prior experience in cryptocurrency investing can understand. To validate these claims, we have

conducted a survey of our early testers to get their opinion of the product.

The results of the tester survey are as follows: 80% of respondents agree that the user interface is easy and straightforward to use (20% strongly agree), 80% of respondents think that accessing a cryptocurrency is simple, 80% of respondents agree that the analysis is easy to read and understand, and 100% of respondents think that the overall experience is smooth. The results of this survey further demonstrate that Retrospect can solve current investors' problems.

## 4.2. Experiment 2: Establishing the best model

Predicting data involves having a well and functional model. To find the ideal model to fit our data, I have conducted tests on different models to obtain the best possible model accuracy. The models I have tested are Linear Regression, Random Forest Regression, Lasso Regression, Elastic Net Regression, and Stochastic Gradient Descent Regression. I have conducted 4 tests in total. The first 2 tests are the model accuracy (R<sup>2</sup> since we are using regressions) of each model on the first data set (no stored previous data was plugged into the model). The next test is the model accuracy (R<sup>2</sup>) of each model on the first data and past stored data (5 packages of saved data from 5 different periods). We can observe that past data help improve the model's accuracy.

The final 2 tests are the model accuracy of each (R<sup>2</sup>) of each model on the first data and stored data (10 packages).

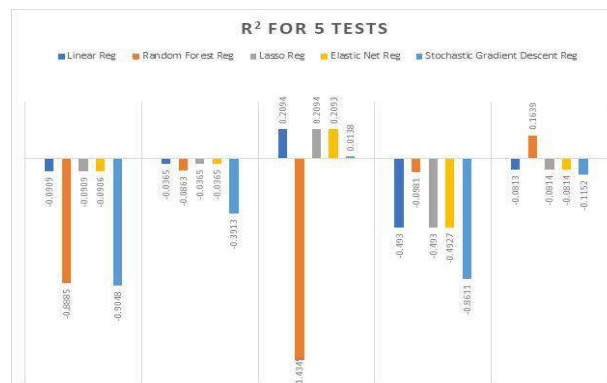


Figure 7. R<sup>2</sup> of each model

The reader should be as surprised as I am. We observe that Linear Regression, Random Lasso Regression, and Elastic Net Regression are more accurate when little data is used. The same case applies to a moderate amount of data to fit a model. However, when more test data is loaded into the model (2000+ test data from different packages from different times), Random Forest Regression seems to hold up better. The Stochastic Gradient Descent Regression is the worst performing model overall. Since we will be conducting large-scale machine learning models with a large amount of train data, Random Forest Regression is our best choice as model. However, if we were to analyze a smaller set of data, Linear, Lasso, and Elastic Net regressions seem to be a better fit.

## 4.3. Experiment 3: Model Accuracy

We have established the model for our cryptocurrency prediction analysis, which is a Random Forest Regression. We have chosen it thanks to its ability to maintain a higher R<sup>2</sup> with more data being plugged into it. Now, we will conduct an analysis to verify the model's accuracy in "real

world application." To do this, we will save the model's prediction analysis for one day and check the next day if the 24h price change corresponds to the change predicted by the model. We will check if the price change corresponds to the prediction tier our model has established (very bearish < -10%, bearish < -5%, somewhat bearish < 0%, neutral = 0%, somewhat bullish > 0%, bullish > 5%, very bullish > 10%). We will be conducting this experiment 5 times at irregular time intervals. Figure 9 shows the experiment's results. The model is accurate if the price change is in the tier the model has placed. The model is a bit off if the price change and the predictions are both positive or negative, and the price change is not in the right tier, but 5-10% off (e.g., the model placed it somewhat bullish, but it is bullish). The model is inaccurate if it does not satisfy the 2 conditions above.

```

1. for crypto in savedData:
2.     pred = float(savedData[crypto])
3.     if crypto in savedChange:
4.         change = float(savedChange[crypto])
5.
6.     if pred < 10 and pred > -10:
7.         if abs(pred - change) <= 5:
8.             accurate+=1
9.         elif abs(pred - change) > 5 and pred * change >= 0:
10.            bitOff+=1
11.        else:
12.            inaccurate+=1
13.    elif pred >= 10 and change <= -10:
14.        accurate+=1
15.    elif pred <= -10 and change <= -10:
16.        accurate+=1
17.    elif pred >= 10 and abs(pred-change) <= 5:
18.        bitOff += 1
19.    elif pred <= -10 and abs(pred-change) <= 5:
20.        bitOff += 1

```

Figure 8. Python code to evaluate each output's accuracy

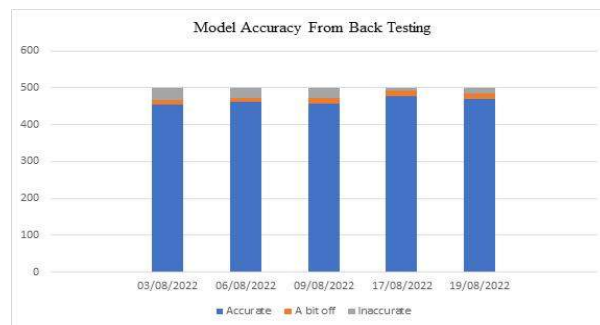


Figure 9. Model Accuracy

After analyzing the results, we conclude that the real accuracy of our model is around 92.8%, which is almost 6x higher than the model's R2 and enough to justify its use.

From Experiment 1, we have successfully concluded that our app help solve the problems our potential users might have: tool availability, usability, and understandability through a survey for our initial testers. From Experiment 2, we have established the ideal model for our needs, Random Forest Regression, thanks to its ability to remain accurate with more data plugged into it. Although the accuracy of our prediction is not top-notch, we offset the margin of error by a

significant amount by indicating whether we are bullish or bearish on the currency instead of indicating the predicted % change, which is the result of our data. That is why the "real" model accuracy is so high, as shown in Experiment 3. This concludes the Evaluation part.

The first related work is written by Johan Bollen, Huina Mao, and Xiao-Jun Zeng and tries to predict the stock market using Twitter mood. Using different libraries for sentiment analysis, like Opinion Finder and a mathematical model, the team concluded that a relationship exists between sentiment on Twitter and the price change, stating that sentiment is reflected after 3 to 4 days. Like my work, our model takes in the sentiment of tweets for a financial product to make predictions. However, our model vastly differs, and my model takes in more variables than just tweets. This work supports the idea that a relationship exists between financial instruments and people's reactions.

The second related work, written by Ross C. Phillips and Denise Gorse, discusses cryptocurrency price drivers using a Wavelet coherence analysis. This work seeks to "demonstrate how factor relationships are prone to strengthen and weaken their correlation with price as cryptocurrency goes through different market regimes." In this work, the authors fetched data from 2018 and back to each cryptocurrency's creation date and only analyzed four cryptocurrencies. In my work, I only used the most recent data to make the model more sensitive to the present. Both works take into consideration social metrics, but M. C. Phillips and M. Gorse's work considers Reddit activity while my work considers Twitter activity. Other metrics used in the related work include Wikipedia and Google searches. The objective of our works is also different. The related works consider the change in the correlation of cryptocurrency price drivers during different market regimes, while my work considers the future price change of the cryptocurrency.

The third related work is written by Stuart Colianni, Stephanie Rosales, and Michael Signorotti and aims to create a trading algorithm using Twitter sentiment analysis. The algorithm did surprisingly well by reporting an accuracy of 95% day-to-day and 76.23% hour-to-hour. Like my work, this work considers tweets polarity as a factor to plug in the model. However, instead of predicting the percentage change for the cryptocurrency, the related work's model, logistic regression, only predicts whether the cryptocurrency will increase or decrease in the next hour or day. Also, only data from Bitcoin was collected, so the model may not be a fit for every cryptocurrency. Nevertheless, the accuracy of the model is still impressive.

## **5. RELATED WORK**

Yuanyi Chen presented a system to remote control Android Mobile Phone by using a computer [11]. In the paper, the author explained that developers need to understand and identify the relationship between the four components, active page, service, content provider and broadcast receiver, of the Android system in order to create a remote control application. Also, the author described how the remote control system works from PC device to the server and mobile device. Our application has a similar approach. We use Wifi to send the information from one device to the other. However, our app uses another mobile device as a controller instead of PC, which makes the controller device be more efficient at the time to help another user since most of the people carry the mobile device.

Sørensen H. et al presented a wireless system to share screens in video calls [12]. They proposed a system that can share both digital content as well as physical artifacts in a video call. Our app is similar to this system, our system mirroring the screen in real time. However, our system is not for a video call and not only for screen share; it also provides remote control.

Bi L. et al proposed a system to remote control power point play in computers without installing any program in mobile devices [13]. It uses Java Native Interface (JNI) technology to control the windows system's function. In our research, we use Android Studio that uses JNI in order to compile our code. As different from this paper, we control the screen of other mobile devices to provide help instead of remote control power point play.

## 6. CONCLUSIONS

In this paper, I have proposed an application to help investors analyze cryptocurrencies and conduct a thorough analysis of their chosen currency. This application solves the major pains investors currently have with the crypto space: lack of analytical tools and difficulty in using current tools. I have shown a demand for my app through online surveys and by explaining the recent surge in popularity of cryptocurrencies. I have outlined the challenges I faced during the creation of Retrospect. I have explained how Retrospect functions and how I built it using python 3, dart, and flutter through Android Studio, Google Firebase, AWS, and different python packages [14]. I have also conducted three experiments to validate the app's utility and model's accuracy. The experiment results show that the app effectively solves the challenges investors are facing. Furthermore, it shows the app's usefulness thanks to its prediction accuracy. Retrospect will be a free-to-use app listed on the google play store and the apple store that will be available to any country.

Although Retrospect solved cryptocurrency investors' problems, its model's  $R^2$  still has a long way to go to become accurate. Indeed, it would be optimal if the model's accuracy never reaches the negative ground. By improving the model accuracy, we can come to better conclusions and simplify investors' jobs even further. Therefore, I aim for 99% accuracy. Furthermore, optimizations can be made to the app's model processing speed. Due to Twitter's API hard rate cap, the model database takes 15 minutes to update. If we break through this time barrier, the app can update more often than two hours.

In the future, I plan to improve the model's accuracy by trying out more models or my mapping my data differently. Furthermore, I could find a solution to Twitter's API cap and not wait for 15 to get my data ready [15]. I also plan to improve the app's user interface and features to make it a better experience. These are the elements I wish to work on in the future.

## REFERENCES

- [1] Chan, Stephen, et al. "A statistical analysis of cryptocurrencies." *Journal of Risk and Financial Management* 10.2 (2017): 12.
- [2] Phillip, Andrew, Jennifer SK Chan, and Shelton Peiris. "A new look at cryptocurrencies." *Economics Letters* 163 (2018): 6-9.
- [3] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349.6245 (2015): 255-260.
- [4] Vranken, Harald. "Sustainability of bitcoin and blockchains." *Current opinion in environmental sustainability* 28 (2017): 1-9.
- [5] Qadan, Mahmoud, David Y. Aharon, and Ron Eichel. "Seasonal and calendar effects and the price efficiency of cryptocurrencies." *Finance Research Letters* 46 (2022): 102354.
- [6] Kalliamvakou, Eirini, et al. "The promises and perils of mining github." *Proceedings of the 11th working conference on mining software repositories*. 2014.
- [7] Lenz, Eva Maria, and Ian D. Wilson. "Analytical strategies in metabonomics." *Journal of proteome research* 6.2 (2007): 443-458.
- [8] Alexander, Carol, and Michael Dakos. "A critical investigation of cryptocurrency data and analysis." *Quantitative Finance* 20.2 (2020): 173-188.
- [9] Glez-Peña, Daniel, et al. "Web scraping technologies in an API world." *Briefings in bioinformatics*

- 15.5 (2014): 788-797.
- [10] Chatterjee, Nilanjan, et al. "Real-time communication application based on android using Google firebase." *Int. J. Adv. Res. Comput. Sci. Manag. Stud* 6.4 (2018).
  - [11] Chen, Yuanyi. "Research on Application System of Remote-Control Computer of Android Mobile Phone." *Journal of Physics: Conference Series*. Vol. 1992. No. 2. IOP Publishing, 2021.
  - [12] Sørensen, Henrik, et al. "Wireless smart phone mirroring in video calls." *IFIP Conference on Human-Computer Interaction*. Springer, Cham, 2015.
  - [13] Bi, Lingyan, et al. "Design and application of remote control system using mobile phone with JNI interface." *2008 International Conference on Embedded Software and Systems Symposia*. IEEE, 2008.
  - [14] Esmaeel, Hana R. "Apply android studio (SDK) tools." *International Journal of Advanced Research in Computer Science and Software Engineering* 5.5 (2015).
  - [15] Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter. "Tampering with Twitter's sample API." *EPJ Data Science* 7.1 (2018): 50.





# A NOVEL SYSTEM FOR REGIONAL TWITTER HATE SPEECH ANALYSIS AND DETECTION USING DEEP LEARNING MODELS AND WEB SCRAPING

Nicole Ma<sup>1</sup>, Yu Sun<sup>2</sup>

<sup>1</sup>Sage Hill School, 20402 Newport Coast Dr, Newport Coast, CA 92657

<sup>2</sup>California State Polytechnic University, Pomona, CA, 91768,  
Irvine, CA 92620

## **ABSTRACT**

*Instances of hate speech on popular social media platforms such as Twitter are becoming increasingly common and intense. However, there still exists a lack of comprehensive deep-learning models to combat Twitter hate speech. In this project, a comprehensive detection and reporting platform, entitled "TweetWatch," was created to solve this issue. A binary classification CNN (Convolutional Neural Network) and a multi-class CNN were created to detect hate speech from real-time Twitter data and classify tweets with hate speech into five categories. The binary classification model has an AUC score of 98.95% and an F1 score of 97.88%. The multi-class classification model has an AUC score of 89.46%. All metrics reached over a targeted 5% increase from previous models in multiple papers, validating the proposed solution. Additionally, the only real-time choropleth map for hate speech in the United States was successfully created.*

## **KEYWORDS**

*Web scraping, Natural language processing, Deep learning, Neural networks*

## **1. INTRODUCTION**

Online instances of hate speech are extremely common on virtually all social media platforms [1][2][3]. Based on previous research, 53% of Americans said they were targeted by hateful speech online and 37% reported severe attacks, but sites like Twitter still rely on an artificial intelligence algorithm that is only around 50% effective. This algorithm often misses instances of hate speech, which are usually targeted towards marginalized groups that already face so much turbulence in real life.

Deep learning methods for hate speech detection are able to outperform state-of-the-art char/word n-gram methods by nearly 18 F1 points. However, despite deep learning being at the forefront of hate speech classification, there still remains a lack of accurate deep learning models that can both detect instances of hate speech on Twitter and categorize them [4][5]. One of the most successful binary hate speech classification models reached an F1-Score of 84.83% and an AUC (Area Under the Receiver Operating Characteristic Curve) score of 90.39% [6][7]. The most successful multi-class toxic sentiment classification attempt reached an AUC score of 82% [8]. Additionally, only 51% of tweets violating Twitter guidelines are flagged by AI, while the other 49% have to be manually reported by other users. The methods behind these models, such as

CNN-LSTMs and the use of F1 and AUC scores as metrics served as inspiration for this project [9]. Furthermore, very little research has been done on the relationship between hate crimes, hate speech, and geographic locations of the incidents, which served as motivation for the choropleth map component of the project.

TweetWatch is a platform that automatically reports tweets marked as hate speech by passing real-time Twitter data through two novel deep learning models: a binary convolutional neural network (CNN) to detect hate speech and a multi-class CNN to classify hate speech into five categories: sexual orientation, special needs, gender, race, and other. Moreover, the solution includes an accessible, interactive choropleth map [10] of the United States created from the collected data. Previously, little effort has been made to find a correlation between geographical location and hate speech frequency, which TweetWatch solves using its innovative choropleth map. Furthermore, the deep learning models created for TweetWatch are significantly (over a 5% improvement) more accurate in terms of AUC and F1 scores.

To prove results, AUC and F1 scores were used to evaluate the accuracy of both models and select the best combination of batch size and epochs. First, we evaluated the reliability of the binary CNN using AUC and F1 Scores – were evaluated for 9 combinations of different batch sizes and epochs. Secondly, we similarly evaluated the reliability of the multi-class network using AUC scores, also for 9 combinations of different batch sizes and epochs.

The rest of the paper is organized as follows: Section 2 illustrates the details of the challenges faced during the span of the experiment; Section 3 focuses on the details of the methodology and the various components of the solution; Section 4 presents an analysis of the accuracy and viability of the solution, following by an evaluation of related works in Section 5. Finally, Section 6 provides concluding remarks, as well as points out possible future developments of this project.

## **2. CHALLENGES**

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

### **2.1. Lack of Annotated Data**

As with many other supervised machine learning algorithms, one of the main challenges was finding sufficient annotated training data. Because of restrictions placed on the Twitter API, there is a lack of a consolidated, complete dataset of hate speech instances on Twitter. This problem is further exacerbated by the lack of annotated categorical hate speech datasets. To circumvent this problem, five different annotated datasets were combined to create a comprehensive dataset, and data points were manipulated to fit into each of the five categories, such as gender-based and sexuality-based discrimination. Data augmentation was also used to expand the set of training and testing data, facilitated by the Python `nlpaug` library.

### **2.2. Eliminating Biases**

Another challenge with hate speech detection is dealing with societal nuances on Twitter. For example, marginalized communities often use demeaning jokes with each other and reclaim slurs for empowerment. Therefore, a common problem while dealing with hate speech detection is differentiating between non-harmful tweets and harmful tweets that often contain similar keywords. To solve this issue, extensive effort was used to make sure that training data included counterexamples of data that include hate speech keywords, such as slurs. This ensures that

context becomes important for the binary CNN as it learns to differentiate between hate speech and non-hate speech based on contextual phrases rather than specific words.

### 2.3. Constructing a Compact and Accessible Visualization Platform

Another one of the main challenges is constructing a compact visualization platform that is able to summarize and analyze the collected data in a compact, readable, and accessible format. Especially because one of the goals is to find a correlation between geographical location and hate speech frequency, a considerable challenge was to present this data in a graphical way. To solve this challenge, a choropleth map was created using Dash by Plotly to utilize color intensity and a continuous logarithmic scale to signify varying levels of hate speech frequency in different states. A pie chart that reconfigures itself based on user interaction was also created to facilitate a compact visualization of hate speech categorical breakdowns in each state.

## 3. SOLUTION

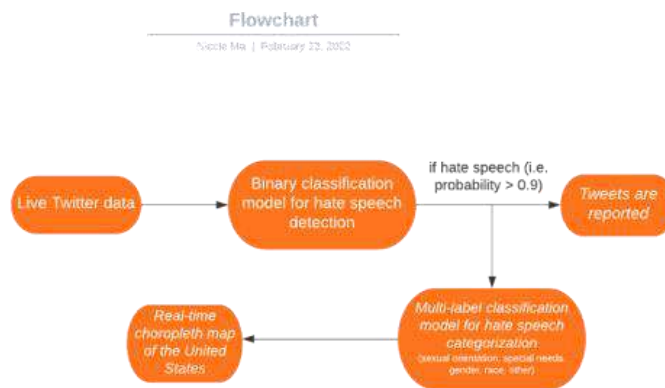


Figure 1. Overview of the solution

TweetWatch is a multi-step hate speech detection and categorization algorithm that automatically reports tweets marked as hate speech by passing real-time Twitter data through two novel deep learning models: a binary convolutional neural network (CNN) to detect hate speech and a multi-class CNN to classify hate speech into five categories: sexual orientation, special needs, gender, race, and other [11]. Natural language processing, such as tokenization and vectorization processes, is utilized to optimize the accuracy and efficiency of the neural networks. Through the use of dual neural networks, TweetWatch is able to go beyond simple identification of hate speech and also provide instant analysis of the frequencies of different categories of hate speech. Live Twitter data is scraped using the Twitter API, and tweets flagged as hate speech by the binary network are passed to Google Firebase [12]. Then, the multi-class network pulls data points from Google Firebase and categorizes the data, allowing TweetWatch to integrate the collected data into a publicly available, interactive choropleth map of the United States.

```

import json
class TweepPrinter(TweepStreamListener):
    def __init__(self, time_limit):
        super(TweepPrinter, self).__init__()
        self.start_time = time.time()
        self.limit = time_limit
        # super(TweepPrinter, self).__init__()

    def on_status(self, tweet):
        if (time.time() - self.start_time) < self.limit:
            # statuses.append(unicode(tweet['text']))
            tweet_text = tweet.text.lower()
            tweet_text = tweet_text.replace("\n", "")
            tweet_text = re.sub(r'[^\w]*', '', str(tweet_text))
            tweet_text = re.sub(r'http://\w+', '', str(tweet_text))
            tweet_text = ''.join(filter(lambda x: x.isprintable(), tweet_text))

            testing1 = model.predict([tweet_text])
            if testing1[0] >= 0.9 and len(tweet_text) >= 50:
                statuses.append(str(tweet_text))
                loc = tweet.user.location
                if loc != None and " " in loc:
                    loc = str(loc)
                    index1 = loc.index(",")
                    twt_location1 = loc[(index1 + 2) :].lower()
                    twt_location2 = loc[: index1].lower()
                    for i in range(0, 99, 2):
                        if twt_location1 == states_all[i] or twt_location1 == states_all[i+1]:
                            location = states_all[i+1].upper()
                            locations.append(location)
                        # locations_full.append(loc_full)
                        elif twt_location2 == states_all[i] or twt_location2 == states_all[i+1]:
                            location = states_all[i+1].upper()
                            locations.append(location)
                        # if there is comma but not the rest
                    else:
                        location = "Invalid Location"
                        if location == "Invalid Location":
                            locations.append(location)
                    else:
                        locations.append("Invalid Location")

```

Figure 2. Web Scraping Live Twitter Data

TweetWatch utilizes the Tweepy library and the Twitter API to scrape real-time tweets, as well as information about the tweets, such as the location of the users. The scraper also standardizes the data, such as by removing links and reconfiguring emojis, to ensure that the format of the real-time data reflects that of the training data used for the convolutional neural networks. The collected data is passed to Google Firebase, where the data points are then sorted by properties such as location. New data from the web scraping algorithm is passed to Google Firebase every two minutes.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 5000, 200)	11897400
conv1d (Conv1D)	(None, 4996, 128)	128128
global_max_pooling1d (GlobalMaxPooling1D)	(None, 128)	0
dropout (Dropout)	(None, 128)	0
dense (Dense)	(None, 10)	1290
dropout_1 (Dropout)	(None, 10)	0
dense_1 (Dense)	(None, 2)	22

=====  
 Total params: 12,026,840  
 Trainable params: 12,026,840  
 Non-trainable params: 0

Figure 3. Binary Classification Model

To create the binary convolutional neural network, a total of 40000 tweets were used to train the model. The data was de-biased by making sure there are counterexamples of data that contain hate speech keywords (e.g. slurs) and standardized by converting to lowercase and removing links, usernames, and non-ASCII characters using regular expression operations. Then standard NLP data pre-processing was utilized by fitting a Keras Tokenizer on collected tweets to split strings into tokens and using spaCy to create text embeddings. The final model compiles the model with the Adam optimizer and binary cross-entropy loss function and uses layers such as Conv1D, pooling, dropout, and dense.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 5000, 96)	4594304
spatial_dropout1d (Spatial1D Dropout1D)	(None, 5000, 96)	0
conv1d (Conv1D)	(None, 5000, 5)	1445
leaky_relu (LeakyReLU)	(None, 5000, 5)	0
max_pooling1d (MaxPooling1D)	(None, 2500, 5)	0
bidirectional (Bidirectional)	(None, 2500, 600)	734400
spatial_dropout1d_1 (Spatial1D Dropout1D)	(None, 2500, 600)	0
conv1d_1 (Conv1D)	(None, 2500, 5)	9005
leaky_relu_1 (LeakyReLU)	(None, 2500, 5)	0
max_pooling1d_1 (MaxPooling1D)	(None, 1250, 5)	0
bidirectional_1 (Bidirectional)	(None, 1250, 600)	734400
spatial_dropout1d_2 (Spatial1D Dropout1D)	(None, 1250, 600)	0
conv1d_2 (Conv1D)	(None, 1250, 5)	9005
leaky_relu_2 (LeakyReLU)	(None, 1250, 5)	0
max_pooling1d_2 (MaxPooling1D)	(None, 625, 5)	0
bidirectional_2 (Bidirectional)	(None, 600)	734400
dense (Dense)	(None, 5)	3005

Total params: 6,319,964  
Trainable params: 2,225,660  
Non-trainable params: 4,094,304

Figure 4. Multi-class Classification Model

To train the multi-class convolutional neural network, more annotated datasets of Twitter hate speech were collected and de-biased. The datasets were manipulated datasets to fit into one or more pre-determined labels (0: sexual orientation, 1: special needs, 2: gender, 3: race, 4: other) and were concatenated horizontally into one Pandas Dataframe. The nlpaug library’s synonym augmentation function was used to individually augment each dataframe to reach 12,000 tweets for each label (60,000 total), and a convolutional neural network was constructed using the leaky ReLU activation function and convolutional layers such as pooling and spatial dropout. The model was compiled with the Adam optimizer and categorical cross-entropy loss function.

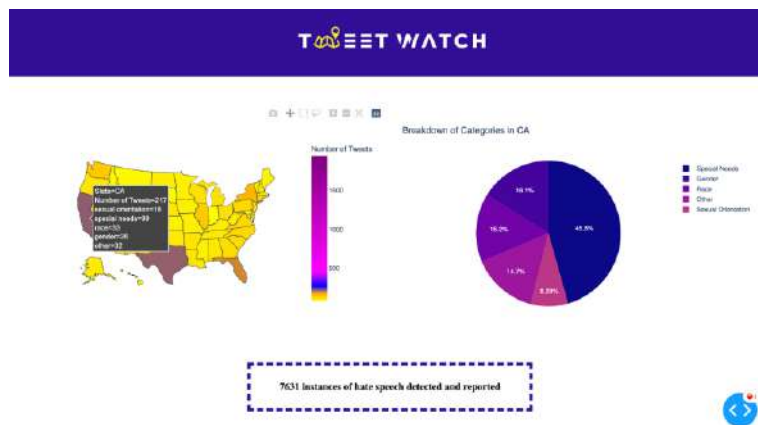


Figure 5. Choropleth Map

Using Dash by Plotly, data is collected from Google Firebase every two minutes and reconfigured into an interactive choropleth map of the United States. The choropleth map uses a logarithmic scale to measure the frequency of online hate speech in each state. By hovering over

a state on the choropleth map, users are able to view the corresponding breakdown of categories of hate speech in the state.

## 4. EXPERIMENT

### 4.1. Experiment 1

To evaluate the reliability of the binary convolutional neural network, two accuracy metrics – AUC and F1 Scores – were evaluated for 9 combinations of different batch sizes and epochs. A grid search was utilized to optimize the efficiency of the evaluation, and the obtained metrics of each epoch were recorded to construct training and validation curves.

Batch Size	Epochs	AUC (%)	F1 Score (%)
128	15	96.63	92.34
128	22	96.7	93.33
128	50	97.43	93.39
256	15	97.13	93.42
256	22	97.24	94.01
256	50	97.35	94.24
512	15	98.42	94.43
512	22	98.95	97.88
512	50	98.91	97.87

Table 1. AUC and F1 Scores for Binary Classification Model Trials

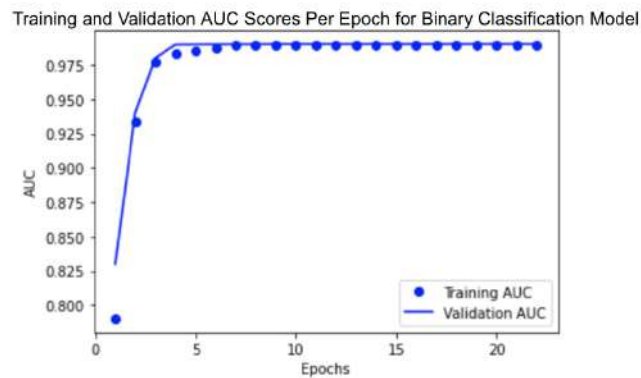


Figure 6. Graph of training AUC vs Validation AUC

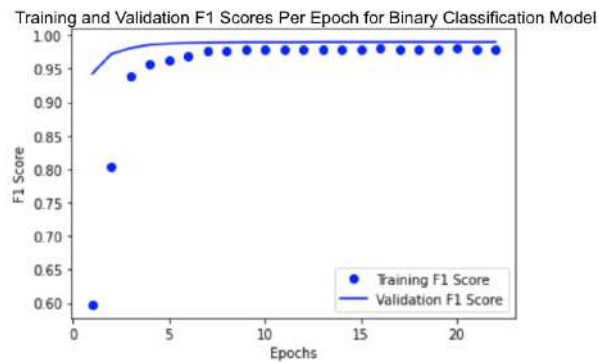


Figure 7. Graph of Training F1 Score vs Validation F1 Score

After grid searching, the most accurate binary classification model had an AUC score of 98.95%, an F1 score of 97.88%, and consisted of a batch size of 512 and 22 epochs. The reasoning behind the higher accuracy for 22 epochs, when compared to 50 epochs, is likely because the model overfitted between 22 and 50 epochs. The training and validation AUC and F1 curves for the best-performing binary model show a dramatic increase per epoch until it converges.

## 4.2. Experiment 2

Similar to the first experiment, 9 combinations of batch sizes and epochs were used to test the accuracy of the multi-class convolutional neural network with different parameters. Once again, a grid search of these combinations was used to analyze the accuracy of the model with respect to its AUC score, and the training and validation AUCs were recorded at each epoch to track the improvement of the model through the course of its training.

Batch Size	Epochs	AUC (%)
128	15	81.95
128	22	88.71
128	50	89.46
256	15	82.83
256	22	85.76
256	50	86.66
512	15	82.88
512	22	84.87
512	50	86.54

Table 2. AUC Scores for Multi-Class Model Trials



Training and Validation AUC Scores Per Epoch for Multi-label Classification Model

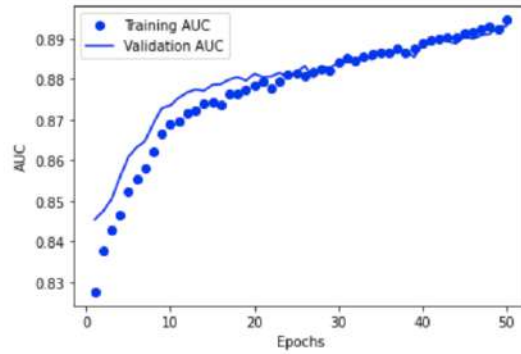


Figure 8. Graph of Training AUC and Validation AUC

After grid searching, the most accurate multi-class model had an AUC score of 89.46% and consisted of a batch size of 128 and 50 epochs. The training and validation AUC curves for the best-performing multi-class model show a steady increase per epoch.



