Machine Learning & Applications

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 4th International Conference on Machine Learning & Applications (CMLA 2022), June 25~26, 2022, Copenhagen, Denmark
- 9th International Conference on Computer Science, Engineering and Information Technology (CSEIT 2022)
- 14th International Conference on Networks & Communications (NeTCoM 2022)
- 3rd International Conference on NLP & Big Data (NLPD 2022)
- 14th International Conference on Applications of Graph Theory in Wireless Ad hoc Networks and Sensor Networks (GRAPH-HOC 2022)
- 14th International Conference on Wireless & Mobile Networks (WiMoNe 2022)
- 4th International Conference on Internet of Things (CIoT 2022)
- 14th International Conference on Network and Communications Security (NCS 2022)

Published By

AIRCC Publishing Corporation

Volume Editors

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

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

ISSN: 2231 - 5403
ISBN: 978-1-925953-70-1
DOI: 10.5121/csit.2022.121101 - 10.5121/csit.2022.121113

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

Preface

4th International Conference on Machine Learning & Applications (CMLA 2022), June 25~26, 2022, Copenhagen, Denmark, 9th International Conference on Computer Science, Engineering and Information Technology (CSEIT 2022), 14th International Conference on Networks & Communications (NeTCoM 2022), 3rd International Conference on NLP & Big Data (NLPD 2022), 14th International Conference on Applications of Graph Theory in Wireless Ad hoc Networks and Sensor Networks (GRAPH-HOC 2022), 14th International Conference on Wireless & Mobile Networks (WiMoNe 2022), 4th International Conference on Internet of Things (CIoT 2022), 14th International Conference on Network and Communications Security (NCS 2022) was collocated with 4th International Conference on Machine Learning & Applications (CMLA 2022). 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 CMLA 2022, CSEIT 2022, NETCoM 2022, NLPD 2022, GRAPH-HOC 2022, WiMoNE 2022, CIoT 2022 and NCS 2022. 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, CMLA 2022, CSEIT 2022, NETCoM 2022, NLPD 2022, GRAPH-HOC 2022, WiMoNE 2022, CIoT 2022 and NCS 2022 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 CMLA 2022, CSEIT 2022, NETCoM 2022, NLPD 2022, GRAPH-HOC 2022, WiMoNE 2022, CIoT 2022 and NCS 2022.

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

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair

Organization

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

Program Committee Members

Abdalhossein Rezai, University of Science and Culture, Iran
Abdallah Makhoul, University of Bourgogne, France
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Abdelhadi Assir, Hassan 1st University, Morocco
Abderrahmane Ez-zahout, Mohammed V University, Morocco
Abdessamad Belangour, University Hassan II Casablanca, Morocco
Abhay Kumar Agarwal, Kamla Nehru Institute of Technology, India
Abhilash, Cyrielle Castle, India
Addisson Salazar, Universitat Politècnica de València, Spain
Adnan Saher Mohammed, Karabuk University, Turkey
Ahmed Farouk AbdelGawad, Zagazig University, Egypt
Ahmed Haitham Najim, Sfax TUNISIA-Imam Al-Adham University, Iraq
Ajay B Gadicha, Amity University, India
Al janabi, Alhikma college University, Baghdad, Iraq
Alexander Gelbukh, Instituto Politécnico Nacional, Mexico
Ali A. Al-Zuky, Mustansiriyah University, Iraq
Ali H. Wheeb, University of Baghdad, Iraq
Aliasghar Tarkhan, University of Washington, USA
Alireza Valipour Baboli, University Technical and Vocational, Iran
Allel Hadjali, LIAS/ENSMA, France
Alper Ugur, Pamukkale University, Turkey
Alyssa Gonzalez, University of Baghdad, Iraq
Amal Azeroual, Mohammed V University, Morocco
Amina El murabet, Abdelmalek Essaadi University, Morocco
Amitabh Mishra, University of West Florida, United States
An Braeken, Vrije Universiteit Brussel, Belgium
Ana Luísa Varani Leal, University of Macau, China
Anala M R, RV College of Engineering, India
Anandakumar Haldorai, Sri Eshwar College of Engineering, India
Anastasios Doulamis, National Technical University of Athens, Greece
Anita Yadav, Harcourt Butler Technical University, Kanpur
Antoni B. Chan, City University of Hong Kong, Hong Kong
António Abreu, ISEL - Polytechnic Institute of Lisbon, Portugal
Anuj Singal, GJU S&T, India
Aparicio Carranza, Computer Engineering Technology, USA
Archit Yajnik, Sikkim Manipal University, India
Aref Wazwaz, Dhofar University, Oman
Aridj Mohamed, Hassiba Benbouali University of Chlef Algeria
Ashraf Elnagar, University of Sharjah, UAE
Ashutosh Bahuguna, Ministry of Electronics & IT, India
Ashwag Albakri, Jazan University, Saudi Arabia
Atanu Nag, IFTM University, India
Ayad Ghany Ismaeel, Al-Kitab University College, Iraq
Azeddine Wahbi, Hassan II University, Morocco

Pavel Loskot,	ZJU-UIUC Institute, China
Prakash Kanade,	LeenaBOT Robotics, USA
Prathap Siddavaatam,	Ryerson University, Canada
Prudhvi Parne,	Bank of Hope and University of Louisiana, USA
Przemyslaw Falkowski-Gilski,	Gdansk University of Technology, Gdansk, Poland
Quang Hung Do,	University of Transport Technology, Vietnam
R Devanathan,	Hindustan Institute of Technology and Science, India
R I. Rauf,	University of Abuja, Nigeria
Radha Raman Chandan,	Banaras Hindu University, India
Rajeev Kanth,	University of Turku, Finland
Rajesh Bose,	Brainware University, India
Ramadan Elaiess,	University of Benghazi, Libya
Raman Chandan,	Banaras Hindu University, India
Ramgopal Kashyap,	Amity University Chhattisgarh, India
Reena Malik,	Chitkara University, India
Riffi Jamal,	Sidi Mohamed Ben Abdellah University, Morocco
Robert Ssali Balagadde,	Kampala international University, Uganda
Rodrigo Pérez Fernández,	Universidad Politécnica de Madrid, Spain
Ruhaidah Samsudin,	Universiti Teknologi Malaysia, Malaysia
S Vijayarani,	Bharathiar University, India
S.Sridhar,	Easwari Engineering College, India
Saad Aljanabi,	Alhikma College University Baghdad, Iraq
Sabah Suhail,	University of Tartu, Estonia
Saeed Iranmanesh,	Shahid Bahounar University of Kerman, Iran
Safawi Abdul Rahman,	Universiti Teknologi MARA, Malaysia
Sahar Saoud,	Ibn Zohr University, Morocco
Sahil Verma,	Chandigarh University, India
Said El Kafhali,	Hassan First University of Settat, Mororcco
Saif aldeen Saad Obayes,	Shiite Endowment Office, Iraq
Salah-ddine Krit,	Ibn Zohr University Agadir, Morocco
Samir Kumar Bandyopadhyay,	University of Calcutta , India
Sasikumar,	Vellore Institute of Technology, India
Sathyendra Bhat J,	St Joseph Engineering College, India
Seppo Sirkemaa,	University of Turku, Finland
Seyed Mahmood Hashemi,	Beijing University of Technology, China
Shadan Sadigh Behzadi,	Islamic azad University, Iran
ShahidAli,	Manukau Institute of Technology, New Zealand
Shahnaz N.Shahbazova,	Azerbaijan Technical University, Azerbaijan
Shahram Babaie,	Islamic Azad University, Iran
Shashikant Patil,	ViMEET Khalapur Raigad MS India, India
Shervan Fekri-Ershad,	Najafabad Azad University, Iran
Shing-Tai Pan,	National University of Kaohsiung, Taiwan
Shin-Jer Yang,	Soochow University, Taiwan
Shoeib Faraj,	Technical And Vocational University Of Urmia, Iran
Siarry Patrick,	Universite Paris-Est Creteil, France
Sidi Mohammed Meriah,	University of Tlemcen, Algeria
Sikandar Ali,	China University of Petroleum-Beijing, China
Sikandar Ali,	The University of Haripur, Pakistan
Simanta Shekhar Sarmah,	Alpha Clinical Systems, USA
Smain Femmam,	UHA University France, France
Sourav Sen,	Upstart Network Inc., USA
Sridhar Iyer,	S.G. Balekundri Institute of Technology, India

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

4th International Conference on Machine Learning & Applications (CMLA 2022)

9th International Conference on Computer Science, Engineering and Information Technology (CSEIT 2022)

14th International Conference on Networks & Communications (NeTCoM 2022)

3rd International Conference on NLP & Big Data (NLPD 2022)

14th International Conference on Applications of Graph Theory in Wireless Ad hoc Networks and Sensor Networks (GRAPH-HOC 2022)

14th International Conference on Wireless & Mobile Networks (WiMoNe 2022)

4th International Conference on Internet of Things (CIoT 2022)

14th International Conference on Network and Communications Security (NCS 2022)

BREXIT: PREDICTING THE BREXIT UK ELECTION RESULTS BY CONSTITUENCY USING TWITTER LOCATION BASED SENTIMENT AND MACHINE LEARNING

James Usher and Pierpaolo Dondio

School of Computing Technological University Dublin, Dublin, Ireland

ABSTRACT

After parliament failed to approve his revised version of the 'Withdrawal Agreement', UK Prime Minister Boris Johnson called a snap general election in October 2019 to capitalise on his growing support to 'Get Brexit Done'. Johnson's belief was that he had enough support countrywide to gain a majority to push his Brexit mandate through parliament based on a parliamentary seat majority strategy. The increased availability of large-scale Twitter data provides rich information for the study of constituency dynamics. In Twitter, the location of tweets can be identified by the GPS and the location field. This provides a mechanism for location-based sentiment analysis which is the use of natural language processing or machine learning algorithms to extract, identify, or distinguish the sentiment content of a tweet (in our case), according to the location of origin of said tweet. This paper examines location-based Twitter sentiment for UK constituencies per country and aims to understand if location-based Twitter sentiment majorities per UK constituencies could determine the outcome of the UK Brexit election. Tweets are gathered from the whisperings of the UK Brexit election on September 4th 2019 until polling day, 12th December 2019. A Naive Bayes classification algorithm is applied to assess political public Twitter sentiment. We identify the sentiment of Twitter users per constituency per country towards the political parties' mandate on Brexit and plot our findings for visualisation. We compare the grouping of location-based sentiment per constituency for each of the four UK countries to the final Brexit election first party results per constituency to determine the accuracy of location-based sentiment in determining the Brexit election result. Our results indicate that location-based sentiment had the single biggest effect on constituency result predictions in Northern Ireland and Scotland and a marginal effect on Wales base constituencies whilst there was no significant prediction accuracy to England's constituencies. Decision tree, neural network, and Naïve Bayes machine learning algorithms are then created to forecast the election results per constituency using location-based sentiment and constituency-based data from the UK electorate at national level. The predictive accuracy of the machine learning models was compared comprehensively to a computed-baseline model. The comparison results show that the machine learning models outperformed the baseline model predicting Brexit Election constituency results at national level showing an accuracy rate of 97.87%, 95.74 and 93.62% respectively. The results indicate that location-based sentiment is a useful variable in predicting elections.

KEYWORDS

BREXIT Election, Twitter, Sentiment, UK Election.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 01-16, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121101

1. INTRODUCTION

One area that has experienced an increase in use of Twitter is that of electoral campaigning and political strategy formulation. With the increasing prominence of Twitter as a political communication tool, politicians and political parties now maintain an active presence on same. Twitter provides an optional static data field in the user profile which allows the user to provide their location. Twitter users have an option to fill in this field thus providing their location. Mobile devices now pick up the user location from GPS coordinates and provide a location coordinate for the user to choose from a dropdown menu also. In Twitter, tweets can be posted with the location address field which identifies the user's current position. These geographical tweets, so to speak, with text content have been utilised to detect real-time events, such as estimating Typhoon trajectory or Earthquake location [1]. Sentiment analysis of Twitter messages, is the act of retrieving opinions from tweets. Twitter users express sentiments about specific topics or entities with different strengths and intensities, where these sentiments are strongly related to their personal feelings and emotions. Computational sentiment analysis methods attempt to measure different opinion dimensions. By classifying polarity estimation using Natural Language Processing (NLP) into three polarity classes namely, positive, negative, and neutral or supervised and unsupervised machine learning algorithms fulfil the objective of classification. Sentiment analysis has been accomplished in a variety of genres of communication, including professional, media-like news articles [2], web forums [3,4] and Facebook [5,6]. The growth in sentiment analysis has projected itself to politics, as political strategist and research firms pursue the valuable opinions of large populations to help formulate political strategies. Twitter sentiment principally has become a widely explored foundation for election forecasting [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21], due to instance data availability in conjunction with the extensive use of Twitter globally.

This paper examines location-based Twitter sentiment per UK constituencies and aims to understand if location-based Twitter sentiment majorities per UK constituencies could determine the outcome of the UK Brexit election from the period of September 4th 2019 until polling day December 12th 2019. We build a model for classifying "tweets" into binary features for classifying positive or negative sentiment based on location. We build three machine learning models a decision tree, neural network and Naïve Bayes model using location-based sentiment and constituency-based data from the UK electorate to predict the Brexit election results by first party. In section II, we present the findings of our literature review. In section III, we examine the political landscape, the datasets, python machine learning libraries, data filtering, visualisation, location modelling, bassline modelling, prediction and evaluation. Section IV looks at the results of each model and section V concludes and looks at future research.

2. RELATED WORK

Election forecasting using a 'Twitter Tracker' for the Irish General Election 2011 was explored by [8] allowed users, and journalists, to tap into the content on Twitter pertaining to the election through an accessible dashboard-style interface. The approach assumed that the percentage of votes that a party receives is related to the volume of related content in social media. Larger parties will have more members, more candidates and will attract more attention during the election campaign. Smaller parties, likewise, will have a much smaller presence. Volume was based on the measure as the proportional share of party mentions in a set of tweets for a given time period. They found that Twitter does appear to display a predictive quality which is marginally augmented by the inclusion of sentiment analysis. [15] election predictions use a similar method whereby the prediction is based on the number of times the name of a candidate is mentioned in tweets prior to elections. The approach was successful in predicting the winner of

the Venezuelan, Paraguayan and Ecuadorian Presidential elections held in Latin America during the months of February through April 2013. These findings contrast severely with [14] who found that simple methods for predicting election results based on sentiment analysis of tweets text are no better than random classifiers. They recommend that, in order to improve the accuracy of sentiment analysis, a method is needed to go beyond the reliance on word polarity alone. Pre-processing techniques such as POS tagging and word sense disambiguation might be necessary, as well as the inclusion of non-lexical features. Similarly, [16,20] found that party mentions had no relevance to the predicting the outcome of the German elections 2009. [9] computed the number of Twitter messages referring to a particular political party as an indication of the eventual winner. The analysis achieved an 86% classification accuracy. [10] analysed the on-line popularity of Italian political leaders throughout 2011, the voting intention of French internet-users in both the 2012 Presidential ballot and subsequent Legislative election, and found a remarkable ability of social-media to forecast on average electoral results. Findings also uncovered sentiment analysis of social media seems to provide more accurate predictions when focusing on the most popular leaders or on mainstream parties.

Twitter sentiment can be used also for other electoral purposes such that [17] examined the political preference of voters using Twitter and found that Twitter-generated content and user behavior during the election campaigns contain useful knowledge that can be used for predicting the political preference of those users. In addition, they showed the predicted preference changes over time and that these changes co-occur with campaign-related events. This type of analysis is quite useful, too, when taking into account the controversial 350m Brexit Bus claim by PM Boris Johnson [25], perceived as a deliberate attempt to swing voters to the support the Conservative mandate of 'Getting Brexit Done'. The benefits of monitoring allows party strategists to measure the reaction of campaign-related events via Twitter sentiment and the polls and tailor responses accordingly. This is further demonstrated by [19]. However, [18] finds that not all Twitter sentiment corresponds to the poll's predictions. Alternative election prediction methods and concepts do exist such that [26] examined the use of forecasting a Conservative Party victory through the pound using ARIMA and Facebook's Prophet. [27] successfully used location-based Twitter sentiment to predict the US presidential elections of 2016 and UK general elections of 2017. The study extracted location data provided by users from tweet meta-data and this was used to plot state-wise subjectivity and polarity on a map of the US. For the UK elections, tweets using two different filters (keywords and geo-location) were plotted for visualisation. Findings showed that sentiment based on location did reflect on-ground public opinion. In UK elections, the Labour party performed better than expected and also had a more positive sentiment on Twitter. The study observed that tweets mentioning Donald Trump had higher subjectivity than ones discussing Hillary Clinton. The study concluded that the ability to map user sentiment provided tremendous benefit in accurately predicting public opinion as this allows distinction of user opinions based on geographical location. [28] performed location-based sentiment analysis on 650,000 tweets in order to understand trends and patterns regarding the Indian elections. Discoveries saw both positive and negative sentiment change from one location to the other and 'social events' can trigger a sharp rise in both negative and positive sentiments regarding a political party.

3. DATA COLLECTION

3.1. Political Climate

In the countdown to the Brexit election on 12th December 2019, UK Prime Minister Boris Johnson and his Conservative Party vouched to leave the EU with the 'Withdrawal Agreement' settled. Labour pledged to renegotiate the 'Withdrawal Agreement' although accepting Brexit

and held a referendum, letting voters choose between the renegotiated Withdrawal deal and remaining in the EU. While Labour's election strategy early on was to emphasise that the vote was about more than Brexit, the party changed its focus. The message was that Labour's leadership was not opposing Brexit. By opposing Mr. Johnson's deal, it wanted to find what it believed to be a better one [22]. The Liberal Democrats guaranteed to revoke Article 50, while the SNP proposed to hold a second Brexit referendum, however, revoking Article 50 if the alternative was a no-deal exit. The DUP and the Brexit party supported the Conservative party stance to 'Get Brexit Done'. Plaid Cymru and the Green Party supported a second Brexit referendum, supporting the belief that the UK should stay in the EU.

The incumbent Conservative government Brexit strategist, Domnic Cummings, was known to have spent a considerable amount of time using artificial intelligence to tackle Brexit activities [23]. In his infamous 'weirdos and misfits' blog, he made reference to statistical and ML forecasting [24]. In what appeared to be a daring move to call a snap election, it appears feasible that Prime Minister Boris Johnson's strategists may have conducted their own location-based sentiment analysis to determine if the country backed his 'Get Brexit Done' mandate. Conventional wisdom would suggest that through user location information, a more accurate understanding of the on-ground location public opinion would be advantageous. By consuming location data, a political strategist could candidly gauge the support levels for candidates and policies for each region or constituency, thus, giving a more detailed picture of public opinion ultimately defining political strategy.

Hypothesis 1: Can Twitter location-based sentiment per constituency predict the outcome of the Brexit Election?

3.2. Brexit Twitter Election Dataset

The dataset is filtered to contain tweets from September 4th until 12th December 2019. The Brexit Election Twitter dataset contains over 7.3 million individual tweets collected daily within said period. Each tweet is identified by a tweet identifier, the date-time-seconds of the submission (GMT), location, verified indictor, the text content, #Hastag, number of followers, twitterhandle, and a sentiment score derived from the Native Bayes machine learning algorithm which ranges from -1 = negative. 0 = neutral and 1= positive. The data is stored on a remote server which houses an SQL LITE database to store the retrieved data. The server environment consists of an 8 Core Intel R Xeon (R) CPU E5, a 2630 v4 2.20 GHz Intel processor with 16 GB RAM, 400GB memory and a 64-bit Windows 10 Operating System.

3.3. Westminster Parliamentary Constituency Dataset

This dataset contains all of the results of the Westminster Parliamentary Constituencies for the Brexit Election 2019 [29]. The dataset contains the ons_id, ons_region_id (which act as an identifier for the constituency within the British Isles), the result by First party (i.e. the party that won the seat with the most votes) using the first past the post voting system, country, region name, constituency name, majority votes, electorate valid_votes, invalid_votes, majority and gender of candidate.

3.4. Twitter API

The most common way to access Twitter data is through the Twitter REST API. Using the secure tokens obtained via the Oauth process, this provides authentication and thus allows the user to receive the requested Twitter data. We utilise the Twitter API, the python Tweepy library, the

python Naïve Bayes textblob library, and we use a SQL LITE database as a repository for the Twitter data.

3.5. Python Tweepy Library

We utilise the "Tweepy" python library to accept the Twitter data by creating a tweepy.py file [30] Twitter offers several streaming endpoints, each customised to certain use cases. These streams are categorised as follows:

- Public Streams: Streams of the public data flowing through Twitter. Suitable for following specific users or topic or data mining.
- User Streams: Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
- Site Streams: The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users.

For this study, we are concerned with Brexit public streams. We customise our Stream Listener API from the Tweepy library to capture the incoming tweets from the Brexit #Hashtags contained in Table I.

Table 1. Hashtags

#Hashtags	***Sample Tweets***
#Brexit	*RT @IsolatedBrit: To avoid a fascist revolution. That's why we're leaving the EU? #Brexit*
#BrexitChaos	*No-deal Brexit could put public at risk, warns Met chief #Brexit #BrexitChaos #BrexitCrisis*
#BrexitShambles	*New Labour Leader desperately needed. Preference would be Yvette Cooper, David Lammy or Chuka #BrexitShambles*

3.6. Python Naives Bayes using TextBlob

Naive Bayes is a straightforward model for classification. It has been proven to be effective in text categorisation. *Text blob* employs a multinomial Naive Bayes classifier, where the assumption is that each feature is conditional independent to other features given the class. Bayes theorem is illustrated in equation (1)

$$P\left(\frac{C}{T}\right) = \frac{P\,(C)\;\; P\,(T/C)}{P\,(T)} \tag{1}$$

where *c* is a specific class, in our context either *positive* sentiment or *negative* sentiment, and *t* is a tweet text we want to classify. *P(c)* and *P(t)* is the prior probabilities of a sentiment class *c* and a text *t*. *P(t |c)* is the probability the text appears given this class. The goal is choosing value of *c* to maximise *P(c|t)*. Using a *bag of words* approach, each text *t* can be represented as a vector features $\{w_1, w_2, .., w_n\}$m, where $w_i$represent the occurrency of each word w_i in the text *t*, usually weighted. Therefore $P(w_i|c)$ is the probability of the i^{th} feature in text *t* appears given class *c*. In the Naive Bayes approach, each feature w_i is independent from each other. Therefore $P(w_i|c)$ is the probability of the i^{th} feature in text *t* appears given class *c*. In order to classify a text *t*, we

need to compute the maximum likelihood estimation of each one. When making prediction for a new text *t*, we calculate the log likelihood $\log P(c) + \sum_i \log P(w_i|c)$ of different classes, and take the class with highest log likelihood as prediction.

3.7. Data Filtering

The objective of data filtering is reducing the noise from the Twitter dataset concerning neutral sentiment and non-UK constituency locations. We have collected 7,332,842 tweets between September 4th 2019 and December 12th 2019. We are only interested in the 650 Westminster Parliamentary Constituencies in the UK. To reduce the convolution between the domestic tweets (users who live in the election constituencies) and non-domestic tweets (tweets outside elections constituencies), we can apply geo-location filtering: users whose tweets originate from the election constituency are included in the prediction model inclusive of positive and negative location-based sentiment, the remainder are ignored. Figure 1 illustrates the unique values from the Twitter dataset and shows that there are 120,530 locations.

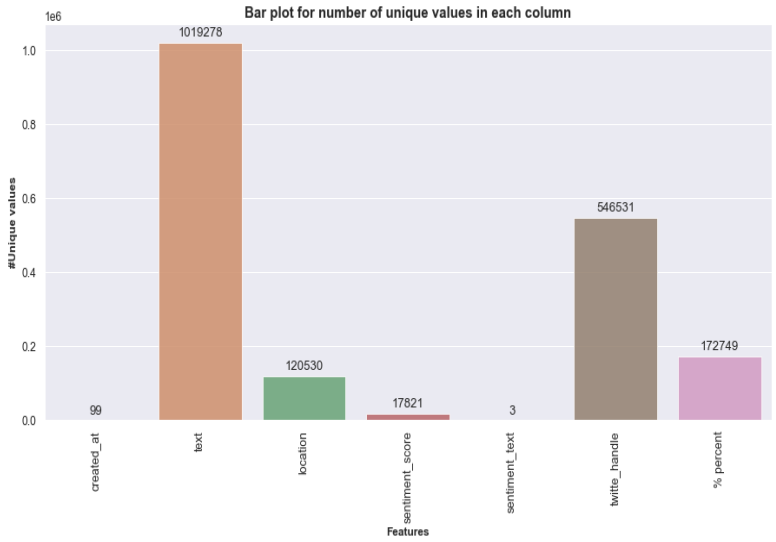


Figure 1. Bar Plot for Unique Values

Looking closer at the ratio on location we can deduce from the Twitter dataset that 61.2% have no location inserted. Furthermore, users have input generic locations such as London, England, Scotland, United Kingdom. Figure 2 illustrates same.

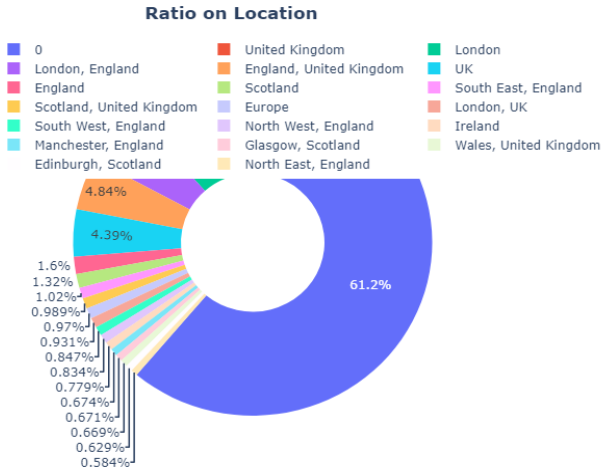


Figure 2. Ratio on Location for No Location

To eliminate non-domestic locations, we merge our Twitter locations with the 'Westminster Parliamentary Constituencies' to filter out the noise, essentially, removing all non-Westminster Parliamentary Constituencies'. Figure 3 shows that we matched 469 constituencies out of 650 for the four UK countries from the Twitter dataset.

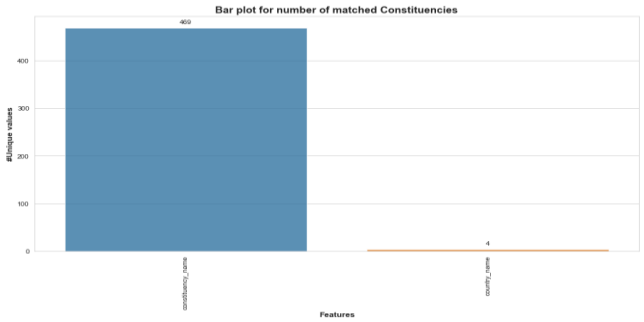


Figure 3. Bar Plot for number of matched constituencies

We show the top 10 ratios by location for Westminster Parliamentary Constituencies in Figure 4 by percentage of tweets after completing the filtering process.

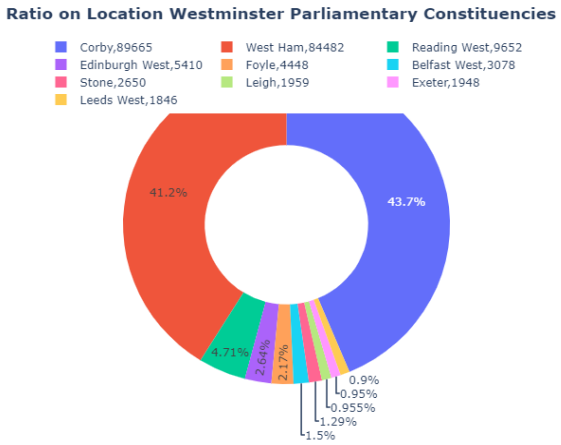


Figure 4. Ratio on Location for Westminster Parliamentary Constituencies.

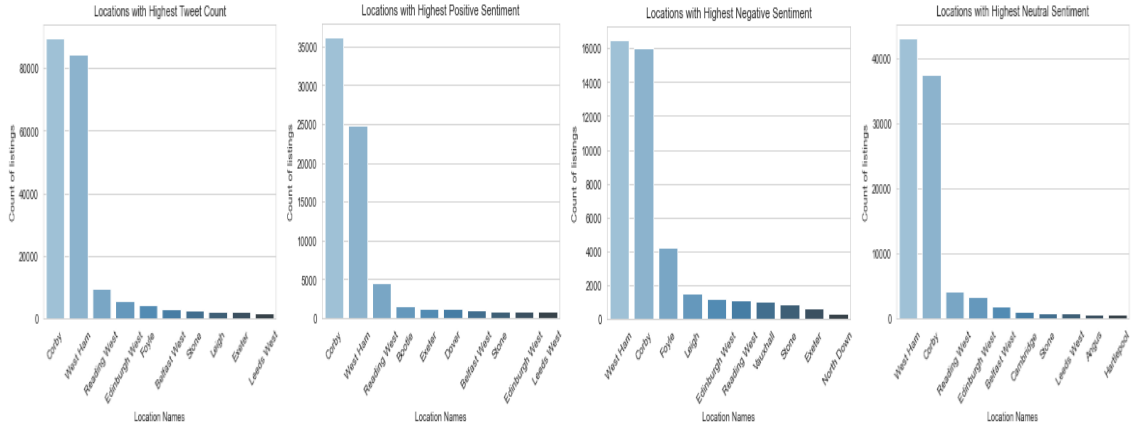


Figure 5. Sentiment Breakdown for Westminster Parliamentary Constituencies

The breakdown of location-based sentiment is illustrated in figure 5. Interestingly, from an English perspective Corby and West Ham feature predominately as locations with high tweet activity; they represent almost 84% of total tweet activity and score highly in positive, negative and neutral tweet sentiment. Corby is a Conversative constituency and West Ham is a Labour constituency. The Scottish constituency of Edinburgh West, which is Liberal Democrats party territory opposed to Brexit, captures almost 0.95% of tweet activity. From a Northern Ireland perspective, Belfast West and Foyle are Sinn Fein and SDLP party respective constituencies both parties opposed to Brexit. Both Northern Ireland constituencies represent 3.67% of the total tweet activity.

3.8. Data Visualisation and Location Modelling

UK choropleths were obtained from the Open Geography portal from the Office for National Statistics (ONS) to create each of the UK maps [31] The matplotlib python library was used to create each visualisation in addition to the 'Westminster Parliamentary Constituencies' data provided by the House of Commons library. The UK choropleths are imported into a static web page which takes the data from the 'Westminster Parliamentary Constituencies' dataset and plots the constituencies. Sentiment is established by computing a value count majority for each constituency from the Twitter dataset. The value count majority is then identified as the majority

sentiment for the particular constituency and applied accordingly to the constituencies for each of the visualisations.

3.9. Baseline Model

Exploratory Data Analysis (EDA) and Linear Discriminant Analysis (LDA), which is a supervised machine learning technique used to find a linear combination of features that separates two or more classes of objects, is undertaken to understand the vote distribution between the UK Brexiteer and Bremain constituencies. The results of which indicate that there is an imbalance, that is, where the class distribution is not equal or close to equal and, is instead, biased or skewed. We see from the target distribution and the linear discriminate analysis that there is an overwhelming majority of UK constituencies in favor of 'Getting Brexit Done'.

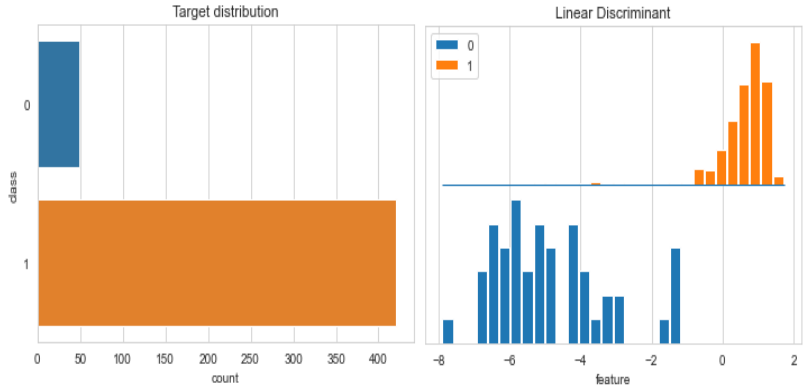


Figure 6. Target and LDA Analysis

Knowing that the EDA and LDA have identified an imbalanced classification, building a baseline metric to evaluate our models would prove meaningful. Particularly in imbalanced classification models, it can appear that the final model isn't really doing much better than guessing. Hence, we need to establish what accuracy is adequate to call our model significant. We import Sklearn dummy classifier library and use the "uniform" strategy which generates predictions uniformly at random. The objective of balancing classes is to either increase the frequency of the minority class or decrease the frequency of the majority class. This is done in order to acquire approximately the same number of instances for both the classes. Baseline accuracy is computed without location-based sentiment and is established for each of the UK countries as illustrated in table 2. Baseline accuracy is also computed for the combined UK constituencies at national level for comparable assessment to all machine learning models.