Figure 9. Graph of AUC and F1 Score base, model, and target

Model and Target Metrics of Multi-label Classification CNN-LSTM

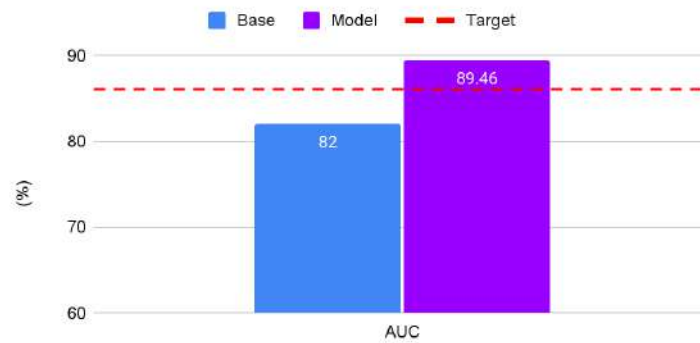


Figure 10. Graph of AUC base, model, and target

Metric	Binary Classification CNN	Multi-Class Classification CNN
AUC Score (%) / Increase from Baseline (%)	98.95 / 9.47	89.46 / 9.10
F1 Score (%) / Increase from Baseline (%)	97.88 / 15.38	N/A

Table 3. Summary of Best-Performing Model Variations

The final metrics for the binary classification model were 98.95% in terms of AUC score and 97.88% in terms of F1 score. These metrics surpassed the performance of models from multiple papers, including a 9.47% increase from the AUC of 84.83% and a 15.38% increase from the F1 score of 90.39% set as the baseline [6]. The final metric for the multi-classification model was 89.46% in terms of its AUC score, showing a 9.10% increase from the performance of the baseline model, which had an AUC Score of 82% [8]. All metrics surpassed the 5% target increase from the baseline models (See Section 1).

## 5. RELATED WORK

Carta et al. (2019) used a dataset of comments from Wikipedia's talk page to classify toxic comments [8]. To facilitate this, they used a supervised multi-class multi-label approach involving the Apache Spark big data framework and word embeddings to create a bag-of-word model. The AUC scores obtained from the model ranged from 0.71 to 0.75. Meanwhile, the multi-class CNN created for TweetWatch reached an AUC score of approximately 0.89, demonstrating an increase from the Carta et al.'s word embedding-based classification model.

Paul et al. (2018) created a set of neural networks to classify tweets as racist, sexist, or neither [6]. The study utilized GloVe embeddings after preprocessing the data by replacing items such as URLs with placeholder tokens. After testing a suite of machine learning models, such as BiLSTMs and CNNs, they found that the CNNs with the greatest reliability had an F1 score of 84.83% and an AUC/AUROC score of 90.39%. The binary CNN created for Tweetwatch was able to reach an F1 score of 97.88% and an AUC score of 98.95%, surpassing the accuracy metrics from the study.

Pereira-Kohatsu et al. (2019) created HaterNet to detect and monitor Spanish Twitter hate speech [13]. HaterNet utilizes the embeddings of words but also emojis and token expressions. Moreover, the analysis phase of HaterNet displays figures such as keyword frequency in tweets classified as hate speech. The best machine learning model from this study achieved an AUC score of 0.828. TweetWatch's binary AUC score of 0.9895 surpasses this value, and is more applicable to English-speaking countries than HaterNet.

## 6. CONCLUSIONS

Despite an increasing amount of hate speech on Twitter, there remains a lack of comprehensive deep-learning models that can both detect and categorize online hate speech. In this project, TweetWatch was created to serve as a comprehensive detection and reporting platform. TweetWatches utilizes two CNNs: a binary classification CNN to detect hate speech from real-time Twitter data and a multi-class CNN to categorize hate speech into five categories: gender, race, sexual orientation, special needs, and others. We used Dash by Plotly to create a real-time choropleth map of the United States with respect to frequency of Twitter hate speech in each American state. By using the AUC and F1 scores as metrics, we show that the novel deep learning networks are more reliable than previous models in multiple papers. The binary classification model had an AUC score of 98.95% and an F1 score of 97.88%. The multi-class classification model returned an AUC score of 89.46%.

Currently, the number of tweets able to be collected through web scraping is limited due to restrictions placed by the Twitter API. Moreover, many users do not have publicly available and accurately labeled locations, making it difficult to obtain a full understanding of region-based hate speech frequency. In contingency with this locational issue, TweetWatch is currently only able to create a choropleth map of the United States instead of the entire world to provide a more comprehensive understanding of global hate speech. Furthermore, the choropleth map shows hate speech frequency relative to each American state, but counties and other regions within each state might have varied frequency.

Future developments of TweetWatch will aim to mitigate these issues. The expansion of TweetWatch into a collaborative, cloud-based website running on users' devices would allow more tweets to be collected through web-scraping. Moreover, by utilizing a translation API and increasing the number of tweets collected through web-scraping, the choropleth map can be expanded to span the entire globe instead of just the United States. Furthermore, by improving upon the algorithm used to extract a user's publicly-available location could be improved on by adding variations of counties within each state.

## REFERENCES

- [1] Weller, Katrin. "Trying to understand social media users and usage: The forgotten features of social media platforms." *Online Information Review* (2016).
- [2] Li, Qing. "Gender and CMC: A review on conflict and harassment." *Australasian Journal of Educational Technology* 21.3 (2005).
- [3] Paz, María Antonia, Julio Montero-Díaz, and Alicia Moreno-Delgado. "Hate speech: A systematized review." *Sage Open* 10.4 (2020): 2158244020973022.
- [4] Mosavi, Amir, Sina Ardabili, and Annamaria R. Varkonyi-Koczy. "List of deep learning models." *International Conference on Global Research and Education*. Springer, Cham, 2019.
- [5] Bisong, Ekaba. *Building machine learning and deep learning models on Google cloud platform*. Berkeley, CA: Apress, 2019.
- [6] Paul, Suvadip, and Jayadev Bhaskaran. "ERASeD: Exposing Racism and Sexism using Deep Learning." (2018).
- [7] Lipton, Zachary Chase, Charles Elkan, and Balakrishnan Narayanaswamy. "Thresholding classifiers to maximize F1 score." *arXiv preprint arXiv:1402.1892* (2014).
- [8] Carta, Salvatore, et al. "A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification." *KDIR*. 2019.
- [9] Wang, Jin, et al. "Dimensional sentiment analysis using a regional CNN-LSTM model." *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. 2016.

- [10] Andrienko, Gennady, Natalia Andrienko, and Alexandr Savinov. "Choropleth maps: classification revisited." *Proceedings ica*. 2001.
- [11] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." *2017 international conference on engineering and technology (ICET)*. Ieee, 2017.
- [12] Chatterjee, Nilanjan, et al. "Real-time communication application based on android using Google firebase." *Int. J. Adv. Res. Comput. Sci. Manag. Stud* 6.4 (2018).
- [13] Pereira-Kohatsu, Juan Carlos, et al. "Detecting and monitoring hate speech in Twitter." *Sensors* 19.21 (2019): 4654.

© 2023 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.



# A SMART MOBILE APPLICATION DESIGNED TO EDUCATE AND AID THE PUBLIC IN COMBATING CLIMATE CHANGE

Kerry Zhang

University High School, Irvine, California

## **ABSTRACT**

*We aim to tackle the issue of improving the global situation regarding climate change by creating a mobile application named Climerry, which educates its users on recent news related to climate on the home screen. Climerry also features a second tab that allows users to view opportunities to improve the climate change situation in the vicinity by typing in a ZIP code or city name. Some examples of opportunities include beach cleanups and tree-planting sessions. By informing and encouraging the general public to become more involved in the effort to preserve our planet, the negative effects of climate change may be much less significant in the future.*

*To prove the effectiveness of this application in encouraging the general public to take action against climate change, one experiment was performed to gauge how much knowledge regarding climate change the participants had gained by using the application. Another experiment tested the reliability of the news API used in the application by testing the accuracy of information in each of the selected articles in the featured news section of the application. The result of the experiments indicated that the application is useful when it comes to providing accurate news and educating its users on the topic of climate change.*

## **KEYWORDS**

*Climate Change, News, Global Warming, Social Issue*

## **1. INTRODUCTION**

Climate change is a highly relevant topic that has a global effect on humanity. It is widely acknowledged by the public, and some governments are taking steps toward improving the situation involving global warming. The benefits of addressing climate change early and reducing carbon footprints are mitigating environmental damage done to certain habitats, which preserves the livelihood of animal and plant species and allows those at lower sea levels to continue going about their lives safely. However, the consequences of leaving climate change unchecked include glaciers, rising sea levels, extreme weather patterns, lowered food supplies, and many other changes caused by the overall higher temperatures that can reduce the quality of life for numerous individuals [1].

Making people aware of this topic is crucial to ensuring a better future for generations to come. Erratic weather events and less suitable land to live in are undesirable outcomes. Many believe that ordinary people are powerless against climate change, but the combined efforts of the general public to take small steps against climate change, such as wasting less food and consuming less water and energy, can make a massive difference for the planet over many years. Using public

transport more frequently whenever possible and using fewer plastics are other beneficial actions for the environment. Therefore, the message that climate change can create a plethora of devastating yet preventable consequences to our planet should be spread to as many people as possible, so that more people will be willing to do their part and take action against climate change.

There are currently several mobile applications that are dedicated to educating their users on the subject of climate change. Three particularly notable mobile applications are Commute Greener, MathTappers: Carbon Choices, and Skeptical Science. Commute Greener focuses primarily on gauging the carbon dioxide emissions that are produced during commutes within the United Kingdom. MathTappers: Carbon Choices extends its reach beyond the previously mentioned application by demonstrating how much carbon dioxide is released by doing everyday activities, such as eating and bathing, to provide its users with an understanding of how their activities play a role in climate change. Lastly, Skeptical Science debunks commonly used arguments that attempt to prove that climate change and global warming do not exist, and this application continues to update itself with new research and new counterarguments.

Despite the helpful features that these applications offer, they still have some downsides. Commute Greener and MathTappers: Carbon Choices is only published on the App Store, which excludes those with Android devices. As Android devices have been increasing in popularity over the last few years, a large portion of the potential userbase is being left out of using these two mobile applications. Commute Greener was mainly developed for calculating commutes within the United Kingdom, and those from other countries will be region-restricted from downloading and using the application, which further reduces the userbase of the app. MathTappers: Carbon Choices has a great concept, but lacks many important lifestyle choices that would more accurately gauge the users' personal carbon footprints, such as measuring the environmental impact of using a bus or train as transportation. While Skeptical Science may serve its purpose well when it comes to debunking misinformation regarding climate change, it has an extremely basic layout that can appear almost unprofessional and unappealing. Looking through paragraphs of black text on a white background may not catch the attention of younger generations. Although this application undoubtedly contains massive amounts of knowledge regarding climate change, potential users may skip over this application without the ability of the application to present information related to climate change to them in engaging and exciting ways.

The issue of climate change is tackled using a mobile application called Climerry, which is published on both the App Store and the Google Play Store. Climerry includes a home screen that provides updated climate news and a second screen to participate in activities that directly or indirectly affect climate change. Users will be able to type their city or ZIP code in a search bar to get more localized results. These in-person opportunities include planting trees, picking up trash, and joining city conferences to contribute ideas. For more sparsely populated areas that may lack any in-person opportunities nearby, online opportunities exist as well, such as attending virtual conferences and learning how to more effectively recycle bottles. Climerry is similar to many other climate change applications in that Climerry informs its users of the current global situation regarding climate change. What makes Climerry stand out from many other applications, however, is the ability of the application to provide its users with opportunities to support their communities and create positive change. Rather than being bystanders and watching as the climate change situation becomes more intense, users can play an active role in the efforts to mitigate worldwide damages caused by climate change.

The remainder of the paper is structured in Sections labeled 2 through 6. Section 2 highlights the obstacles that had to be overcome during the development of the mobile application and

performing experiments with the application. Section 3 describes the general implementation of the application as well as details on specific parts of the application. Section 4 provides a thorough description of the experiments that were performed to prove the effectiveness of the application. In section 5, related works are summarized and compared to our current work. Section 6 offers a conclusion in the form of a summary of the application, the application's limitations, and the steps that can be taken in the future to address these limitations.

## 2. CHALLENGES

There existed several challenges while developing the application, one being the difficulty in finding a reliable method to pinpoint volunteer opportunities specifically to combat climate change around the exact location the user hypothetically inputs. Due to the requirement for extremely localized responses from the application and the need to generate concurrent opportunities during the time the application is used, there needs to be a database that collects all volunteering opportunities surrounding a specific area at a given time. In addition, even through online means, narrowing opportunities to just those involving climate change is a complex process. Moreover, many locations, especially rural areas or thinly-populated communities, do not possess as many of these specific opportunities. Therefore, in addition to a collection of in-person events, the database needed to provide online and virtual opportunities as well, such as lessons on how to properly recycle. With these adjustments in place, the application has become more practical to a larger userbase.

The second challenge while developing the application was the complexity of finding a reliable source of data for the data section of the home page. Because this section needs to update itself consistently, a source of live and accurate data is needed. Many sources, however, including articles, reports, and analyses, all provide static information. One of the only sources available that met this requirement, the Live Climate Scoreboard of Bloomberg [6], possessed an anti-scraping mechanism in its website's AI, preventing it from being used as a reference. The second source found was NASA's Global Climate Change Dashboard [7], which was accessible, but possessed data that was less actively updated. NASA's HTML writings were also more complex, and thus it was difficult to adjust the application's format to scrape the data embedded. Lastly, a system was necessary to consistently perform the scraping action to keep the application up to date without triggering any security systems on the NASA website.

Our final challenge is coming up with experiments that gauge the performance of the application and the effectiveness of the application in educating its users on climate news. To test the performance of the application, a specific feature of the application that plays a crucial role in the application's purpose while being able to be tested for performance would need to be chosen. The news section was selected to be experimented on, and measuring how quickly a freshly published article took to reach the news section of Climerry was the chosen method to test its performance. For testing how effective the application was at educating users on climate, the original idea was to have the participants rate their extent of knowledge of climate news after using the application for a set period. However, everyone has different backgrounds and different levels of knowledge in certain areas prior to using the application. Therefore, an updated experiment was implemented that asked the participants to score their knowledge level in climate news before using Climerry as well.



### 3. METHODOLOGY

#### 3.1. Solution Overview

The application is currently composed of two main pages: the home page and the volunteering page. The landing page includes a frequently updated news section, which gives people an easy, fast, and accessible source of information regarding climate change. All the news in the application is retrieved from one API. This serves to inform people more about the current status of climate change and the events surrounding this subject. The home page also includes a statistics page that uses quantitative data live from NASA to portray the worsening severity of the global situation, hopefully convincing people to heed the warnings of climate change. The user can press the three bars in the top left of the screen to open a slider, then press the second tab in the slider to transfer to a page for volunteering. On the volunteering page, users can search for local volunteering opportunities to combat change. For the application to inform the users of local volunteering opportunities, the application requires the user to input a city name or a ZIP code in a text box at the top of the screen. Based on this information, the application returns opportunities that are within the area which are retrieved from a database, meaning that users can contribute without having to travel unnecessarily long distances. This in turn leads to more community involvement tackling the issue, thus incorporating people into the effort at a faster rate. By experiencing a volunteering event, people are more passionate and aware of the situation.

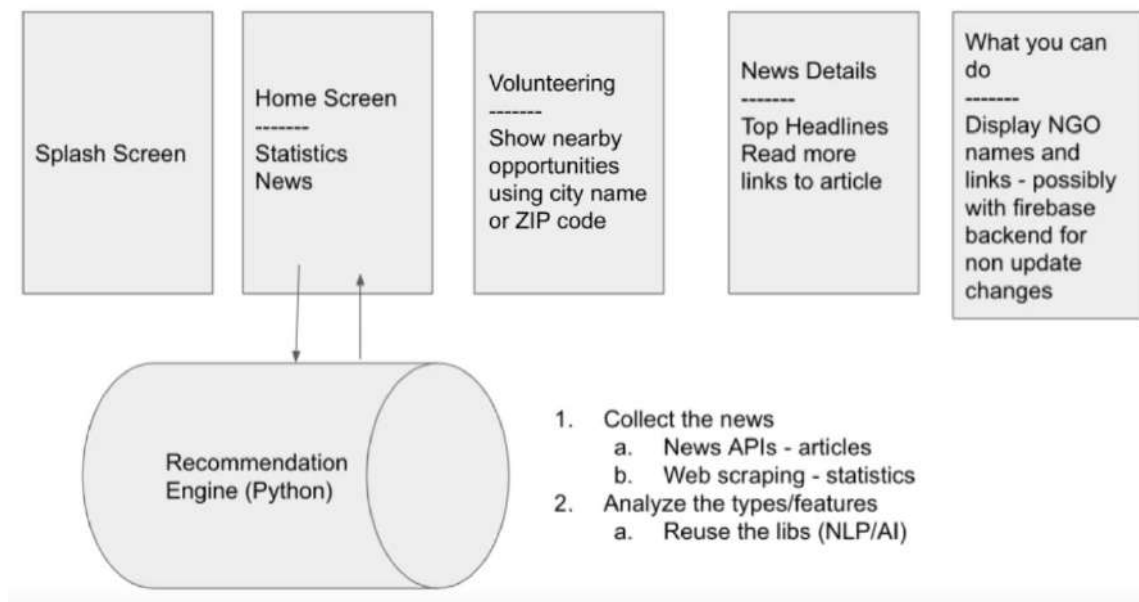


Figure 1. Frontend and Backend Design

#### 3.2. Implementation

Flutter is used as the front end of the application to provide a clean and intuitive user interface, while Python acts as the back end. Since Python and Flutter are different programming languages, a framework called Flask is used to connect the two different sections of code together. In the Python code, several methods are used to provide the functionality to the application, and they are routed with the help of Flask to return the information to Flutter.

The application is composed of two main screens, which are the home screen and the volunteer screen. As shown in the left of Figure 2, the home screen has a news section, which is managed with the help of Python. Using an API key, a News API Client generates every kind of news with a query that contains the words “climate”, “change”, “mitigation”, “conference”, and “IPCC” inside of it. From the dictionary object that is generated, only the articles are selected. The articles are then converted into a JSON file and sent to Flutter to display. Within the home screen, a statistics page is also shown that informs the user about the current carbon dioxide level, global temperature anomaly, arctic sea ice minimum extent, ice sheets, sea level, and ocean warming level. To implement this into the application, the information from the official NASA climate website was scraped in Python using HTTP get requests with the website links. Then, each HTML would be parsed using BeautifulSoup. To individually pick out the desired statistic from each link, the HTML elements with the class tag that contained the statistic were collected. Then, the first element of its contents attribute was stripped, leaving only the number left. From here, the relevant numbers are added to a dictionary. The dictionary of statistics is delivered as a JSON file back to Flutter, where the statistics can be displayed in the application.

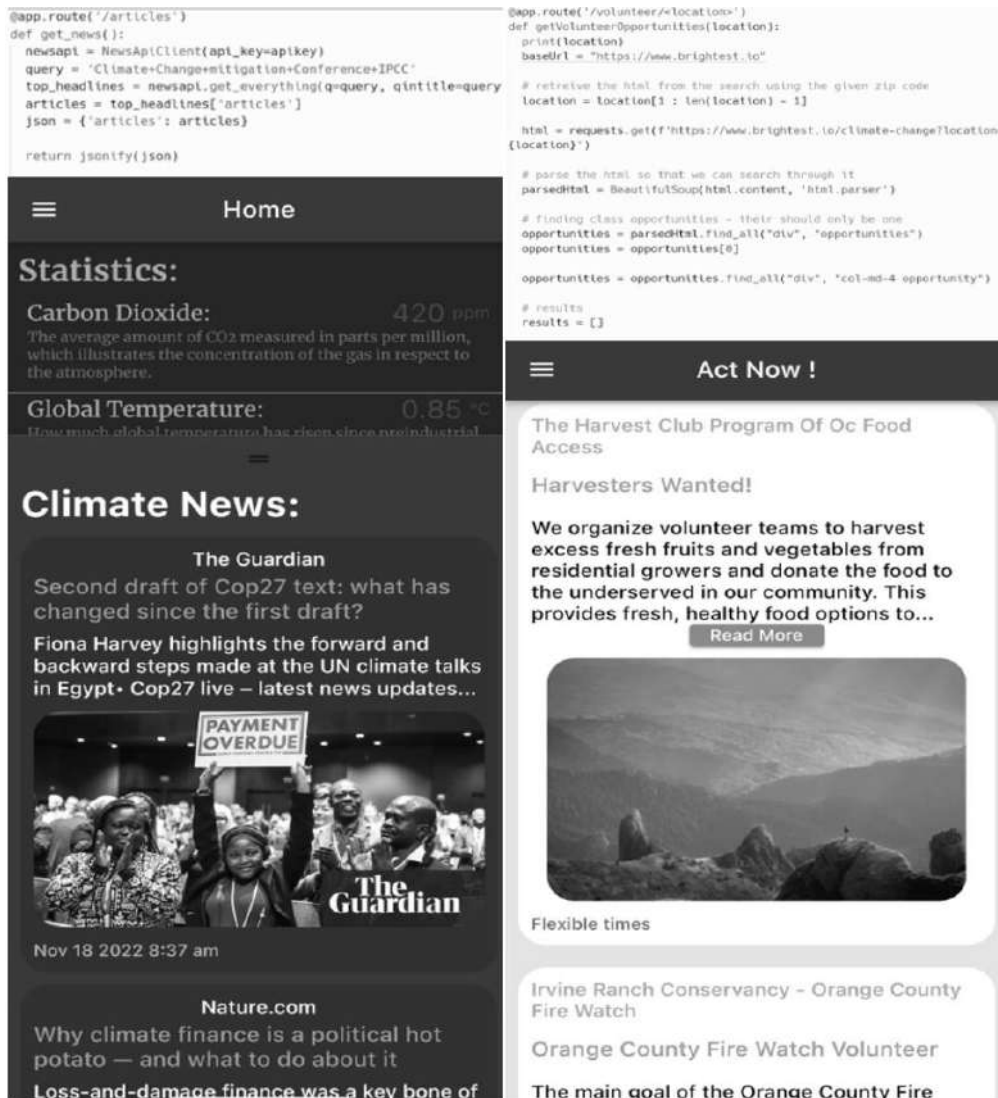


Figure 2. Screenshots of application and corresponding code

Shown in the right of Figure 2 is the volunteer screen, which features a search bar that prompts the user with the message “Search by City State or Zip”. When the user properly inputs their location information, the application will display local, nearby opportunities that the user can partake in. This is achieved by making an HTML get request for Brightest’s website. Brightest is a company that is dedicated to providing sustainable solutions. The HTML is parsed and all HTML elements with the desired div tag are retrieved. The HTML elements are then processed and passed back to Flutter to be displayed. However, this will not be instantaneous, and the user will have to wait before receiving these opportunities. To indicate to the user that the application is not frozen or faulty and the opportunities are still being loaded in, the Flutter code uses a boolean variable called `is Loading` to keep track of when a process is happening. If the opportunities are processing, the `is Loading` variable’s value is true. Because `is Loading` is set to true, the volunteering screen will display a circular progress indicator object slightly below the other elements that are currently on the screen (Reference Figure 3).

## **4. RESULTS**

### **4.1. Experiment 1**

To evaluate whether the application is effective at educating the public on climate change, a survey will be conducted among all participants to test how much knowledge the participants believe they have gained from using this application. Each participant will rate what they believe their current level of knowledge is regarding current climate news from one to ten, then they will scroll through the current climate news section of the application’s home screen for ten minutes. Finally, they will rate their current level of knowledge regarding climate change again and leave any optional feedback they have in a free-response section. By recording scores from both before and after, the current climate news can be tested for whether the information within the news is helpful and is not too basic or common among the general public. Because there will be 32 participants in total, the sample size will be large enough to account for any variability.

Table 1. Climate Awareness (Before vs. After using Climerry)

<b>Participant</b>	<b>Knowledge Before Using Climerry</b>	<b>Knowledge After Using Climerry</b>
1	7	8
2	5	5
3	3	5
4	1	6
5	4	4
6	3	5
7	1	4
8	6	9
9	8	8
10	5	6
11	4	6
12	3	7
13	5	6
14	7	7
15	3	6
16	4	4
17	5	6
18	7	9
19	5	10
20	3	5
21	2	5
22	5	8
23	6	8
24	5	8
25	5	7
26	4	7
27	4	8
28	5	5
29	6	6
30	5	7
31	6	7
32	6	8
<b>Average</b>	<b>4.625</b>	<b>6.5625</b>

## Self-Reported Knowledge of Current Climate News



Figure 3. Climate Awareness (Before vs. After using Climerry)

Among all of the participants, the self-reported knowledge level of current climate news after using Climerry was either greater than or equal to the self-reported knowledge level of current climate news before using Climerry. The knowledge levels before using Climerry ranged from 1 to 8, and the knowledge levels after using Climerry ranged from 4 to 10. Furthermore, according to the data collected from the participants, the self-reported knowledge of current climate change had a significant average improvement of almost 2 points after using Climerry. The majority of the improvements were somewhat moderate, but the largest improvement was 5 points, in which a participant went from having a self-reported knowledge level of 1 to a self-reported knowledge level of 6. In the free-response sections, participants generally admitted that they would likely have never seen many of the articles that they did from Climerry, and they would be willing to keep the application installed and occasionally check the application for future climate news.

#### 4.2. Experiment 2

To combat the possible issue of the application spreading misinformation through faulty news articles, the application is experimented on for its news articles' reliability and accuracy of information related to climate change. The application will be opened and the articles that appear on the will be scrolled through. Articles within the application will be selected at random, and two other articles on each article topic will be manually read through and compared with the original article displayed from the application. If the original article's information seems to match the information of the others, the article will be marked as accurate. However, if the original article states information that directly contradicts the information of both related articles, the article will be marked as not accurate instead. This experiment's design was implemented because it is significantly less likely for both related articles to be false than it is for the original article to be false. Twenty articles will be used as the sample size in the experiment, which is enough to mitigate the effects of variability. The results will be recorded in a table.

Table 2. Accuracy of News Sources

Article	Source	Accuracy of Information
1	24/7 Wall St.	Accurate
2	Business Wire	Accurate
3	Yahoo Entertainment	Accurate
4	Desmog.com	Accurate
5	Archinet	Accurate
6	AllAfrica	Accurate
7	Thechronicle	Accurate
8	The Times of Israel	Accurate
9	New York Times	Accurate
10	Otago Daily Times	Accurate
11	Phys.org	Accurate
12	24/7 Wall St.	Accurate
13	Forbes	Accurate
14	Business Wire	Accurate
15	Yahoo Entertainment	Accurate
16	Archinect	Not Accurate
17	The Chronicle	Accurate
18	Otago Daily Times	Accurate
19	Plos.org	Accurate
20	Euronews	Accurate

Based on the results of the experiment, an overwhelming majority of the news articles selected by the application contains accurate information. Nineteen out of twenty tested articles were compared to other articles on the same topic and were found to have consistent information, which indicates that approximately 95% of future articles selected by the application are trustworthy. Only one article that was labeled as inaccurate had a discrepancy between two other articles on a statistic, but the rest of the information inside the article appeared to be accurate when compared to the related articles. One possible reason why this percentage of accuracy is so high is that the API that selects the articles for the news section only chooses from trustworthy sources. Another possible reason is that articles on a similar topic may reference each other, and the information from the articles would be very similar as a result.

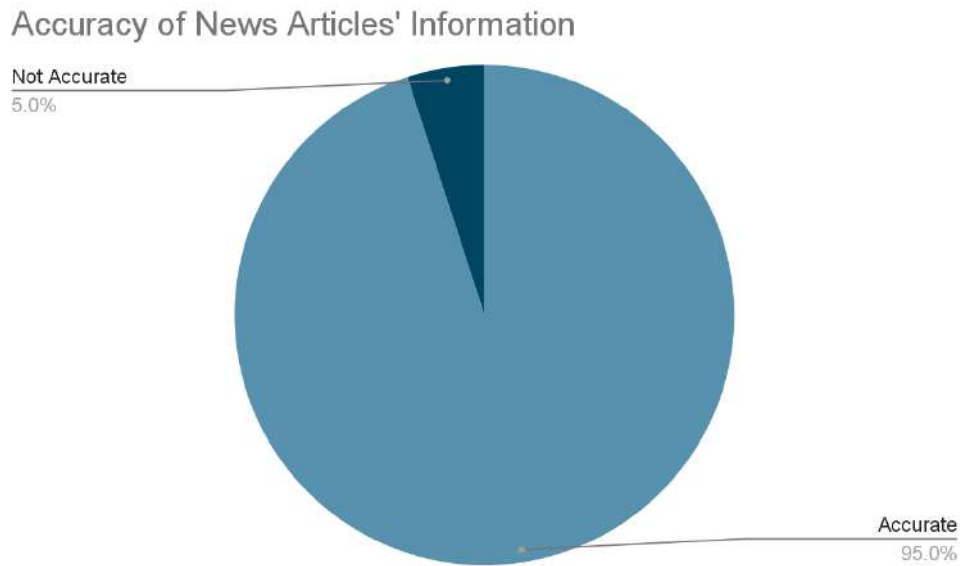


Figure 4. News Source Accuracy

### 4.3. Analysis

The results from the first experiment indicate that the general public can greatly expand their knowledge involving current climate change news by using this mobile application. Since the majority of participants reported an increase in their self-measured level of knowledge regarding climate change, the application appears to be very effective at educating the general public. The application has been proven to serve its purpose well, which will potentially result in more long-time users of the application.

The second experiment's results reveal that the articles that are displayed in the news section of the application are generally very accurate and contain information that remains consistent across other related articles. Therefore, when users gather information regarding current climate change by perusing through the news section of the application, they can feel fairly confident that they are consuming accurate news. Overall, the two experiments prove that users of Climerry can be constantly updated with relevant and true climate change information.

## 5. RELATED WORKS

A study has been conducted on promoting behavior that helps the environment through the implementation of game design principles. The study has found that games and other applications that incorporate game design principles are more effective than other methods, although it cannot provide a conclusive explanation behind this [2]. This study is similar to our work in the sense that both deal with the primary focus of climate change and the utilization of applications. However, the study also dives into the effectiveness of gamification when it comes to behavioral changes in people. On the other hand, our work provides a larger focus on improving the situation with climate change itself.

Another work describes methods to reduce the effects of climate change and the effectiveness of each method, ranging from the addition of nutrients in the ocean to promote biological activity and absorb more carbon dioxide to the injection of reflective aerosol particles into the atmosphere

to reduce global temperatures. The work ends by providing a call to action for how incentives and strong encouragement from governments can improve the situation regarding climate change [3]. This work shares a major similarity with our work, which is the strong emphasis on climate change. While this related work places a stronger focus on possible strategies to tackle climate change, our work incorporates the spreading of news regarding climate change and the possible opportunities to take part in improving the situation regarding climate change.