Table 2. Baseline Metrics

Country	Method	Baseline Accuracy
NI	Uniform	.40
Scotland	Uniform	.44
Wales	Uniform	.32
England	Uniform	.45
UK	Uniform	.54

3.10. Predicting the Brexit Election

To determine the sentiment of a constituency *k*, each tweet that contains the constituency's location, as derived from the location field, is considered a vote for that constituency *k*. If a tweet contains positive or negative sentiment, it counts as a vote towards constituency *k* or, otherwise, it is ignored. The defining vote per constituency *k* is defined by the majority sentiment indicator be that positive or negative sentiment such that positive sentiment is a vote to 'Get Brexit Done' as defined by the 'Brexiteer' parties. This represents the Brexit political mandates of the Conversative, Labour, DUP and the Brexit Party (also known as UKIP). Negative sentiment is a vote to reject Brexit as defined by the 'Bremainer' parties. This is indicative of the Liberal Democrats, Green Party, Plaid Cymru, Sinn Fein, SNP and the smaller independent parties. We pre-process the tweets by removing emoticons. The tweets are turned into word vectors and a standard Naive Bayes classifier setup is employed for classification as outlined in section 3.6. Each visualisation illustrates the Brexit constituency sentiment per 'First Party' result (i.e., the first party past the post or the first party to get the required constituency vote majority, coloured by Brexit stance) and the associated location-based sentiment predicted result split out per country such that the Brexiteer parties are denoted in blue and Bremainer parties are denoted in red.

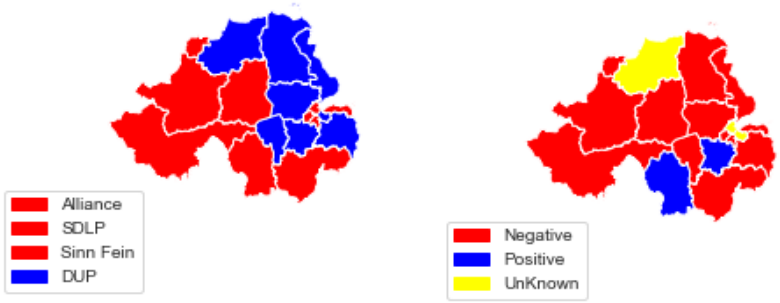


Figure 7. Northern Ireland Location based sentiment

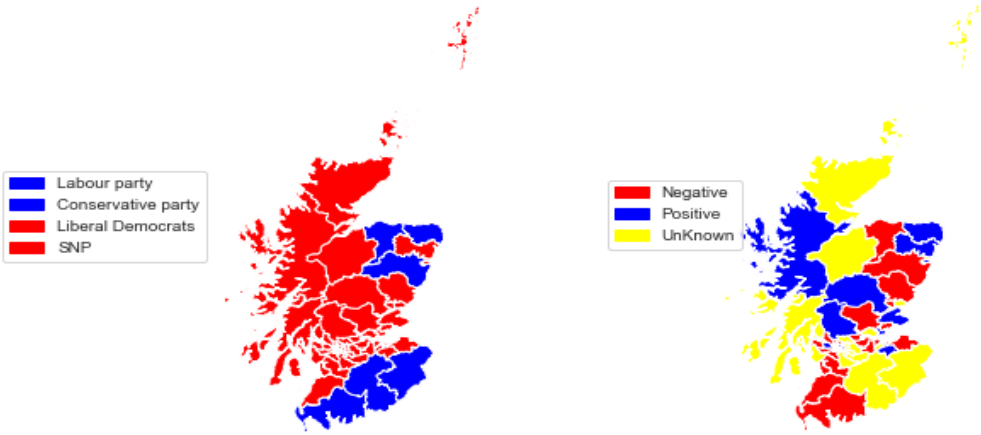


Figure 8. Scotland Location based sentiment

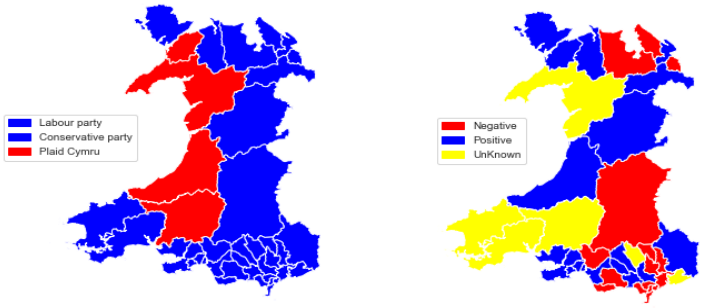


Figure 9. Wales Location-based Sentiment

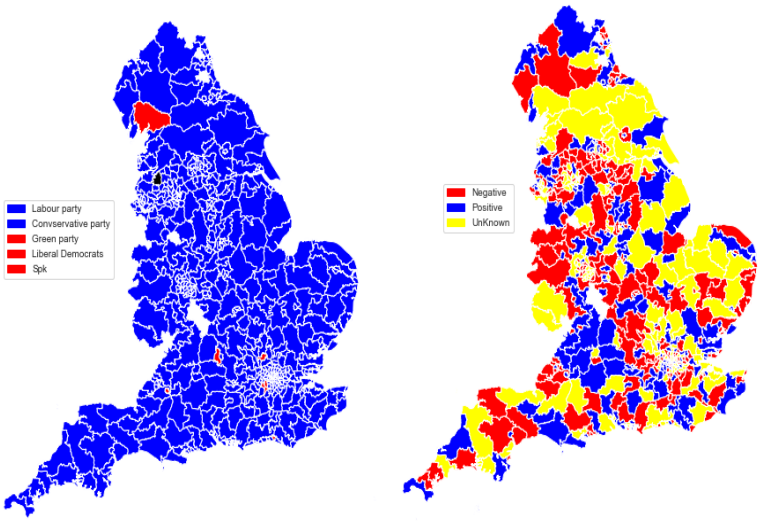


Figure 10. England Location-based Sentiment

3.11. Evaluating the Forecast

We evaluate the location-based sentiment forecast for all constituencies per UK country in Table 3. To measure this, we allocate binary values to the First Party where 1 refers to the Brexiteer parties and 0 refers to the Bremainer parties. We also allocate binary values to the location-based sentiment prediction for the constituency where 1 refers to a predicted vote for the Brexiteer parties and 0 is a vote for the Bremainer parties. The Pearson correlation coefficient is used to detect the degree of linear correlation between two continuous variables, in this case the political party stance and the location-based sentiment. The Pearson correlation coefficient values range from -1 to 1. Positive values mean the selected variables have a positive correlation with the target. Negative value means the selected variable has a negative correlation with the target. The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between two variables. The equation defines the population correlation coefficient and is denoted by:

$$p = \frac{cov(x_1, X_2)}{(\sigma x_1, \sigma X_2)} \quad (2)$$

The Northern Ireland and English constituencies have a positive correlation between location-based sentiment and the First Party. This means that an increase or decrease in the value of the location-based sentiment variable is generally followed by an increase or decrease in the value of the First Party variable. Scotland and Wales show a negative correlation meaning that an increase (decrease) in the value of the location-based sentiment variable is generally followed by a decrease or (increase) in the value of the First Party variable. We use a confusion matrix to compute the performance measurement for our location-based sentiment machine learning classification. Where P = Pearson, B= Baseline= Accuracy, P= Precision, R= Recall and F1 =F1. Accuracy indicates that our highest number of constituencies that were predicted correctly was Northern Ireland 60%, followed by Scotland 58%, Wales 44% and then England at almost 38%.

Table 3. Accuracy of Predictions per UK Country

Country	P	B	A	P	R	F1
NI	0.204	.40	**.60**	.50	.167	.250
Scotland	-0.121	**.**44	**.58**	.143	.40	.21
Wales	-0.271	.32	**.441**	.993	.438	.596
England	0.068	.45	.378	.973	.377	.543

3.12. Machine Learning Models

Hypothesis 2: Can UK Twitter location-based sentiment per constituency combined with Westminster Parliamentary Constituencies' data increase the accuracy of the Brexit Election Prediction baseline? Decision tree, neural network and Naive Bayes models are created.

3.12.1. Sequential Neural Network Model

Sequence classification is predictive modelling where you have some sequence of inputs over space or time and the task is to predict a category for the sequence. In this instance we are predicting the election result per constituency such that the vote can be to 'Get Brexit Done' or to reject Brexit and remain in the European Union. We use both the Twitter and Westminster Parliamentary Constituencies dataset. The combined dataset consists of 32 features and we need to predict the results by constituency. The following categorical values constituency_name, country_name, region_name, country_name, constituency_type, mp_gender and first party are converted into numerical values as neural networks algorithms expect numerical values to achieve cutting-edge results. We use one-hot encoding with Pandas and pass the data into the get_dummies Panda's function; this converts the text or categorical data into numerical data with which the model expects and perform better. The new columns are column binded into the preexisting dataset. The dataset is now split into training and testing. The training data will have 90% samples and test data will have 10% samples. The neural network is build using Keras and Tensorflow. Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. It was developed with a focus on enabling fast experimentation [32] Tensorflow is an end-to-end, open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets

researchers push the state-of-the-art in ML and developers easily build and deploy ML-powered applications [33]. The dataset has an input layer of 32 and output layer of 1. The Dense layer is used to specify the fully connected layer to the neural network. The arguments of the Dense layer are the output dimension which is 32 in the first case, and an input dimension of 479. The activation function used is relu. The loss function and the optimiser are binary_crossentropy which specifies that we have binary classes. The optimiser is Adam. It is an adaptive learning rate optimisation algorithm that's been designed specifically for training deep neural networks [34] The neural network is trained using 100 epochs. The model's accuracy score is computed by the sklearn metrics accuracy score library.

3.12.2. Decision Tree Model

A decision tree is a supervised machine learning algorithm. Decision tree builds a classification model in the form of a tree structure with a root node (the top node) and underlying branches. It breaks the dataset into smaller and smaller subsets whilst simultaneously creating and developing a tree structure. Once the tree is finalised it will have a number of branches also known as decision nodes and each branch will have an underlying node also commonly known as Leaf nodes which represents the classification or decision. Sklearn's decision tree library is imported to compute the decision tree using the same training and test data ratios used to compute the neural network

3.12.3. Naive Bayes

We use a Naive Bayes supervised machine learning algorithm as referenced in section 3.6.

3.13. Evaluating the Models

The results indicate that the decision tree model performed better in terms of accuracy than the neural network and the Naive Bayes models. All three models were significant in terms of the calculated baseline of 54%.

Table 4. Machine Learning Modelling Results

Model	Country	Accuracy
Decision Tree	UK	. 9787
Neural Network	UK	.9574
Naive Bayes	UK	.9362
Baseline	UK	.545

4. RESULTS

Table 3 illustrates the results at the individual country constituency level using location-based sentiment versus the calculated baseline models. Where X_1, X_2, X_3, and X_4, represent the location-based sentiment from the respective UK constituencies at country level "*Northern Ireland", "Scotland", "Wales"* and *"England"*. We can reject the null hypothesis that Twitter location-based sentiment does not predict the Brexit election with a reasonable degree of accuracy such that $\beta_1, \beta_2, \beta_3, . \eta \neq 0$. Evidence presented herein confirms a relationship relates to location-based sentiment for predicting the Brexit election in the case of the constituencies for X_1, X_2, X_3, where X_1 *("Northern Ireland")* exhibits the highest accuracy, followed by

X_2,(‘*Scotland*) and X_3, (‘*Wales’*) respectively. X_4 (‘*England’*) is not indicative of any significant predictive relationship. Table 4 shows the decision tree, neural network and Naive Bayes machine learning models’ accuracy using location-based sentiment at national level exceeds the baseline accuracy result significantly with a 43%, 41% and 39% improvement from the calculated baseline accuracy.

5. CONCLUSIONS

In this paper, we aimed to establish if location-based Twitter sentiment could predict the Brexit Election results at constituency level for each of the four UK countries and, similarly, at a national level. Our results indicate that Twitter location-based sentiment had the single biggest effect on constituency result predictions in Northern Ireland and Scotland and a marginal effect on Wales-based constituencies whilst there was no significant prediction accuracy to England’s constituencies. We further established that Twitter location-based sentiment improves machine learning prediction accuracy for our decision tree, neural network and Naive Bayes models of up to 43%, 41% and 39% respectively from the calculated baseline accuracy. Owing to the constraints of Twitter location-based sentiment, there were 181 constituencies not represented within the dataset. That may have been as a result of non-user location input or the possibility that there was actually no representation from said constituencies on Twitter. In some cases, users opted to shorten their constituency name rather than use the full constituency name for example ‘Cities of London and Westminster’ is inserted as ‘Westminster’. The issue here is that the constituencies would not match to the constituency naming convention contained in the Office for National Statistics choropleth mapping thus providing reconciliation differences. To combat this, our future research will look to counteract this shortcoming by identifying same and produce a model to compensate for said constraints. In the case of non-user location input, further research is warranted on local language detected within the Twitter text to infer said missing locations as a result of non-user location input. Essentially, Twitter text at the word level can be extracted and word selections (based on identified word distributions per known constituencies) can be aligned to the missing constituencies to yield the location taken from our original Twitter dataset. This would help overcome the constituency location sparsity problem and allow for a probabilistic framework for estimating a Twitter user’s constituency-level location based purely on the content of the user’s tweets in the absence of geospatial cues. This would be heavily reliant on a classifier which identifies words in tweets with a local geographic scope such that the observed geographical distribution of the words in tweets correlates to the geo-locations. [35] created a similar type of model using city-level locations. With continuing chunter intensifying of the breakup of the Union and Scottish Independence, location-based sentiment would prove a very useful tool in strategising against the breakup of what once was a sense of British identity that bound the United Kingdom together and now appears to be disintegrating.

ACKNOWLEDGEMENTS

My thanks to the Technological University Dublin School of Computing for their support on this paper.

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, WWW ’10, pages 851–860, New York, NY, USA, 2010. ACM.

[2] P. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. J. Finance 62 (2007), 1139–1168.