A third work highlights those who are skeptical about the existence of climate change and utilize the media to further their movement of discrediting science. The collective new service community has been demonstrated to be a dominating source of news pertaining to climate science, and the attack on climate science bears a striking resemblance to past attacks on other fields of science, such as the pesticide and chemical industries [4]. This work is incredibly similar to our work since both of the works heavily emphasize the general public's knowledge regarding climate change. Our work goes more into spreading information, while the related work explores the spread of misinformation.

## 6. CONCLUSIONS

Our mobile application, Climerry, aims to prevent unnecessary damage to the environment by spreading knowledge of the global situation regarding climate change and encouraging others to take action. To achieve these goals, Climerry contains a current news section on its home screen that is related to climate change, which updates its users on the latest information. Climerry also offers opportunities for its users to actively take part in the effort against climate change by inputting a ZIP code or city name to narrow down the nearest ones.

To test how effective the application is at educating its users and outputting relevant and popular articles to the users, two experiments were performed. The first experiment involved gathering thirty-two participants to label their current knowledge of climate change from one to ten, use the application for ten minutes, then label their knowledge of climate change again using the same scale. The second experiment involved retrieving the articles in the news section of the application and comparing the information in these articles to the information in other articles on the same topic to test the application's ability to output reliable news. According to the results of the experiments, the news section is incredibly effective at helping users gather recent and accurate information regarding climate change. Participants in the experiments generally had a significant self-reported increase in their current knowledge of climate change, and tests comparing articles selected by the application to related articles indicate that the information from the selected articles is very consistent.

## REFERENCES

- [1] "Global Warming Effects." *Environment*, National Geographic, 16 Feb. 2022, <https://www.nationalgeographic.com/environment/article/global-warming-effects>.
- [2] Lab, Brauer Group, and Benjamin Douglas. "Gamification to Prevent Climate Change: A Review of Games and Apps for Sustainability." *Current Opinion in Psychology*, vol. 42, Dec. 2021, pp. 89–94., <https://doi.org/10.31219/osf.io/3c9zj>.
- [3] Fawzy, Samer, et al. "Strategies for Mitigation of Climate Change: A Review." *Environmental Chemistry Letters*, vol. 18, no. 6, 30 July 2020, pp. 2069–2094., <https://doi.org/10.1007/s10311-020-01059-w>.
- [4] Antilla, Liisa. "Climate of Scepticism: US Newspaper Coverage of the Science of Climate Change." *Global Environmental Change*, vol. 15, no. 4, Dec. 2005, pp. 338–352., <https://doi.org/10.1016/j.gloenvcha.2005.08.003>.



**AUTHOR**

**Kerry Zhang**, University High School Class of 2024. Passionate about Economics and Political Science and plans to pursue such a career in the near future.



# A NLP-LEARNING POWERED CUSTOMIZABLE APPROACH TOWARDS AUTO-BLOCKING DISTRACTING WEBSITES

Yulin Zhang<sup>1</sup>, Yu Sun<sup>2</sup>

<sup>1</sup>University High School, 4771 Campus Drive. Irvine, CA 92612

<sup>2</sup>California State Polytechnic University, Pomona, CA, 91768,  
Irvine, CA 92620

## **ABSTRACT**

*Over the past few decades, the problem of distraction and its accompanying side effects has taken its root deeply in all parts of our daily life and extended its ever-increasing influences among young generations [2]. In addition to its alarming prevalence, another characteristic of distraction that raises most concerns is how easily we can get distracted from our tasks at hand while using the electronic devices as a means of solving problems [3]. This paper attempts to address this society-wide problem thoroughly and universally through a technical approach of detecting, analyzing, and blocking the websites intelligently. Our design highlights the applications of machine learning and natural language processing, and is implemented purely in Python, Javascript, and several other web development languages. After retrieving the web content from the target websites through the web scraping process, summarizing the data to a number of short paragraphs via the use of NLP, we were able to perform data analysis on the result and finally block the websites accordingly [4]. With the help of this extension, students and those who wish to improve their concentration in work will be able to put more focus on the tasks at hand and thus boost their work efficiency under any working conditions.*

## **KEYWORDS**

*NLP-learning, distraction, Auto-block*

## **1. INTRODUCTION**

Nowadays, the problem of distraction has become a society-wide problem that appears to be especially prevalent among teenage internet users. For decades, it hindered people's productivity at work and encouraged inattention and daydreaming, rather than focusing and solving the task at hand. In addition, the result of a research project conducted in UCI shows that being able to return to the concentrating state after distraction takes, on average, 23 minutes and 15 seconds, which demonstrates the influential effects distraction had on one's work mode [1]. On the other hand, being able to focus on the task will help students and employees to work more efficiently and systemically. Furthermore, according to an experiment focusing on how different levels of concentration impact the likelihood of being distracted, the data suggests that highly concentrated individuals are, in general, less likely to be distracted by irrelevant affairs in comparison to less concentrated groups.

Several approaches had been invented and widely used to combat this problem, including the practice of mediation, temporarily turning off the device, and using reward-and-punishment mechanism to motivate oneself; However, these techniques often failed to fully address the

problem due to various reasons, some of them being the many forms the distraction may take on and the constant danger of not able to Concentration it poses. Such obstacles made distraction a seemingly unsolvable issue, and any successful attempts to address it to be rather rewarding and valuable.

Although there are already a number of existing anti-distraction techniques available, most of them failed to take into account users' potential actions after blocking the websites.

Some of the existing techniques and systems that have been proposed to address the problem of distraction can be loosely grouped into either manual approaches as discussed above, or "focus apps" that aim to help users to better concentrate on the tasks depending on their needs. However, most of these Apps/extensions either assume that users would never regret their choice of blocking a certain website and come back to unblock it during working hours, or provide little functionalities to prevent similar situations from happening. For example, one of the distraction-free extensions I tried allowed me to enter the websites I blocked at any time after an hour. In practice, chances are this is often the case among groups of users. Their implementations are also not intelligent enough to detect and block websites of certain categories that are specific to each user; Instead, they block a wide range of websites disregarding the information they contain, which may cause the inclusion of unnecessary websites that can be of value to users. A second practical problem is that some services block all but a number of tabs the user is currently working on, and many users find it hard to make a choice between having to deal with the constant distractions from their favorite websites and losing the access to any other websites aside from their relevance.

The goal of our extension is to significantly improve individuals' concentration skills and help them to cultivate a better study/working habit that will benefit them for a lifetime [5]. Our method is inspired by a number of existing programs that aim to address similar issues related to distraction; however, while most of these programs do not taking into account users' specific needs and how they may adjust the program to fit these needs, our extension allows users to update the preferences anytime and anywhere to blocking new categories of websites. In addition, different from those paid services, our extension is always free and openly accessible to users of all ages and backgrounds. Although we have made a number of important improvements from the existing methods, our extension still shares some major similarities with these methods, one of them being the same process of scraping down the web content, making relevant analysis with it, and blocking the website based on the result [6]. By using our extension, students who often find themselves struggling with the desire of looking at websites they like during the study sessions will be able to focus on the actual work and solve them more efficiently.

In three application scenarios, we demonstrate the consistency and accuracy of the extension's capability at blocking distracting websites based on the user input [7]. Each of these scenarios involves experimentally input a keyword into the extension and access 3 websites that are relevant to the keyword and should be blocked, as well as 2 websites that are irrelevant to the keyword and should not be blocked. The experiment was conducted at 5 different points in time, and we recorded whether Concentrate correctly blocked the page in 3 tables that correspond to 3 keywords. We then analyze the results and discover that the extension achieves on average an accuracy score of 98% in blocking relevant websites and 100% in letting irrelevant websites pass across 3 tables, indicating that the extension could effectively identify the similarity between the website content and keywords and block the website accordingly.

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 project, a few challenges have been identified as follows.

### **2.1. Organizing not only Text Summarization but Subject Summarization**

Being able to perform both text and subject summarization is a difficult problem to handle as it is hard to parse over unstructured web content, find an appropriate NLP package, extract only relevant information from large chunks of text and compare it against the possible categories the website may fall into [8]. Completing all these requirements requires a way to efficiently scrape over the website, summarize its content, and make data transmissions between front and back end. Through the use of several well-known libraries, we were able to scrape over the content with the requests module, summarize the web content under the help of the NLTK library, and finally use that page contents to block the website according to the result we received.

### **2.2. Managing Events through Content Scripts and Background Scripts in the Chrome API is Often Difficult**

It is rather challenging to organize and maintain the bidirectional communication and data transmission between the background and content script due to the various restrictions chrome API applies to protect users' privacy. In order for the summarization and blocking process to function well, the content script has to undergo a multi-steps process, including collecting the web content, passing the data to the background script for analysis and summarization, then waiting patiently for the background script to pass back the decision of blocking or allowing the website [9]. During the process, each script may get triggered at unexpected times due to the shift of chrome API's permission, which requires careful examination of the issue and a once-and-for-all solution to allow for successful communications between two scripts. To address the issue, we added listeners to the background script part in order to regulate the transmission and keep relevant events under monitoring.

### **2.3. Designing a Clean User Interface to Allow the user to Make Changes to their Account info with Little Work**

Given the small and confined area of the extension's pop-up, it's extremely difficult to design a clean and simple user interface that also includes comprehensive functionalities. Generally, when facing similar situations, some extension developers would choose to include an additional website with detailed descriptions on how to use the extension, or pay for professional designers to plan out and organize the interface for them; However, since neither of these solutions is applicable to our case due to the lack of resources and funding, we had to take a different approach. To exploit the given room to its fullest potential, we included a friendly welcome message, implemented the core function of taking inputs from users with the minimum amount of space required, and left the rest in either complete white or light green to give a clean visualizing effect.

### 3. SOLUTION

The implementation of our extension involves the acceptance of user inputs, finding the appropriate websites based on the user's need, maintaining data transmission between frontend and backend, summarizing and analyzing the web content gathered via web scraping, and finally using the result to decide whether or not to block the web pages [10]. To use the extension, the user first needs to go to the website that has the potential of posing a threat to their work efficiency, then enable the extension to allow the content script scraping down the website content. After receiving the content, the content script would pass the data to the background script for further processing. Inside the background script, various NLP and related machine learning techniques were applied to make a precise summarization of the textual data. With the result of summarization, the script would fetch the user's preference from the Firebase database system, compare it against the result, and make the final decision between blocking the website or allowing it to continue to exist, then hand the decision to the content script where it would come into practice. The extension also provides users with the right to adjust and update their preferences of category-based blocking at any moment: In order to make the adjustment, users need to enter a message that indicates their new preference in the extensions pop-up, which will be transferred to the Firebase system to replace the old preference. After the update, the background script will fetch the new preference from the database each time it's needed to make a new comparison.

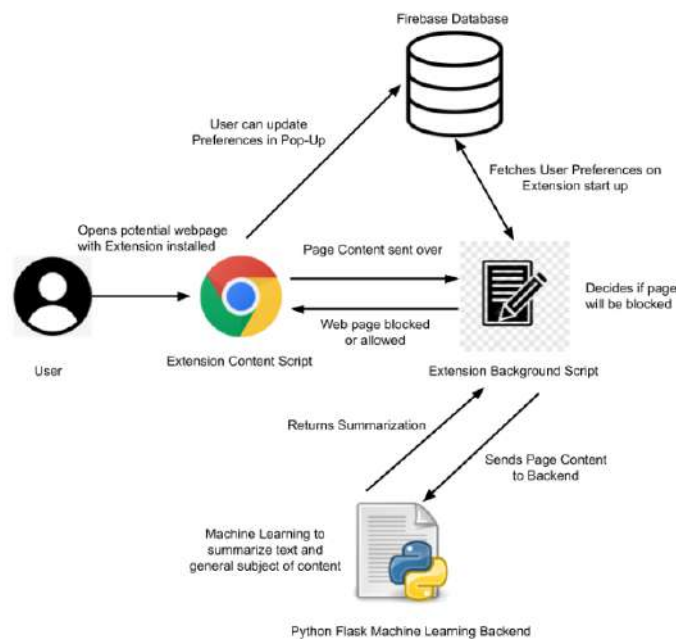


Figure 1. Overview of the solution

```

@app.route("/getSummary/<data>")
def api_call(data):
    site_content = []
    page = data
    soup = BeautifulSoup(page, 'html.parser')
    articles = soup.find_all("p")

    for data in articles:
        if data is not None:
            data = data.text
            while "<" in data and ">" in data:
                index1 = data.find("<")
                index2 = data.find(">")
                tag = data[index1:index2+1]
                data = data.replace(tag, "")
            while "[" in data and "]" in data:
                index1 = data.find("[")
                index2 = data.find("]")
                tag = data[index1:index2+1]
                data = data.replace(tag, "")
            site_content.append(data)
    site_content = " ".join(site_content)
    site_content2 = site_content[(len(site_content)//2):]
    site_content = site_content[:len(site_content)//2]

    generate_summary(site_content, 2)
    return generate_summary(site_content2, 2)

```

Figure 2. Route code

```

def generate_summary(data, top_n = 5):
    print("Generating the summary")
    stop_words = stopwords.words('english')
    summarize_text = []

    sentences = read_sentences(data)
    sentence_similarity_matrix = build_similarity_matrix(sentences, stop_words)
    sentence_similarity_graph = nx.from_numpy_array(sentence_similarity_matrix)
    scores = nx.pagerank(sentence_similarity_graph, max_iter=1000)
    ranked_sentence = sorted(((scores[i], s) for i, s in enumerate(sentences)), reverse = True)

    for i in range(top_n):
        summarize_text.append(" ".join(ranked_sentence[i][1]))

    print("Summarized Text: \n", " ".join(summarize_text) + ".")
    return "Summarized Text: \n", " ".join(summarize_text) + "."

```

Figure 3. Summary code

After users input a keyword to the extension popup, the content script will get executed and trigger the main body of the extension. First, the extension will scrape down the content on the current webpage and pass it to the backend part for summarization using natural language processing. Inside the back end, various NLP and ML techniques will be applied to the data in order to make a precise summarization. First, the script will read the data into separate sentences and filter out special characters. Then the sentences will be passed into another function to make a similarity matrix, which is based on the textual similarity and sentence ranking. Because of the function's ability to find out the similar words and put them into the matrix, the result will directly reflect the main topics discussed in the webpage. The matrix will then be used to rank each sentence and return the top selected ones as the summarization of the entire chunk of data. After having the summary, the extension will compare it against the keyword to see if it contains the word. If so, the current webpage will be categorized as a potentially distracting website and get blocked at the user end. However, if the result is negative, users will still have access to the website content. During this multi-step process, a number of third-party libraries are being used to assist with the precision of the summarization, including NLTK, Numpy, NetworkX, etc. The backbone of the extension lies on the web server created using the Flask library. It is responsible for not only setting up multiple routes for the users to navigate and explore, but also connecting the functionalities of the aforementioned NLP techniques. In addition, the server also takes the role of interacting with the database of the extension. Each time the users update a field in their

setting, the change will get passed back into the Firebase system through the API defined inside the web server [12]. On the user level, if they want to modify, add or remove the categories of websites they want the extension to block, they will be able to do so by clicking the extension popup and making their changes by interacting with it. In this way, the Flask server essentially connects the front end, back end, and the database system for user convenience. The parental control feature also allows parent users to set up a password right after the installation. After the setup of the password, each time the user attempts to modify the block list or update the password, they will be asked to re-enter the password to validate their identity as parents. Through this authentication process, if the user fails to prove their identity, their attempt to make changes will get rejected until the correct password is being entered.

```

IRR_SITES = [
    "The Useless Web." The Useless Web, n.d. https://theuselessweb.com/.

    "Yahoo | Mail, Weather, Search, Politics, News, Finance, Sports & Videos." Yahoo! Yahoo!, n.d.
    https://www.yahoo.com/.
]

GAMES_SITES = [
    "Games - Free Online Games at Addicting Games." Addicting Games, n.d. https://www.addictinggames.com/.

    Jonathan Bolding published 5 November 22, Kerry Brunskill published 5 November 22, Mollie Taylor published
    5 November 22, Ted Litchfield published 4 November 22, Christopher Livingston published 4 November 22, Andy
    Chalk published 4 November 22, Imogen Mellor published 4 November 22, et al. "PC Gamer." pcgamer, November 5,
    2022. https://www.pcgamer.com/.

    "Your Favorites PC/Mac Games up to 70% off! Digital Games, Instant Delivery, 24/7!" Instant, n.d.
    https://www.instant-gaming.com/.

]

MOVIE_SITES = [
    "Ratings, Reviews, and Where to Watch the Best Movies & TV Shows." IMDb. IMDb.com, n.d.
    https://www.imdb.com/.

    "Unlimited Movies, TV Shows, and More." Netflix, n.d. https://www.netflix.com/.

    "Watch Free TV & Movies Online: Stream Full Length Videos." Tubi, n.d. https://tubitv.com/.

]

CAT_SITES = [
    Vicente, Rogério. "HTTP CATS." HTTP Status Cats API. Accessed November 5, 2022. https://http.cat/.

    TheCatSite, n.d. https://thecatsite.com/.

    "Great Artists' Mews." FatCatArt. Accessed November 5, 2022. https://fatcatart.com/.

]

```

Figure 4. Screenshot of code

## 4. EXPERIMENT

### 4.1. Experiment 1

To evaluate the efficiency of our approach of website blocking, we have performed 5 trials at different points of time. In each trial we enter a specific keyword into Concentration, then proceed to access 3 sets of websites (each set consists of 3 should-be-blocked websites that are relevant to the keyword and 2 should-not-be-blocked websites that are irrelevant to the keyword) and record the results (blocked/not blocked) in the tables below. Each table corresponds to a specific keyword as indicated by the table name.

Table 1. Testing websites with the keyword “gaming”

Testing Websites with the keyword "gaming"					
Timestamp	Irreverent Website 1 [2]	Irreverent Website 2 [3]	Relevant Website 1 [4]	Relevant Website 2 [5]	Relevant Website 3 [6]
7am, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
1pm, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
5pm, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
7am, 10/18/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
9pm, 10/18/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked

Table 2. Testing websites with the keyword “movie”

Testing Websites with the keyword "movie"					
Timestamp	Irreverent Website 1 [2]	Irreverent Website 2 [3]	Relevant Website 1 [7]	Relevant Website 2 [8]	Relevant Website 3 [9]
7am, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
1pm, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
5pm, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
7am, 10/18/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
9pm, 10/18/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked

Table 3. Testing websites with the keyword “cats”

Testing Websites with the keyword "cats"					
Timestamp	Irreverent Website 1 [2]	Irreverent Website 2 [3]	Relevant Website 1 [10]	Relevant Website 2 [11]	Relevant Website 3 [12]
7am, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
1pm, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	FN; Not Blocked	TP; Blocked
5pm, 10/17/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
7am, 10/18/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked
9pm, 10/18/22	TN; Not Blocked	TN; Not Blocked	TP; Blocked	TP; Blocked	TP; Blocked

The experiment shows highly consistent and accurate results among 3 website sets. The data from the 1st and 2nd table indicates that Concentration correctly blocked every relevant website and did not block every irrelevant website, showing 100% accuracy in website blocking when the keywords "gaming" and "movie" were entered. The 3rd table shows that Concentration failed to block 1 relevant website and inaccurately blocked 1 irrelevant website, leading to an 93.3% accuracy in correctly blocking relevant websites and 100% accuracy in not blocking irrelevant websites when the keyword "cats" was entered. Overall speaking, Concentration performed fairly well in determining whether to block the page as judged by the high accuracy it achieves in the experiment, which is in alignment with my previous expectation that Concentration will yield accurate and reliable results on most user inputs.



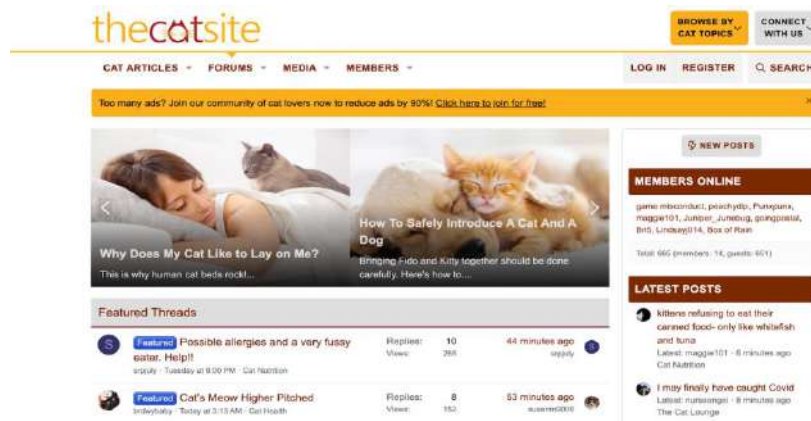


Figure 5. Screenshot of the website that Concentration fails to block when the keyword is "cats"

## 5. RELATED WORK

N K Nagwani [13] performed text summarization on large chunks of textual data through the use of MapReduce framework. By comparison, our work combines summarization and categorization of website contents in relatively smaller sizes. Our algorithm is capable of monitoring real-time websites status, taking note of the specific categories the websites fall into, and blocking the sites accordingly based on the result, while Nagwani's work focuses solely on implementation of efficient means of text summarization and relevant information extraction.

Igor Kotenko et al. [14] presented an automated approach to detect, evaluate, and block websites of inappropriate content intelligently. In their model, they used techniques such as text and html analysis, as well as methods of machine learning and data mining to construct a systematic algorithm of identifying and blocking various web pages and -sites. Our method shares several major similarities with this approach, including the content summarization, category evaluation, and denial of access on the user's end. As different from their model which implements the system using F-Secure platform, ours exploits natural language processing (NLP) technique and Flask server to compose our main functionalities in Python.

Zhang et al. [15] conducted an experiment on the suitability of users' feedback for Web content summarization across several representative social services and concluded the superiority of bookmarking service over others. In addition, they further proved the experimental result by implementing the SSNote system and comparing its output with manual summaries. The paper put the focus on the evaluation of potential impacts of user's activities to the quality of auto-summarization, while ours takes only the website content into account and leaves alone with user's preferences.

## 6. CONCLUSIONS

In this project, we proposed an intelligent approach to auto-block distracting websites based on user inputs using web scraping and natural language processing. The application is implemented as a Chrome extension that scrapes down and summarizes the web content, obtains the keyword that the user previously inputted, compares the summary to the keyword and finally blocks the website if the summary is similar to the keyword. Experiment indicates that the extension performs well across most user inputs, achieving 100% accuracy for input keywords "gaming" and "movie" and  $\geq 93\%$  accuracy for input keyword "cats" and showing high consistency across all blocking results.

After successfully retrieving the web content from the user's browser, it is rather challenging to perform accurate text summarizations on the web content. As a result, the prediction of the exact category the website falls into is not always entirely accurate, especially when a website covers many topics. In addition, since we have to design our UI in limited spaces, it's also difficult to come up with a clean design that provides all necessary functionalities the users may need to use. The project's UI does not increase usability or practicability as much as it could. Finally, the extension takes longer-than-usual time on summarizing for websites that consist of a lot of textual content.

Given these limitations at hand, we can do more research on how we may use more performant NLP libraries and algorithms to achieve more accurate results of summarization in a more timely manner [11]. We can also use users' feedback and suggestions on how we may further improve our UI appearance as the guidelines of our future designs.

## REFERENCES

- [1] Sörqvist, Patrik, and John E Marsh. "How Concentration Shields Against Distraction." *Current directions in psychological science* vol. 24,4 (2015): 267-272. doi:10.1177/0963721415577356
- [2] Connelly, S. Lisa, Lynn Hasher, and Rose T. Zacks. "Age and reading: the impact of distraction." *Psychology and aging* 6.4 (1991): 533.
- [3] Dontre, Alexander J. "The influence of technology on academic distraction: A review." *Human Behavior and Emerging Technologies* 3.3 (2021): 379-390.
- [4] Thomas, David Mathew, and Sandeep Mathur. "Data analysis by web scraping using python." 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2019.
- [5] Monsma, Eva, Melanie Perreault, and Robert Doan. "Focus! Keys to developing concentration skills in open-skill sports." *Journal of Physical Education, Recreation & Dance* 88.7 (2017): 51-55.
- [6] Herring, Susan C. "Web content analysis: Expanding the paradigm." *International handbook of Internet research*. Springer, Dordrecht, 2009. 233-249.
- [7] Livingston, Samuel A., and Charles Lewis. "Estimating the consistency and accuracy of classifications based on test scores." *Journal of educational measurement* 32.2 (1995): 179-197.
- [8] Cambria, Erik, and Bebo White. "Jumping NLP curves: A review of natural language processing research." *IEEE Computational intelligence magazine* 9.2 (2014): 48-57.
- [9] Kikuchi, Yukio. "Numerical simulation of the blocking process." *Journal of the Meteorological Society of Japan*. Ser. II 47.1 (1969): 29-54.
- [10] Abdullah, Hanin M., and Ahmed M. Zeki. "Frontend and backend web technologies in social networking sites: Facebook as an example." 2014 3rd international conference on advanced computer science applications and technologies. IEEE, 2014.
- [11] Cheng, Xinyun, et al. "A combined method for usage of NLP libraries towards analyzing software documents." *International Conference on Advanced Information Systems Engineering*. Springer, Cham, 2020.
- [12] Gu, Xiaodong, et al. "Deep API learning." *Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering*. 2016.
- [13] Nagwani, Naresh Kumar. "Summarizing large text collection using topic modeling and clustering based on MapReduce framework." *Journal of Big Data* 2.1 (2015): 1-18.
- [14] Kotenko, Igor, et al. "Analysis and evaluation of web pages classification techniques for inappropriate content blocking." *Industrial Conference on Data Mining*. Springer, Cham, 2014.
- [15] Park, Jaehui, et al. "Web content summarization using social bookmarks: a new approach for social summarization." *Proceedings of the 10th ACM workshop on Web information and data management*. 2008.



# EVALUATING AND IMPROVING CONTEXT ATTENTION DISTRIBUTION ON MULTI-TURN RESPONSE GENERATION USING SELF-CONTAINED DISTRACTIONS

Yujie Xing and Jon Atle Gulla

Norwegian University of Science and Technology

## ABSTRACT

*Despite the rapid progress of open-domain generation-based conversational agents, most deployed systems treat dialogue contexts as single-turns, while systems dealing with multi-turn contexts are less studied. There is a lack of a reliable metric for evaluating multi-turn modelling, as well as an effective solution for improving it. In this paper, we focus on an essential component of multi-turn generation-based conversational agents: context attention distribution, i.e. how systems distribute their attention on dialogue's context. For evaluation of this component, we introduce a novel attention-mechanism-based metric: DAS ratio. To improve performance on this component, we propose an optimization strategy that employs self-contained distractions. Our experiments on the Ubuntu chatlogs dataset show that models with comparable perplexity can be distinguished by their ability on context attention distribution. Our proposed optimization strategy improves both non-hierarchical and hierarchical models on the proposed metric by about 10% from baselines.*

## KEYWORDS

*Natural Language Processing, Response Generation, Dialogue System, Conversational Agent, Multi-Turn Dialogue System*

## 1. INTRODUCTION

In recent years, generation-based conversational agents have shown a lot of progress, while multiturn generation-based conversational agents are still facing challenges. Most recent work ignores multiturn modelling by considering a multi-turn context as a 1-turn context [1, 2]. Some works try to deal with multi-turn modelling using modified attention mechanisms, hierarchical structures, utterance tokens, etc. [3, 4, 5]. The main difference between multi-turn conversational agents and regular (1-turn) conversational agents is that instead of dealing with an utterance in a context on the word-level, multi-turn models deal with a dialogue on the utterance-level, so that models can understand an utterance as a whole and focus on important utterances rather than important words.

An example of important/unimportant utterances existing in the same context is given by Table 1.

Table 1: An example of important utterances and unimportant utterances under the same context in the Ubuntu chatlog dataset [6]. Unimportant utterances are marked in red.

User	Utterances
Taru	Haha sucker.
Kuja	?
Taru	Anyways, you made the changes right?
Kuja	Yes.
Taru	Then from the terminal type : sudo apt - get update
Kuja	I did.

In this example, the first two utterances (“Haha sucker.” and “?”) are unimportant utterances that are irrelevant to the main topic of the context. Human dialogues naturally contain many of these unimportant utterances. These utterances do not distract humans from understanding the main idea of the context, since humans can easily ignore them and focus instead on important utterances; however, a model usually lacks this capability and can be distracted by these utterances, resulting in a lower performance in generating relevant responses to the main topic of a context. Therefore, it is crucial that a multi-turn model can decide which utterances in the context are important and which are unimportant, and distribute its attention accordingly. In this paper, we define the research topic as context attention distribution, which denotes how much attention is distributed respectively to important and unimportant utterances in a context. A model with a good performance on context attention distribution should pay more attention to important utterances and less attention to unimportant utterances.

Recent work lacks a measurement for the performance of multi-turn modelling. Common metrics rely on general evaluation metrics such as BLEU [7], which measures the quality of generated responses. These metrics cannot directly describe a model’s ability on dealing with multi-turn contexts, since the quality of generated responses is influenced by many aspects. Better performance in dealing with multi-turn context may result in better general performance; however, a better general performance does not necessarily mean that the model has a better ability on dealing with multi-turn contexts. Thus, as a supplementary to general evaluation metrics like BLEU, we propose a metric that measures a conversational agent’s performance on context attention distribution, which is specifically designed for evaluating a model’s performance on multi-turn modelling. Since most multi-turn conversational agents have the attention mechanism and rely on it to distribute attention to different utterances in a context, we propose distracting test as the evaluation method to examine if a model pays more attention to the important utterances. The test adds unrelated utterances as distractions to the context of each dialogue and compares the attention scores of distracting utterances (i.e., unimportant utterances) and original utterances (i.e., important utterances). The ratio of the average attention score of distracting utterances and original utterances is defined as the distracting attention score ratio (DAS ratio). We use DAS ratio as the evaluation metric for a model’s performance on context attention distribution. A model with good capability on context attention distribution should have higher scores on original utterances and lower scores on distracting utterances, thus a lower DAS ratio.

Furthermore, we propose a self-contained optimization strategy to improve a conversational agent’s performance on context attention distribution. For each dialogue, we randomly pick some utterances from the training corpus outside the current dialogue as self-contained distractions, and insert them into the current dialogue with different levels of possibilities. The attention paid to these distractions is minimized during the training process through multi-task learning. With this

optimization strategy, a model learns to distribute less attention to unimportant utterances and thus more attention to important utterances.

In this paper, we examine the following research questions: 1) How do existing multi-turn modeling structures perform on context attention distribution? 2) Can the proposed optimization strategy improve a model’s performance on context attention distribution? 3) Which probability level is the best for inserting distractions in the proposed optimization strategy?

Our contributions are as follows:

- (1) We deal with a less studied problem: evaluating and improving context attention distribution for multi-turn conversational agents.
- (2) We propose a novel evaluation metric for multi-turn conversational agents: DAS ratio. It measures a model’s performance on context attention distribution, i.e. the capability of distributing more attention to important utterances and less to unimportant ones.
- (3) We propose an optimization strategy that minimizes the attention paid to self-contained distractions during the training process, and thus makes the model try to pay less attention to unimportant utterances. The strategy can easily be added and adapted to existing models.

Extensive experiments on 23 model variants and 9 distracting test sets show an overall improvement in the performance on context attention distribution for the proposed strategy. We will share our code for reproducibility (in the final version, a Github link will be provided). Related work is introduced in Section 2. In Section 3, we introduce our base models and proposed methods. We show our experiments settings in Section 4 and results in Section 5. Finally, we give a conclusion in Section 6.

## 2. RELATED WORKS

Common evaluation metrics for conversational agents measure the similarity between the generated responses and the gold responses. Liu et al. [8] summarizes commonly used metrics: word overlap-based metrics (e.g. BLEU) and embedding-based metrics. Bruni et al. [9] propose an adversarial evaluation method, which uses a classifier to distinguish human responses from generated responses. Lowe et al. [10] propose a model that simulates human scoring for generated responses. Zemlyanskiy et al. [11] examine the quality of generated responses in a different direction: how much information the speakers exchange with each other. Recently, Li et al. [5] propose a metric that evaluates the human-likeness of the generated response by measuring the gap between the corresponding semantic influences. Different from the above, our proposed evaluation metric is based on the attention mechanism and is intended to measure a model’s performance on context attention distribution.

Most generation-based conversational agents apply simple concatenation for multi-turn conversation modelling [2, 1], which regards a multi-turn context as a 1-turn utterance. Some works try to model multi-turn conversations through the hierarchical structure: Serban et al. [3, 4] first introduce the hierarchical structure to dialogue models. Tian et al. [12] evaluate different methods for integrating context utterances in hierarchical structures. Zhang et al. [13] further evaluate the effectiveness of static and dynamic attention mechanism. Gu et al. [14] apply a similar hierarchical structure on Transformer, and propose masked utterance regression and distributed utterance order ranking for the training objectives. Different from hierarchical models, Li et al. [5] encode each utterance with a special token [C] and apply a flow module to train the model to predict the next [C]; then they use semantic influence (the difference of the predicted and original tokens) to support generation. In our paper, instead of modelling the relations of

inter-context utterances as [14] or the dialogue flow as [5], our optimization strategy improves multi-turn modelling by n distinguishing important/unimportant utterances directly on the attention mechanism.

### 3. METHODS

Our proposed evaluation metric and optimization strategy can work on attention mechanisms including Transformers. In this paper, we choose an LSTM Seq2Seq model with attention mechanism [15, 16, 17] as the base model, since most hierarchical structured multi-turn conversational agents are based on LSTM [3, 4, 12, 13] while few are based on Transformers.

The basic task of generation-based conversational agents is to predict the next token given all the past and current tokens from the context and response, and to make the predicted response as similar to the original response as possible. Formally, the probability of response  $Y$  given context  $X$  is predicted as:

$$P(Y|X) = \prod_{t=1}^n p(y_t|y_1, \dots, y_{t-1}, X), \quad (1)$$

Where  $X = x_1, \dots, x_m$  and  $Y = y_1, \dots, y_n$  are a context-response pair.

#### 3.1. LSTM Seq2Seq Model with Attention

We simplify an LSTM unit as  $LSTM$ , and we denote the attention version of an LSTM with an asterisk ( $LSTM^*$ ). They are well introduced in previous work [18]. We calculate the hidden vector  $h_t$  at step  $t$  as:

$$h_t = LSTM^*(h_{t-1}, E(z_t), c_{t-1}), \quad (2)$$

where  $h_{t-1} \in R^{dim}$  is the hidden vector at step  $t-1$ ,  $dim$  is the dimensionality of hidden vectors, and  $E(z_t)$  is the word embedding for token  $z_t \in \{x_1, \dots, x_m, y_1, \dots, y_{n-1}\}$ .  $c_{t-1}$  is the context vector at step  $t-1$ , and it is input to the next step  $t$  only in the decoder. Each  $h_t$  and  $c_t$  of the current step  $t$  are combined through a linear layer and an activation to predict the next token.

#### 3.2. Attention Mechanism & Utterance Integration (UI)

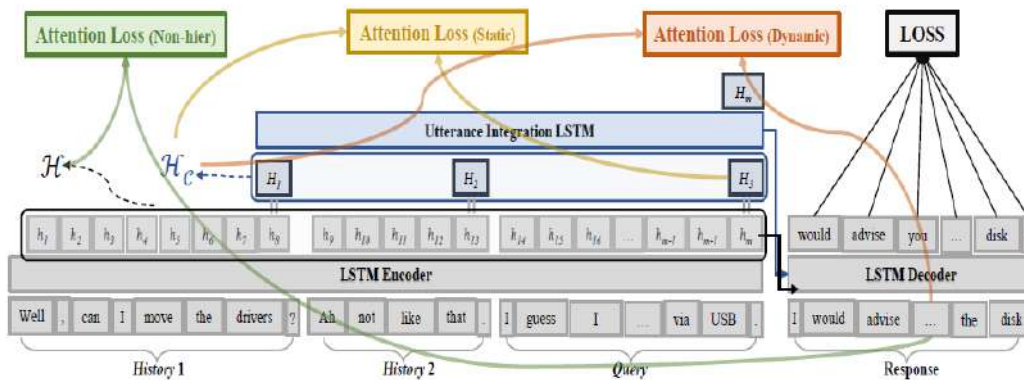


Figure 1: Structure of non-hierarchical, static and dynamic attention loss.

We examine both non-hierarchical and hierarchical structures. For hierarchical structures, following [13], we develop two attention mechanisms: static and dynamic. Following [12], we develop models that are both with and without utterance integration LSTM units. For the non-hierarchical structured model, there are no hidden vectors for utterances. All hidden vectors of tokens in the encoder are concatenated and used in the attention mechanism. Denoting the concatenated vector  $H = [h_1, h_2, \dots, h_m]$ , we calculate the context vector  $c_t$  for each decoding Step  $t$  as:

$$c_t = H \cdot (\text{softmax}(H^\top \cdot h_t)). \quad (3)$$

For the hierarchical models, we use the hidden vector of each utterance's last token as the hidden vector of the utterance, and we discard the hidden vectors for the other tokens. Thus, compared to the non-hierarchical structured model, we have much fewer hidden vectors from the encoder.

The context vector of static attention mechanism is calculated based on the utterance-level concatenated vector and the hidden vector of the last utterance in the context. Denoting the hidden vector of  $k$  th utterance as  $H_k$ , and the hidden vector of the last utterance in the context as  $H_q$ , we have the context's concatenated vector  $\mathcal{H}_C = [H_1, H_2, \dots, H_q]$ . We calculate the context vector  $c_t$  for static attention mechanism as:

$$c_t = \mathcal{H}_C \cdot (\text{softmax}(\mathcal{H}_C^\top \cdot H_q)), \quad (4)$$

where it is easy to see that the static context vector remains unchanged by the decoder.

The context vector of dynamic attention mechanism is calculated based on the utterance-level concatenated vector and the hidden vector of each token in the decoding step. We calculate the context vector  $c_t$  for dynamic attention mechanism as:

$$c_t = \mathcal{H}_C \cdot (\text{softmax}(\mathcal{H}_C^\top \cdot h_t)). \quad (5)$$

Compared to the static attention mechanism, the context vector  $c_t$  varies at each decoding step. Finally, with the utterance integration LSTM unit, we calculate  $H_m$  from  $H_1, H_2, \dots, H_q$ :

$$H_m = LSTM(H_1, H_2, \dots, H_q). \quad (6)$$

For models with utterance integration (UI),  $H_m$  is input to the first step of the decoder, while for models without UI, regular  $h_m$  is input instead.

### 3.3. Distracting Test & Attention Score (AS)

We examine if a multi-turn conversational agent distributes more attention to important utterances through the distracting test and attention scores.

In the distracting test, for each dialogue before the end of the context, we insert several utterances that are irrelevant to the main idea of the dialogue as distractions. These utterances are named *distracting utterances*, and they can be randomly picked utterances from the training corpus (**random**), be formed by frequent words from the training corpus (**frequent**), or be formed by rare words from the training corpus (**rare**). We compare the attention scores of the distracting utterances with the attention scores of the original utterances. A well-performing model should distribute less attention to the distracting utterances while more attention to the original utterances. For an utterance  $H_k$ , the corresponding attention score  $AS(H_k)$  is calculated as:



$$AS(H_k) = \begin{cases} \frac{m}{q} \cdot \text{mean}_t \left( \frac{\sum_{h_t \in H_k} \exp(h_i^\top \cdot h_t)}{\sum_{i=1}^m \exp(h_i^\top \cdot h_t)} \right) & \text{Non-hierarchical} \\ \frac{q \cdot \exp(H_k^\top \cdot H_q)}{\sum_{k=1}^q \exp(H_k^\top \cdot H_q)} & \text{Static attention} \\ \text{mean}_t \left( \frac{q \cdot \exp(H_k^\top \cdot h_t)}{\sum_{k=1}^q \exp(H_k^\top \cdot h_t)} \right) & \text{Dynamic attention} \end{cases} \quad (7)$$

$h_i$  denotes hidden vectors from the encoding steps and  $h_t$  denotes hidden vectors from the decoding steps.  $m$  is the number of tokens in a context, and  $q$  denotes the number of utterances in a context. Note that for non-hierarchical models we multiply by an  $m$  in each  $AS(H_k)$  to avoid bias caused by the total number of tokens in different contexts. Similarly for hierarchical models, we multiply by a  $q$  in each  $AS(H_k)$  to avoid bias caused by the number of total utterances in different contexts. As a result, for an utterance  $H_q$ ,  $AS(H_q)$  will be 100% (or approximately 100% for non-hierarchical models) if the model assigns  $H_q$  an about average attention score among all utterances. We denote the last utterance in a context as *Query* and the rest of utterances in the context as *History*. Since different models have different scalars on attention scores, we calculate the average AS for all distracting utterances and all *History* in each dialogue, and use the ratio of them for evaluation. This ratio is denoted as distracting attention score ratio (**DASratio**), which measures a model's ability on context attention distribution:

$$\text{DAS ratio} = \text{mean}_{d \in D} \left( \frac{\text{mean}(AS(H_{\text{Distraction}}))}{\text{mean}(AS(H_{\text{History}}))} \right), \quad (8)$$

where  $d$  means a single dialogue, and  $D$  denotes all dialogues in a test set.  $H_{\text{Distraction}}$  denotes distracting utterances, and  $H_{\text{History}}$  denotes utterances in *History*.

### 3.4. Optimization with Self-Contained Distractions on Attention Mechanism

To train a conversational model to distribute more attention to important and less attention to unimportant utterances, we propose the following optimization strategy: 1) For each dialogue, we select some random utterances from other dialogues in the training corpus as self-contained distractions. We decide whether to insert these distractions into the current dialogue or not stochastically by a probability level. We denote the probability level as the training inserting probability. The locations of inserting distractions are randomly decided, while the locations are always before Query (the last utterance of the context). 2) We create a bitmask  $M$  to track whether an utterance is original (0) or distracting (1). During the training period, the model uses the bitmask to calculate the attention loss  $\mathcal{L}_{\text{attention}}^t$ , which is summed up with the loss from the response generator. For each decoding step  $t$ , the attention loss is calculated as:

$$\mathcal{L}_{\text{attention}}^t = \begin{cases} \text{MSE}(\text{softmax}(\mathcal{H}^\top \cdot h_t) \circ M, 0) & \text{Non-hierarchical} \\ \text{MSE}(\text{softmax}(\mathcal{H}_C^\top \cdot H_q) \circ M, 0) & \text{Static attention} \\ \text{MSE}(\text{softmax}(\mathcal{H}_C^\top \cdot h_t) \circ M, 0) & \text{Dynamic attention} \end{cases} \quad (9)$$

where  $\circ$  means Hadamard product, or element wise multiplication. As shown in Equation (9), our goal is to minimize the attention assigned to all the self-contained distractions. During the distracting test, no bitmask is offered to the model. The illustration of attention loss on both non-hierarchical and hierarchical models is shown in Figure 1.

## 4. EXPERIMENTS

In this section, we introduce the setups of the experiment.

### 4.1. Dataset

We use the Ubuntu chatlogs data set [6] as the training and testing corpus, which contains dialogues about solving technical problems of Ubuntu. We choose this dataset because the dialogues have both technical topics and casual chats, meaning that it is easier to distinguish important/unimportant utterances than datasets whose topics are consistent. We use about 0.48M dialogues for training, 20K dialogues for validation, and 10K dialogues for testing. These are the original settings of the Ubuntu chatlogs dataset. We removed all single-turn dialogues.

### 4.2. Training

Our methods are built on an LSTM Seq2Seq model with attention mechanism. We used Pytorch [19] for implementation. The LSTM model has 4 layers and the dimension is 512. The training procedure was with a batch size of 256, a learning rate of 1.0, and a gradient clip threshold of 5. The vocabulary size is 25000 and the dropout rate is 0.2. The learning rate is halved when the perplexity stops dropping, and the training is stopped when the model converges.

### 4.3. Examined Models

We examine our proposed evaluation metric on 5 models: non-hierarchical LSTM (Non-hier), static attention without utterance integration LSTM unit (Static), static attention with utterance integration LSTM unit (StaticUI), dynamic attention without utterance integration LSTM unit (Dynamic), and dynamic attention with utterance integration LSTM unit (DynamicUI). In addition, we examine our proposed optimization strategy on these 5 models with 3 training inserting probabilities—0.5, 0.7, and 1.0. Models with a training inserting probability of 0 are regarded as baselines. For comparison, we pick the best overall model and train the model with self-contained distractions but without training on the attention loss (Non-atten-loss), i.e. the model does not know which utterances are distractions. In total, we train and evaluate 23 model variants.

### 4.4. Evaluation

Table 2: Examples of distracting test sets. Distracting utterances are marked red.

	<b>Random: 0.5</b>	<b>Random: 0.7</b>	<b>Random: 1.0</b>
<i>History</i>	\	Well, can I move the drives?	<b>Yes.</b>
	<b>Or kill all speedlink.</b>	<b>Anyways, you made the Changes right?</b>	Well, can I move the drives?
	Well, can I move the drives?	Ah not like that.	<b>Then from the terminal type: sudoapt-get update.</b>
	Ah not like that.	<b>I did.</b>	Ah not like that.
	<b>Frequent: Begin</b>	<b>Frequent: Middle</b>	<b>Frequent:End</b>
<i>History</i>	<b>Why should I help you?</b>	Well, can I move the drives?	Well, can I move the drives?
	<b>I have my right.</b>	<b>Why should I help you?</b>	Ah not like that.
	Well, can I move the drives?	<b>I have my right.</b>	<b>Why should I help you?</b>
	Ah not like that.	Ah not like that.	<b>I have my right.</b>
	<b>Rare:Begin</b>	<b>Rare:Middle</b>	<b>Rare:End</b>
<i>History</i>	<b>Would you have lunch?</b>	Well, can I move the drives?	Well, can I move the drives?
	<b>I should have lunch.</b>	<b>Would you have lunch?</b>	Ah not like that.
	Well, can I move the drives?	<b>I should have lunch.</b>	<b>Would you have lunch?</b>
	Ah not like that.	Ah not like that.	<b>I should have lunch.</b>
<i>Query</i>	<b>I guess I could just get an enclosure and copy via USB.</b>		
<i>Response</i>	<b>I would advise you to get the disk.</b>		

For the distracting test, we set the number of distracting utterances for each dialogue to 2. We chose 2 to make the distracting utterances a complete turn and to make the number of distracting utterances the minimum, since dialogues from the corpus normally have only 4 to 8 utterances in the contexts. We have 3 distracting test sets. 1) Random distracting test set: distracting utterances in this test set are randomly picked from the training corpus (outside the current dialogue), and they are randomly picked in every evaluation step, which means that there is no pre-prepared random distracting test set. 2) Frequent distracting test set: distracting utterances in this test set are formed by frequent words in the training corpus, but these utterances do not appear in the training corpus. In our experiments, we use “why should I help you” and “I have my right” as examples of distracting utterances with frequent words. 3) Rare distracting test set: distracting utterances in this test set have words that are rare in the training corpus, and these utterances do not appear in the training corpus. In our experiments, we use “would you have lunch?” and “I should have lunch” as examples of distracting utterances with rare words.

In the distracting test, we insert distracting utterances into different locations. For 1) random, we insert utterances to a random location before Query in each context. Similar to the optimization strategy, we use different probability levels to decide whether a distracting utterance is to be inserted or not. We denote these as testing inserting probability. In our experiments, we set the probability levels to be 0.5, 0.7, and 1.0. We expect the model to perform stably on all different probability levels. For 2) frequent and 3) rare, we have three kinds of inserting locations: at the beginning of a context (marked as Begin), in the middle of the context (marked as Middle), and at the end of the context (before Query and after History, marked as End). In total, we have 9 test sets for evaluation. See Table 2 for the example of each test set.

## 5. RESULTS AND DISCUSSIONS

Table 3 illustrates the main results on DAS ratios. It shows the DAS ratios of 23 trained model variants on 9 distracting test sets. Figure 2 shows the DAS ratios of 3 example model variants

(StaticUI with training inserting probability of 0.0 as the baseline, Non-atten-loss StaticUI with training inserting probability of 0.7, and StaticUI with training inserting probability of 0.7) on 9 distracting test sets. Table 4, Table 5 and Table 6 show the detailed results on average Attention Score (average AS) of distracting utterances and average AS of History.

In Table 3, we show the perplexity and History’s average AS of each model on the non-distracted test set under the “Original” column. Since perplexity scores on the distracting test sets are similar, we show the perplexity scores on the non-distracted test set only. We show the DAS ratios of each model on each of the distracting test sets under the “DAS ratio for distracting test set” column. A lower DAS ratio means that a model distributes less attention to distracting utterances (unimportant utterances) and more attention to the original utterances in History (important utterances), from which it can be inferred that the model has better performance on context attention distribution. Both perplexity and DAS ratio are the lower, the better.

### 5.1. Perplexity and Average AS on Non-Distracted Test Set

Perplexity scores are shown in the “Perp.” column, under the “Original” column in Table 3. Perplexity scores of the examined 23 models are similar; the Static models trained with our proposed optimization strategy and a higher training inserting probability level achieves slightly better performance than other models. Average AS are shown in the “Avg.” column, under the “Original” column in Table 3. The average AS of History tells about a model’s attention distribution for History and Query. A higher score indicates that less attention is distributed to Query. Recall that AS of an utterance is 100% (or approximately 100% for non-hierarchical models) if the utterance is paid about average attention among the dialogue. Overall, the models distribute attention of lower than average to History, especially for models with static attention (i.e. the Static model and StaticUI model), which distribute more attention to Query than non-hierarchical models and models with dynamic attention.

This is apparent from the structure of static attention. We also show the results of a StaticUI model without training on the attention loss (Non-atten-loss StaticUI model) as a comparison. The StaticUI model trained with our optimization strategy distributes more attention to query than the Non-atten-loss StaticUI model. This is because the optimization strategy decreases the model’s attention distributed to distracting utterances in History, thus decreasing the overall attention distributed to History.

### 5.2. Distracting Test: Random

Results of the random distracting test with different testing inserting probabilities (0.5, 0.7, and 1.0) are shown in the “Random” column in Table 3. Models with training inserting probabilities of 0.0 (shown in the row where “Prob” is 0.0) are baseline models to which our proposed optimization strategy is not applied. In general, our proposed optimization strategy with training inserting probabilities of 0.5 or 0.7 achieves better performance on DAS ratios (i.e. the models achieve lower DAS ratios) on random distracting test sets of all 3 testing inserting probabilities. The Static model and the DynamicUI model achieves the best performance with a training inserting probability of 0.5, while the Non-hier model, the StaticUI model and the Dynamic model achieve the best performance with a training inserting probability of 0.7. A training inserting probability of 1.0 leads to worse performance. One reason is that it assumes there must be some distracting utterances in a context, while that is not always the case.

Table 3: Results of perplexity (Perp.) and average AS of History (Avg.) on the original test set (%) are shown in the “Original” column. We also show results of DAS ratios on 9 distracting test sets and 23

model variants.

Prob	Model	Original		DAS ratio on distracting test sets											
		Perp.	Avg.	Random			Frequent			Rare					
	Structure			0.5	0.7	1.0	Begin	Middle	End	Begin	Middle	End	Begin	Middle	End
0.0	Non-hier	43.2	91.3	0.93	0.93	0.93	0.75	0.80	0.84	0.80	0.92	1.01	0.80	0.92	1.01
	Static	44.1	61.4	0.82	0.82	0.79	0.37	0.80	1.31	0.80	0.77	1.21	0.37	0.77	1.21
	StaticUI	44.6	57.5	0.79	0.76	0.76	<b>0.32</b>	0.75	1.32	0.75	0.75	1.22	0.30	0.75	1.22
	Dynamic	45.4	81.4	0.89	0.89	0.88	0.65	0.86	1.02	0.86	0.89	1.06	0.66	0.89	1.06
	DynamicUI	44.7	91.6	0.94	0.94	0.93	0.72	0.84	0.86	0.84	0.93	0.97	0.73	0.93	0.97
0.5	Non-hier	43.4	87.2	0.84	0.83	0.81	0.63	0.74	<b>0.76</b>	0.74	0.81	0.86	0.69	0.81	0.86
	Static	44.5	66.5	0.70	0.69	0.67	0.42	0.78	1.12	0.78	0.71	0.99	0.34	0.71	0.99
	StaticUI	44.3	47.7	0.74	0.74	0.70	0.39	0.71	1.08	0.71	<b>0.69</b>	0.96	0.40	<b>0.69</b>	0.96
	Dynamic	44.6	81.9	0.79	0.78	0.77	0.64	0.74	0.84	0.74	0.77	0.85	0.61	0.77	0.85
	DynamicUI	43.9	86.7	0.82	0.81	0.80	0.60	0.84	0.87	0.84	0.80	0.83	0.61	0.80	0.83
0.7	Non-atten-loss StaticUI	44.7	71.1	0.73	0.73	0.72	0.39	0.68	0.93	0.68	0.80	1.11	0.40	0.80	1.11
	Non-hier	43.2	86.9	0.84	0.82	0.80	0.72	0.82	0.82	0.82	0.85	0.87	0.71	0.85	0.87
	Static	<b>44.0</b>	57.6	0.73	0.72	0.69	0.40	0.70	1.08	0.70	0.70	0.98	0.41	0.70	0.98
	StaticUI	44.9	43.7	<b>0.67</b>	<b>0.67</b>	<b>0.65</b>	0.36	<b>0.66</b>	1.02	<b>0.66</b>	0.70	0.99	0.36	0.70	0.99
	Dynamic	44.3	82.0	0.76	0.75	0.73	0.58	0.71	0.86	0.71	0.73	0.83	0.58	0.73	0.83
1.0	DynamicUI	44.8	85.3	0.93	0.93	0.93	0.45	0.78	0.80	0.78	0.80	<b>0.81</b>	0.60	0.80	<b>0.81</b>
	Non-atten-loss StaticUI	44.1	55.4	0.72	0.70	0.69	0.45	0.70	0.98	0.70	0.73	0.97	0.43	0.73	0.97
	Non-hier	47.3	95.9	0.91	0.90	0.90	0.84	0.86	0.85	0.86	0.87	0.88	0.85	0.87	0.88
	Static	<b>44.0</b>	65.4	0.70	0.70	0.68	0.49	0.74	1.08	0.74	0.71	0.88	0.46	0.71	0.88
	StaticUI	49.6	73.5	0.96	0.95	0.94	0.66	0.86	1.53	0.86	<b>0.21</b>	1.50	<b>0.21</b>	0.86	1.50
1.0	Dynamic	44.7	88.8	0.79	0.78	0.77	0.63	0.75	0.82	0.75	0.77	0.82	0.65	0.77	0.82
	DynamicUI	45.2	90.2	0.87	0.86	0.85	0.73	0.81	0.83	0.81	0.88	0.88	0.75	0.88	0.88
	Non-atten-loss StaticUI	44.1	76.5	0.72	0.71	0.69	0.49	0.74	0.98	0.74	0.77	0.98	0.49	0.77	0.98



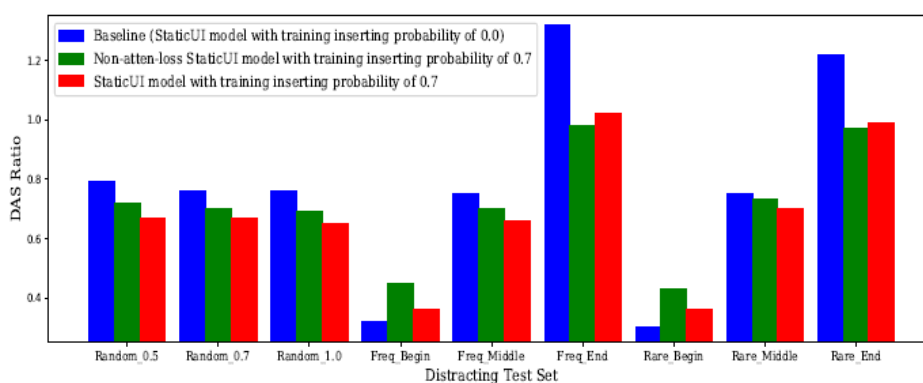


Figure 2: DAS ratios of 3 example model variants on 9 distracting test sets. The lower the DAS ratio, the better the performance.

The StaticUI model with a training inserting probability of 0.7 achieves the best overall performance on DAS ratio. As shown in Figure 2, on all the random distracting test sets (probabilities of 0.5, 0.7, and 1.0), the StaticUI model is better than the baseline StaticUI model and the Non-atten-loss StaticUI model. The baseline model is not trained with any self-contained distractions (training inserting probability is 0.0), and it gets the worst performance. The Non-atten-loss model is trained with self-contained distractions (with a training inserting probability of 0.7) while not knowing which utterances are distractions, and it achieves a better performance than the baseline. The StaticUI model with a training inserting probability of 0.7 is trained to minimize the attention loss of self-contained distractions and it achieves the best performance. Naturally since the optimization strategy minimizes the attention loss of distractions, the StaticUI model distributes less attention to History and more attention to Query (refer to the “Avg” column in Appendix 4 for more details); nevertheless, a lower DAS ratio shows that the model distributes even less attention to the distracting utterances compared to the original utterances in History.

Note that even if both our proposed strategy and the random distracting test use the same trick: insert random distracting utterances among original utterances in History, the random utterances inserted in the distracting test are different from those inserted in the training process, thus it is difficult for the test to be biased in favor of models with our proposed strategy. It is apparent that less attention is distributed to History, while DAS ratio calculates the ratio between the distracting utterances and the original utterances in History, so it shows the attention distributed to the distracting utterances regardless of the total attention distributed to History. Moreover, we adopt three testing inserting probability levels to ensure stable evaluation results for each model.

### 5.3. Distracting Test: Frequent and Rare

Results of the frequent and the rare distracting test are shown in the “Frequent” and “Rare” columns in Table 3. Different from the random distracting test, the inserting locations of these two tests are decided manually. As a nature of LSTM model, all models distribute more attention to utterances near Query and less attention to utterances far away from Query, as can be seen in Table 3 and Figure 2 that DAS ratios are higher for End test set (near Query) and lower for Begin test set (far away from Query). Since the results on Begin and End test sets are biased by the structure of LSTM, we mainly analyze the results on Middle test sets. For the Middle test sets of both the frequent and rare distracting test, the best models are still those trained with our proposed optimization strategy. StaticUI models with training inserting probabilities of 0.5 and 0.7 achieve the best performance (lowest DAS ratios) on the Frequent Middle and Rare Middle

test sets. The Non-atten-loss models can be better than the models trained with a wrong training inserting probability. Telling from similar DAS ratios, the frequent distracting test set is as difficult for the trained models to distinguish as the rare distracting test set, although for humans, the rare distracting utterances are much easier to distinguish than the frequent ones.

Table 4: Results of perplexity (Perp.) and average AS of History (Avg.) on the original test set (%) are shown in the ‘‘Original’’ column. Besides, we show the results on the random distracting test of: DAS ratio, average AS of distracting utterances (DAS) (%), and average AS of original utterances in History (Avg.) (%).

Probability	Model	Original		Distracting Test Set											
		Perp.	Avg.	Random 0.5			Random 0.7			Random 1.0					
	Structure			DAS ratio	DAS	Avg.	DAS ratio	DAS	Avg.	DAS ratio	DAS	Avg.	DAS ratio	DAS	Avg.
0.0	Non-hier	43.2	91.3	0.93	86.8	93.1	0.93	87.1	93.8	0.93	87.6	94.5	0.93	87.6	94.5
	Static	44.1	61.4	0.82	52.6	64.5	0.82	53.5	65.6	0.79	53.4	67.6	0.79	53.4	67.6
	StaticUI	44.6	57.5	0.79	47.4	60.2	0.76	46.3	61.3	0.76	47.7	62.7	0.76	47.7	62.7
	Dynamic	45.4	81.4	0.89	74.9	84.4	0.89	75.6	85.2	0.88	76.2	86.6	0.88	76.2	86.6
	DynamicUI	44.7	91.6	0.94	87.5	93.4	0.94	88.0	93.7	0.93	88.2	94.5	0.93	88.2	94.5
	Non-hier	43.4	87.2	0.84	77.2	91.6	0.83	77.1	93.3	0.81	77.0	95.5	0.81	77.0	95.5
0.5	Static	44.5	66.5	0.70	50.3	71.5	0.69	50.5	73.5	0.67	51.1	76.5	0.67	51.1	76.5
	StaticUI	44.3	47.7	0.74	38.1	51.2	0.74	39.2	53.1	0.70	39.1	55.5	0.70	39.1	55.5
	Dynamic	44.6	81.9	0.79	68.3	86.6	0.78	69.1	88.2	0.77	69.4	90.8	0.77	69.4	90.8
	DynamicUI	43.9	86.7	0.82	74.5	91.1	0.81	75.2	92.5	0.80	75.8	94.8	0.80	75.8	94.8
	Non-labelled	44.7	71.1	0.73	55.5	75.6	0.73	56.6	77.1	0.72	57.4	79.6	0.72	57.4	79.6
	StaticUI	43.2	86.9	0.84	76.5	91.3	0.82	75.9	93.1	0.80	75.9	95.4	0.80	75.9	95.4
0.7	Static	<b>44.0</b>	57.6	0.73	45.5	62.2	0.72	45.8	64.0	0.69	46.4	66.9	0.69	46.4	66.9
	StaticUI	44.9	<b>43.7</b>	<b>0.67</b>	<b>32.4</b>	<b>48.1</b>	<b>0.67</b>	<b>33.2</b>	<b>49.9</b>	<b>0.65</b>	<b>34.0</b>	<b>52.1</b>	<b>0.65</b>	<b>34.0</b>	<b>52.1</b>
	Dynamic	44.3	82.0	0.76	66.3	87.2	0.75	66.6	89.1	0.73	67.2	91.8	0.73	67.2	91.8
	DynamicUI	44.8	85.3	0.93	86.8	93.1	0.93	87.1	93.8	0.93	87.6	94.5	0.93	87.6	94.5
	Non-labelled	44.1	55.4	0.72	43.3	59.9	0.70	43.4	62.0	0.69	44.3	64.4	0.69	44.3	64.4
	StaticUI	47.3	95.9	0.91	88.7	98.0	0.90	89.3	98.7	0.90	89.5	99.9	0.90	89.5	99.9
1.0	Static	<b>44.0</b>	65.4	0.70	49.7	71.1	0.70	51.3	73.1	0.68	51.8	76.4	0.68	51.8	76.4
	StaticUI	49.6	73.5	0.96	74.8	77.8	0.95	75.2	79.4	0.94	76.7	81.2	0.94	76.7	81.2
	Dynamic	44.7	88.8	0.79	74.2	93.4	0.78	74.4	95.2	0.77	75.4	97.4	0.77	75.4	97.4
	DynamicUI	45.2	90.2	0.87	81.3	93.6	0.86	81.5	94.9	0.85	81.9	96.5	0.85	81.9	96.5
	Non-labelled	44.1	76.5	0.72	59.5	82.1	0.71	59.7	84.4	0.69	60.3	87.6	0.69	60.3	87.6
	StaticUI	44.1	76.5	0.72	59.5	82.1	0.71	59.7	84.4	0.69	60.3	87.6	0.69	60.3	87.6







#### 5.4. Detailed Results on the Distracting Tests

In addition to DAS ratio, Table 4 shows the average AS of distracting utterances and of original utterances in *History*. Table 5 and Table 6 additionally show the AS of the first or last utterances in *History*. Note again that an attention score of 100% for a utterance indicates that this utterance receives an average attention score, e.g. for a dialogue containing 10 utterances, an attention score of 100% indicates that the utterance receives 10% attention out of all.