[3] S. Das and M. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Manage. Sci. 53, 9 (2007), 1375–1388
[4] Abbasi, H. Chen, and A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Info. Syst. 26, 3 (2008)
[5] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro. 2013. Sentiment analysis of Facebook statuses using naïve bayes classifier for language learning. In Proceedings International Conference in Information, Intelligence, Systems and Applications (IISA'13). 1–6.
[6] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs", Proceedings of International Conference on Weblogs and Social Media, 2007.
[7] N. Beauchamp. Predicting and interpolating state-level polling using Twitter textual data. In New Directions in Analysing Text as Data Workshop, 2013.
[8] A. Bermingham and A. F. Smeaton. On using Twitter to monitor political sentiment and predict election results. In SAAI '11, 2011
[9] A. Boutet, H. Kim, E. Yoneki, et al. What's in Your Tweets? I Know Who You Supported in the UK 2010, General Election. In ICWSM '12, pages 411{414, 2012.
[10] A. Bruns, J. Burgess, et al. # ausvotes: How Twitter covered the 2010 Australian federal election. Communication, Politics & Culture, 44(2):37{56, 2011.
11] A. Ceron, L. Curini, S. M. Iacus, and G. Porro. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society, 16(2):340{358, 2014
[12] M. Choy, M. Cheong, M. N. Laik, and K. P. Shung. US presidential election 2012 prediction using a census corrected Twitter model. arXiv preprint arXiv:1211.0938, 2012
[13] M. Choy, M. L. Cheong, M. N. Laik, and K. P. Shung. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520, 2011
[14] J. E. Chung and E. Mustafaraj. Can collective sentiment expressed on Twitter predict political elections? In AAAI '11, pages 1770{1771, 2011.
[15] M. Gaurav, A. Srivastava, A. Kumar, and S. Miller. Leveraging candidate popularity on Twitter to predict election outcome. In SNA-KDD Workshop, 2013.
[16] A. Jungherr, P. Jurgens, and H. Schoen. Why the pirate party won the German election of 2009 or the trouble with predictions: A response to ... Social Science Computer Review, 30(2):229{234, 2012.
[17] A. Makazhanov, D. Ra_ei, and M. Waqar. Predicting political preference of Twitter users. Social Network Analysis and Mining, 4(1):1{15, 2014.
[18] Y. Mejova, P. Srinivasan, and B. Boynton. GOP primary season on Twitter: popular political sentiment in social media. In WSDM '13, pages 517{526, 2013
[19] F. Nooralahzadeh, V. Arunachalam, and C. Chiru. 2012 Presidential Elections on Twitter - An Analysis of How the US and French Election were Reflected in Tweets. In CSCS '13, pages 240{246, 2013.
[20] E. T. K. Sang and J. Bos. Predicting the 2011 Dutch Senate Election Results with Twitter. In Workshop on Semantic Analysis in social media, pages 53{60, 2012.
[21] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In ICWSM '10, pages 178 -185, 2010.
[22] General election 2019: Brexit - where do the parties stand? https://www.bbc.com/news/uk-politics-48027580
[23] https://www.theguardian.com/politics/2020/jul/12/revealed-dominic-cummings-firm-paid-vote-leaves-ai-firm-260000
[24] https://dominiccummings.com/tag/machine-learning/
[25] https://www.bbc.com/news/uk-42698981
[26] J. Usher and P. Dondio. 2020. BREXIT Election: Forecasting a Conservative Party Victory through the Pound using ARIMA and Facebook's Prophet. In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020). Association for Computing Machinery New York, NY, USA, 2020. ACM
[27] Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. 2020. Location-based Sentiment Analyses and Visualization of Twitter Election Data. Digit. Gov.: Res. Pract. 1, 2, Article 14 (April 2020), 19 pages
[28] Maima Almatrafi, Suhem Parack, and Bravim Chavan. 2015. Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In Proceedings of

the 9th International Conference on Ubiquitous Information Management and Communication (IMCOM '15).
[29] https://commonslibrary.parliament.uk/research-briefings/cbp-8749/
[30] Tweepy Documentation, http://docs.tweepy.orglen/v3.S.0/
[31] https://geoportal.statistics.gov.uk/datasets/ons::westminster-parliamentary-constituencies-december-2019-boundaries-uk-bfc-v2/explore?location=55.450000%2C-2.000000%2C5.78
[32] https://keras.io/about/
[33] https://www.tensorflow.org/
[34] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. 2014. arXiv:1412.6980v9
[35] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). Association for Computing Machinery, New York, NY, USA, 759–768.

AUTHORS

James Ushe is an Irish native and dwells in Dublin when not spending his time in the countryside of Co Meath. He spends a lot of his free time uncloaking patterns in geopolitical events such as Brexit. Currently he has produced four published papers on Brexit. When not conducting Brexit experiments you will find him enjoying music.

ANTI-VIRUS AUTOBOTS: PREDICTING MORE INFECTIOUS VIRUS VARIANTS FOR PANDEMIC PREVENTION THROUGH DEEP LEARNING

Glenda Tan Hui En[1*], Koay Tze Erhn[1*] and Shen Bingquan[2]

[1]Raffles Institution, Singapore
[2]DSO National Laboratories, Singapore

ABSTRACT

More infectious virus variants can arise from rapid mutations in their proteins, creating new infection waves. These variants can evade one's immune system and infect vaccinated individuals, lowering vaccine efficacy. Hence, to improve vaccine design, this project proposes Optimus PPIme – a deep learning approach to predict future, more infectious variants from an existing virus (exemplified by SARS-CoV-2). The approach comprises an algorithm which acts as a "virus" attacking a host cell. To increase infectivity, the "virus" mutates to bind better to the host's receptor. 2 algorithms were attempted – greedy search and beam search. The strength of this variant-host binding was then assessed by a transformer network we developed, with a high accuracy of 90%. With both components, beam search eventually proposed more infectious variants. Therefore, this approach can potentially enable researchers to develop vaccines that provide protection against future infectious variants before they emerge, pre-empting outbreaks and saving lives.

KEYWORDS

Virus Variants, Transformers, Deep Learning.

1. BACKGROUND AND PURPOSE OF RESEARCH AREA

1.1. The Emergence of More Infectious Virus Variants

Background and Motivation: A small proportion of rapid mutations in viral genomes can significantly increase infectivity, leading to waves of new infections, which in turn create opportunities for more viral mutations to occur. This mechanism has prolonged the devastating pandemic caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As variants of concern emerge, including Beta, Delta and Omicron [1], evidence points towards mutations in the SARS-CoV-2 spike glycoprotein that increase its binding affinity towards the human angiotensin-converting enzyme 2 (hACE2) receptor [2], thus raising transmissibility.

Challenges and Objectives: Currently, vaccines are the solution to reducing virus transmissibility by training one's immune system to develop a response against the virus. However, emergent virus variants can evade one's immune system and even infect vaccinated individuals, lowering vaccine efficacy. For instance, the Covid-19 Delta variant drastically lowered Pfizer BioNTech vaccine's efficacy from 93.7% to 39%, triggering global infection waves in 2021 [3, 4]. Unless vaccines are designed to combat both current and future more infectious variants, our pandemic battle will be a prolonged cat-and-mouse game where we struggle to develop booster shots to

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 17-31, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121102

catch up with ever-mutating variants. Hence, we aim to use deep learning to predict future, more infectious virus variants. This can help researchers to prepare for vaccine production against these variants before they arise, pre-empting outbreaks and saving lives.

Contributions: In this paper, our contributions include:

a. Developing a deep learning approach, Optimus PPIme, that generates mutations from an existing virus protein (exemplified by SARS-CoV-2) to predict future, more infectious variants.
b. Developing a protein-protein interaction (PPI) transformer neural network, that can score the binding affinity between a virus protein and host receptor with a high test accuracy of 90%. Only protein primary sequences are needed as input.

In section 1, we introduce our Optimus PPIme approach and cite related work. Section 2 describes our research question and hypothesis while Section 3 documents the development of Optimus PPIme. Our results are shown in Section 4, while implications of our approach and future work are covered in Sections 5 and 6 respectively.

1.2. Our Deep Learning Approach – Optimus PPIme

Consider the following: A virus attacking a host cell aims to discover mutations that maximize its binding affinity to the host receptor, thereby increasing its infectivity. This is akin to a game character deciding on an optimal strategy to maximize its long-term reward –any action made at one time-step affects subsequent rewards. In the first context, our agent (the virus) can substitute any amino acid (AA) in its sequence of length L (L = 1273 for SARS-CoV-2) with 1 of the 20 AAs, giving rise to an action space of 20L. We exclude insertion and deletion mutations as these are less common in nature [5]. The environment (PPI transformer network) then outputs a PPI score for the proposed mutated protein (new state) and host receptor. The reward received is the change in PPI score (final – initial score of original virus protein S_0 and host receptor).

1.3. Related Work

PPIs are usually determined via tedious and costly high-throughput experimental methods, such as isothermal titration calorimetry and nuclear-magnetic resonance [6]. This necessitates the use of computational PPI models. However, many require 3D protein structures, which are harder to obtain than primary sequences. Even methods such as MuPIPR [7] –that only require primary sequences as inputs– fail to generalize to novel proteins. To address these, we propose a PPI transformer network that uses only primary sequences and generalizes to novel proteins.

Primary sequences can be represented as strings of letters denoting AAs. Protein motifs and domains (conserved functional patterns) [8] are also analogous to words and phrases. Furthermore, information is contained in primary sequences and natural sentences [9]. Such similarities make natural language processing (NLP) ideal for extracting protein features.

NLP tasks have seen state-of-the-art performance with the rise of transformers. These encoder-decoder networks adopt the self-attention mechanism, which relates different positions of a sequence to compute a rich representation of features [10]:

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

Where Q, K, and V are the query, key and value matrices while d_K is the dimension of K. Given that interactions between different AAs in a protein sequence give rise to the protein's structure and properties, transformer encoders are suitable protein feature extractor networks.

Lastly, we define a similarity measure between S_0 and a generated variant protein with sequence alignment scores derived from Block Substitution Matrix 62 (BLOSUM62) [11]. The BLOSUM distance between 2 sequences, $S = s_1 \ldots s_L$ and $S' = s'_1 \ldots s'_L$, is given by [12]:

$$D(S, S') = \sum_{i=1}^{L} (B_{s_i s_i} - B_{s_i s'_i})$$

2. HYPOTHESIS OF RESEARCH

Our research question is: Is a PPI predictor environment sufficient for virus agents to predict future, more infectious variants?

We hypothesize that given an environment that scores the binding affinity for a virus-host PPI with high accuracy, agents can predict future, more infectious variants.

3. METHODOLOGY

3.1. Dataset Collection

We train our PPI transformer on 31,869 experimentally-validated positive virus-human (V-H) interactions from BioGRID [13] and VirHostNet [14], excluding PPIs with SARS-CoV-2. We reference 34,836 negative V-H interactions from [15], giving rise to 66,705 training examples.

For the 50 novel virus protein test PPIs, we fix hACE2 as the second input protein. Positive PPIs include the original SARS-CoV-2 spike and its 23 variants listed in CoVariants [16], while 26 negative PPIs were sampled from unseen non-coronavirus viral proteins in [14]. All sequences were extracted from UniProt [17] and tokenized with KerasTextVectorization. We added start, end of sentence (SOS, EOS) and class (CLS) tokens, padding up to 1,300 tokens.

3.2. The Environment – PPI Transformer Network

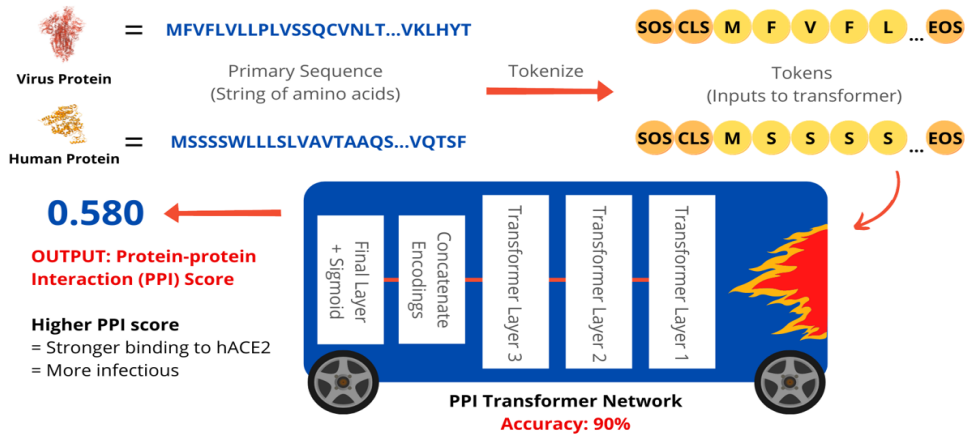


Figure 1. Overview of the PPI transformer network.

The inputs to our PPI transformer are the primary sequences of a virus and human receptor protein (represented by strings of amino acids). These strings of amino acids are then tokenized before they are fed to the transformer network. The transformer network then outputs a PPI score measuring how well the virus protein binds to the human receptor protein. The higher the PPI score, the stronger the virus-host binding and the more infectious the variant.

3.2.1. Experiment 1: Masked Language Modeling (MLM) Pre-training vs No MLM Pre-training

Initially, our transformer predicted the same value of 0.99 for all test data (see Appendix), indicating its inability to learn. This may be due to unbalanced dependencies and amplified output perturbations caused by layer normalization and aggravated by the Adam optimizer. Since learning semantics of input sequences improves downstream task performance (transfer learning) [18, 19], we pre-train our transformer on Masked Language Modeling (MLM) and fine-tune on PPI. In each sequence, we mask a random 15% of tokens with a mask token 90% of the time, or a random token otherwise. MLM attempts to reconstruct the original sequence.

3.2.2. Experiment 2: Sharpness-Aware Minimization (SAM) vs No SAM

Today's neural networks optimize on training loss, but this is inadequate in generalizing to unseen data (novel proteins for our PPI task) due to complex and non-convex training loss landscapes. SAM overcomes this by simultaneously minimizing loss value and loss sharpness. This algorithm enables a model to learn parameters whose entire neighbourhoods have uniformly low training loss values (low losses and curvature) rather than those with solely low loss values (low losses but complex curvature) [20, 21]. Hence, we determine the effects of SAM on generalizability. Using the model pre-trained with MLM in Experiment 1 as our baseline, 3 new models were trained with the addition of SAM on MLM pre-training only, PPI training only, and for both tasks (see Appendix for our implementation).

3.2.3. Experiment 3: Data Augmentation vs No Data Augmentation

Given that image augmentation improves image classifiers' robustness [22], we aim to determine if augmenting protein sequences could similarly boost our PPI test accuracies. The models were trained with 3 different augmentation techniques [23] during MLM pre-training: substituting a random AA with alanine – a generally unreactive AA (Alanine Sub) – or an AA of the highest BLOSUM62 similarity score –the most likely AA that can replace the original AA (Dict Sub), and reversing the whole protein sequence (Reverse). 25% of proteins were augmented and 20% of amino acid positions were replaced for substitution augmentations. Augmentation was not applied to PPI training as it distorts the proteins' structure and properties.

Experimental Setup: All PPI models adopted the same architecture (see Appendix) and were trained for the same number of epochs (50 for MLM pre-training, 15 for PPI training). They were evaluated on the same novel virus test set, with test accuracy and F1 scores as metrics.

3.3. The Agent – Proposing Future More Infectious Virus Variants

Initially, we attempted a Deep Q-Learning Network (DQN) agent (see Appendix). However, it required heavy computation and converged slowly due to our substantial search space of 20L (25,460 actions for SARS-CoV-2). Thus, we explore 2 more efficient algorithms, Greedy Search and Beam Search, to search for the variant with the highest PPI infectivity score. These algorithms rely on the greedy concept: making the optimal choice(s) at each time-step.

Greedy Search Algorithm
Inputs: current sequence S = original spike sequence S_0, actions A = 25460
while BLOSUM distance <= 40 **do**:
Perform mutations on S within A to create a batch of variants with shape (25460, 1)
Tokenize, compute PPI scores and store in array P
Select the sequence with the highest PPI score, S_{best} = argmax(P)
Update A = A – 20 actions for mutated position & S = S_{best} if BLOSUM distance with $S_0 \leq 40$
return S

Algorithm 1. Greedy Search algorithm.

Beam Search Algorithm
Inputs: S = {S_0}, A = 25460, beamwidth = 10, no. sequences with BLOSUM distance > 40, (η) = 0
While η < *beamwidth* **do**:
For *sequence s in S* **do**:
a. Perform mutations on s to create a batch of variants with shape (25460, 1)
b. Tokenize, compute PPI scores and store in array P
Select 10 best sequences with the highest PPI scores, S_{best} = $argmax_{10}$(P)
Update S = {for s in S_{best} if BLOSUM distance with $S_0 \leq 40$} and η = 10 - length(S)
return S

Algorithm 2. Beam Search algorithm.

We used Phyre2 [24] to predict the generated variant sequences' 3D structures. Then, possible binding modes of the variants and hACE2 were proposed by the HDOCK server, a hybrid of template-based modeling and free docking which achieved high performance in the Critical Assessment of Prediction of Interaction [25]. We used docking scores as a further metric to validate our proposed variants, where negative scores indicate spontaneous PPIs.

4. RESULTS AND DISCUSSION

4.1. PPI Transformer Network Results

Table 1. Performance of the PPI models across all 3 transformer experiments.