From Table 4 it is clear that the average AS of the original utterances in *History* varies by model variants. A higher average AS for *History* indicates a lower AS for *Query*. Some models distribute most of the attention to *Query* while some models distribute the attention evenly to both *History* and *Query*. Normally, *Query* contains more relevant information, so we expect a lower average AS for *History*; however, the average AS for *History* is not the lower the better, since there are still some utterances in *History* that are important for the context. A lower average AS for *History* comes together with a lower average AS for distracting utterances (or a lower DAS), so DAS ratio is better suited for evaluating a model's capability on context attention distribution, since it takes the average AS for original utterances in *History* into account. In Table 4, the models with the lowest DAS ratio also have the lowest average AS for distracting utterances and original utterances, while in Table 5 and Table 6, it is not always the case.

In Table 5 and Table 6, for the distracting test sets where distracting utterances are put in the beginning/end of the context, we show AS for the first/last utterance in *History* to have a clearer comparison. We can see in columns of Frequent: Begin and Rare: Begin that the distracting utterances usually receive lower attention than the first utterance in *History*, while the other original utterances in *History* receive more attention than the first utterance. This indicates a good performance of the model variants. Utterances far away from *Query* are normally distributed lower attention, so in a normal case, it is natural that the utterances that come after the first utterance receive more attention; however, these distracting utterances receive less attention, regardless of the fact that they are placed after the first utterances. It can thus be inferred that most model variants can distinguish distracting utterances as unimportant and distribute less attention to them. Similarly, the last utterances in *History* usually get more attention, while as the columns of Frequent: End and Rare: End show, distracting utterances receive less attention compared to other original utterances in *History*, regardless of that the distracting utterances are placed closer to *Query*.

#### 5.5. Summary of Results

DAS ratio can distinguish conversational agents with similar perplexity on their ability of context attention distribution. In general, models trained with our proposed optimization strategy focus less on distracting utterances and more on original utterances in *History*. For most models, DAS ratios decrease by about 10% when trained with our proposed strategy with a 0.5 or 0.7 probability level. 0.7 is generally the best option for a training inserting probability.

### 6. CONCLUSIONS AND FUTURE WORKS

We have studied context attention distribution, an essential component of multi-turn modelling for open-domain conversational agents. We have proposed an evaluation metric for context attention distribution based on the distracting test: DAS ratio. We have also improved the performance of context attention distribution for common multi-turn conversational agents through an optimization strategy via reducing the attention loss of self-contained distracting utterances. Extensive experiments show that our proposed strategy achieves improvements on

most models, especially with a training inserting probability level of 0.7. Future works can focus on adapting the proposed evaluation metric and optimization strategy to transformer-based conversational agents.

## ACKNOWLEDGEMENTS

This paper is funded by the collaborative project of DNB ASA and Norwegian University of Science and Technology (NTNU). We also received assist on computing resources from the IDUN cluster of NTNU [20]. We would like to thank Benjamin Kille and Peng Liu for their helpful comments.

## REFERENCES

- [1] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-demos.30>
- [2] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3377–3390. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.272>
- [3] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957>
- [4] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567>
- [5] Z. Li, J. Zhang, Z. Fei, Y. Feng, and J. Zhou, "Conversations are not flat: Modeling the dynamic information flow across dialogue utterances," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 128–138. [Online]. Available: <https://aclanthology.org/2021.acl-long.11>
- [6] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2015, pp. 285–294. [Online]. Available: <http://aclweb.org/anthology/W15-4640>
- [7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: <https://aclanthology.org/P02-1040>
- [8] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 2122–2132. [Online]. Available: <http://aclweb.org/anthology/D16-1230>
- [9] E. Bruni and R. Fernandez, "Adversarial evaluation for open-domain dialogue generation," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 2017, pp. 284–288. [Online]. Available: <http://aclweb.org/anthology/W17-5534>
- [10] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

- Association for Computational Linguistics, 2017, pp. 1116–1126. [Online]. Available: <http://aclweb.org/anthology/P17-1103>
- [11] Y. Zemlyanskiy and F. Sha, “Aiming to Know You Better Perhaps Makes Me a More Engaging Dialogue Partner,” in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2018, pp.551–561. [Online]. Available: <http://aclweb.org/anthology/K18-1053>
- [12] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao, “How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2017, pp. 231–236. [Online]. Available: <http://aclweb.org/anthology/P17-2036>
- [13] W. Zhang, Y. Cui, Y. Wang, Q. Zhu, L. Li, L. Zhou, and T. Liu, “Context- Sensitive Generation of Open-Domain Conversational Responses,” in *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 2437–2447. [Online]. Available: <http://aclweb.org/anthology/C18-1206>
- [14] X. Gu, K. M. Yoo, and J.-W. Ha, “DialogBERT: Discourse-aware response generation via learning to recover and rank utterances,” in *In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*, 2021.
- [15] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: <http://dx.doi.org/10.1162/neco.1997.9.8.1735>
- [16] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>
- [17] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [18] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, “A Persona-Based Neural Conversation Model,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016, pp. 994–1003. [Online]. Available: <http://aclweb.org/anthology/P16-1094>
- [19] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *NIPS-W*, 2017.
- [20] M. Sjölander, M. Jahre, G. Tufte, and N. Reissmann, “EPIC: An energy-efficient, highperformance GPGPU computing research infrastructure,” 2019.



# IMPLEMENTATION OF A NEW E-VOTING SYSTEM BASED ON BLOCKCHAIN USING ECDSA WITH BLIND SIGNATURES

Lina Lumburovska, Vesna Dimitrova, Aleksandra Popovska-Mitrovikj

Ss. Cyril and Methodius University of Skopje, Faculty of Computer Science and Engineering, Skopje, North Macedonia

## **ABSTRACT**

*The latest research shows the benefits, the impact, and the usage of Blockchain and decentralized systems with a high confidence. Its popularity becomes even higher with the electronic voting systems based on the technology itself. In this paper we propose a new implementation of an electronic voting system based on Blockchain using ECDSA with blind signatures. Additionally, the system is compared with other electronic voting systems based on Blockchain technology. Mainly these types of systems hardly ever fulfill the scalability. Nevertheless, our system has an advantage in comparison with the other systems. Since the idea of the Blockchain technology is to show the flexibility and equal privileges to all nodes, this implementation with Angular and Spring Boot shows that, so everyone can track the chain. To sum up, this implementation can have a good usage in smaller departments, because of the performances and all mathematical operations.*

## **KEYWORDS**

*Blockchain technology, ECDSA, e-voting, blind signatures*

## **1. INTRODUCTION**

The rapid development of the digitalization and technology increases the need for replacing many processes of everyday life from their traditional way of working to an electronic version of it. The motivation for building an electronic voting system stem from this replacement and can be stated as a direct consequence. On the other side, the popularity of decentralization was the idea behind the Blockchain technology [1] and that is how the centralized systems were also replaced. Having one central node that operates the entire process in one system is not the best approach because in the history there were examples where the data has been changed or abused. Replacing the traditional system with a decentralized system where all elements have the same privileges to it is the main structure of the Blockchain systems [2, 3].

Building an electronic voting system based on Blockchain is definitely a great example how this technology can be used in practice and how the decentralized systems are constructed. In this certain example each vote in the election can be represented as one block in the system. The Blockchain technology is consisted of four elements: peer-to-peer network, cryptography, consensus algorithm and punishment or reward. As soon as these requirements are fulfilled, the system can be created [4].

In the peer-to-peer network all elements are called nodes and they have the same role in the system, where they can asynchronously communicate from different places and time zones. The

choice of the cryptography has a major part in the security of the system, so choosing a right cryptographic algorithm has a huge impact on the behavior of the system. Every new node that wants to join the system must prove that it is valid and valuable to the system by solving a given problem. The consensus algorithm can be of different types, but the most used is the proof of work. Based on this third element, the new node can get a reward or punishment if the problem is solved or not [5]. Having these elements in mind, the Blockchain technology can be implemented and used in practice. The Blockchain technology can be used in different sectors such as: medicine, law, IoT (Internet of things), artificial intelligence, cyber security, electronic voting etc. For this paper, the last usage of the above listed is researched.

The last usage of the Blockchain technology can contribute a lot in time when voting is happening [6], since the counting of the votes is not done physically, but instead it is automated. The counting of the votes is done in real time, and the admin can see the statistics over the whole voting process. The hardest part in this case is obtaining the anonymity, so there is no way how the admin or the other users can see which user voted for which candidate. It is important to notice that when we are talking about voting, we do not think always about political voting but instead it is more meant as part of one organization such as: dean elections at the faculty, choosing the project manager within one department and so on.

Implementing an electronic voting system based on Blockchain is a modern topic that professionals work on in the last years. The hardest part in this implementation is keeping the anonymity of the voters, where there is no way how the admin or the other users in the system can find which user voted for which candidate and the vote is still counted in the main statistics. For that purpose, the implementation of the electronic voting system based on Blockchain is done using ECDSA (Elliptic Curve Digital Signature Algorithm) with blind signatures. Additionally, interactive zero-knowledge proof is used to prove the security of this algorithm. There are many electronic voting systems, and this system is also compared with other electronic voting systems based on Blockchain. The main problem that most of these systems have is the problem with the scalability [7]. Our newly proposed system is compared with different systems that are made by companies and individuals. Based on our roles, we are comparing the system in the second group, and we concluded that the system has a huge advantage that the other systems do not have. Since Blockchain technology is everything about equal roles and privileges, we decided to include the chain in the application, so the users can see the flow of the voting process. We believe that the UX changes in the application can contribute to the usage of the system. Apart from that, the ECDSA with blind signatures fulfills the security issues for the system and the main disadvantage is the performance, due to the long-lasting mathematical operations. For that purpose, we limit the usage of our system in smaller organizations and departments.

The structure of the paper starts with the introduction to the topic, followed by a section for explaining the implementation into technical details (the technologies behind it, the ECDSA algorithm itself, different types of zero-knowledge proof). The third section is focused on the comparison of this system and other systems that are build using the Blockchain technology (some of the systems are created by companies and some of them are created by individuals). The paper ends with a conclusion an acknowledgment of the authors.

## **2. TECHNICAL IMPLEMENTATION OF ELECTRONIC VOTING SYSTEM BASED ON BLOCKCHAIN USING ECDSA WITH BLIND SIGNATURES**

### **2.1. Technology in Use**

The system is implemented as a web application which is made using Angular and Java (Spring Boot). The front-end of the system is implemented in Angular [8] which is a modern Typescript framework and uses Bootstrap library for a beautiful styling. On the other side, the back-end is implemented in Java with Spring Boot [9]. The database used in this system is a relational H2 base where the data is queried using SQL language. In these cases, the usage of Spring Boot reduces the need for SQL queries, so the basic CRUD (Create-Read-Update-Delete) operations are already implemented using the Repository of Spring Boot with no additional code. The anonymity implementation for hiding the voter's choice is calculated in the back-end. Basically, the combination of ECDSA with blind signature is not that straight forward and does not work without any additional changes. This combination only works with a modified Paillier cryptosystem. ECDSA uses SHA256 as a cryptographic hash function and this is used as the second element of the Blockchain technology. One of the reasons why Angular is chosen for the front-end is the increase scalability that provides this framework which covers to some extent the disadvantages of the electronic voting systems on a global level. Spring Boot is mainly used for microservices and in this case is also a good solution because the module for the ECDSA is implemented separately of the application [10].

### **2.2. Architecture of the System**

The architecture of the whole application is separated into three parts: the front-end, the back-end and the module for the ECDSA (which is also part of the back-end). Each of these must be initialized separately. The front-end architecture is based on the principle of lazy loading which means that not all modules are loaded together and, on every page, but as the user navigates through the page, only the needed/necessary modules for that specific page are loaded. Another part of the architecture is one of the best practices in Angular: separation of the code into components, modules, pages, views, and services which helps into the code maintenance. Additionally, the code shows an example where the same code is used on two places (only with different content) and instead of coping the code, a shared component is used where a dynamic parameter is sent. The components on the front-end have three parts: template (an html file that shows the structure of the page), style (a sass file that has the classes for the style) and logic (a typescript file that contains the logic for the component) [11, 12].

Every service in the front-end is connected with a controller on the back-end, which also helps in the code maintenance and best practices. For example, the Typescript file that represents the service for the candidate (candidate.service.ts) has the functions that are connected with the controller for it (CandidateController.java) and so on. Each entity from the back-end is directly connected with the interface on the front-end where both have the same fields with the same types [13, 14].





Figure 1. Architecture of the front-end

The above-described architecture is shown on Figure 1, where the core of the application is built on separate modules. On the other side, Figure 2 gives an example how the back-end architecture is structured.

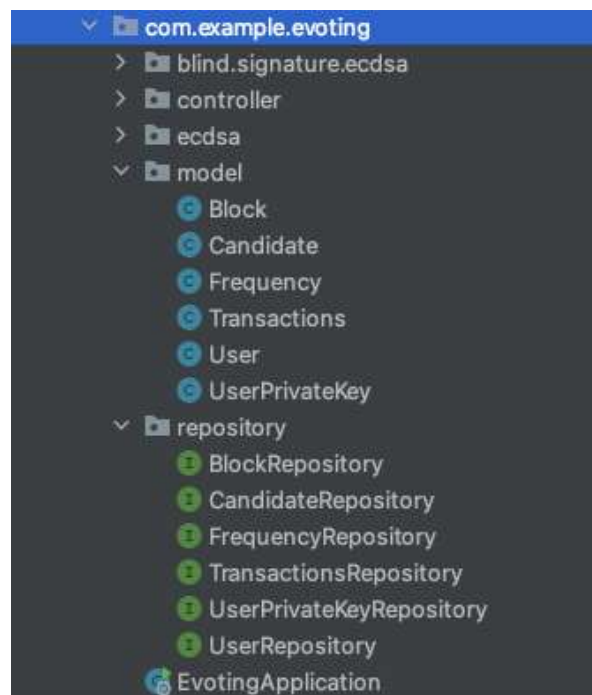


Figure 2. Architecture of the back-end

The application has two roles: user and admin and both have different view of it. The admin is already an existing entity that is created with the creation of the database and the user must be registered first via the registration form. The admin can add, delete, edit the users, and can see the frequency per candidates (only number of votes and nothing else for the voting). The user has the screen with the list of candidates, but he can only click on the vote button (no operations for the candidates). Once the user clicks on the button, he gets a confirmation modal. On this page the

user not only chooses which candidates he wants to vote for, but he can also see the chain of the blocks in the current Blockchain.

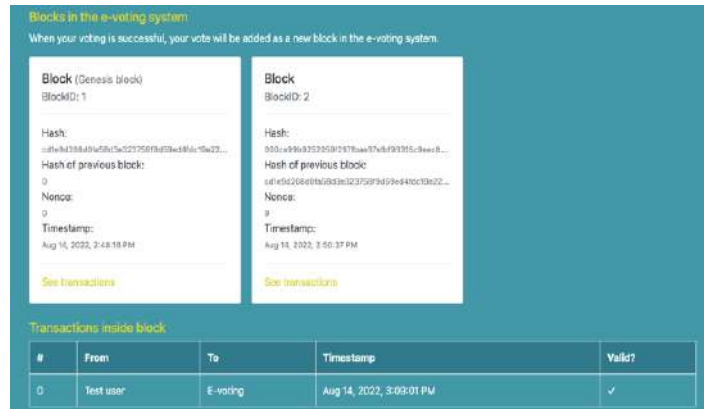


Figure 3. User panel with blocks and transactions

This is a really good view, because the users enjoy the benefits of the Blockchain technology and their contribution to the system is equal. By clicking on the blocks, the user can see the transactions of each block. Figure 3 show the chain of the blocks with the transactions (this is only part of the page, the table with the users is omitted in this figure). The other part of the voting is done on the second page where the user adds a transaction that is sent to the admin panel. First, the user must enter his private key and once it is done, the user can perform his voting. When the user comes to this step the ECDSA with blind signatures is calculated and he clicks on the ‘Sign Vote and Confirm’ button.

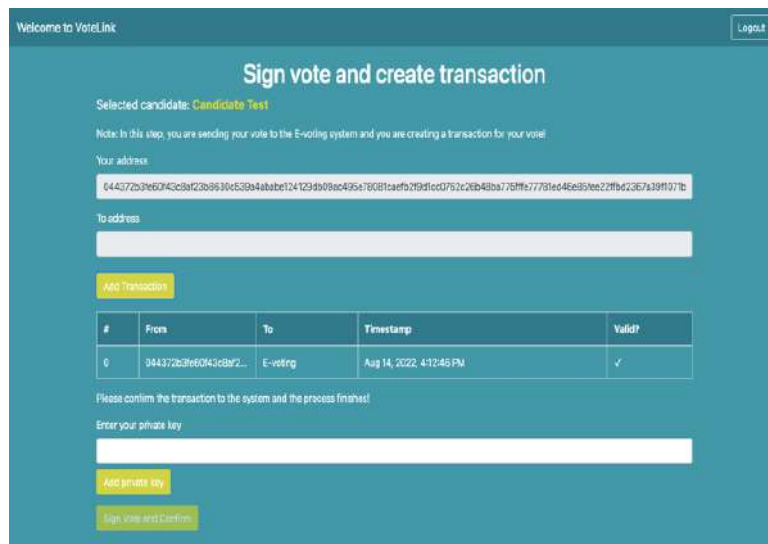


Figure 4. Form for submitting the vote

Figure 4 shows the last part of the user panel where the user enters the private key and starts the sign and verify process with ECDSA and blind signatures.

### 2.3. ECDSA with Blind Signatures

The definition of a digital signature may be listed as an electronic, encrypted stamp that is used for authentication of digital information where the integrity of the information is being fulfilled. The module of the ECDSA implementation consists of the following elements: Point (one point on the elliptic curve that is defined with the coordinates  $x$  and  $y$ ), Curve (the curve that is defined with the equation:

$$y^2 = x^3 + A * x + B(\text{mod } P)$$

where  $A$ ,  $B$ , and  $P$  are parameters. The curve has the order and the group of the curve. The class Math contains the functions for executing mathematical operations such as: the sum of two points of the curve and others. The private and the public key are known and used elements for every cryptography. The class Signature is calculated based on the mathematical equations of ECDSA algorithm but for the purpose of this paper, those equations are omitted. Those equations give the modified Paillier cryptosystem [15] that has the sign and verification processes. All these elements together are the core of the ECDSA algorithm with blind signatures. The definition of a blind signatures comes directly from its name and shows that the message is blinded before it is signed.

Another advantage of this system is that it uses additional check if the message is created correctly. Using zero-knowledge proof the processes of signing and verifying can be checked if they have been constructed correctly. There are two ways how to do this: interactive and non-interactive zero-knowledge proof. The current implementation of the system uses the interactive version because of the birthday attack. The interactive version requires two parts to make an interaction to prove this and the non-interactive requires only one way to execute the calculations. The main problem for choosing the interactive zero-knowledge proof is the possibility of a birthday attack. The non-interactive zero-knowledge is done using another hash function and with the birthday attack a duplicate may be found. If this happens, the system loses its security and with that the electronic voting will fail. On the other hand, the performances with the non-interactive zero-knowledge proof will be better since there is no waiting time from the opposite side. In this paper, the focus is on security and that is why the interactive zero-knowledge proof is used [16].

### 3. COMPARISON TO OTHER SYSTEMS FOR ELECTRONIC VOTING

There are many electronic voting systems that do the same function, but they are implemented on different ways. There is no right or wrong way how to do this, but it depends on many factors. The most meaningful characteristics for comparison of the systems are anonymity, audit, integrity, accessibility, scalability and so on. The systems can be implemented by companies or independent individuals. The most famous systems that are created by companies are: Follow My Vote, Voatz, Polyas, Luxoft, Polys, Agora and others. These are systems that are mainly stable, tested and used for years in practice. The research has shown that these systems have one disadvantage and that is the scalability (and the programming language). Our new system needs to be compared with the group of systems built from individuals. One of the most famous systems in this group are: Schema OVN, Schema DATE and the Schema BES. The anonymity, audit, verification by voter, accessibility and affordability are positive for all systems and that is something very important. On the other hand, accuracy and scalability are characteristics that are not fulfilled for none of the systems [17]

The new system for electronic voting, proposed in this paper, shows a characteristic that is not common for the other systems and that is the “block-chain”. Since the Blockchain technology put the emphasis on the equal nodes and privileges, none of the systems shows this in the user interface. That is what our system shows, and the users can see visually how the whole chain is structured. Nevertheless, this has technical issues and one of them is the DOM tree which now is constructed with so many elements. This can be optimized by implementing a pagination with only  $n$  elements per request or implementing a search filter that will only focus on filtering the elements by certain parameter.

#### 4. CONCLUSIONS

Blockchain technology and decentralized system shows that every node in the chain have the same privileges and role. According to its structure, the technology does not have one central unit and with this modification the changes for abuse and changing data is not the case anymore. Most of the electronic voting system suffer from the scalability, and this is even the case with the bigger companies. Building electronic voting system can have various of advantages and disadvantages, but in the end the most important parts that every system must have are obtaining the security and keeping the vote as anonymous.

In this paper, we propose a new system of electronic voting based on Blockchain, which represents a stable system with a good architecture and high code maintenance. After some modifications, the ECDSA with blind signatures can be used with zero-knowledge proof and together they build a secure chain that users can use to vote easily. The analysis of this system shows that due to limited resources and performance issues, the best practice where to use this system is on smaller target group and that is how also the technical issues may occur less. The system has an easy-to-use user interface and shows some hints for the user. The implementation with Angular and Spring Boot has detailed and structured architecture which uses the best practices. When the system is compared to other systems, it shows that our system has an advantage that is not present in the others. The advantage is that our system shows the user how the chain of the voting is constructed and that is exactly one of the benefits of the Blockchain. This comes with a disadvantage in the DOM tree of the application, but we propose a way with using pagination to solve the issue.

In the end, we can conclude that the Blockchain technology has a positive impact in the world of technology and the advantages of it can be used to build helpful systems for the whole environment.

#### ACKNOWLEDGEMENTS

This research was partially supported by Faculty of Computer Science and Engineering at "Ss Cyril and Methodius" University in Skopje.

#### REFERENCES

- [1] Easttom, C, (2015) “Modern cryptography”, Applied mathematics for encryption and information security. McGraw-Hill Publishing.
- [2] Yaga, D., Mell, P., Roby, N. & Scarfone, K, (2019) “Blockchain technology overview”, arXiv preprint arXiv:1906.11078.
- [3] Jafar, U., Aziz, M. J. A. & Shukur, Z, (2021) “Blockchain for electronic voting system—review and open research challenges”, *Sensors*, Vol. 21, No. 17, pp5874.
- [4] Joshi, A. P., Han, M. & Wang, Y, (2018) “A survey on security and privacy issues of Blockchain technology”, *Mathematical foundations of computing*, Vol. 1, No. 2, pp121.

- [5] Zheng, Z., Xie, S., Dai, H., Chen, X. & Wang, H, (2017) "An overview of Blockchain technology: Architecture, consensus, and future trends", In 2017 IEEE international congress on big data (BigData congress), pp557-564.
- [6] Kohno, T., Stubblefield, A., Rubin, A. D. & Wallach, D. S, (2004) "Analysis of an electronic voting system", In IEEE Symposium on Security and Privacy, 2004, Proceedings. 2004, pp27-40.
- [7] Song, J. G., Moon, S. J. & Jang, J. W, (2021) "A scalable implementation of anonymous voting over Ethereum blockchain", Sensors, Vol. 21, No. 12, pp 3958.
- [8] <https://www.edureka.co/blog/advantages-and-disadvantages-of-angular/#AdvantagesDisadvantages> accessed: 10.11.2022
- [9] <https://bambooagile.eu/insights/pros-and-cons-of-using-spring-boot/> accessed: 10.11.2022
- [10] Rajesh, R. V, (2016) "Spring Microservices", Packt Publishing Ltd.
- [11] <https://angular.io> accessed: 10.11.2022
- [12] Gutierrez, F, (2021) "Spring Boot. In Spring Cloud Data Flow", Apress, Berkeley, CA, pp9-31.
- [13] Moiseev, A. & Fain, Y, (2018) "Angular Development with TypeScript", Simon and Schuster.
- [14] <https://spring.io/projects/spring-boot> accessed: 10.11.2022
- [15] Yi, X. & Lam, K. Y, (2019) "A new blind ECDSA scheme for bitcoin transaction anonymity", In Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, pp613-620.
- [16] Dumas, J. G., Lafourcade, P., Miyahara, D., Mizuki, T., Sasaki, T. & Sone, H, (2019) "Interactive physical zero-knowledge proof for Norinori", In International Computing and Combinatorics Conference, Springer, pp166-177.
- [17] Enguehard, C, (2008). "Transparency in electronic voting: the great challenge. In IPSA International Political Science Association RC 10 on Electronic Democracy", Conference on "E-democracy-State of the art and future agenda", pp.édition-électronique.

## AUTHORS

**Vesna Dimitrova**, is a professor at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. She obtained the Ph.D. in 2010. She participated/coordinated more than 30 projects. She was a chair of one International Conference and a member of program/scientific committee at more than 40 conferences. She has published over 70 scientific papers. Her research areas are cryptography, information security and application of ML, DL and NLP in security, cryptography and cryptanalysis.



**Aleksandra Popovska-Mitrovikj**, is an associate professor at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. She obtained the Ph.D. in 2014. She participated in several scientific research projects, many international scientific conferences, and is a co-author of many scientific research papers published in journals and proceedings of international conferences. Her research areas are coding theory, cryptography and application of ML in security, blockchain technology.



**Lina Lumburovska** is a Master student at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University of Skopje. She is currently finishing her Master thesis in the field of cryptography, coding, and security. She has been author and co-author of 8 scientific papers in national and international conferences. She works as a full-stack developer in kern.ai (a start-up that builds an open-source data-centric IDE for NLP).



# PRIVACY-PRESERVING ONLINE SHARING CHARGING PILE SCHEME WITH DIFFERENT NEEDS MATCHING

Zhiyu Huang<sup>1,2</sup>

<sup>1</sup>School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China

<sup>2</sup>Hunan Key Laboratory for Service Computing and Novel Software Technology, Xiangtan, China

## ABSTRACT

*With the development of electric vehicles, more and more electric vehicles have difficulties in parking and charging. One of the reasons is that the number of charging piles is difficult to support the energy supply of electric vehicles, and a large number of private charging piles have a long idle time, so the energy supply problem of electric vehicles can be solved by sharing charging piles. The shared charging pile scheme uses Paillier encryption scheme and improved scheme to effectively protect user data. The scheme has homomorphism of addition and subtraction, and can process information without decryption. However, considering that different users have different needs, the matching is carried out after calculating the needs put forward by users. This scheme can effectively protect users' privacy and provide matching mechanisms with different requirements, so that users can better match the appropriate charging piles. The final result shows that its efficiency is better than the original Paillier scheme, and it can also meet the security requirements.*

## KEYWORDS

*Private charging pile sharing service, Privacy protection, Demand analysis, Homomorphic encryption, Internet of things*

## 1. INTRODUCTION

With the advancement of "carbon peak and carbon neutral" goals and the development of electric vehicles (EVs), EVs have the potential to effectively reduce air pollution caused by daily transportation [1]. As the number of EVs on the road increases, the demand for charging infrastructure also increases. However, the current prevalence and coverage of charging stations are insufficient to meet this demand[2-3]. Surveys have shown that between 2015 and 2020 in Table 1, the number of EVs and charging stations has been steadily increasing, with a particularly notable increase in the number of private charging stations, from 8,000 in 2015 to 874,000 in 2020[16]. However, compared to the growth rate of EVs, the number of charging stations is still far from adequate. As a result, for those EV users who are unable to install charging stations, the problem of charging difficulties is becoming increasingly apparent. In 2020, the Chinese government proposed to include charging stations as one of the fields of the nation's "new infrastructure", with an estimated investment of approximately 10 billion to build charging stations. According to international data surveys, it is expected that by 2030, there will be 5 million EVs on the road in California alone, and 12-24 million private charging stations and 10-20 million public charging stations globally. Charging facilities have become an indispensable

infrastructure in new energy development planning [4-5]. Considering the high installation cost of charging piles [6], other technologies are needed to make up for the shortcomings of charging piles.