No.	MLM	SAM	Data Augmentation	Test Accuracy / %	Loss	F1 Score
1	✓	x	x	44.0	1.080	0.417
2	x	x	x	50.0	4.288	0.658
3	✓	MLM	x	52.0	3.690	0.667
4	✓	MLM + PPI	x	72.0	0.707	0.774
5	✓	PPI	x	74.0	0.642	0.787
6	✓	PPI	Reverse	74.0	0.786	0.787
7	✓	PPI	Dict Sub	78.0	0.910	0.814
8	✓	**PPI**	**Alanine Sub**	**90.0**	**0.438**	**0.906**

Experiment 1: Although Model 2 (without MLM pre-training) achieved a higher test accuracy and F1 score than Model 1 (with MLM pre-training), it outputted the same PPI value for all test data (0.99), indicating its inability to learn. In contrast, MLM pre-training helped Model 1 to learn relevant protein features and it outputted different PPI scores for test data. Model 1 was thus used as the baseline for subsequent models to improve upon.

Experiment 2: MLM pre-training with SAM (Models 3 and 4) causes transformer layers that are nearer to the output to learn parameters which improve its predictions of the original AAs being masked (the MLM task). However, these MLM-specific parameters may not be best suited for our PPI task, which uses natural proteins without masking. Thus, SAM on PPI (Model 5) is essential for optimizing the parameters in our PPI task.

Experiment 3: Alanine Sub (Model 8) improved PPI test accuracy the most as it does not drastically alter the protein syntax as compared to Reverse and Dict Sub. This is likely due to alanine's nonpolar methyl side-chain (see Appendix), giving rise to alanine's unreactivity [26]. Model 8 was therefore chosen as the optimal PPI transformer network environment.

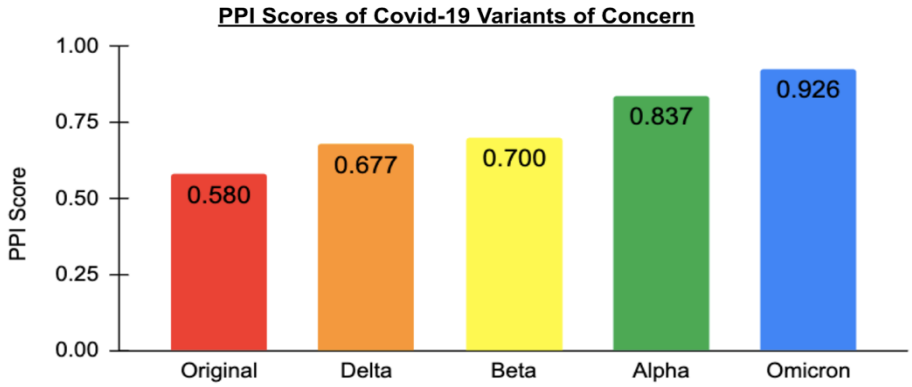


Figure 2. Graph showing PPI scores for different Covid-19 variants of concern.

All variants of concern achieved higher PPI scores than the original spike protein (0.580). The Delta variant (0.677) achieved a higher PPI score than the original, although lower than Alpha (0.837) and Omicron (0.926). These results reflect that our PPI transformer network can make real-world predictions which corroborate well with current Covid-19 research data [27, 28].

4.2. Virus Agent Algorithm Results

Figure 3. 3D structures of original, greedy and beam search spike proteins bound to hACE2.
Legend: blue: hACE2 receptor, orange: spike glycoprotein.

Table 2. hACE2 binding metrics for the spike glycoprotein variants generated by greedy and beam search.

Algorithm	PPI Score	Docking Score	RMSD with S_0 / Å
Greedy Search	0.99969	–214.76	0.870
Beam Search	0.99973	–218.90	1.095

Based on Table 2, the spike variant proposed by Beam Search attained higher PPI and more negative docking scores than that for Greedy Search, reflecting its greater hACE2 binding affinity. Unlike greedy search which always exploits the best immediate action, beam search strikes a balance between exploitation and exploration (trying a new action). Since AAs in a sequence interact with one another, by considering 10 sequences at each time-step, beam search is likelier to find mutations that may not maximize short-term rewards but will optimize long-term rewards due to future AA interactions. From Figure 3, the proposed variants' structures also have little deviation from the original protein, with RMSDs close to those of current variants (see Appendix). Therefore, an agent armed with a PPI transformer network can propose future more infectious variants, proving our hypothesis. The variants can then be validated experimentally.

5. IMPLICATIONS AND CONCLUSION

We discovered that given an accurate PPI transformer network that measures the infectivity of a proposed variant, our Optimus PPIme approach can effectively predict possible more infectious variants. This narrows the scope of mutated virus proteins for docking and wet-lab testing to validate the variants' infectivity and feasibility.

With only knowledge of the virus and receptor protein sequences, our Optimus PPIme approach can be applied to other dangerous viruses to expedite vaccine development before infectious variants become widespread.

6. LIMITATIONS AND FUTURE WORK

Currently, our Optimus PPIme approach does not consider insertion or deletion mutations in the virus protein, which are also likely to occur in nature. Besides that, the ability of a virus to evade vaccine antibodies is another metric for infectivity, which we did not consider in our Optimus PPIme approach.

Hence, future work can be done to generate insertion or deletion mutations in the virus variant, and to use the evasion of antibodies as a further metric for infectivity.

7. ACKNOWLEDGEMENTS

We would like to thank our mentor, Dr Shen, for his invaluable guidance and advice throughout this research project!

REFERENCES

[1] World Health Organization. 2021. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[2] Barton et al. 2021. Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. Elife. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8480977/
[3] Lovelace, B. 2021. Israel says Pfizer Covid vaccine is just 39% effective as delta spreads, but still prevents severe illness. https://www.cnbc.com/2021/07/23/delta-variant-pfizer-covid-vaccine-39percent-effective-in-israel-prevents-severe-illness.html
[4] Towey, R. 2021. WHO says delta variant accounts for 99% of Covid cases around the world. https://www.cnbc.com/2021/11/16/who-says-delta-variant-accounts-for-99percent-of-covid-cases-around-the-world.html
[5] Sanjuán et al. 2010. Viral Mutation Rates. J Virol: 9733-9748. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2937809/
[6] Zhou, M., Li, Q. and Wang, R. 2016. Current Experimental Methods for Characterizing Protein-Protein Interactions. Wiley Public Health Emergency Collection: 738-756. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7162211/
[7] Zhou et al. 2020. Mutation effect estimation on protein-protein interactions using deep contextualized representation learning. NAR Genomics and Bioinformatics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7059401/
[8] Xiong, J. 2006. Protein Motifs and Domain Prediction. Essential Bioinformatics: 85-94. https://www.cambridge.org/core/books/abs/essential-bioinformatics/protein-motifs-and-domain-prediction/E17046CB1CD04184A828D8BAC2D222AF
[9] Ofer, D., Brandes, N. and Linial, M. 2021. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J: 1750-1758. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050421/
[10] Vaswani et al. 2017. Attention Is All You Need. https://arxiv.org/pdf/1706.03762.pdf
[11] Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA: 10915-10919. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50453/
[12] Jha et al. 2021. Protein Folding Neural Networks Are Not Robust. https://arxiv.org/pdf/2109.04460.pdf
[13] Stark et al. 2006. BioGRID: a general repository for interaction datasets. Nucleic acids research, 34(Database issue), D535–D539. https://pubmed.ncbi.nlm.nih.gov/16381927/
[14] Guirimand et al. 2015. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic acids research, 43(Database issue), D583–D587. https://pubmed.ncbi.nlm.nih.gov/25392406/
[15] Kshirsagar et al. 2021. Protein sequence models for prediction and comparative analysis of the SARS−CoV−2 −human interactome: Pacific Symposium on Biocomputing, 26, 154−165. https://pubmed.ncbi.nlm.nih.gov/33691013/
[16] Hodcroft, E. 2020. CoVariants. https://covariants.org/
[17] UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic acids research, 49(D1), D480–D489. https://pubmed.ncbi.nlm.nih.gov/33237286/
[18] Lanchantin et al. 2020. Transfer Learning for Predicting Virus-Host Protein Interactions for Novel Virus Sequences. https://doi.org/10.1101/2020.12.14.422772
[19] Devlin et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805
[20] Foret et al. 2021. Sharpness-Aware Minimization for Efficiently Improving Generalization. https://arxiv.org/abs/2010.01412
[21] Chen et al. 2021. When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations. https://arxiv.org/abs/2106.01548
[22] Krizhevsky et al. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
[23] Shen et al. 2021. Improving Generalizability of Protein Sequence Models with Data Augmentations. https://www.biorxiv.org/content/biorxiv/early/2021/02/18/2021.02.18.431877.full.pdf
[24] Kelley et al. 2015. The Phyre2 web portal for protein modelling, prediction and analysis. Nature Protocols: 845-858. https://www.nature.com/articles/nprot.2015.053

[25] Yan et al. 2017. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Research, Volume 45: 365-373. https://academic.oup.com/nar/article/45/W1/W365/3829194

[26] Lefèvre, F., Rémy, M. H. and Masson, J. M. 1997. Alanine-stretch scanning mutagenesis: a simple and efficient method to probe protein structure and function. Nucleic Acids Research, Volume 25: 447-448. https://academic.oup.com/nar/article/25/2/447/1204328

[27] Mannar et al. 2021. SARS-CoV-2 Omicron Variant: ACE2 Binding, Cryo-EM Structure of Spike ProteinACE2 Complex and Antibody Evasion. https://www.biorxiv.org/content/10.1101/2021.12.19.473380v1.full.pdf

[28] Kim et al. 2021. Differential Interactions Between Human ACE2 and Spike RBD of SARS-CoV-2 Variants of Concern. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8328061/

AUTHORS

Bingquan Shen is currently a Senior Researcher in DSO National Laboratories. He received his PhD from the National University of Singapore in Mechanical Engineering (Control and Mechatronics) in 2014. His current research interests include limited label learning, adversarial machine learning, and deep learning applications.

Glenda Tan is currently a student at Raffles Institution. Her current research interests include computer vision, transformers and generative adversarial networks.

Koay Tze Erhn is currently a student at Raffles Institution. Her research interests include natural language processing, transformers and computational biology.

8. APPENDIX

8.1. PPI Dataset Breakdown

Table 3. Breakdown of train and test datasets for the PPI transformer network.

Dataset	Train	Test
BioGRID	11,612 (+)	N/A
VirHostNet	20,257 (+)	26 (–)
Kshirsagar et al.	34,836 (–)	N/A
CoVariants	N/A	24 (+)
Total	66,705	50

In order for our PPI transformer to generalize to unseen virus proteins and make unbiased predictions of future virus variants' infectivity, the training set does not contain V-H interactions involving SARS-CoV-2. Instead, only the PPI transformer's test set contains V-H interactions involving Covid-19 variants and hACE2.

8.2. Our SAM Implementation

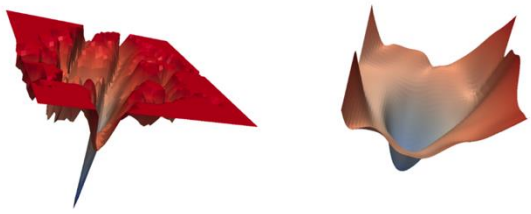


Figure 4. Loss landscapes without SAM training (left) and with SAM training (right)[20].

PPI SAM Algorithm
Inputs: neighbourhood (ρ) = 0.05, batch size = 8, learning rate (α) = 0.001, PPI model's weights w_0, timestep t = 0
while*not converged***do**
1. Sample batch B = {(x_1, y_1), …(x_8, y_8)}, where x = [V_{token}, H_{token}]
2. Backpropagation 1: Compute training loss and gradient g
3. Scale gradient by factor ρ / (|| g || + 1e-12) and update weights
4. Backpropagation 2: Compute training loss and final gradient G
5. Update weights: $w_{t+1} = w_t - \alpha * G$

t = t + 1
return w_t

Algorithm 3. PPI SAM algorithm.

From Figure 4, a loss landscape without SAM training has a sharp global minimum (left) and is difficult to converge, whereas SAM training results in a loss landscape with a wide minimum (right) that is easier to converge [20]. From Algorithm 3, our SAM implementation involves 2 backpropagation steps with a scaling step between them.

8.3. Visualization of Protein Augmentation Techniques

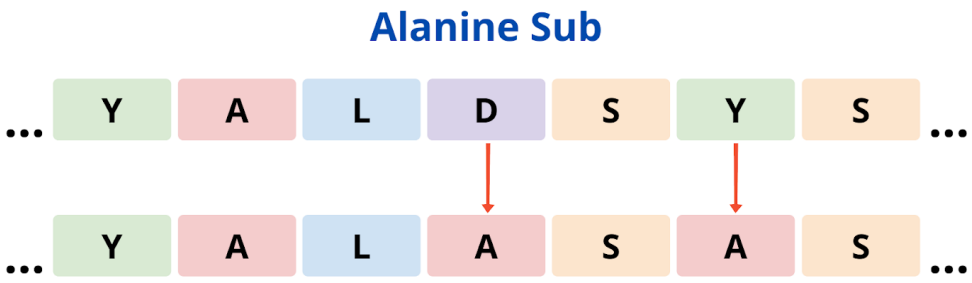


Figure 5a. Visualization of Alanine Sub augmentation.

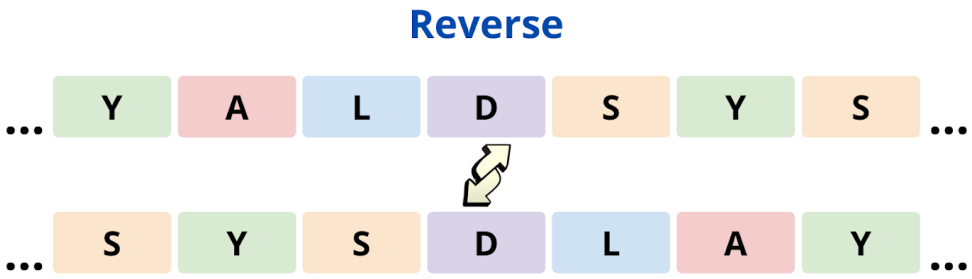


Figure 5b. Visualization of Reverse augmentation.

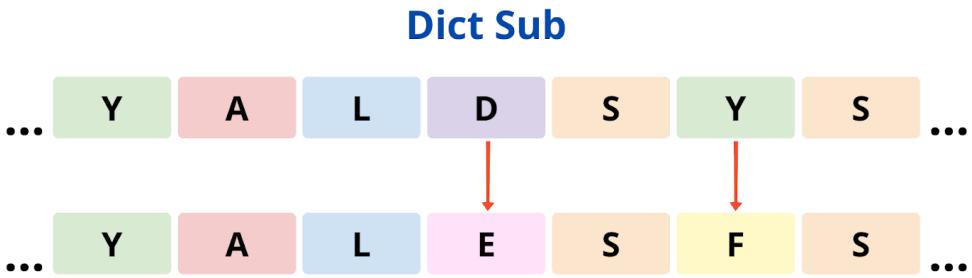


Figure 5c. Visualization of Dict Sub augmentation.

The augmentation techniques were implemented by performing string manipulations on the protein sequence and subsequently feeding the augmented proteins to the transformer network during training. Combinations of augmentations (e.g. all 3, Alanine Sub & Reverse, Alanine Sub & Dictionary Sub, Reverse & Dictionary Sub) were also attempted but they produced models with lower accuracies.

8.4. Skeletal Structure of Amino Acid Alanine

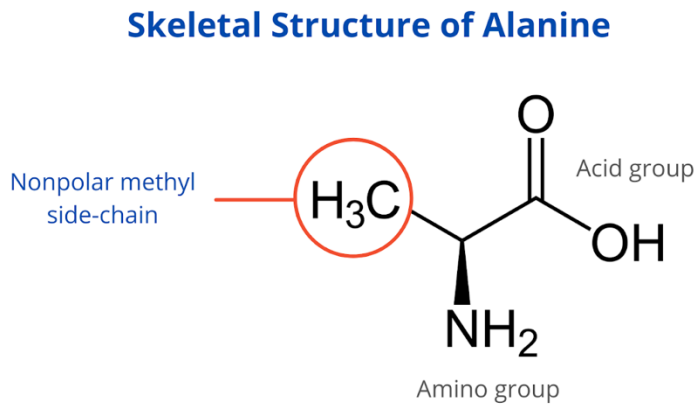


Figure 6. Skeletal structure of alanine with its nonpolar methyl side-chain.

Alanine is a generally unreactive amino acid, attributed to its nonpolar methyl side-chain. This makes Alanine Substitution a suitable candidate for data augmentation.

8.5. Model Architectures

8.5.1. Architecture of the PPI Transformer Network

The PPI transformer network was implemented using the tf.keras framework. The network consists of an encoder to extract features of the virus and receptor proteins, a Concatenate layer to concatenate the encodings of the 2 proteins and finally a 1-unit Dense layer with sigmoid activation which outputs the PPI score. No decoder is used as only an encoder is required for extracting the protein features.

Encoder: The encoder is adapted from the Bidirectional Encoder Representations from Transformers (BERT) encoder and it comprises 3 identical layers. Each layer is made up of a MultiHeadAttention component and a fully-connected feed-forward network. A residual connection is applied around the 2 components, followed by LayerNormalization.

Hyperparameters: After experimenting with different values for the hyperparameters, we arrive at this combination of hyperparameters which produced the best-accuracy model of 90%.

Table 4. Hyperparameters for the PPI transformer encoder.

Hyperparameter	**Value**
Number of transformer layers in encoder	3
Number of heads for MultiHeadAttention	8
Number of embedding dimensions	128
Dropout rate	0.1
Adam Optimizer learning rate	0.001
Batch size	8
Maximum length of proteins	1,300
MLM pre-training epochs	50
PPI training epochs	15

8.5.2. Architecture of the DQN agent

The inefficient DQN agent adopts the same encoder architecture as the PPI transformer network, and the encoder is connected to a 25,460-unit Dense layer with softmax activation. 25,460 represents the number of possible mutations the SARS-CoV-2 spike protein can perform at each time-step (1,273 AA positions x 20 amino acids). The DQN agent was trained for 1000 episodes with a learning rate of 0.7, a batch size of 128 and a minimum replay size of 500.

8.6. Failed Results

```
[[0.8290639 ]
 [0.48207778]
 [0.82868207]
 [0.8670187 ]
 [0.831499  ]
 [0.7862657 ]
 [0.40302208]
 [0.7863891 ]
 [0.43530652]
 [0.48911947]
```

```
[[0.9999083 ]
 [0.999647  ]
 [0.99973285]
 [0.9996846 ]
 [0.99968445]
 [0.99991024]
 [0.99972934]
 [0.99992645]
 [0.99972963]
 [0.999717  ]
```

Figure 7. (from left to right) PPI scores outputted by Model 1 (with MLM pre-training) and Model 2 (without MLM pre-training).

```
0 , FR: -0.0024803877 , FP: 0.5405822
Updating target network weights
1 , FR: -0.0040413737 , FP: 0.5390212
Updating target network weights
2 , FR: 0.0059351325 , FP: 0.5489977
Updating target network weights
3 , FR: 0.00032663345 , FP: 0.5433892
Updating target network weights
4 , FR: 0.0019819736 , FP: 0.54504454
Updating target network weights
```

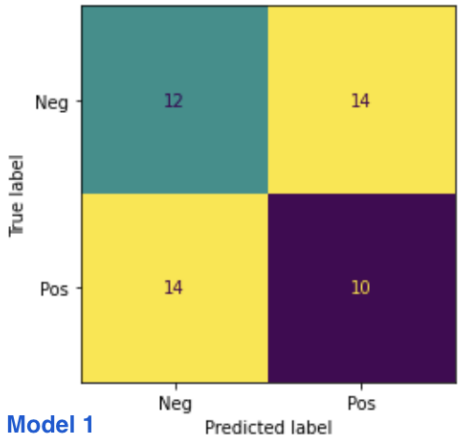
```
Updating target network weights
16/16 [==============================]
16/16 [==============================]
16/16 [==============================]
819 , FR: 0.0035093427 , FP: 0.5465719
Updating target network weights
16/16 [==============================]
16/16 [==============================]
16/16 [==============================]
820 , FR: 0.0015875101 , FP: 0.5446501
Updating target network weights
16/16 [==============================]
16/16 [==============================]
16/16 [==============================]
821 , FR: -0.0005326867 , FP: 0.5425299
Updating target network weights
16/16 [==============================]
16/16 [==============================]
16/16 [==============================]
822 , FR: -0.0005326867 , FP: 0.5425299
Updating target network weights
16/16 [==============================]
16/16 [==============================]
16/16 [==============================]
823 , FR: -0.0011529922 , FP: 0.5419096
Updating target network weights
```

Figure 8. (from left to right) First and last 5 episodes of DQN training.
Legend: FR: final reward (final - initial PPI score of S_0 and hACE2) at the end of each episode, FP: final PPI score at the end of each episode.

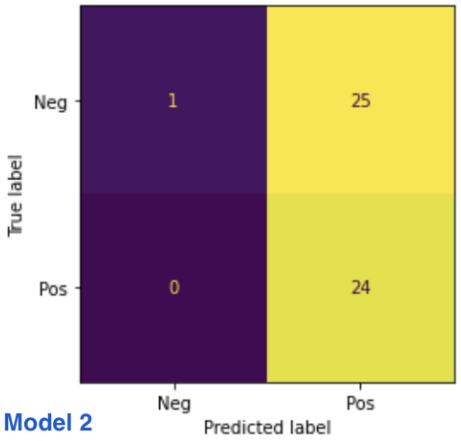
From Figure 7, Model 2 outputs the same value of 0.99, indicating its inability to learn. In contrast, MLM pre-training helps Model 1 to output different PPI values, showing that the model has successfully learned the features of proteins. From Figure 8, DQN's final reward does not

increase and oscillates between being positive and negative, reflecting the agent's inability to converge.

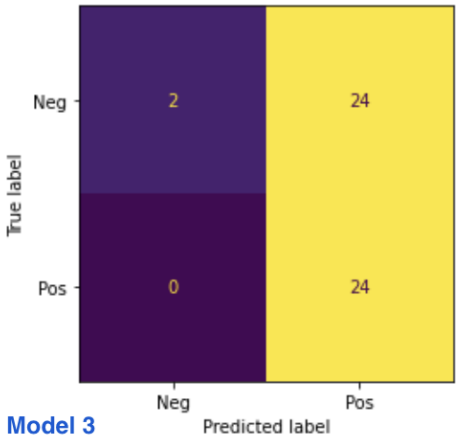
8.7. Confusion Matrices of PPI Transformer Networks



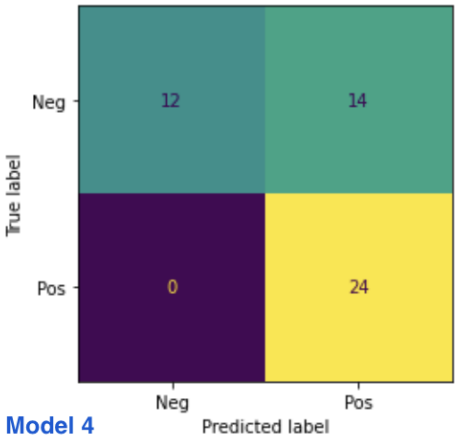
Model 1



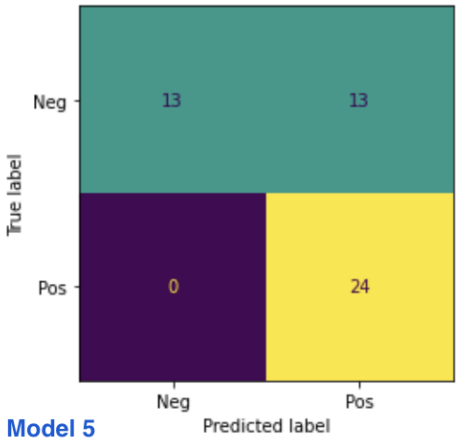
Model 2



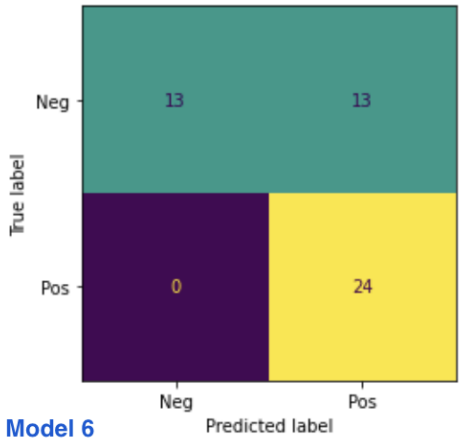
Model 3



Model 4



Model 5



Model 6

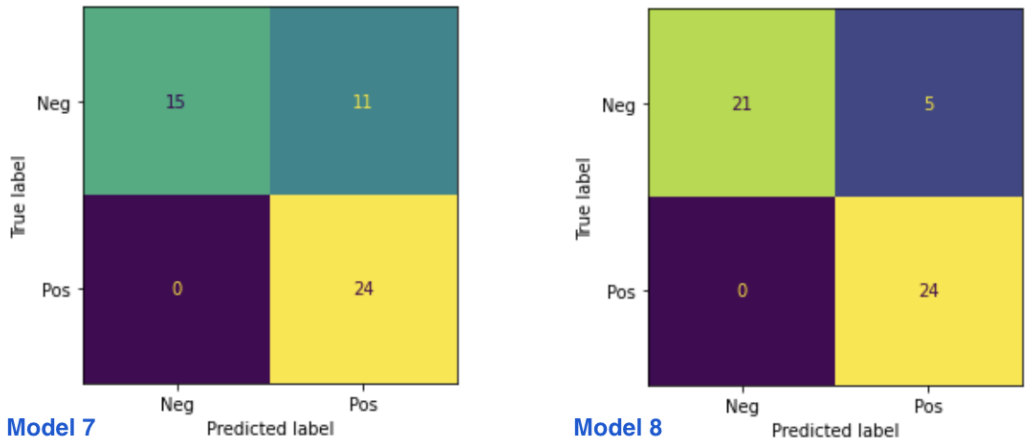


Figure 9. Confusion matrices of PPI transformer networks 1 to 8.

Based on the confusion matrices, Model 8 achieved the highest test accuracy. It correctly predicted all 24 positive interactions and 21 out of 26 negative interactions in our test set.

8.8. Comparison of Our Proposed Variants with Covid-19 Variants of Concern

Table 5. Number of spike mutations and RMSDs of the different variants.

Variant	Number of Spike Mutations	RMSD with S_0 / Å
Alpha	4	0.897
Beta	8	0.901
Delta	8	1.091
Omicron	17	NA
Greedy (Proposed)	6	0.870
Beam (Proposed)	6	1.095

The RMSD values were derived by matchmaking the variants' Protein Data Bank (PDB) files with the original spike protein in ChimeraX. At the time of submission, Omicron's PDB file and hence its RMSD were not yet available due to the variant's recency.

STUDY ON EMOTIONAL STATE CHANGE BASED ON DYNAMIC EXPRESSION SIMILARITY

Yan Zhang, Xiangyang Feng and Ming Zhu

College of Computer Science and Technology,
Donghua University, Shanghai, China

ABSTRACT

Facial expressions can express different emotions. Similar facial expressions usually correspond to the same emotions, and the changing process of emotional states is reflected in the dynamic changes of facial expressions. However, existing studies mainly focus on instantaneous emotional states, which cannot reflect the intensity of emotions. This paper proposes a method to study the process of emotional change based on dynamic expression similarity, which can assess not only the change of emotional state but also the change of emotional intensity. First, the features of dynamic expressions are extracted based on the VGG16 network model. Then, the cosine similarity of the expression features is calculated to match the corresponding emotions. At the same time, the expression intensity of each frame is calculated to evaluate the change in emotional intensity. The experimental results show that the similarity calculated in this paper is increased by 9.7% on average, which can be used for the study of emotional states.

KEYWORDS

Dynamic Expression Similarity, The Emotional State Change, Emotional Intensity.

1. INTRODUCTION

The external manifestations of emotions are often called expressions. It is a quantified form of movement of body parts in response to emotional changes. Because the facial expression is a pattern made up of all facial muscle changes, it can express a wide range of emotions and is the primary indicator for identifying human emotions. Emotion research based on facial expressions has been an active research topic in the past ten years. Ekman and Friesen [1] defined six basic expression categories: happy, angry, surprised, fearful, disgusted, and sad, corresponding to six basic human emotions. Facial expressions are recognized and classified according to the expression categories defined by Ekman. On the one hand, low-latitude facial expression features are extracted, such as Gabor wavelet transform, LBP local binary mode [2], HOG direction gradient histogram [3], etc. On the other hand, people classify expressions into pre-defined categories by seeking classifiers with better stability, such as BP neural networks and clustering [4].

However, most current facial expression studies aim to distinguish between different expressions and use static expressions to study instantaneous emotional states. However, in many scenarios, such as performance imitation, lie detection, driver fatigue detection, etc., people often do not pay attention to the category of expressions but pay more attention to the emotional information transmitted in the process of expression changes, so as to judge the current emotional state. Since similar expressions often express the same emotional information, it can be judged whether they

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 33-45, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121103

are expressing the same emotion by calculating the similarity degree of the expressions through the similar information between the expressions. At the same time, the dynamic process of facial expressions contains more abundant and accurate emotional information, and the similarity of dynamic expressions is closer to the state of human emotional changes. On the other hand, in addition to the type of expression, the degree of expression is also an indicator of human emotions. Expressions of the same category, such as smiling and laughing, displeasure and anger, have different levels of expression. The degree of expression corresponds to the intensity of emotion, and the change in emotion reflects the dynamic change process of emotional state. However, the current research on expression similarity is limited to the similarity between static expressions, while static expressions are only limited to instantaneous emotions and cannot reflect changes in emotional states. As a result, the ability to describe the changing state of emotions is poor, and the expression intensity has not been used in the evaluation of emotional intensity.

In response to the above problems, this paper proposes a dynamic expression similarity evaluation method to study the process of emotional change. First, the VGG16 network is used to extract the expression features, and then the KPCA principal component analysis method is used to perform dimensionality reduction analysis on the high-dimensional expression features, and finally the cosine similarity is used to reduce the dimension. The matching method compares the expression features of each frame, calculates the similarity, and calculates the intensity estimate for the expression of each frame, and combines feature similarity and intensity similarity to study the change of emotion.

1.1. Our Contributions

1. In this paper, we study the changes in emotional states from dynamic expressions and propose the use of the VGG16 network to extract expression features, which solves the problem of inaccurate feature extraction for face feature point localization. Based on this, the KPCA-based feature dimensionality reduction method and cosine similarity matching method are used to improve the similarity results.
2. We proposed to use expression intensity for the study of emotional states and to study the changes in emotional states by fusing intensity curves based on the results of dynamic expression similarity.

2. RELATED WORK

Facial expressions can intuitively reflect the emotional state of human beings and are the most direct way to express emotions. From the state of emotions, it can be divided into instantaneous emotions and dynamic emotional changes. Expression is the carrier of emotions. Instantaneous emotions are mainly expressed through static expressions [5], and dynamic changes in emotions can be reflected through dynamic expression changes [6]. Nasuha et al [7]. proposed a CNN emotion classification model that reduces parameters by separating convolutional layers and studied instantaneous emotions through the classification accuracy of seven basic emotions but did not consider the fusion of expression intensity and emotion recognition. Sui et al [8]. proposed the PSO-ELM model to recognize dynamic expressions; Fisher [9] studied emotion recognition through dynamic emotional expressions, which measure the understanding of others' emotions in everyday life but did not consider different levels of emotion.

Many scholars study expressions from the perspective of expression similarity. Vemulapalli [10] proposed a parsimonious space that is closer to human visual preferences to describe facial expressions and created a large-scale facial expression comparison dataset to obtain this simple

space. This article argues that if the other two expressions are visually like the third, then the distance between these two expressions in the parsimony space will be much smaller than the distance between them and the third expression. However, only the similarity between static expressions is described. For example, if a static smile is displayed, whether it is a smile, a big laugh, a wry smile, a fake smile, or a sneer, it will be judged as similar. That is, a smile and a big laugh will also be judged to be similar in emotion, but in fact, the intensity of the two emotions is different. Unfortunately, little research on emotions takes emotional intensity into account. Schroff et al. [11] proposed a new method, FaceNet, which can directly learn the mapping from images to points in Euclidean space. The distance between the points in the Euclidean space of the features corresponding to the two images directly corresponds to whether the two images are similar. Similarly, Schroff et al. [11] only studied the similarity of static emotions, that is, transient emotions. And whether it is from the perspective of emotion recognition or similarity, it ignores the dynamics of facial expressions and the importance of expression intensity. Expression intensity can reflect the intensity of emotions and deeply understand the psychology of characters. This paper proposes an evaluation method based on dynamic expression similarity and emotional intensity to study the change process of emotion.