Table 1 Approximate number of electric vehicles and charging piles in the world

Year	2015	2016	2017	2018	2019	2020
Number of electric vehicle	570	1280	1840	2740	3890	4840
Number of public charging piles	58	149	240	387	516	807
Number of private charging piles	8	63	232	477	703	874

With the rapid development of Internet of Things technology, Internet of Things devices connect everything and have gradually entered the mode of Internet of Everything [7-9]. As one of the applications of the Internet of Things, the Internet of Vehicles can realize the information exchange between vehicles and provide certain research value and commercial value. The application of V2X technology of the Internet of Vehicles in cloud (edge) computing is the cornerstone of building a smart city and smart transportation [10-12]. At present, the research on shared charging pile is still in the initial stage. The traditional charging pile sharing scheme generally consists of three entities, including charging pile provider, electric vehicle and matching server, in which both buyers and sellers upload their own information to the server for matching calculation, and the server returns the matching results to both buyers and sellers, as shown in Fig 1. However, in the traditional charging pile sharing scheme, the user's information is published or uploaded to the server through simple encryption, and the server needs to decrypt the participants' information to get their plaintext information. Therefore, in the traditional scheme, the user's privacy may be attacked and leaked. In the traditional charging pile sharing system, all information will be published directly on the Internet. One of the biggest problems faced by the system is that users will expose their private information to the public platform when they apply for it. For example, a malicious user has used a certain charging station. The charging pile is marked and recorded, and he may not use it directly through the platform when he knows that the charging pile is unmanaged for a period of time.

At the same time, there is also the possibility that the shared service platform exposes the privacy of customers. Because the location information of electric vehicles may include workplaces, home addresses, special hospitals or frequently visited entertainment venues, buyers' hobbies and health status information are leaked, and the privacy of charging pile sellers will also be greatly affected threaten. On the other hand, once the information of buyers and sellers is obtained by malicious attackers, not only will there be profitable and targeted advertisements, but also related work and home addresses will be threatened, and may even lead to personal safety. Therefore, in order not to disclose the private information of customers, it is necessary to design a secure service platform. This paper proposes the use of homomorphic encryption technology to protect the privacy of users.

In order to meet the above challenges, the main contributions of this paper are summarized as follows:

- 1) We use homomorphic encryption technology to encrypt user information, and at the same time use homomorphic characteristics to process ciphertext, and match the obtained results in the cloud server. In the public service platform, users' effective information will not be exposed, and matching can be completed efficiently.

2) For users with different needs, we designed the demand parameter  $\omega$ . Through matching calculation, we can get the matching index parameter  $W$ . By comparing  $W$ , we can get the most suitable buyers and sellers. This requirement parameter can better match users with different requirements.

3) We use chinese remainder theorem (CRT) to speed up the modular exponentiation in the decryption process of cloud server, CRT is used to convert  $a^b$  from  $Z_{n^2}$  to  $Z_{p^2}$  and  $Z_{q^2}$  for calculation. We use Paillier scheme with optimized parameters, which can speed up the encryption calculation although it loses homomorphism.

The remainder of this paper is organized as follows: In Section 2, we introduce the homomorphic encryption, parameter optimization of Paillier scheme and china remainder theorem. In Section 3, we introduce the system model and present the proposed scheme. In Section 4, we describe the performance evaluation results. Finally, we conclude the paper in Section 5.

## 2. RELATED WORK

In this section, homomorphic encryption technology, parameter optimization of Paillier scheme and Chinese remainder theorem are introduced.

### 2.1. Homomorphic Encryption

Encryption technology is often used to protect privacy, among which homomorphic encryption is a special encryption method, which has the characteristics of directly calculating encrypted data, such as addition and multiplication, and will not reveal any information of the original text during the calculation process. And the calculated result is also encrypted, and the result obtained after decrypting the processed ciphertext with the key is exactly the result obtained after processing the original text. Paillier scheme has the homomorphism of addition/subtraction. For plaintext  $m_1$  and  $m_2$ , there is a function  $E()$  that makes  $E(m_1+m_2)=E(m_1)\cdot E(m_2)$ . Paillier scheme satisfies the standard semantic security of encryption scheme[13], that is, the ciphertext is indistinguishable (IND-CPA) under the attack of selected plaintext, that is, the information about plaintext will not be leaked in ciphertext. Its security is proved by the hypothesis of deterministic composite residue. So far, no algorithm can be cracked in polynomial time, so Paillier encryption scheme is considered to be safe. The detailed process includes the following steps.

- **KeyGen()** : Pick two prime numbers  $p$  and  $q$  compute  $n = p * q$  and  $\lambda = \text{lcm}(p-1, q-1)$ . Choose a random number  $g$ , and  $\text{gcd}(L(g^\lambda \text{ mod } n^2), n) = 1$ , computer  $\mu = (L(g^\lambda \text{ mod } n^2))^{(-1) \text{ mod } n}$ , where  $L(x) = (x-1)/n$  the public and private keys are  $pk = (n, g)$  and  $sk = (\lambda, \mu)$ , respectively.
- **Encrypt()** : Enter the plaintext message  $m$  and select the random number  $r$ . Encrypt plaintext:

$$c = g^m \cdot r^n \text{ mod } n^2, \quad (1)$$

- **Decrypt()** : Enter ciphertext  $C$ . Calculate plaintext message:



$$m = L(c^\lambda \bmod n^2) \cdot \mu \bmod n. \quad (2)$$

## 2.2. Parameter Optimization of Paillier scheme

In order to simplify computation without affecting the algorithm's correctness, the algorithm may take  $g=n+1$  during the key generation phase[14]. This allows for the simplification of the calculation of  $g^m$  during the encryption process.

For  $g^m=(n+1)^m$ , using the binomial theorem, we can express  $g^m$  as the sum of the product of the binomial coefficients and the corresponding powers of  $n$  and  $1$ , where each term of the sum can be calculated efficiently.

$$(n+1)^m \bmod n^2 = \binom{m}{0} n^m + \binom{m}{1} n^{m-1} + \dots + \binom{m}{m-2} n^2 + mn + 1 \bmod n^2, \quad (3)$$

As the previous  $m-1$  terms are multiples of  $n$ , under the condition of modulo  $n^2$  operation, they can all be eliminated, thus this modulo exponentiation operation can ultimately be simplified to one modulo multiplication operation, thus accelerating the encryption process.

$$c = (1 + mn) \cdot r^n \bmod n^2, \quad (4)$$

Decrypt ciphertext  $c$ :

$$m = \frac{c^k - 1 \bmod n^2}{n}. \quad (5)$$

## 2.3. Chinese remainder theorem

The Chinese Remainder Theorem, also known as the Sunzi Theorem, originates from the ancient Chinese mathematical treatise "Sunzi Suanjing" and describes the isomorphism of two algebraic spaces. Specifically, an algebraic space can be decomposed into several mutually orthogonal subspaces and the original space corresponds one-to-one to the decomposed space, similar to two forms of the same space. Specifically, when  $n = pq$  and  $p, q$  are relatively prime, there exists the algebraic isomorphism property:  $a \bmod n = a \bmod p + a \bmod q$ , thus the operations under  $\bmod n$  can be transformed into operations under  $\bmod p$  and  $\bmod q$ . By converting to this form, the calculation efficiency is higher. Therefore, this property can be utilized to accelerate modular exponentiation operations under  $\bmod n$ .

## 3. THE PROPOSED SCHEME

### 3.1. System Model

The matching scheme of shared charging piles consists of multiple electric vehicle buyers, multiple charging pile sellers, multiple edge proxy servers, a cloud server and a certificate certification center. All entities communicate through the mobile network. The Figure.1 describes our system model.

**Electric vehicle (EVs):** As a user of shared charging piles, when charging piles are needed, a charging request will be sent out, and the EV set is expressed as  $\{1, \dots, i, \dots, I\}$ . After receiving the response, you will get the public key of information encryption, and the terminal equipment

of the Internet of Things will encrypt the information to be sent with the public key and send it to the nearest proxy server.

**Private charging piles (PCPs):** As a provider of shared charging piles, there will be  $J$  private charging piles in a given area, and the collection of charging piles is denoted as  $\{1, \dots, j, \dots, J\}$ . Each PCP is managed by the owner and is equipped with a socket for EV charging. When each PCP has free time, it will issue an application for energy supply, and the provided information will be encrypted on the Internet of Things terminal equipment with the provided public key, and then the encrypted information will be sent to the nearest proxy server.

**Proxy server:** The proxy server has certain computing power, and is mainly responsible for collecting the encrypted information provided by nearby electric vehicle buyers and charging pile sellers who apply for matching, and using homomorphic characteristics to calculate the encrypted information. In the calculation process, important information is protected by Paillier, and the edge proxy server will not get useful information.

**Cloud server:** A cloud server is a server with powerful computing power, which can process encrypted information sent by proxy servers. After processing the information, use the matching scheme provided to match the buyer and the seller. After the matching, the best matching object will be obtained, and then the next round of matching will be carried out.

**Certificate certification (CA):** The certificate certification is the only authoritative identity certification institution and is completely reliable. All user entities need to be registered and authenticated by the certificate authority, and when the user sends an application, the corresponding public and private key pairs are generated by the key management center of the certificate authority and sent to the corresponding users. The information of the certificate certification center is absolutely confidential, and there is no possibility of collusion.

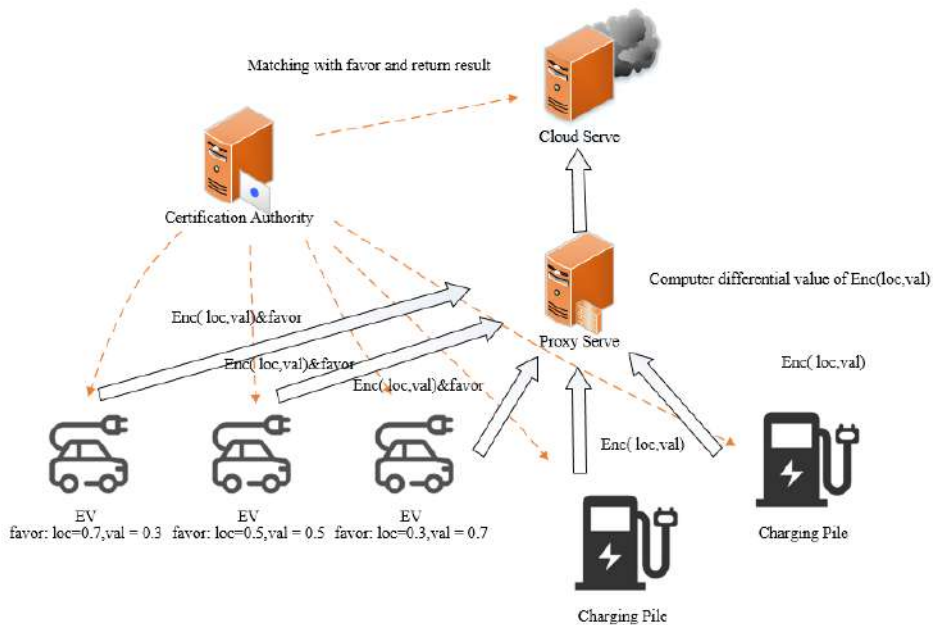


Figure.1 System model.

### 3.2. Requests and Information Encryption

In the shared charging pile matching system, the certificate certification center is responsible for managing and issuing public keys and maintaining public key information. In the system, electric vehicle buyers and charging pile sellers provide information, including location, price, demand and other information. Buyers of electric vehicles need to provide their own location  $x_i, y_i$ , proposed price ( $r_i$ ), farthest acceptable distance  $d_{i\max}$ , demand  $\alpha_{i=1,2,\dots,n}=\{0, 1\}$  and the price proposed by the demand  $r^{\alpha_{i=1,2,\dots,n}}$ . The seller of charging piles needs to provide the location  $(x_j, y_j)$ , the price ( $r_j$ ), the demand  $\alpha_{j=1,2,\dots,n}=\{0, 1\}$  of charging piles and the price  $r^{\alpha_{j=1,2,\dots,n}}$ . After the user sends a request, the certificate certification center will send the public key  $pk$  to the user, and the electric vehicle buyer and the charging pile seller will select the random numbers  $r_i$  and  $r_j$  respectively, and use the public keys  $pk$ ,  $r_i$  and  $r_j$  to encrypt the provided information.

$$C_{m_i} = \text{Enc}(m_i) = g^{m_i} \cdot r_i^n \bmod n^2, \quad (6)$$

$$C_{m_j} = \text{Enc}(m_j) = g^{m_j} \cdot r_j^n \bmod n^2, \quad (7)$$

After encryption,  $C_m$  are obtained. The buyers of electric vehicles package  $C_m$  and  $w$  and send them to the proxy server. The sellers of charging piles send  $C_m$  to the proxy server. Except at the user end, the private information of users is all the information obtained after encryption and processing.

### 3.3. Ciphertext Processing

The proxy server has a certain computing power, and can use the homomorphic addition/subtraction characteristics of pailier encryption scheme to process the ciphertext information as well as the plaintext. The process is as follows.

$$C_{a_{m_{ij}}} = C_{m_i} \cdot C_{m_j} = g^{m_i+m_j} \cdot (r_i \cdot r_j)^n \bmod n^2, \quad (8)$$

$$C_{d_{m_{ij}}} = C_{m_i}/C_{m_j} = g^{m_i-m_j} \cdot (r_i/r_j)^n \bmod n^2, \quad (9)$$

The demand of buyers and sellers is encrypted by homomorphic addition, and the information such as location, price and demand price is encrypted by homomorphic subtraction to get the sum of the processed information and the difference of the information. Among them, the main function of (8) is to judge whether the demand of electric vehicle buyers can be met. Information difference is a difference comparison between the information provided by buyers and sellers, which can reflect the information similarity of both parties. The smaller the information difference, the closer the information provided by the charging pile seller is to the preference of the electric vehicle buyer, which is more suitable for matching and has a higher matching probability. On the contrary, the larger the information difference between the two parties means that the user's matching probability will be smaller.

### 3.4. Information Decryption

The cloud server owns the private key  $sk$  issued by CA, including  $p$  and  $q$ . In Paillier cryptosystem, the main cost of decryption is modular exponentiation under  $Z_{n^2}$ . With the private key (decomposition  $p, q$  of  $n$ ), the modular exponentiation under  $Z_{n^2}$  can be converted into  $Z_{p^2}$  and  $Z_{q^2}$  by CRT.

The optimization function using CRT is expressed as  $L_{p(x)}=(x-1)/p$  and  $L_{q(x)}=(x-1)/q$  respectively, and the decryption process needs to be divided by using the following mathematical principles.

$$h_p = L_p(g^{p-1} \bmod p^2)^{-1} \bmod p, \quad (10)$$

$$h_q = L_q(g^{q-1} \bmod q^2)^{-1} \bmod q, \quad (11)$$

$$m_p = L_p(c^{p-1} \bmod p^2)h_p \bmod p, \quad (12)$$

$$m_q = L_q(c^{q-1} \bmod q^2)h_q \bmod q, \quad (13)$$

$$m = \text{CRT}(m_p, m_q) \bmod pq, \quad (14)$$

$\text{CRT}(m_p, m_q \bmod pq)$  is to use CRT to calculate the modulus index, and the detailed process is as follows. For the modulus index  $a^b \bmod n$ ,  $n=pq$ , CRT is used to convert  $a^b$  from  $Z_n$  to  $Z_p$  and  $Z_q$  for calculation. Calculate the mapping  $a^b$  of  $Z_p$  on  $m_p=a_p^{b_p}$ , where  $a_p=a \bmod p$  can be obtained from Euler theorem, where  $\phi(p)=p-1$  is Euler function. Calculate the mapping  $m_q=a_q^{b_q}$  of  $a^q$  on  $a^b$ , which is the same as the calculation process of  $m_p$ . Calculate  $m_p$  and  $m_q$  separately and then aggregate them back.

$$m = m_p \cdot q^{-1}(\bmod p) \cdot q + m_q \cdot p^{-1}(\bmod q) \cdot p, \quad (15)$$

Because  $p$  and  $q$  are coprime, there are  $q^{-1}(\bmod p)q+p^{-1}(\bmod q)p=1$ . Substituting into the formula (15) gives:

$$m = m_p + (m_q - m_p)p^{-1}(\bmod q) \cdot p, \quad (16)$$

In Paillier scheme, CRT is used to speed up decryption of plaintext to get plaintext  $m$ . After receiving the processed information, the modular operation under  $Z_{n^2}$  is converted into modular operation under  $Z_{p^2}$  and  $Z_{q^2}$  by using private key  $sk$  by using China remainder theorem, and then decrypted.

$$am_{ij} = \text{Dec}(Cam_{ij}) = m_i + m_j, \quad (17)$$

$$dm_{ij} = \text{Dec}(Cdm_{ij}) = m_i - m_j. \quad (18)$$

### 3.5. System Matching

After decryption by using CRT-optimized decryption scheme, the sum of information and the difference between information are obtained. After obtaining the decrypted information, first calculate the distance between the buyer  $i$  and the seller  $j$ :

$$d_{dij} = \sqrt{d_{xij}^2 + d_{yij}^2}, \quad (19)$$

The maximum acceptable distance of the EV buyer  $i$  is also encrypted and decrypted, because the maximum acceptable distance does not carry specific information such as location and price, so it is not regarded as important privacy information, so the cloud server gets the same clear text  $d_{imax}$  as the user. To meet the matching conditions of buyer  $i$ , we must first compare the direct distance between buyer and seller. If  $d_{dij} < d_{imax}$ , it means that the distance between buyer  $i$  and seller  $j$  is less than the maximum distance accepted by buyer  $i$ , which meets the matching conditions. If  $d_{dij} > d_{imax}$ , the distance between users  $i$  and  $j$  does not meet the conditions, the seller  $j$  cannot match the buyer. Remove the unqualified sellers by comparison before proceeding to the next step.

Demand analysis is an interesting part of this paper. We consider that different buyers of electric vehicles may have different needs. Specifically, the distance and price are the information that the buyer  $i$  and the seller  $j$  must provide. Besides, other related demands  $\alpha_i$  and  $\alpha_j$  can be set, and the sum of them can be obtained through information and calculation. There are three situations:

Case1:  $\alpha_{ij}=0$ , indicates that neither user has this requirement.

Case2:  $\alpha_{ij}=1$ , it means that only one of the buyer  $i$  or the seller  $j$  owns the demand, and in case1 and case2, the corresponding  $i$  and  $j$  are removed because the demand cannot be provided.

Case3:  $\alpha_{ij}=2$ , it means that  $i$  have the demand, and  $j$  can also provide the demand. At this time, whether to use it can be judged by the demand price difference  $d_{rij}$ .

i) when  $d_{rij} < 0$ , it means that the price proposed by  $i$  is less than that proposed by  $j$ . At this time,  $i$  and  $j$  cannot match. ii) When  $\alpha_{ij}=2$  and  $d_{rij} > 0$  are met at the same time, it means that  $i$  and  $j$  meet the matching conditions, and the corresponding  $i$  and  $j$  are added to the matching set.

After getting the matching set that meets the distance and demand, the cloud server will make the final price matching. For buyer  $i$ , all sellers  $j$  who meet the demand will calculate the demand price difference  $d_{rij}$  provided by  $i$  and  $j$  through formula (9). Because there is no information in the buyer's preferences that can lead to the leakage of location information,  $w$  are also unimportant information that can be obtained in the cloud server only after being decrypted by the private key  $sk$ . At this time, the information in the cloud server includes the location information difference  $d_{dij}$ , demand and value  $\alpha_{ij}$ , price information difference  $d_{rij}$ , demand price difference  $d_{rij}$ , buyer's preferences  $w$  and the number  $k$  of sellers  $j$  in the number matching set. When matching the buyer  $i$  of the electric vehicle, the seller  $j$  in the matching set is calculated respectively to obtain  $W_{ij}$ .

$$W_{ij} = d_{dij} * w_{di} + d_{rij} * w_{ri} + \sum_{n=1}^k d_{\alpha_{ijn}} * w_{\alpha_{ijn}}, \quad (20)$$

For buyer  $i$ ,  $W_{ij}$  is used as a matching evaluation index to judge the suitability of matching with seller  $j$ . So we sort  $W_{ij}$  in descending order. The smallest  $W_{ijmin}$  is obtained, which means that the current sellers  $j$  and  $u$  are the most suitable matching objects in terms of price and demand, so  $i$  match  $j$ .

### 3.6. Matching Result Return

After the matching of buyer  $i$  is completed, the cloud server will send a request to the successfully matched  $i$  and  $j$ . User  $i$  and  $j$  use the Paillier scheme with optimized parameters to generate public keys  $pk_i$  and  $pk_j$  and send them to the cloud server. The cloud server generates a random number  $r$ , and encrypts the private key  $sk$  with  $pk_i$  and  $pk_j$ .

$$C = \text{Enc}(sk)_{pk} = (1 + mn) \cdot r^n \bmod n^2, \quad (21)$$

And the matched result is packaged and sent to the proxy server. The proxy server stores the encrypted address information uploaded by the user. After receiving the packaged result and the encrypted private key, the proxy server finds the corresponding encrypted address information  $C_{Loc_i}$  and  $C_{Loc_j}$  and the encrypted price information  $C_{r_j}$  of the seller  $j$  through the matching results  $i$  and  $j$ . The proxy server packages and sends the ciphertext of  $C_{Loc_j}$ ,  $C_{r_j}$  and private key  $sk$  encrypted with public key  $pk_i$  to buyer  $i$ , and packages and sends the ciphertext of  $C_{Loc_j}$  and private key  $sk$  encrypted with public key  $pk_j$  to seller  $j$ . The buyer  $i$  and the seller  $j$  use their own private keys to decrypt and get  $sk$ .

$$sk = \text{Dec}(C)_k = \frac{c^k - 1 \bmod n^2}{n}, \quad (22)$$

Then use  $sk$  to decrypt the encrypted address information  $C_{Loc_i}$ ,  $C_{Loc_j}$  and  $r_j$ .

$$m = L(c^\lambda \bmod n^2) \cdot \mu \bmod n, \quad (23)$$

At this time, the buyer  $i$  is matched, and the  $i+1$ th buyer is matched in the next round, and the matched seller  $j$  is eliminated from the matching set until the last seller set is empty, indicating that the current round of matching has ended. Re-apply to CA, get a new public and private key pair, and start the next round of matching.

## 4. PERFORMANCE EVALUATION RESULTS

### 4.1. Number Analysis

In this section, we consider a 3km\*3km scene with different numbers of  $I$  buyers and  $J$  sellers. All the simulations are done on python and on 2.5 GHz Inter Core i5-7300HQ CPU and 32G RAM. Finally, all the simulation results are averaged in 50 simulations, and the consistent results are finally obtained.

Paillier encryption algorithm is a public key encryption algorithm based on number theory, which has high performance in security and time cost.

The following is the time cost of operating on a piece of data of the original Paillier encryption algorithm:

Randomly generate public key and private key:  $O(1)$

Encryption operation:  $O(\log n)$

Decryption operation:  $O(\log n)$

Addition/subtraction operation (adding/subtracting two ciphertext numbers):  $O(1)$

Where  $n$  refers to the length of the public key (the number of digits of the modulus).

In our simulation experiment, for  $I$  buyers and  $J$  sellers, the time cost from issuing an application to obtaining the corresponding public key is  $O(1)$ . For all users, because the encryption operation uses the original Paillier encryption scheme, its time cost corresponds to  $O(\log n)$ . Our setting is that all users encrypt on the terminal equipment of the Internet of Things, and each user does it independently, so the time cost is fixed regardless of the number of matching users. When a user encrypts  $n$  data, the time cost is  $n \cdot O(\log n)$ . When the terminal equipment of the Internet of Things encrypts the information to be sent, the proxy server will add/subtract the ciphertext data, and the corresponding operation is  $O(1)$ . Every buyer  $I$  in the proxy server needs to add and subtract with the seller  $J$ . For  $J$  sellers and  $K$  pieces of information, the time cost is  $J \cdot k \cdot O(1)$ . For  $I$  buyers, the total time cost required for calculation in the proxy server is  $I \cdot J \cdot k \cdot O(1)$ . Decrypt each calculated result in the cloud server. Under the condition that all sellers meet the requirements, the time cost of each decryption operation is  $O(\log n)$  corresponding to  $I \cdot J \cdot k$  calculation results. After decryption, we execute the matching algorithm, calculate the matching index  $w_{ij}$  for  $M$  users who satisfy user  $I$ , and then get the minimum matching index  $w_{ijmin}$  for user  $I$  after sorting it, with the time cost of  $j \cdot O(1)$ . At this time, the buyer  $I$  and the seller  $J$  are successfully matched. The time cost corresponding to the above process is shown in Table 2.

Table 2 Paillier scheme time cost in this paper

	I buyers	J sellers	n data	Buyer matching	Index ranking
Encrypt	$O(\log n)$	$O(\log n)$	$k \cdot O(\log n)$	/	/
Process	/	/	$k \cdot O(1)$	$I \cdot J \cdot k \cdot O(1)$	$j \cdot O(1)$
Decrypt	$O(\log n)$	$O(\log n)$	$k \cdot O(\log n)$	/	/

In the process of decryption, we use CRT to speed up the calculation process. The original Paillier encryption algorithm needs to do modular exponential operation under  $Z_{n^2}$ . However, when the cloud server knows the private key  $sk$  and the corresponding coefficients  $p$  and  $q$ , the modular exponentiation under  $Z_{n^2}$  is transformed to  $Z_p^2$  and  $Z_q^2$ , thus improving the encryption and decryption efficiency. The time required to decrypt the ciphertext with the original Paillier encryption scheme and the time required to accelerate the calculation with CRT are shown in the figure. As can be seen from Figure.2, the decryption time is about 1/3 of that of Paillier encryption scheme after accelerated calculation with CRT. Compared with DJN scheme [15], the decryption time is basically the same as that of Paillier scheme. Therefore, using CRT to speed up the decryption process can effectively improve efficiency.

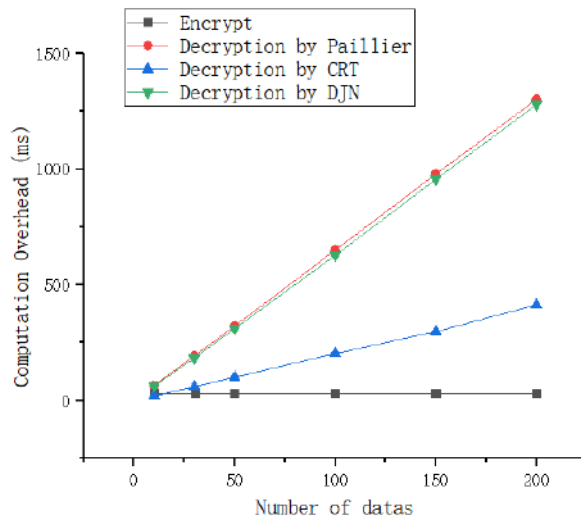


Figure.2 Computational overhead of data encryption and decryption with different schemes

After the cloud server gets the matching result, it sends out a successful matching application, and the buyer I and the seller J call the Paillier encryption algorithm with optimized parameters to generate a public key pair, and send the public key n to the cloud server, and the private key is stored at the user end. Compared with the original Paillier encryption algorithm, the encryption scheme with parameter optimization simplifies the modular exponential operation into a modular multiplication, which speeds up the encryption process. Its time efficiency is shown in Figure.3.

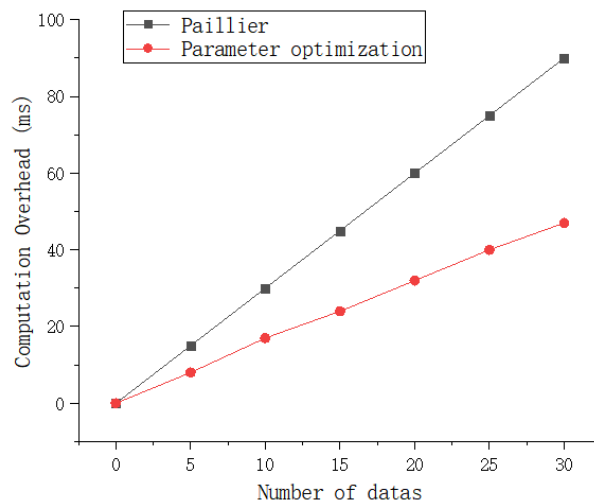


Figure.3 Comparison of encryption time cost between Paillier scheme and parameter optimization

#### 4.2. Correctness Analysis

For ciphertext C in paillier scheme with optimized parameters, its correctness is expressed as:

$$Dec(C) = \frac{c^x - 1 \text{ mod } n^2}{n}$$



$$\begin{aligned}
&= \frac{((1+n)^{mr^n})^{\tau\lambda} - 1 \pmod{n^2}}{n} \\
&= \frac{(1+n)^{m\tau\lambda} - 1 \pmod{n^2}}{n} \\
&= \frac{nm\tau\lambda \pmod{n^2}}{n} = \frac{nm}{n} = m,
\end{aligned} \tag{24}$$

For ciphertext  $C$  in Paillier scheme which uses CRT optimization for decryption, its correctness is expressed as:

$$\begin{aligned}
m &= m_p \cdot q^{-1} \pmod{p} \cdot q + m_q \cdot p^{-1} \pmod{q} \cdot p \\
&= m_p \cdot (1 - p^{-1} \pmod{q} \cdot p) + m_q \cdot p^{-1} \pmod{q} \cdot p \\
&= m_p + (m_q - m_p) \cdot p^{-1} \pmod{q} \cdot p,
\end{aligned} \tag{25}$$

Where  $h_p$ ,  $h_q$ ,  $m_p$ ,  $m_q$  and are obtained from formula (10), formula (11), formula (12) and formula (13) respectively.

### 4.3. Security Analysis

First of all, we assume that there are curious buyers and sellers, denoted as  $B$ , who want to attack through some information on the network to obtain other users' private information. In this paper, the original Paillier encryption scheme was adopted before the matching was completed. This scheme has been fully studied, so far there is no polynomial time algorithm to break it, so the security of Paillier encryption scheme is considered to be reliable. When  $B$  obtains the ciphertext message and the processed ciphertext message in the proxy server through attack, it cannot obtain effective information because there is no corresponding private key  $sk$ . So in the proxy server, we think the information is safe and reliable. When the processed information is sent to the cloud server, the cloud server needs to use the private key  $sk$  to decrypt the information. Suppose that  $B$  obtains the sum and difference of the decrypted information in the cloud server through special means attack, and these  $B$  have their own information, they will infer other useful information through the difference between the existing information and the information obtained by the attack in the cloud server. When inferring the position of other sellers through the difference between their own position information and the obtained position information, because there is only a straight distance, the inferred information cannot locate the specific position of the seller. From the above, even if the attack obtains the encrypted information in the proxy server or the decrypted information in the cloud server,  $B$  cannot infer the valid information. Therefore, for curious buyers and sellers, our scheme is safe and effective.

Secondly, for the premeditated attacker  $C$ , in our scheme, in order to prevent  $C$  from eavesdropping, all the communication between entities is encrypted. In our scheme, the random number  $r$  generated each time is different, and the encrypted result is also different. After attacking the information obtained by the proxy server and the cloud server, the user's location information and price information cannot be obtained through calculation. And we will refresh the key after each round of user matching. In this case, we think the scheme is also safe and effective.