3. DYNAMIC FACIAL EXPRESSION SIMILARITY MATCHING

The same emotional state often has similar facial expressions, and the dynamic changes in facial expressions directly reflect a person's emotional changes. The dynamic face similarity matching algorithm proposed in this paper mainly includes expression feature extraction, dimensionality reduction of high-dimensional expression features based on KPCA, emotional intensity estimation, and expression similarity matching based on cosine similarity and determine emotional states using similarity results. Its experimental flow chart is shown in Figure 1.

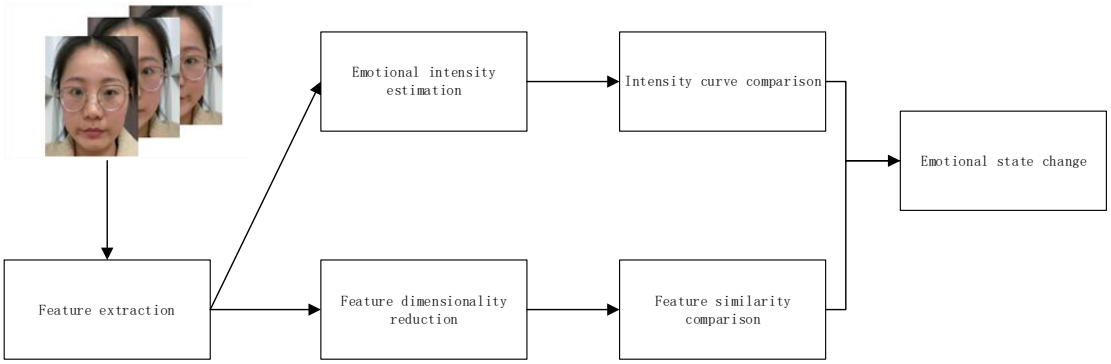


Figure 1. Emotional state change matching flow chart.

3.1. Feature Extraction Network

Facial expression feature extraction is mainly to locate and extract facial organ features, texture regions, and predefined feature points. General facial expression feature extraction methods are divided into two categories: traditional feature extraction and feature extraction based on deep learning. Traditional feature extraction can be based on geometric features, gray features, motion features, frequency features, etc., using the LBP local binary mode [2] and the HOG direction gradient histogram feature [3] [21]. The method based on deep learning [12] mainly uses convolutional neural networks to extract high-dimensional features of facial expressions in combination with the current popular network, which solves problems such as ignoring local feature information, feature loss, and inaccurate feature extraction in traditional manual extraction of facial features.

As one of the most representative networks in deep learning, the convolutional neural network has achieved great success in the field of image processing, relying on several convolutional layers for feature extraction, and many successful tasks such as image classification and recognition are based on CNN [13]. Compared with traditional image processing algorithms, CNN has the advantage of being able to directly input the original image to extract image features, avoiding the complex image preprocessing process. When using neural networks to extract expression features, the output of the last layer of the network is not directly used as expression features, because each value in the feature vector output by the last layer represents the confidence that the expression may be a certain expression. If the expression does not appear in the classified expression, its value will be suppressed to a very low level, which is not conducive to the subsequent similarity calculation.

In this paper, VGG16 is selected as the feature extraction network of facial expressions. VGG16 won the runner-up to ImageNet in 2014, and its basic mechanism follows the traditional CNN architecture. VGG16's network has a total of 16 layers, including 13 convolutional layers and 3 full connection layers [14]. In this experiment, the output value of the last fully connected layer of VGG16 was selected as the feature vector of the expression. According to the structure diagram of VGG16, the dimension of the feature vector obtained from this layer is 4096.

3.2. KPCA Feature Dimensionality Reduction

Dimensionality reduction of facial expression features is to select the most representative features from the original facial expression features and reduce the time complexity of the algorithm while preserving the facial expression features. Existing dimensionality reduction methods are mainly divided into linear and nonlinear dimensionality reduction [15]. Nonlinear dimension reduction [16] mainly includes local linear embedding (LLE) that preserves local features and isometric feature mapping (Isomap) that preserves global features. Linear dimension reduction mainly includes Principal Component Analysis (PCA) [12], Linear Discriminant Analysis (LDA), etc. PCA dimension reduction mainly maps high-dimensional data features to new low-dimensional feature spaces and transforms them into several new comprehensive features. The new low-dimensional comprehensive features are linear combinations of the original high-dimensional features. However, PCA dimension reduction only uses the global information of the face image, and the effect is not very good under different facial expressions and postures. Scholkopf [17] proposed kernel principal component analysis (KPCA) as a nonlinear extension of PCA, which performs linear principal component dimension reduction in the high-dimensional feature space by mapping the data features of the input space to the high-dimensional feature space. Therefore, this paper adopts the KPCA feature dimension reduction method to reduce the dimension of the extracted features.

The facial expression feature vector extracted based on VGG16 has 4096 dimensions, which is a deeper facial expression feature, and the computational complexity of facial expression sequence similarity is too high. Feature dimension reduction can reduce the complexity of facial expressions while preserving their original features. The KPCA feature dimensionality reduction method used in this article is a nonlinear principal component analysis method that introduces a kernel function based on PCA dimensionality reduction and uses a nonlinear mapping to map the feature to a high-dimensional or even infinite dimension. The dimensional space is converted to linear, and principal component analysis is performed in the mapped kernel function space [18,19], to achieve the purpose of feature dimensionality reduction. Its commonly used kernel functions generally have linear kernel function, Q-order polynomial kernel function, Sigmoid kernel function, Gaussian radial basis kernel function, and Laplace kernel function. The Gaussian radial basis kernel function is shown in Equation 3-1:

$$K(X_i, X_j) = \exp(-\frac{||X_i - X_j||}{\sigma^2}) \quad (3\text{-}1)$$

Where X_i, X_j can be represented as an eigenvector of input space, $||X_i - X_j||$ can be viewed as the square Euclidean distance between $X_i$and X_j,σ is a free parameter. Since the Gaussian radial basis kernel function can map the feature data to the feature space, the value of each element is between (0,1] and there is only one kernel parameter, which can alleviate the nonlinear component between the features, so this paper chooses Gaussian the radial basis kernel function is used as the kernel function of KPCA. The process of reducing the dimensionality of the extracted high-dimensional features using KPCA is as follows:

(1) Based on the high-dimensional features extracted from the VGG network, the feature values are centralized, and then the features are mapped to the high-dimensional feature space using the Gaussian radial basis kernel function.
(2) According to the eigen values mapped to the higher-dimensional space, the mean value of the features is calculated, and the covariance matrix of facial features is constructed.
(3) The eigen values and eigenvectors were calculated according to the constructed covariance matrix of facial expression features. The feature vector corresponding to the maximum eigen value is selected as the low-latitude feature after dimensionality reduction of the high-dimensional feature, namely the principal component feature.
(4) The eigenvectors corresponding to the eigen values constitute the subspace after dimensionality reduction, which is the expression feature after dimensionality reduction.

3.3. Matching Facial Expression Similarity

The common similarity match mainly includes distance measures and correlation measures. A distance measure is a metric that compares the definitions of two images using a distance function. The more similar the images, the smaller the distance. Distance measures include Euclidean distance, Manhattan distance, and Chebyshev distance. Correlation measures include cosine similarity, Pearson correlation coefficient, and so on.

The Euclidean distance measures the absolute distance between each point in multi-dimensional space [20]. Two n-dimensional eigenvectors $X = (X_1, X_2, X_3, \ldots\ldots, X_n)$ and $Y = (Y_1, Y_2, Y_3, \ldots\ldots, Y_n)$.Then the Euclidean distance formula of the two feature vectors is as follows:

$$D(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (3\text{-}2)$$

To reduce the amount of calculation in the experiment, the following Euclidean distance calculation formula is generally adopted, without the square root of the Euclidean distance:

$$D(X, Y) = \sum_{i=1}^{n}(X_i - Y_i)^2 \quad (3\text{-}3)$$

Cosine similarity evaluates the similarity of two vectors by calculating the cosine of the angle between them. Two N-dimensional feature vectors X, Y, the similarity formula of the two feature vectors is as follows:

$$\cos(\theta) = \frac{\sum_{i-1}^{N} X_i Y_i}{\sqrt{\sum_{i-1}^{N} X_i^2}\sqrt{\sum_{i-1}^{N} Y_i^2}} \quad (3\text{-}4)$$

The cosine value is in the range [-1, 1]. The cosine value of two coincident vectors is 0, the cosine value of two vectors in the same direction is 1, and the cosine value of two vectors in the

opposite direction is -1. Compared with Euclidean distance, cosine distance pays more attention to the difference in direction between two vectors.

By comparing several groups of experiments, this paper finally decided to use the cosine similarity distance function as the similarity judgment function, because the cosine distance is used to measure the consistency of vector dimension direction, and the cosine value of the vector, including angle, is used to reflect similarity. In this paper, the low-latitude facial features obtained by KPCA dimensionality reduction are treated as vector features, and the degree of similarity is reflected by calculating the cosine value. The Euclidean distance function, on the other hand, does not pay attention to the difference in dimension. The greater the absolute distance value, the greater the similarity; the greater the absolute distance value, the less similar it is. And the algorithm complexity is higher than the cosine function.

According to the similarity results, it is judged that it is the same emotion, and then the relationship between its intensity changes is studied according to the intensity of emotion.

3.4. Emotion Intensity Estimation

For the task of expression recognition and classification, it only needs to identify what kind of expression it is, but for emotions, emotions are a dynamic process that requires not only similar emotion categories but also similar intensity of emotions. Expression intensity is a measure of emotional intensity that can reflect the degree of emotional change and is an indispensable part of emotional research based on dynamic expression similarity. However, there are relatively few studies on expression intensity at present. On the one hand, it is because the recognition and classification of static expressions and the research on similarity do not need to involve the knowledge of intensity, and on the other hand, there is no uniform definition of expression intensity. Prkachin et al. [22] used facial motion units to define expression intensity, but it took a lot of time and manpower; Hess et al. [23] defined expression intensity through the difference between expression images in different frames in the video, but the above the studies on expression intensity were all separate studies, and none of them combined studies on expression intensity and emotion.

Different levels of emotions reflect the psychology of different people. For dynamic emotions, expression intensity can dynamically reflect the process of emotional intensity changes and then reflect the state of emotions. Expression intensity estimation is the process of dividing each type of facial expression into degrees so that it better reflects the changes in facial emotion. For example, smiling and laughing represent two different degrees of happy emotions. In this paper, the expression intensity values are estimated based on the VGG16 network and Softmax, and the threshold value after the output of Softmax is defined as the intensity value of the expression. From the output of the Softmax function, the intensity value ranges from 0 to 1, which exactly corresponds to the process of intensity from none to climax. In this paper, the intensity curve of each group of emotions is obtained based on the intensity estimation value, and the state of the emotion is judged based on the matching degree of the intensity curve based on feature similarity.

4. EVALUATION EXPERIMENT

This paper conducts experiments based on dynamic expression similarity and emotional intensity to verify the change in an emotional state. First, use VGG16 to train an expression feature extraction model and obtain the emotional intensity value based on the model; secondly, use the cosine function to calculate the feature similarity in the similarity calculation. It is judged

whether it is the same emotional state according to the feature similarity and emotional intensity of dynamic expressions.

4.1. Dataset

The data set for this experiment includes the Extended Cohn-Kanade Dataset (CK+), which is used for the training of the VGG16 expression feature extraction model, and the collected pictures of dynamic expression changes to evaluate the results of emotional changes. The CK+ dataset includes 123 subjects and 593 image sequences, of which 327 sequences have expression labels. And the labeled expression sequence contains the change of the subject's expression from calm to peak, which can be used for the study of emotional changes. The change sequence of surprise emotions of a subject in the data set is shown in Figure 2, and its surprise emotion gradually reaches its peak emotion.

Figure 2. A sequence of images in CK+ dataset.

4.2.Model Training

Model training uses TensorFlow framework, GPU model is NVIDIA GTX1650, 4G memory. The weights are initialized through TensorFlow's variable_initializer, the initial learning rate is set to 0.1 using the Adam optimizer, and the learning rate is optimized using cross-entropy with a step size of 0.0001. When the loss of the training set for three epochs is no longer reduced, the learning rate is reduced by 10 times, and the batch_size is set to 50 to prevent the model from overfitting. The training process of the model is shown in Figure 3, and its accuracy and loss values are saved every 100 iterations. On the left is the accuracy curve for the training set. The X-axis represents the iteration epoch, and the Y-axis is the change in accuracy. The figure on the right is the loss curve of the training set. The X-axis represents the iteration epoch, and the Y-axis is the loss value for the training set. From the 150th epoch, the accuracy has been infinitely close to 1. The model gradually converges.

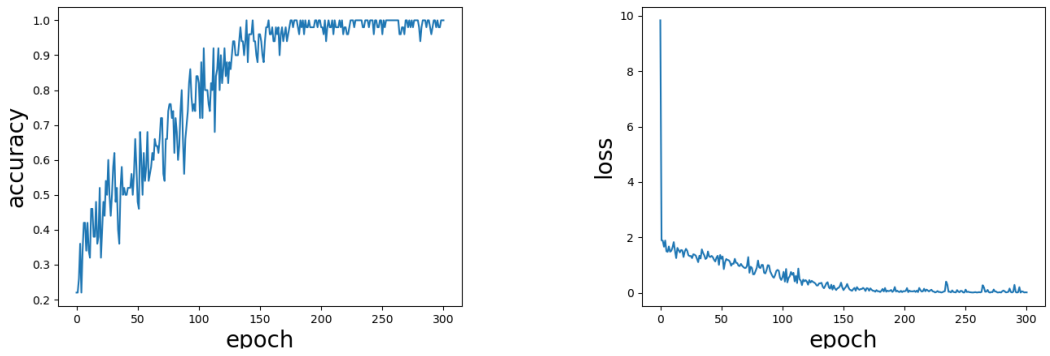


Figure 3. Loss and accuracy changes.

4.3. Expression Similarity Calculation

In this paper, the expression similarity calculation is proposed. Firstly, the expression features are extracted based on the full connection layer of the VGG16 model, and the extracted features have 4096 dimensions, which are used as the feature vector of the expression. The similarity was calculated based on two facial expression sequences, one of which was used as the standard sequence and the other as the test sequence. The feature vectors of each frame of the facial expression sequence were combined to form a high-dimensional feature matrix, and the kernel principal component dimension reduction algorithm was used to reduce the dimension of the feature matrix.

The dimensionality reduction calculation of an extracted single frame of facial expression was used to effectively reduce the latitude of facial expression while keeping the original features unchanged to the maximum extent. The cosine function was used to calculate the comparison of similarity between a single frame of facial expression after dimensionality reduction and before dimensionality reduction. The data for dimension accuracy before and after dimensionality reduction is shown in Table 1.

Table 1. Comparison of similarity of different dimensions

Dimension	4096	1024	256	64	32	16	8
Precision	0.9036	0.9037	0.9038	0.9111	0.9025	0.8256	0.7898

According to the results in Table 1, when the dimensionality of facial features is reduced by nuclear principal component analysis, its accuracy goes through a process of first increasing and then decreasing. When it is reduced to 64 dimensions, it has the highest similarity to the original feature and can retain the facial feature information to the maximum extent. As can be seen from the accuracy after dimensionality reduction in Table 1, there is only a 0.0086 difference between the accuracy of 32 dimensions and 64 dimensions. However, considering the complexity of dimensionality reduction, the complexity of facial features reduced to 32 dimensions is higher than that of 64 dimensions. Therefore, when We were reducing the dimensions of facial features, we chose to reduce the high-dimensional facial features to 64 dimensions.

4.4. Experimental Results

Table 2 shows the comparison of the calculation results of the similarity of expressions, where VGG16 indicates that the expression feature extraction method uses a deep learning model, and no indicates that the expression features are extracted according to the facial geometric features. PCA, KPCA, and Isomap represent three different dimensionality reduction methods, and the cosine function and Euclidean distance represent two different similarity matching algorithms. Figures 4 and 5 show the sequence of two groups of emotions, respectively. Figures 7 and 8 show the sequence of two groups of emotions, respectively. Emotional intensity is the process of having no emotion to experiencing the peak of emotion. Figures 6 and 8 show a comparison of the two sets of intensity curves.