## 5. CONCLUSIONS

In this paper, we solve the security problem of shared charging pile scheme through homomorphic encryption technology. In order to protect the privacy of users' location and provide matching strategies for users with different needs, we have formulated a privacy protection shared charging pile scheme based on users with different needs. First of all, we use the public key to encrypt the information in the terminal equipment of the Internet of Things, which effectively protects the privacy information such as location. Through homomorphism, the ciphertext matching the user is calculated in the proxy server, and CRT is used in the cloud server to accelerate the encryption process. We design the matching rules, calculate the matching index  $W$  and compare them to get the most suitable matching result. When we return the results, we use Paillier scheme with optimized parameters to effectively speed up the encryption process. Finally, our numerical analysis results show that the decryption time after CRT optimization is about 1/3 of the original Paillier scheme and DJN scheme. The encryption time after parameter optimization is 1/3 faster than that of the original Paillier scheme. At the same time, we also analyzed the security of the scheme, and the attacks of both curious users and malicious attackers are safe and reliable in the scheme on the public platform.

## REFERENCES

- [1] J. Zhang, H. Yan, N. Ding, J. Zhang, T. Li and S. Su, "Electric Vehicle Charging Network Development Characteristics and Policy Suggestions," 2018 International Symposium on Computer, Consumer and Control (IS3C), 2018, pp. 469-472.
- [2] S. Qiao, "Technical Analysis and Research on DC Charging Pile of Electric Vehicle," 2021 International Conference on Smart City and Green Energy (ICSCGE), 2021, pp. 89-93.
- [3] Y. Zhang, Y. Wang, F. Li, B. Wu, Y. -Y. Chiang and X. Zhang, "Efficient Deployment of Electric Vehicle Charging Infrastructure: Simultaneous Optimization of Charging Station Placement and Charging Pile Assignment," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 10, pp. 6654-6659, Oct. 2021.
- [4] A. J. Qarebagh, F. Sabahi and D. Nazarpour, "Optimized Scheduling for Solving Position Allocation Problem in Electric Vehicle Charging Stations," 2019 27th Iranian Conference on Electrical Engineering (ICEE), 2019, pp. 593-597.
- [5] H. Hu, S. Ni and L. Zhang, "Analysis of the carrying capacity of charging station based on regional charging demand," 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020, pp. 1688-1691.
- [6] S. Mallapuram, N. Ngwum, F. Yuan, C. Lu and W. Yu, "Smart city: The state of the art, datasets, and evaluation platforms," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, pp. 447-452.
- [7] I. M. Nafi, S. Tabassum, Q. R. Hassan and F. Abid, "Effect of Electric Vehicle Fast Charging Station on Residential Distribution Network in Bangladesh," 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2021, pp. 1-5.
- [8] C. Liu, K. T. Chau, D. Wu and S. Gao, "Opportunities and Challenges of Vehicle-to-Home, Vehicle-to-Vehicle, and Vehicle-to-Grid Technologies," in Proceedings of the IEEE, vol. 101, no. 11, pp. 2409-2427, Nov. 2013.
- [9] Y. Wang, Z. Su and K. Zhang, "A Secure Private Charging Pile Sharing Scheme with Electric Vehicles in Energy Blockchain," 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2019, pp. 648-654.
- [10] Zhao Tong, Feng Ye, Ming Yan, Hong Liu, Sunitha Basodi. A Survey on Algorithms for Intelligent Computing and Smart City Applications[J]. Big Data Mining and Analytics, 2021, 4(03): 155-172.
- [11] Anagnostopoulos T , Luo C , Ramson J , et al. A multi-agent system for distributed smartphone sensing cycling in smart cities[J]. Journal of Systems and Information Technology, 2020, ahead-of-print(ahead-of-print).

- [12] Vidal S . Intelligent transport system in smart cities: aspects and challenges of vehicular networks and cloud[J]. Computing reviews, 2019(7):60.
- [13] Paillier P . Public-Key Cryptosystems Based on Composite Degree Residuosity Classes[C]// Advances in Cryptology - EUROCRYPT '99, International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic, May 2-6, 1999, Proceeding. Springer, Berlin, Heidelberg, 1999.
- [14] Catalano D , Gennaro R , Howgrave-Graham N , et al. Paillier's Cryptosystem Revisited. 2002.
- [15] Ivan, Damgrd, Mads, et al. A generalization of Paillier's public-key system with applications to electronic voting[J]. International Journal of Information Security, 2010, 9(6):371-385.
- [16] <https://baijiahao.baidu.com/s?id=1691272446545480912&wfr=spider&for=pc>

# RESEARCH ON CONSTRUCTION OF RDF WITH HBASE

Hui Hu

College of Computer Science and Technology, Nanjing University of  
Aeronautics and Astronautics, Nanjing, China

## **ABSTRACT**

*Resource Description Framework (RDF) is designed as a standard metadata model for data interchange on the Internet. Because of machine comprehensibility, it has been successfully used in many areas, such as the intelligent processing of numerous data. While the generation of RDF with relational database (RDB) receives much attention, little effort has been put into the automatic construction of RDF with HBase due to its flexible data structure. Since more data is stored in HBase, it is necessary to extract useful information from HBase. In this paper, we are devoted to construction of RDF with HBase. We put forward formal definitions of RDF and HBase and propose our strategy for generating RDF with HBase. We develop a prototype system to create RDF, and test results demonstrate the feasibility of our method.*

## **KEYWORDS**

*Semantic Web, RDF, HBase, Construction*

## **1. INTRODUCTION**

The vision of the Semantic Web is to establish a semantic framework that allows sharing and reusing data across applications, businesses, and communities [1]. As the core layer of the Semantic Web, RDF is the standard for data representation and exchange recommended by the World Wide Web Consortium. RDF is a metadata model with good machine readability because it adds semantics to the data representation when describing data on the web. RDF allows web applications to share, exchange, and integrate data without losing data semantics [2]. Because of the widespread use of RDF, large amounts of RDF are proliferating. For example, Best Buy and the New York Times use RDF for data storage [3]. But the efficient generation of RDF data remains an open problem. While there have been many efforts to convert relational databases, generating RDF from HBase, a NoSQL database, currently needs more attention.

It is critical to process data in a standard way dealing with massive data with heterogeneous data formats. Because RDF provides a standard, unified semantic manner for information processing, it is necessary to study generating RDF with these data. Numerous studies have been done on RDF generation, primarily concentrating on relational databases and XML. However, as an essential player in the era of Big Data, NoSQL databases have yet to receive much research. Consequently, extracting semantic information from NoSQL databases and constructing RDF is necessary. HBase is an integral part of the NoSQL database. According to Google research, many companies are beginning to choose HBase as their data store. It is an excellent idea to generate RDF from HBase for information processing.

This paper concentrates on constructing RDF with HBase. We give formal representations of HBase and RDF and propose an approach to convert HBase data to RDF based on these representations. And then, We developed an RDF construction program with the proposed transformation method and proved the feasibility of our approach through experiments.

The remaining portions of this essay are structured as follows. The relevant work of RDF construction is introduced in Section 2. Section 3 provides some preliminaries of RDF and HBase. Section 4 details the mapping rules of converting HBase to RDF. The designed system and the processing consequence are shown in Section 5. Section 6 presents the conclusion of this thesis and our future work.

## 2. RELATED WORK

Many methods exist to build RDF by extracting information from data with another format. This section presents various RDF construction approaches with databases, XML, and JSON.

Due to the widespread use of relational databases (RDB) in many fields, numerous attempts to build RDF using relational databases have been made. In [4], the authors proposed mapping rules and created RDF from MySQL tables. After specifying the mapping relationship between the RDB and RDF models, the authors in [5] proposed the mapping rules to generate RDF with RDB. A recent review [6] specified some tools for translating RDB into RDF. The authors assessed the capabilities of a total of 17 tools. The W3C recommends two methods for creating RDF from RDB. One is direct mapping, and the other is indirect mapping using a mapping language. The construction methods of RDF from RDB have primarily followed these two methods. Some efforts also exist to construct RDF with other database models, like MongoDB [7, 8].

The second is constructing RDF with XML. In [9], the authors propose XSPARQL language, which combines XQuery and SPARQL. With XSPARQL, people can query both XML and RDF and then implement data conversions between the two formats. The authors of [10] suggest using keywords or graphical queries to convert XML into RDF. In [11], the authors suggest a declarative method based on Scala programming language for converting XML to RDF. In [12], to convert XML to RDF, the authors created a template language based on XPath, which is helpful for users unfamiliar with XPath and RDF triples. For converting XML Schema to Shape Expressions (ShEx), an RDF validation language, some mappings from XML Schema to ShEx are proposed in [13].

JSON is utilized frequently because of its benefits of quick parsing and high transmission effectiveness. Therefore, the conversion of JSON data to RDF has received some attention. In [14], to map JSON to RDF, the authors identify the metadata of JSON and align it with domain vocabulary terms. With OWL, the authors of [15] propose the mapping from JSON to RDF and develop a mapping tool. The authors of [16], who focus on the output from SPARQL queries, convert JSON to JSON-LD, a compact RDF format. Concentrating on coverage data representing spatiotemporal data, the authors of [17] convert the data with JSON format to RDF. In [18], the authors encode coverage data in many formats, such as RDF and JSON. For representing RDF, they provide a mapping between the JSON and RDF via JSON-LD.

There have been few attempts to construct RDF using the HBase database in addition to the above. This paper formally defines HBase and RDF and concentrates on RDF generation from the HBase. Furthermore, our method differs from suggestions in [19] that only support the transformation from simple HBase to RDF.

### 3. PRELIMINARIES

In this section, according to the characteristics of RDF and HBase, we offer formal definitions about HBase and RDF. On the basis of the definitions, it is easy to describe the process of mapping from HBase to RDF.

#### 3.1. Data Model of RDF

RDF is a domain-independent universal description language that does not define domain semantics. The RDF data model consists of a group of RDF statements, represented by the triple denoted as (subject, predicate, object). The triple's first element denotes the resource, the second is its property, and the third is the attribute value of the resource. RDF uses Universal Resource Identifier (URI) to denote resources.

Depending on the predicate type, the object can be a literal value or resource. To overcome the defect that RDF does not define domain semantics, people use the RDF Schema to define domain semantics. The RDF Schema defines a set of modeling primitives with fixed semantics, such as `rdfs: domain` [20].

Following is the formal definition of the RDF data model drawn on the research of [21].

**Definition 1 (RDF data model):** An RDF data model is defined as a 5-tuple:  $RW = (V, E, \Sigma, L, A)$  where:

$V = \{V_1, V_2, V_3, \dots, V_n\}$  is a finite vertex set in the RDF graph, which can be IRI, literals, or blank nodes.

$E = \{E_1, E_2, E_3, \dots, E_n\} \subset V_i \times V_j, i \neq j$  is a finite edge set in the RDF graph.

$\Sigma = \{C, DP, OP, D, T\}$ ,  $C = \{C_1, C_2, C_3, \dots, C_n\}$  is finite set of RDF class resource tags,

$DP = \{DP_1, DP_2, DP_3, \dots, DP_n\}$  is finite set of RDF datatype property tags,

$OP = \{OP_1, OP_2, OP_3, \dots, OP_n\}$  is finite set of RDF object property tags,

$D = \{D_1, D_2, D_3, \dots, D_n\}$  is finite set of RDF data type tags,  $T = \{T_1, T_2, T_3, \dots, T_n\}$  is finite set of RDF instance tags.

$L = \{L_V, L_E\}$  is used to assign labels for vertices and edges of an RDF graph.  $L_V : V \rightarrow \Sigma$  is used to assign labels for vertices, and  $L_E : E \rightarrow \Sigma$  is used to assign labels for edges.

$A$  is an axiom set containing class, attribute, and instance axioms, as shown in Table 1.

Table 1. Axiom set for  $A$ .

RDF triple	RDF axiom
$(L_V(V_i) \in \Sigma.C, \text{rdf:type}, \text{owl:Class})$	$Type(L_V(V_i), \text{Class})$
$(L_V(V_i) \in \Sigma.C, L_E(V_i \times V_j) \in \Sigma.OP, L_V(V_j) \in \Sigma.C)$	$ObjectProperty(L_E(V_i, V_j), \text{domain}(L_V(V_i)), \text{range}(L_V(V_j)))$
$(L_V(V_i) \in \Sigma.C, L_E(V_i \times V_j) \in \Sigma.DP, L_V(V_j) \in \Sigma.C)$	$DatatypeProperty(L_E(V_i, V_j), \text{domain}(L_V(V_i)), \text{range}(L_V(V_j)))$
$(L_V(V_i) \in \Sigma.T, \text{rdf:type}, L_V(V_j) \in \Sigma.C)$	$Individual(L_V(V_i), L_V(V_j))$

An RDF vertex  $V_i \in V$  has a corresponding label denoted as  $L_V(V_i)$  to indicate the subject or object. The vertex label can be IRI, literal or empty nodes. An edge  $(V_i, V_j) \in E$  indicates a directed edge between two vertexes, and its label  $L_E(V_i, V_j)$  denotes the triple's predicate.

There are three types of axioms in Table 1, which are class axioms, attribute axioms, and instance axioms. Let the RDF graph vertex  $V_i, V_j \in V$ , and its labels are  $L_V(V_i), L_V(V_j)$ . The vertex with class resource label means RDF Class concept. When edge  $E_m = (V_i, V_j)$  exists, its label  $L_E(V_i, V_j)$  indicates the RDF predicate that can be a datatype property or object property.

### 3.2. HBase Data Model

As an open-source, scalable, distributed database, HBase is a column-oriented database that differs from relational databases. We can consider the HBase table a multidimensional map indexed by row key, timestamp, and column.

Following content of this section introduce the formal representation of the HBase.

Definition 2 (HBase database model): HBase database model  $HM$  is represented as tuple  $(B, RE, HI)$ , where:

$B = TN \cup CF \cup CQ \cup D$  is the set of essential elements of HBase.

$TN = \{TN_1, TN_2, TN_3, \dots, TN_n\}$  is the set of HBase table names.

$CF = \{CF_1, CF_2, CF_3, \dots, CF_n\}$  is the set of HBase column families.  $cf(tn)$  specifies the column family in table  $tn$ .

$CQ = \{CQ_1, CQ_2, CQ_3, \dots, CQ_n\}$  is the set of HBase column qualifiers.  $cq(tn, cf)$  specifies the column qualifier of column family  $cf$  in table  $tn$ .

$C = \{C_1, C_2, C_3, \dots, C_n\}$  is the set of HBase table columns specified by column family and qualifier.  $c(tn)$  specifies the column of table  $tn$ .

$D$  is a finite set of distinct HBase table column data type.  $D(c)$  denotes data type of HBase column.

$RE$  is a finite set of constraint relation defined by user about HBase database model.

$RE(TN_1, C, TN_2)$  denotes reference relation from table  $TN_1$  to table  $TN_2$  by column  $C$ .

$RE(TN_1, CF, TN_2)$  denotes reference relation from table  $TN_1$  to table  $TN_2$  by column family  $CF$ .

$RE(TN_1, CF)$  denotes embed relation in table  $TN_1$  through column family  $CF$ . Users embed entity information into column families rather than storing it in tables.

$HI$  is a set of HBase instances. An HBase instance is a 6-tuple containing a table name, row key, column family, column qualifier, cell value, and cell timestamp. For a HBase instance  $hi \in HI$ ,  $hi.tn$  suggests its table name,  $hi.rk$  means its row key,  $hi.cf$  denotes its column family,  $hi.cq$  implies its column qualifier,  $hi.v$  suggests its cell value and  $hi.ts$  means its timestamp.

#### 4. MAPPING HBASE DATABASE MODEL TO RDF

Here Function  $\varphi$  is used to map the elements of HBase to RDF.

Rule 1:  $\forall tn \in TN \rightarrow L_v(V_i) = \varphi(tn) \in \Sigma.C \wedge (L_v(V_i), rdf:type, owl:Class)$

Rule 1 maps the HBase table to the RDF vertex with the class label. For example, a table named “employee” is mapped to RDF class  $\langle ns:employee \rangle$ .

Rule 2:  $\forall c \in C(t) \rightarrow \varphi(c) = L_E(V_i \times V_j) \in \Sigma.DP \wedge \varphi(t) = L_v(V_i) \in \Sigma.C$ .

An HBase column in the HBase database model is mapped to an RDF edge with a datatype property label. Table columns in Table 2 such as “personal:name” and “office:phone” are mapped to RDF Datatype Property in Table 3.

Table 2. A sample data with HBase table employee.

Row Key	Time Stamp	Column Family: personal	Column Family: office
00001	timestamp1	personal:name=“John”	office:phone=“415-212-5544”
00001	timestamp1	personal:residence_phone=“415-111-1111”	office:address=“1021 Market St”
00001	timestamp2	personal:residence_phone=“415-111-1234”	

Table 3. Mapping Result of Rule2.

$\langle ns:personal:name \rangle$ a owl:DatatypeProperty . $\langle ns:personal:residence:phone \rangle$ a owl:DatatypeProperty . $\langle ns:office:phone \rangle$ a owl:DatatypeProperty . $\langle ns:office:address \rangle$ a owl:DatatypeProperty .
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Rule 3:

$\forall tn1, tn2 \in TN, c \in c(tn1), RE(tn1, c, tn2) \rightarrow L_E(V_i \times V_j) = \varphi(c) \in \Sigma.OP \wedge L_v(V_i) = \varphi(tn1) \in \Sigma.C$

$\wedge L_v(V_j) = \varphi(tn2) \in \Sigma.C$ .

Rule 3 maps an HBase relation  $RE(tn1, c, tn2)$  between table  $tn1$  and table  $tn2$  to an RDF edge with an object property label. For example, in Table 4, a user follows another user specified by the column value, then the column is mapped to Object Property as shown in Table 5.



Table 4. A sample data with table user.

Row Key	Column Family: follows		
AK	follows:1=foo	follows:2=bar	follows:3=baz
foo	follows:1=bar	follows:2=AK	

Table 5. Mapping Result of Rule 3.

```

<ns:user> a owl:Class .
<ns:follows:1> a owl:ObjectProperty .
<ns:follows:1> rdfs:domain <ns:user> .
<ns:follows:1> rdfs:range <ns:user> .
<ns:follows:2> a owl:ObjectProperty .
<ns:follows:2> rdfs:domain <ns:user> .
<ns:follows:2> rdfs:range <ns:user> .
<ns:follows:3> a owl:ObjectProperty .
<ns:follows:3> rdfs:domain <ns:user> .
<ns:follows:3> rdfs:range <ns:user>

```

Rule 4:

$$\forall tn1, tn2 \in TN, cf \in cf(tn1), RE(tn1, cf, tn2) \rightarrow L_E(V_i \times V_j) = \varphi(cf) \in \Sigma.OP \wedge L_V(V_i) = \varphi(tn1) \in \Sigma.C \wedge L_V(V_j) = \varphi(tn2) \in \Sigma.C.$$

An HBase relation  $RE(tn1, cf, tn2)$  between table  $tn1$  and table  $tn2$  is mapped to the RDF edge with an object property label. For example, in Table 6, one user points to another user by column family qualifier, then the column family “follows” is mapped to the object property as shown in Table 7.

Table 6. A sample data with table user.

Row Key	Column Family: follows		
AK	follows:foo=1	follows:bar=1	follows:baz=1
foo	follows:bar=1	follows:AK=1	

Table 7. Mapping Result of Rule 4.

```

<ns:user> a owl:Class .
<ns:follows> a owl:ObjectProperty .
<ns:follows> rdfs:domain <ns:user> .
<ns:follows> rdfs:range <ns:user>

```

Rule 5:

$$\forall tn1 \in TN, cf \in cf(tn1), RE(tn1, cf) \rightarrow L_V(V_i) = \varphi(tn1) \in \Sigma.C \wedge L_V(V_j) = \varphi(cf) \in \Sigma.C \wedge L_E(V_i \times V_j) = \varphi(ref - cf) \in \Sigma.OP.$$

An HBase relation  $RE(tn1, cf)$  is mapped to the RDF edge with an object property label. This relation means the entity is embedded in the table column family. For example, in Table 8, entity “department” is embedded in the column family “department”, and then the column family is mapped to the RDF class as shown in Table 9.

Table 8. There are sample data of table student.

Row Key	Column Family: student	Column Family: department
001	student:name= "bob"	department:dno=5001
001	student:sex="M"	department:dname="Computer"
001		department:header="Alice"

Table 9. Mapping Result of Rule 5.

```

<ns:student> a owl:Class .
<ns:department> a owl:Class .
<ns:ref-department> a owl:ObjectProperty .
<ns:ref-department> rdfs:domain <ns:student> .
<ns:ref-department> rdfs:range <ns:department> .
<ns:dname> a owl:DatatypeProperty .
<ns:dname> rdfs:domain <ns:department> .
<ns:dname> rdfs:range <xsd:string> .
<ns:header> a owl:DatatypeProperty .
<ns:header> rdfs:domain <ns:department> .
<ns:header> rdfs:range <xsd:string> .

```

Rule 6:

$$(1) \forall hi \in HI \rightarrow L_V(V_i) = \varphi(hi.rk) \in \Sigma.T \wedge L_V(V_j) = \varphi(hi.v) \in \Sigma.D \wedge L_E(V_i \times V_j) = \varphi(hi.cf : hi.cq) \in \Sigma.DP$$

$$(2) \forall hi \in HI, RE(hi.tn, hi.cf : hi.cq, tn2) \rightarrow L_V(V_i) = \varphi(hi.rk) \in \Sigma.T \wedge L_V(V_j) = \varphi(hi.v) \in \Sigma.T \wedge L_E(V_i \times V_j) = \varphi(hi.cf : hi.cq) \in \Sigma.OP$$

$$(3) \forall hi \in HI, RE(hi.tn, hi.cf, tn2) \rightarrow L_V(V_i) = \varphi(hi.rk) \in \Sigma.T \wedge L_V(V_j) = \varphi(hi.cq) \in \Sigma.T \wedge L_E(V_i \times V_j) = \varphi(hi.cf) \in \Sigma.OP$$

$$(4) \forall hi \in HI, RE(hi.tn, hi.cf) \rightarrow L_V(V_i) = \varphi(hi.rk) \in \Sigma.T \wedge L_V(V_j) = \varphi(hi.cf) \in \Sigma.T \wedge L_E(V_i \times V_j) = \varphi(ref - cf) \in \Sigma.OP \quad \mathbf{R}$$

Rule 6 maps a table instance to an RDF instance based on the schema about classes and properties created by the rule 1 to rule 5. The (1) maps cell values to the RDF datatype properties from rule 2. The (2), (3), and (4) add values to the properties mapped from rule 3 to 5. For example, the table "employee" instance with row key "00001" in Table 2 is mapped to RDF, as shown in Table 10.

We all know that HBase instances contain not only values but also timestamps of the cell, so how to map the timestamp of HBase instances is a problem worth investigating. Much research has been done with temporal RDF modeling, such as [22, 23]. In this paper, to be compatible with traditional RDF, our method use the reification mechanism provided by RDF to represent the time of HBase. The reification mechanism uses the three predicates (rdf:subject, rdf:predicate, and rdf:object) provided by RDF to describe the three parts of the statement. Mapping the timestamp of an HBase instance is actually mapping the timestamp to the corresponding statement, as depicted in Table 10.

Table 10. Mapping result of Rule 6.

```

<ns:00001> a <ns:employee> .
<ns:00001> <ns:personal:name> "John" .
<ns:st1> a <rdf:Statement> .
<ns:st1> <rdf:subject> <ns:00001> .
<ns:st1> <rdf:predicate> <ns:personal:name> .
<ns:st1> <rdf:object> "John" .
<ns:st1> <ex:timestamp> timestamp1 .
<ns:00001> <ns:personal:residence_phone> "412-111-1111" .
<ns:00001> <ns:office:phone> "412-212-5544" .
<ns:00001> <ns:office:address> "1021 Market St" .
...

```

## 5. SYSTEM IMPLEMENTATION AND EXPERIMENTAL RESULTS

### 5.1. System Architecture

To validate our construction method of RDF with HBase, we developed a prototype system to extract the HBase data and map these data to RDF. We develop this system with OpenJDK 17.0.1 and JavaFX on a PC (Intel i5-5200U (4) @ 2.700GHz, RAM 8 GB, and ArchLinux system).

The system obtains data by connecting the database and then outputs RDF data. The database parsing, RDF construction, and display modules comprise most of the system, as depicted in Figure 1. Database parsing module analyses the semantic relationships and data information of data based on the HBase database formal definition. The RDF construction module constructs RDF Schema with obtained semantic relationships from the parsing module. And then, this module constructs RDF data with RDF Schema and data information by mapping rules introduced in Section 4. Finally, the display module exhibits the RDF data created by the RDF construction module.

Figure 2 displays the system's screen snapshot. The GUI of the system contains three display areas. The TextArea on the left exhibits information about the HBase data model, including tables and columns. Moreover, the TextArea (in the upper right) below the label "RDF Schema" shows the corresponding RDF Concept, such as Classes and Properties, in the form of the RDF Turtle. The TextArea (in the lower right) below the label "RDF Individual" exhibits the corresponding RDF individuals.

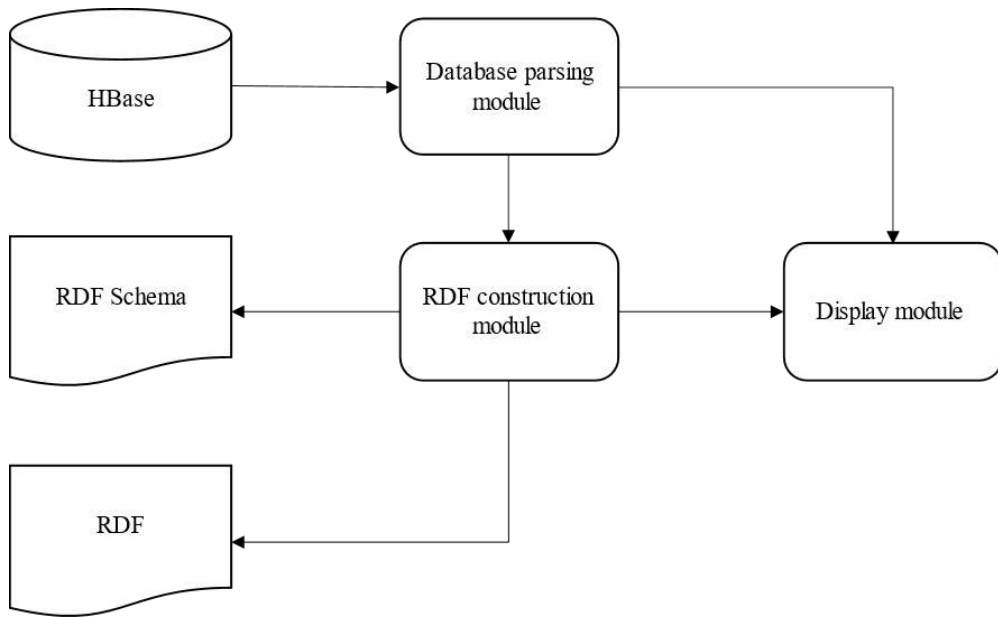


Figure 1. The architecture of System.

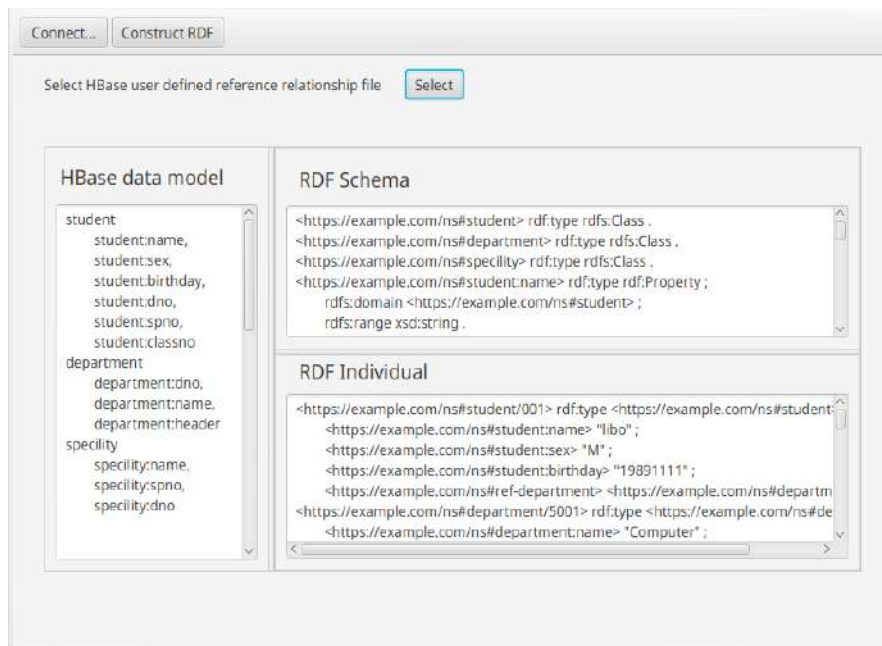


Figure 2. The screen snapshot of System.

## 5.2. Experimental Results and Discussions

We conducted experiments with the HBase data source from the article “Introduction to Hbase Schema Design” created by a Cloudera engineer. Since the paper [19] does not provide the source code of the system for building RDF from HBase data, this section simulates the implementation of the corresponding construction system according to the method given in the paper.

The results of the RDF construction experiments based on HBase data source are shown in Table 11. The first column of the table represents the test metrics of the RDF construction experiment, including the time consumed for construction, the number of RDF classes generated, the number of datatype properties, the number of object properties, the number of domain axioms, the number of range axioms, and the number of RDF triples generated. and the number of RDF triples generated. The second to third columns are the experimental results of the method in this paper (denoted as H2R) and the method proposed in the paper [19] (denoted as OBDI).

Table 11. Experimental results.

<b>Test metrics</b>	<b>H2R</b>	<b>OBDI</b>
Construction time (ms)	8934	10142
RDF classes	5	4
Datatype properties	9	16
Object properties	5	0
Domain axioms	14	0
Range axioms	14	0
RDF triples	680047	190020

Table 11 shows that our method can retain more database information for RDF creation compared to another method. And also, our method has advantages in terms of construction time. Our method constructs more RDF triples per unit time when compared with the other method. While paper [19] only considers the mapping of basic data attributes in HBase database, our approach not only considers the mapping of basic data attributes in HBase, but also pays attention to implicit reference relationships in HBase database, such as column-value based references, column-qualifier based references, and embedded references. Implicit reference relationships in the data are mapped to RDF by parsing user-defined implicit reference relationship documents. Meanwhile, our approach only traverses the HBase data source once to resolve the semantic relationships of HBase data, while the paper [10] traverses the data source twice to resolve the schema information and data information of the HBase data source. Therefore, our method not only retains more database information for constructing RDF compared with another method, but also has advantages in construction efficiency.

## 6. CONCLUSIONS AND FUTURE WORK

Because of the growth of the Web, more and more people select NoSQL databases to solve the data management problems that come from big data. HBase occupies a part of the market share in NoSQL databases. Hence, the study of transforming HBase data to RDF is conducive to data use and solving the problem of insufficient available RDF data. In this paper, we design a method to convert HBase semantic relationships and data to RDF and develop a system to confirm its feasibility.

In the future, optimizing the implementation system to improve efficiency is the first task. We will also pay attention to expanding the method for more column-oriented databases.

## REFERENCES

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.
- [2] J. Hendler, T. Berners-Lee, and E. Miller, "Integrating applications on the semantic web," *Journal of the Institute of Electrical Engineers of Japan*, vol. 122, pp. 676–680, 01 2002.

- [3] Z. Ma, M. A. M. Capretz, and L. Yan, “Storing massive resource description framework (RDF) data: a survey,” *Knowl. Eng. Rev.*, vol. 31, no. 4, pp. 391–413, 2016.
- [4] E. Bytyçi, L. Ahmedi, and G. Gashi, “RDF mapper: Easy conversion of relational databases to RDF,” in *Proceedings of the 14th International Conference on Web Information Systems and Technologies, WEBIST 2018, Seville, Spain, September 18-20, 2018* (M. J. Escalona, F. J. D. Mayo, T. A. Majchrzak, and V. Monfort, eds.), pp. 161–165, SciTePress, 2018.
- [5] J. F. Sequeda, M. Arenas, and D. P. Miranker, “On directly mapping relational databases to RDF and OWL,” in *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012* (A. Mille, F. Gandon, J. Misselis, M. Rabinovich, and S. Staab, eds.), pp. 649–658, ACM, 2012.
- [6] F. Michel, J. Montagnat, and C. Faron Zucker, “A survey of RDB to RDF translation approaches and tools,” research report, I3S, May 2014. ISRN I3S/RR 2013-04-FR 24 pages.
- [7] H. Abbes and F. Gargouri, “Mongodb-based modular ontology building for big data integration,” *J. Data Semant.*, vol. 7, no. 1, pp. 1–27, 2018.
- [8] N. Soussi and M. Bahaj, “Exploiting nosql document oriented data using semantic web tools,” in *International Conference on Advanced Intelligent Systems for Sustainable Development*, pp. 110–117, Springer, 2018.
- [9] S. Bischof, S. Decker, T. Krennwallner, N. Lopes, and A. Polleres, “Mapping between RDF and XML with XSPARQL,” *J. Data Semant.*, vol. 1, no. 3, pp. 147–185, 2012.
- [10] M. Kharrat, A. Jedidi, and F. Gargouri, “A semantic approach for transforming XML data to RDF triples,” in *14th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2015, Las Vegas, NV, USA, June 28 - July 1, 2015* (T. Ito, Y. Kim, and N. Fukuta, eds.), pp. 285–290, IEEE Computer Society, 2015.
- [11] J. P. McCrae and P. Cimiano, “LIXR: quick, succinct conversion of XML to RDF,” in *Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 19, 2016* (T. Kawamura and H. Paulheim, eds.), vol. 1690 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016.
- [12] J. Huang, C. Lange, and S. Auer, “Streaming transformation of XML to RDF using xpath-based mappings,” in *Proceedings of the 11th International Conference on Semantic Systems, SEMANTiCS 2015, Vienna, Austria, September 15-17, 2015* (A. Polleres, T. Pellegrini, S. Hellmann, and J. X. Parreira, eds.), pp. 129–136, ACM, 2015.
- [13] H. Garcia-Gonzalez and J. E. L. Gayo, “Xmlschema2shex: Converting XML validation to RDF validation,” *Semantic Web*, vol. 11, no. 2, pp. 235–253, 2020.
- [14] F. Freire, C. Freire, and D. Souza, “Enhancing JSON to RDF data conversion with entity type recognition,” in *Proceedings of the 13th International Conference on Web Information Systems and Technologies, WEBIST 2017, Porto, Portugal, April 25-27, 2017* (T. A. Majchrzak, P. Traverso, K. Krempels, and V. Monfort, eds.), pp. 97–106, SciTePress, 2017.
- [15] S. J. R. Méndez, A. Haller, P. G. Omran, J. Wright, and K. Taylor, “J2RM: an ontology-based json-to-rdf mapping tool,” in *Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020 (UTC)* (K. L. Taylor, R. S. Gonçalves, F. Lécuyer, and J. Yan, eds.), vol. 2721 of *CEUR Workshop Proceedings*, pp. 368–373, CEUR-WS.org, 2020.
- [16] P. Lisena and R. Troney, “Transforming the JSON output of SPARQL queries for linked data clients,” in *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23-27, 2018* (P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis, eds.), pp. 775–780, ACM, 2018.
- [17] J. D. Blower and M. Riechert, “Coverages, JSON-LD and RDF data cubes,” in *Proceedings of the Workshop on Spatial Data on the Web (SDW 2016) co-located with The 9th International Conference on Geographic Information Science (GIScience 2016), Montreal, Canada, September 27-30, 2016* (K. Janowicz, J. Lieberman, K. Taylor, G. McKenzie, S. Gao, S. J. D. Cox, and E. Parsons, eds.), vol. 1777 of *CEUR Workshop Proceedings*, pp. 9–16, CEUR-WS.org, 2016.
- [18] P. Baumann, E. Hirschorn, J. Masó-Pau, V. Meticariu, and D. Misev, “All in one: Encoding spatio-temporal big data in xml, json, and RDF without information loss,” in *2017 IEEE International Conference on Big Data (IEEE BigData 2017), Boston, MA, USA, December 11-14, 2017* (J. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, and M. Toyoda, eds.), pp. 3406–3415, IEEE Computer Society, 2017.

- [19] V. Kiran and R. Vijayakumar, "Ontology based data integration of nosql datastores," in 2014 9th International Conference on Industrial and Information Systems (ICIIS), pp. 1–6, IEEE, 2014.
- [20] G. Antoniou, P. Groth, F. van Harmelen, and R. Hoekstra, A Semantic Web Primer, 3rd Edition. MIT Press, 2012.
- [21] T. Fan, L. Yan, and Z. Ma, "Mapping fuzzy RDF(S) into fuzzy object-oriented databases," Int. J. Intell. Syst., vol. 34, no. 10, pp. 2607–2632, 2019.
- [22] F. Zhang, Z. Li, D. Peng, and J. Cheng, "RDF for temporal data management - a survey," Earth Sci. Informatics, vol. 14, no. 2, pp. 563–599, 2021.
- [23] D. Yang and L. Yan, "Transforming XML to RDF(S) with temporal information," J. Comput. Inf. Technol., vol. 26, no. 2, pp. 115–129, 2018.

## AUTHORS

**Hui Hu** is currently a master's candidate in the College of Computer Science and Technology at the Nanjing University of Aeronautics and Astronautics, China. His research interests include RDF and Semantic Web.



# ETHICAL ALGORITHMS IN HUMAN-ROBOT-INTERACTION. A PROPOSAL

Jörg H. Hardy

Free University of Berlin, Germany, Department of Philosophy,  
AkakiTsereteli State University of Kutaisi, Georgia,  
Faculty of Business, Law and Social Sciences

## **ABSTRACT**

*Autonomous robots will need to form relationships with humans that are built on reliability and (social) trust. The source of reliability and trust in human relationships is (human) ethical competence, which includes the capacities of practical reason and moral decision-making. As autonomous robots cannot act with the ethical competence of human agents, a kind of human-like ethical competence has to be implemented into autonomous robots (AI-systems of various kinds) by way of ethical algorithms. In this paper I suggest a model of the general logical form of (human) meta-ethical arguments that can be used as a pattern for the programming of ethical algorithms for autonomous robots.*

## **KEYWORDS**

*AI Algorithms, Ethical Algorithms, Ethics of Artificial Intelligence, Human-Robot-Interaction*

## **1. INTRODUCTION: ETHICAL COMPETENCE OF ROBOTS – A CHALLENGE FOR HUMAN-ROBOT-INTERACTION**

Over the past two decades, robots with high levels of autonomy have become members of the human society. Autonomous robotics has made tremendous progress with significant advances in areas such as driverless cars, home assistive robots, robot-assisted surgery, and unmanned aerial vehicles. However, incidents such as the fatal Tesla crash show the risks from improper use of this technology. The progress in the development and deployment of robots (and any kind of autonomous AI-systems) sets experts the task to design algorithms that can generate reliable, trustworthy robots that possess a human-like ethical competence.

Autonomous robots will need to form relationships with humans. Human relationships are built on (social) trust, which is a key influence in decisions whether an autonomous agent, be it a human or a robot, should act or not. Studies of reliability and trust in automation have shown that trust depends on reliability and predictability: trust increases slowly if the system behaves as expected, but drops quickly if we experience failure [1]. Autonomous robots are independent decision-makers, and may therefore exhibit unpredictable behaviour. Reasoning with trust and norms is necessary to justify and explain a robots' decisions and to draw inferences about accountability for failures, and hence induce meaningful communication with autonomous robots [2].

Trust is a specific human stance, namely the expectation of beneficial behaviour. The stance of trust is a complex both cognitive, and volitional process, informed by the broader context of



morality, ethics, and social norms. Trust in human interaction is based on the ethical competence of human persons. Autonomous robots too would need a certain kind of a human-like ethical competence in order to be trusted [2, 3].

A reliable, trustful human-robot-interaction has to be guided by an ethics of algorithms. For the various aspects of the ethics of algorithms see Bostrom & Yudkowsky [4], Arnold & Scheutz [5], Malle, B. F. [6], Aggarwal & Mishra [7], Nida-Rümelin & Weidenfeld [8], and Tsamados & Aggarwal & Cowls & Morley & Roberts & Taddeo & Floridi [9]. The realm of meta-ethical questions is described by Beauchamp & Childress 2013 [10] and Siep [11]. In this paper, I confine myself to a proposal for a model of meta-ethical arguments and a corresponding logical pattern for ethical algorithms. My proposal is based on a minimalist theory of ethical competence that is not committed to a particular meta-ethical theory.

## 2. THE META-ETHICAL FRAMEWORK: THE MORAL STANCE

The source of (human) ethical competence is the moral stance [12]. The moral stance is a person's capacity and enduring motivation to *accept moral demands* for their own sake, regardless of any socio-economic reward for moral behaviour.

Human beings pursue happiness, and all our happiness-conducive rational and deliberate *social* (interpersonal) activity is intrinsically desirable. This assumption about the *condition humana*, which I take to be uncontroversial, has an important consequence for the understanding of morality. Morality prevails our entire social life. Throughout our whole life, we make moral demands on other people, and we are faced with their moral demands. If we accept those demands, we act in other person's interests and thus put *constraints* on some particular self-interests. Having taken the moral stance, we permanently *want* to put constraints on *particular self-interests*. If moral actions are nevertheless *desirable*, they are desirable for their own sake, and if a persistent social behaviour is desirable for its own sake, then it is part of our pursuit of happiness, that is, part of an *overall* desirable life. We can maintain the moral stance for a lifetime only if morality is intrinsically desirable and part of our pursuit of happiness [13].

People have various motivating reasons for moral actions, such as, for example, the interest in successful social cooperation, the desire for social recognition, religious belief, altruism. When we have taken the moral stance, we take morality to be intrinsically desirable. A social action is intrinsically desirable if we take it to be desirable for its own sake, regardless of its consequences. By contrast, a social action is extrinsically desirable when being done for gaining a certain socio-economic success. It is true, we take many (and probably the most) social actions to be both, intrinsically and extrinsically desirable. However, if an action is intrinsically desirable, its desirability (and its value) does not *depend* on any external social success or reward.

The moral stance has three aspects:

(i) If we have taken the moral stance, we have certain *moral beliefs*, which we express in specifically *moral demands*. As moral agents, we accept moral demands and act in other person's interests.

(ii) When we act in other person's interests, we put constraints on some of our self-interests. Thus, moral agents are capable of having second order volitions. The moral stance includes the particular capacities of practical reason and moral decision making. Through practical reasoning we form intentions, which consist of a belief and a corresponding desire. Practical reasoning therefore is both a cognitive, and volitional capacity—for an intention without a belief would lack propositional content, and an intention without a desire would lack motivating force.

(iii) Moral experiences create a specifically moral familiarity between persons. Moral agents consider each other not merely as contracting parties who agree upon certain terms of contract, but they have certain attitudes towards each other, such as gratitude, respect, recognition, solidarity or moral resentment and even indignation. In other words: Morality is a mode of people's encountering with each other. As moral agents, we *share* the desire for the common experiences of respect, solidarity, sincerity, and trust.

## 2.1. Moral Demands

Moral demands have four features:

(i) Moral demands aim to protect *common goods*, such as bodily integrity and (personal and social) autonomy, and the (logically speaking) *particular instances* of a common good, above all human beings, with certain vulnerable properties. For this reason, arguments for ethical claims have to rely on general evaluative assumptions about common goods, which ideally every person can agree upon. Since the acceptance of a moral demand expresses the will of a person, the general evaluative assumptions of ethical arguments—which we might also call ethical principles—are *common agreements upon common goods*.

(ii) Moral demands are evident: What we owe to each other is obvious because we all know the common goods, which moral demands aim to protect. Morality is, as Immanuel Kant says, a matter of fact of reason (*Critique of Practical Reason*, 5:31). Thus, we do not need complex and fallible reasonings in order to understand the content of moral demands. However, we need complex reasonings in order to find solutions to particular ethical problems.

(iii) Moral demands are universal; they hold for any person and any action in any situation—regardless of any particular property of an individual person.

(iv) Moral demands are categorical (or unconditional, respectively); moral actions do not depend on any particular condition, and they are not primarily a means for achieving a certain end, but they are rather an end in itself (intrinsically desirable). It is true; we very often do moral actions for their own sake (intrinsically desirable) as well as for the sake of social advantages (extrinsically desirable), because we seek social recognition and want to avoid blame and punishment. Having taken the moral stance, we, however, do moral actions for their own sake because they contribute to our pursuit of happiness.

Knowing these features of moral demands is part of our understanding of the human condition. Let me briefly explain: If we accept moral demands for their own sake, we then follow moral norms (or laws), which aim to protect common goods. All human beings are equal in seeking happiness; we all have the desire for conducting a good life. And we all share the same vulnerabilities, we all know that we all have the same vulnerable properties, such as the fact that we all can suffer from pain. Once we are aware of the fact that moral norms are made to protect the vulnerable properties of human beings, we know that moral norms are universal. Everyone can suffer from pain and no one wants to suffer from pain. If we accept an individual person's demand not to be hurt because we consider it a *moral* demand, then we accept *everyone's* demand not to be hurt. For example, if I am sure that not inflicting pain on a human being is morally right, I expect everyone else to think the same way. The very idea that there is a moral obligation only for me—or a particular group of people, respectively—to perform moral actions does not make sense. When we keep in mind that moral norms are made for protecting common goods, we also know that moral norms must be categorical (or unconditional, respectively). If we seriously respect the happiness-conducive interests of other persons, we want to do this under any possible conditions—even though we might sometimes fail to perform morally right actions

through negligence. It would not make sense to accept moral demands and to do moral actions merely as a means for achieving a certain particular end that we would not want to achieve under some other conditions. When we want to protect common goods, we consider moral demands universal and categorical (or unconditional, respectively). In this respect, taking the moral stance means to achieve what Lawrence Kohlberg (“Study of Moral Development”, New York, Garland 1994) describes as the sixth and highest stage of moral development.

There is a possible objection against the idea of categorical moral demands: Consider a situation in which someone hurts an assassin in order to prevent him from attacking defenseless people. Actions of that kind are undoubtedly morally right. In some cases, in which a person is faced with a conflict of moral norms, she/he has to break a certain moral law and to impair a certain good in order to protect a higher good. The fact of moral conflicts shows that we need to agree upon a hierarchy of goods in order to solve those conflicts, but it does not conflict the assumption that moral demands are categorical.

## 2.2. Common Goods and the Awareness for Humanity

Human beings share various common goods, to which general evaluative premises of ethical arguments refer to, such as human dignity, bodily integrity, and (personal and social) autonomy. Particular instances of common goods have vulnerable and valuable properties that moral demands aim to protect. A certain property is vulnerable because it can be impaired or even destroyed, and we consider such a property valuable because we want to protect it.

The insight into the value of common goods is the motivating reason for moral obligation: We know that everyone can suffer from bodily pain and from losing the authority over her/his own life, and we do not want anyone to suffer from pain or to lose authority over her/his life. Let us call this insight the *awareness for humanity*. The awareness for humanity is both a certain kind of *understanding* and *empathy*. We all know what it means to be hurt or to lose authority over one’s own life. These experiences are common ones—we did not have them without *sharing* them with other persons. Empathy provides the awareness for humanity with its volitional, motivating force. There is, however, no universal empathy, for only propositional attitudes can be generalized. The awareness for humanity therefore requires empathy and understanding, that is, the understanding of the *human condition*. If we have taken the moral stance, we know that we all share the same vulnerability and the desire for common goods like bodily integrity and (personal and social) autonomy, and we therefore want to take care of each other’s happiness-conducive interests. Sharing is caring, and caring is sharing.

The *recognition* of a common good is a *motivational belief*, which conjoins the insight that a certain vulnerable and valuable thing is in fact a good with the intention to protect such a good for its own sake—regardless of any other particular interest that one might also have for protecting a certain good. We just would not *have* the belief that something *is* a good without having the desire to *protect* it.

According to a widespread view, moral obligations (duties) and moral norms are objective obligations and norms since they should not depend on an individual person’s (contingent) wanting. This idea seems to be an implication of the assumption that moral demands are universal and categorical (in the sense explained above). But we have to be careful with the assumption of objective norms and obligations. It is true, having taken the moral stance, we want to protect the vulnerability of every person, and we always have this intention, not merely in a particular situation. We may consider moral obligations and norms *as if* they were objective, because we want to accept moral demands as being universal and categorical. Yet, moral intentions are subjective because of the simple reason that individual persons want to accept

moral obligations and to establish moral norms. There is no obligation and no social norm without the corresponding wanting of individual persons. What we ought to do, is what we *want* to do. Moral obligations and moral norms don't fall like stars from heaven, but they are the result of the *common agreements* and *intentions* of *individual* persons.

We recognize common goods and accept universal, categorial moral demands through our *own* moral thinking and wanting, that is, through our personal autonomy. If we accept moral demands for their own sake, our *individual* (self-guiding) intention to do moral actions coincides with our acceptance of a quasi-objective moral obligation. Universal moral norms are those, which we (ideally) *all can agree* upon.

As robots cannot take the human moral stance, they cannot possess *human* ethical competence. However, autonomous robots are independent decision-makers. As robots make decisions and act by virtue of algorithms, technologists have to implement ethical algorithms into autonomous robots.

### 3. THE LOGICAL FORM OF META-ETHICAL ARGUMENTS AND ETHICAL ALGORITHMS

Meta-ethical arguments, that is, arguments with a normative, ethical conclusion have both evaluative, and descriptive premises that refer to common goods, particular instances of common goods, moral agents, intentions, and moral actions (or a certain kind of moral actions, respectively). When we *argue* for ethical claims, we agree on *general evaluative premises*, which express assumptions about common goods whose particular instances have certain vulnerable properties  $\{V_1, \dots, V_n\}$ .

For example, the human body's vulnerable property is the fact that it can suffer from pain. A person's mind can be manipulated. A person's dignity can be humiliated. Those are the vulnerable properties ethical arguments typically refer to. More precisely: When we argue for ethical claims, we have to make

- (i) general *evaluative* assumptions about common goods that we want to protect, which ideally all moral agents can agree upon,
- (ii) general and particular *descriptive* assumptions about the vulnerable properties of a particular instance of a common good,
- (iii) general and particular descriptive assumptions about the intentions and obligations of moral agents, and finally
- (iv) general and particular descriptive assumptions about particular actions (or a set of actions, respectively), which is necessary and adequate for protecting the vulnerable properties of a particular instance of a common good.

A particular moral action *A* is *necessary* for protecting a vulnerable property *V* of a particular instance of a common good if only action *A* will prevent a particular instance of a common good in a given particular situation from being impaired or even destroyed. A particular moral action *A* is *adequate* if and only if an agent is in the position to do *A* and doing *A* does *not* impair her/his own well-being.

Meta-ethical arguments have this general form:

- (1) ( $\forall$ common good CG,  $\forall$ particular instance ICG of a common good,  $\forall$ vulnerable property V,  $\forall$ moral agent MA,  $\forall$ an agent's intention I to protect V): If an abstract entity CG is a common good and if (logically speaking) a particular instance ICG of a common good, for example, an

individual person, has the vulnerable property V, then every moral agent MA has the intention I to protect the vulnerable property V of any particular instance ICG, for example the vulnerable property V of any other person. (The *antecedens* of this assumption contains an evaluative as well as a descriptive statement.)

(2) ( $\forall$ common good CG,  $\forall$ particular instance ICG of a common good,  $\forall$ vulnerable property V): The entity CG is a common good and every particular instance ICG of CG (for example, every individual person) in fact has the vulnerable property V.

(3) ( $\forall$ moral agent MA,  $\forall$ an agent' s intention I to protect V): Therefore, every moral agent MA has the intention I to protect everyone' s vulnerable property V.

(4) ( $\forall$ moral agent MA,  $\forall$ an agent' s intention I to protect V,  $\forall$ an agent' s obligation O): If an agent MA has the intention I to do actions of the kind A, which are necessary and adequate for the protection of a vulnerable property V of ICG, then she/he *ought*—has the (self-guiding) obligation O—to perform particular actions of the kind A (and must not do opposing actions of the kind non-A).

(5) ( $\forall$ moral agent MA): An agent MA has the intention I to do actions of the kind A, which are necessary and adequate for the protection of a vulnerable property V of ICG.

Conclusion: Therefore, every moral agent ought to do—has the obligation to do—actions of the kind A.

Meta-ethical arguments of this kind have the following elementary logical form:

Let ICG be a *particular instance* of a *common good*, V a *vulnerable property* of an ICG, MA a *moral agent*, I an agent' s *intention* to do an *action* A, which is necessary and adequate for the protection of a vulnerable property V of ICG, and to avoid any action that would *impair* an ICG, respectively, and O an agent' s (self-guiding) *obligation* to do a certain action A, which is necessary and adequate for the protection of a vulnerable property V of an ICG.

(1)  $\forall(x, y) [ICG(x) \& V(x)] \supset [MA(y) \& I(y)]$

(2)  $\forall(x) ICG(x) \& V(x)$

(3)  $\forall(y) [MA(y) \& I(y)] \supset [MA(y) \& O(y)]$

(4)  $\forall(y) MA(y) \& I(y)$

Conclusion:  $\forall(y) MA(y) \& O(y)$

Here is an example:

If bodily integrity is a common good and if every individual person—as being (logically speaking) a particular instance of the common good bodily integrity—has the vulnerable property that she/he can be hurt, then every moral agent has the intention to protect everyone from being hurt.

Bodily integrity is a common good and every individual person can suffer from being hurt.

Therefore, every person has the intention to protect everyone from being hurt.

If we (and every moral agent) have the intention to protect everyone from being hurt, then we ought to act the way that we don't hurt someone.

Therefore, we (and every moral agent) ought to act the way that we don't hurt someone.

Notice that arguments of this kind are not vulnerable to the objection of the so-called naturalistic fallacy since the premises entail the entire evaluative information of the conclusion.

Ethical algorithms based on this logical form of meta-ethical arguments will include

- logical and semantic information about common goods and particular instances of a common good, such as human beings, with certain vulnerable and valuable properties such as bodily integrity and personal autonomy
- logical and semantic information about the intentions of moral agents
- logical and semantic information about sets of actions that are necessary and adequate for protecting the vulnerable properties of a particular instance of a common good.

In order to implement ethical competence into autonomous robots, we have to combine ethical algorithms of the above kind with *technical tools* that enable an autonomous system to *identify* a morally right (necessary and adequate) *particular action under particular circumstances*. Very useful methods for this complex task are model checking and rational verification.

#### 4. MODEL CHECKING AND ETHICAL ALGORITHMS

Scholars of computer science have identified a logic language and method to evaluate reliability and trust in human-robot interactions through *model checking*, which is capable of reasoning about concepts such as epistemic dependence between agents [14]. This formalism can be extended by ethical and normative reasoning and, in particular, ethical algorithms for tool support [2, 4, 15]. Most relevant to this aim is the paradigm of *rational verification*, which is based on model checking, that enables analyzing human-robot behaviours under the assumption of agents behaving rationally, and allowing for incentives and preferences [16].

#### 5. CONCLUSION

The study of the ethical aspects of human-robot-interaction is an interface between computer science and philosophical ethics. The source of human ethical competence is the moral stance, that is, a person's capacity and enduring motivation to accept moral demands for their own sake. As robots cannot act with human ethical competence, reliable and trustful human-robot-interaction requires the implementation of the capability of ethical decision-making into autonomous robots by way of ethical algorithms that are based on the general logical form of meta-ethical arguments. Meta-ethical arguments, that is, arguments with a normative, ethical conclusion have both evaluative, and descriptive premises that refer to common goods, particular instances of a common good—above all human beings—with certain vulnerable properties, intentions of moral agents, and moral actions, which aim at protecting particular instances of a common good. In this paper, I have proposed a model of the general logical form of meta-ethical arguments and ethical algorithms. Computer scientists and technologists can use this model and specific applications of this model in order to implement ethical competence and ethical decision-making factors into algorithms and software tools with support for autonomous AI-systems.

#### ACKNOWLEDGEMENT

The model of ethical algorithms that I propose in this paper is a revised version of a model that I have presented at the FLoC 2018 – workshop “Robots, Morality, and Trust through the Verification lens” in Oxford, July 2018, [http://qav.cs.ox.ac.uk/robots\\_morality\\_trust/](http://qav.cs.ox.ac.uk/robots_morality_trust/). I am most grateful to Morteza Lahijanian, Lu Feng and Nils Jansen for critical comments that helped me improving my model.

**REFERENCES**

- [1] Huang, X. &Kwiatkowska, M. (2017) “Reasoning about Cognitive Trust in Stochastic Multiagent Systems”, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*, AAAI Press.
- [2] Lahijanian, M. &Kwiatkowska, M. (2016) “Social Trust: A Major Challenge for the Future of Autonomous Systems”, *AAAI Fall Symposium on Cross-Disciplinary Challenges for Autonomous Systems*, AAAI Press.
- [3] Kuiper, B.(2016)“Human-like Morality and Ethics for Robots”, *Proceedings of the Association for the Advancement of Artificial Intelligence*, Ann Arbor, Michigan.
- [4] Bostrom, N. &Yudkowsky, E. (2014)“The Ethics of Artificial Intelligence”, in: K. Frankish & W. M. Ramsey (eds.), *The Cambridge Handbook of Artificial Intelligence*, Cambridge, pp 316-334.
- [5] Arnold, T. &Scheutz, M. (2016)“Against the moral Turing test: accountable design and the moral reasoning of autonomous systems”, *Ethics and Information Technology*, Vol. 18, 2, pp 103-115, Springer.
- [6] Malle, B. F. (2016) “Integrating robot ethics and machine morality: the study and design of moral competence in robots”, *Ethics and Information Technology*, Vol. 18, pp. 243–256, Springer.
- [7] Aggarwal, S. &Mishra, S. (2021) *Responsible AI: Implementing Ethical and Unbiased Algorithms*, Springer.
- [8] Nida-Rümelin, J. & Weidenfeld, N. (2018)*Digitaler Humanismus: Eine Ethik für das Zeitalter der Künstlichen Intelligenz*, München.
- [9] Tsamados, A. & Aggarwal, N. &Covels, J. & Morley, J. &Roberts, H. & Taddeo, M. &Floridi, L. (2022)“The ethics of algorithms: key problems and solutions”, *AI & Society*, Vol. 37, pp 215–230. <https://doi.org/10.1007/s00146-021-01154-8>
- [10] Beauchamp, Tom L. & Childress, J. F. (2013) *Principles of Biomedical Ethics*, Oxford, Oxford University Press.
- [11] Siep, L. (2004) *Konkrete Ethik. Grundlagen der Natur- und Kulturethik*, Frankfurt am Main, Suhrkamp.
- [12] Hardy, J. (2017) “Understanding Ethical Reasoning”, Hoesch, M. & / Laukötter, S. (eds.). *Natur und Erfahrung. Bausteine zu einer praktischen Philosophie der Gegenwart*. Festschrift für Ludwig Siep, Münster, mentis.
- [13] Hardy, J. (2011) *Jenseits der Täuschungen – Selbsterkenntnis und Selbstbestimmung mit Sokrates*, Göttingen, V & R unipress (ch.XIV).
- [14] Kwiatkowska, M. (2007) “Quantitative verification: models, techniques and tools”, *Proceedings of ESEC/SIGSOFT FSE 2007*, pp 449-458, IEEE CS Press.
- [15] Alechina, N. & Halpern, J. Y. &Kash, I. A. & Logan, B. (2017) “Incentivising Monitoring in Open Normative Systems”, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*, AAAI Press.
- [16] Wooldridge, M. & Gutierrez, J. &Harrenstein, P. &Marchioni, E. &Perelli, G. &Toumi, A. (2016) “Rational verification: from model checking to equilibrium checking”, *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence*, AAAI Press, pp 4184-4190.

**AUTHOR**

**Dr. Jörg H. Hardy** is Senior Lecturer (PD) at the Free University of Berlin, Germany, Department of Philosophy, and Professor of Philosophy and Linguistics at the Akaki Tsereteli State University of Kutaisi, Georgia, Faculty of Business, Law and Social Sciences.

## **AUTHOR INDEX**

<i>Aleksandra Popovska-Mitrovikj</i>	145
<i>Anishka Duvvuri</i>	21
<i>Bill Xu</i>	79
<i>Chunhei Zhu</i>	37
<i>Hui Hu</i>	167
<i>Jon Atle Gulla</i>	127
<i>Jörg H. Hardy</i>	179
<i>Kerry Zhang</i>	105
<i>Laura L. Pullum</i>	59
<i>Lina Lumburovska</i>	145
<i>Muhamad Adib Bahari</i>	49
<i>Muhammad Kamal Abdul Kiram</i>	49
<i>Navya Kovvuri</i>	21
<i>Nicole Ma</i>	93
<i>Randy Klepetko and Ram Krishnan</i>	01
<i>Rebecca Victor</i>	21
<i>Shah Runnizam Mohd Salleh</i>	49
<i>Sneka Kumar</i>	21
<i>Tanush Kaushik</i>	21
<i>Vesna Dimitrova</i>	145
<i>Yu Sun</i>	93,79,117
<i>Yujie Xing</i>	127
<i>Yulin Zhang</i>	37,117
<i>Zhiyu Huang</i>	153
<i>Aleksandra Popovska-Mitrovikj</i>	145
<i>Anishka Duvvuri</i>	21
<i>Bill Xu</i>	79
<i>Chunhei Zhu</i>	37
<i>Hui Hu</i>	167
<i>Jon Atle Gulla</i>	127
<i>Jörg H. Hardy</i>	179
<i>Kerry Zhang</i>	105