Table 2. Comparison of similarity results of different dimensionality reduction methods

Method	VGG16	Cosine function（%）	Euclidean distance	Increase（%）
KPCA	√	**92.5**	**0.29**	**9.7**
	×	86.5	0.65	3.7
PCA	√	91.8	0.33	9
	×	85.7	0.87	2.9
Isomap	√	90.6	1.07	7.8
	×（standard）	82.8	2.17	0

According to the experimental results in Table 2, after extracting facial features based on geometric features, the benchmark method adopted isomap feature dimension reduction, and the cosine function was used to calculate the similarity of 0.828 and the European distance of 2.17. When VGG16 is used to extract facial features, the cosine similarity is 0.906 after isomap feature dimensionality reduction. Compared with facial features extracted without VGG16, the cosine function similarity improved by 7.8%, and the Euclidean distance is 1.07, which is also smaller than 2.17, and its similarity is higher. It shows that the facial features extracted by the deep learning model are more comprehensive and accurate than the facial geometric features. Based on the facial features extracted from VGG16, the feature dimension reduction method was changed. After PCA dimension reduction, cosine similarity was 0.918, and KPCA dimension reduction cosine similarity was 0.925, which were 1.2% and 1.9% higher than isomap similarity, respectively. The results show that KPCA can retain the high-dimensional facial features to a greater extent after dimensionality reduction. VGG16 was used to extract facial expression features, and the similarity of KPCA features after dimensionality reduction was 0.925, which was about 9.7% higher than that of the benchmark method.

Figure 4 shows a set of standard happy emotion sequences that contain 9 frames of emotion change. From the first frame in the upper left corner to the ninth frame in the lower right corner, the happy emotion frame has experienced expressionless to the highest climax of happy emotion. Its happy mood intensity goes from nothing to a climax. Figure 5 is a set of emotional frames imitating the happy emotion sequence in Figure 4, and the feature similarity between the two sets of emotions is 92.5%. Figure 6 shows the curve of the intensity change between the two sets of happy emotion sequences. From the template emotion sequence and the imitator's emotion

sequence, the expressions in the two sequences have experienced the process from having no emotion on their faces to reaching a climax emotion. The emotional intensity curve is consistent with the change of the emotional sequence. The emotional intensity increases gradually and then peaks. If the comparison between the first frame in the emotional sequence 1 and the first frame in the emotional sequence 2 is considered alone, and the intensity curve is not considered, the interpreted emotion is not happy and cannot reflect the emotional change process. According to the similarity results combined with the emotion intensity curve to judge the emotion change process, it can be concluded from the trend of the intensity curve that the dynamic characteristics of the two emotion sequences are similar, and that the emotion change process is also similar.

Figure 4. Happy emotion sequence 1

Figure 5. Happy emotion sequence 2

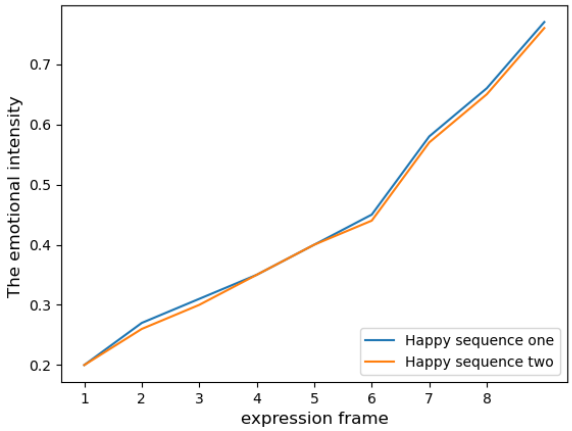


Figure 6. Comparison of the intensity curves of happy emotions

Figure 7 shows a sequence of dynamic changes of a group of template surprise emotions from the first frame in the upper left corner to the twelfth frame in the lower right corner. Figure 8 is a sequence of surprise emotions imitating Figure 7, and the feature similarity between the two sets of emotions is 92.48%. Figure 9 depicts a comparison of the intensity of the two surprised emotion sequences depicted in Figures 7 and 8. The two surprised emotion sequences are from no expression to a surprised climax expression, and their intensity has also experienced a change

process from 0 to close to 1, which is the same as the change trend of expression. It can be concluded that the method of dynamic expression similarity combined with emotion intensity estimation proposed in this paper can be used to effectively evaluate the process of emotion change.

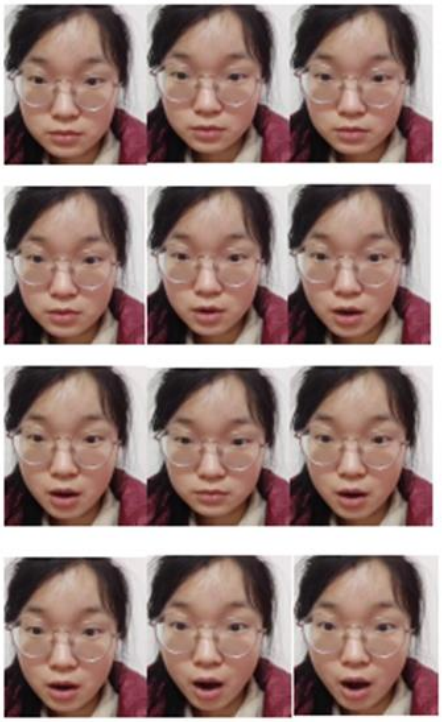


Figure 7. Surprise emotion sequence 1

Figure 8. Surprise emotion sequence 2

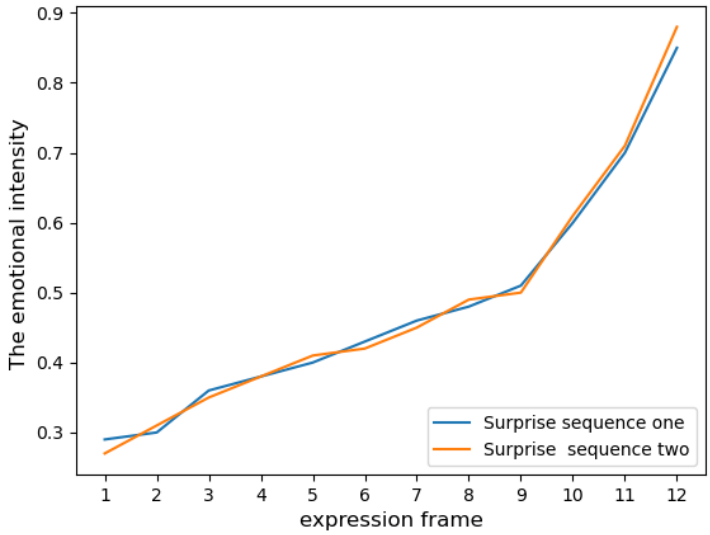


Figure 9. Comparison of surprise intensity curves

5. CONCLUSION

This paper proposes a method to study the emotion process based on dynamic facial expression similarity. Firstly, facial expression features are extracted through the VGG16 network to improve the inaccuracy of facial expression feature extraction based on facial geometric features. Secondly, facial expression feature similarity and facial expression intensity similarity are combined to make the results more accurate. Compared with facial feature extraction based on geometric features, the proposed algorithm can effectively extract facial features and then calculate the similarity of facial expression sequences. On average, compared with facial feature extraction based on geometric features, the similarity of facial expression sequences is improved by 9.7% on average, which can better evaluate the process of emotional change. However, the expression similarity measurement method proposed in this paper is mainly based on positive or near-positive expressions without considering the influence of side and facial occlusion on expression. Therefore, the future research direction is mainly to study the influence of head posture on the emotion change process.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system (facs): a technique for the measurement of facial actions," *Rivista Di Psichiatria*, vol. 47, no. 2, pp126-38, 1978.

[2] J. ma, "Facial expression recognition based on feature fusion," *Industrial control computer*, vol. 33, no. 11, p4, 2020.

[3] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *International Journal of Computer Vision*, vol. 76, no. 1, pp93-104, 2008.

[4] T. Zavaschi, A. S. Britto, L. Oliveira, and A. L. Koerich, "Fusion of feature sets and classifiers for facial expression recognition," *Expert Systems with Applications*, vol. 40, no. 2, pp 646-655, 2013.

[5] Y. Tang, X. Zhang, X. Hu, S. Wang, and H. Wang, "Facial expression recognition using frequency neural network," *IEEE Transactions on Image Processing*, vol. 30, pp 444-457, 2021.

[6] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial expression recognition in videos using dynamic kernels," *IEEE Transactions on Image Processing*, vol. 29, pp 8316-8325, 2020.

[7] A. Nasuha, F. Arifin, A. S. Priambodo, N. Setiawan, and N. Ahwan, "Real time emotion classification based on convolution neural network and facial feature," *Journal of Physics Conference Series*, vol. 1737, no. 1, pp 012008, 2021.

[8] X. Sui, L. Xue, and D. Li, "Dynamic Expression Recognition Based on Hybrid Features and Optimized Extreme Learning Machine Model," *IOP Conference Series*: Materials Science and Engineering, vol. 719, no. 1, p. 012080(6pp), 2020.

[9] H. Fisher, "Emotion recognition from realistic dynamic emotional expressions cohere with established emotion recognition tests: A proof-of-concept validation of the emotional accuracy test." *Journal of Intelligence*. vol. 9, 2021.

[10] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in 2019 IEEE/CVF *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp 5676-5685.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp815-823.

[12] B. Yang and Y. Yang, "Pca is used to reduce dimension of deep learning image feature extraction," *Computer system application*, vol. 28, no. 01, pp 281-285, 2019.

[13] L. Rampasek and A. Goldenberg, "Tensorflow: Biology's gateway to deep learning?" *Cell Systems*, vol. 2, no. 1, pp12-14, 2016.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[15] J. Liu and F. Zhao, "High dimensional data reduction technology and research progress," *Electronic technology*, vol. 031, no. 003, pp 36-38, 43, 2018.

[16] J. Du, X. Wang, and L. Hu, "Nonlinear dimensionality reduction technology and visualization application," *Journal of Donghua University (Natural Science edition)*, vol. 46, no. 4, pp 6, 2020.

[17] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp1299-1319, 1998.
[18] M. R. Galiaskarov, V. V. Kurkina, and L. A. Rusinov, "Online diagnostics of time-varying nonlinear chemical processes using moving window kernel principal component analysis and fisher discriminant analysis," *Journal of Chemometrics*, pe2866, 2017.
[19] A. A. Joseph, T. Tokumoto, and S. Ozawa, "Online feature extraction based on accelerated kernel principal component analysis for data stream," *Evolving Systems*, vol. 7, no. 1, pp15-27, 2016.
[20] L. Gao, "Research and application of text clustering algorithm," Ph.D. dissertation, *University of Electronic Science and Technology of China*, 2013.
[21] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *International Journal of Computer Vision*, vol. 76, no. 1, pp 93-104, 2008.
[22] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp267-274, 2008.
[23] U. Hess, S. Blairy, and R. E. Kleck, "The intensity of emotional facial expressions and decoding accuracy," *Journal of Nonverbal Behavior*, vol. 21, no. 4, pp 241-257, 1997.

MINING BIOMEDICAL LITERATURE TO DISCOVER NATURAL CURE FOR RECURRENT DISEASE

Farhi Marir[1], Hussein Fakhry[1] and Aida J. Azar[2]

[1]College of Technological Innovation, Zayed University,
Academic City, Dubai, UAE
[2]College of Medicine, Mohammed Bin Rashid University of
Medicine and Health Sciences (MBRU), Dubai Health Care City,
Dubai, United Arab Emirates.

ABSTRACT

The advances in digital data collection and storage technology allows the storage of a huge amount of medical publications in MEDLINE. This database contains more than 25 million references to journal articles and abstracts in life sciences and biomedicine. This research work builds on Swanson use of mathematical association between A and C concepts/terms through a list of B concept/terms retrieved from large medical literature databases that contain either A&B or B&C terms links A to C. Swanson discovered evidence that fish oil (A) may cure vessel blood disorder (C) and that magnesium (A) may be effective against migraine headache (C), which were clinically proven two years later. We present a cooccurrence mining algorithm and an A&C pre-defined domain Knowledge Base (containing for instance Garlic Composition and Blood pressure causes) to filter and reduce the exponential number of shared B terms retrieved from MEDLINE articles using Swanson's Arrowsmith machine. The reduced number of relevant B terms makes it easier to build scientific evidence to validate publicly known remedies for recurrent diseases for instance establishing whether an important association exists between garlic and its impact on blood pressure.

KEYWORDS

Co-occurrence Text Mining, ABC Arrowsmith Discovery Machine, Dietary Aliments & Disease Knowledge Base, and MEDLINE medical database.

1. INTRODUCTION

During the eighties, Swanson [1] developed a new bibliographic-based approach that associates medical literature and articles from MEDLINE database for creating new scientific knowledge. Based on the mathematical associative relationship, he stated that if one publication states a relationship between two phenomena A and B while another publication reports on the relationship between B and C phenomena a number B_1, B_2, B_3...B_n term connections can be made, and new scientific knowledge could be generated. In a decade, he identified seven examples of complementary non interactive structures in the biomedical literature. Figure 1 shows three A to C linkages: Blood viscosity (B_1), Platelet Aggregation (B_2), and Vascular Reactivity (B_3). The inference that fish oil may benefit Reynaud's disease was clinically proven two years later [2] In a similar way, he discovered several associative relationships like magnesium deficiency and the occurrence of migraine, which was also clinically validated [3]. These scientific knowledge discoveries are generated using Arrowsmith computer-aided tool [4]

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 47-57, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121104

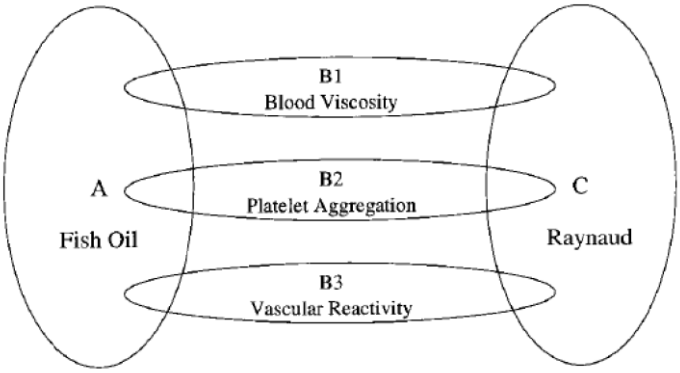


Figure 1. Diagram representing Swanson's first discovery: Fish oil cures Reynaud's diseases [5]

The potential of Swanson's research has been widely acknowledged by the scientific community, but likewise, its complexity concerns the vast information space and the possible exponential number of Bi connections given that MEDLINE contains more than twenty-five million medical articles. In addition to this complexity is that most information is represented in natural language, and this requires advanced natural language processing. For these reasons, it took Swanson ten years to discover seven natural cures to recurrent diseases.

This research aimed at enhancing Swanson's research work, particularly enhancing the Arrowsmith tool with advanced Text Mining techniques such as the co-occurrence Text Mining approach [6, 7] and proper Knowledge Base regarding the components of the dietary aliment and the sources of disease to undertake further bibliographic-based medical discoveries. These will be saved in a medical knowledge repository consisting of three entries the remedy/medicine, the disease, and scientific evidence that validate or refute the link between remedies and a given disease. Such a knowledge repository will motivate the medical research community to further investigate the clinical validity of the Text Mining induced findings of links between diseases and their cures.

The paper is organized as follows Section 2 gives the literature review on Text Mining and knowledge discovery in biomedical literature. Section 3 presents the process of knowledge discovery through co-occurrence Text Mining and domain Knowledge Base, and its implementation on a specific case of "Garlic" and its impact on "Blood Pressure". The last two sections are devoted to the conclusion and references.

2. LITERATURE REVIEW

In the last two decades, the advances in digital data collection devices and data storage technology allowed the storage of huge amounts of academic publications e.g., MEDLINE which contains more than 25 million references to medical journal articles in life sciences with a concentration on biomedicine abstracts (https://www.nlm.nih.gov/pubs/factsheets/medline.html). Underneath these empirical data are great opportunities for knowledge discovery using artificial intelligence and machine learning particularly Data Mining and Knowledge Discovery (DMKD). DMKD utilizes methods, algorithms, and techniques from a variety of disciplines to extract useful information, emerged during the 1990s, and grew rapidly, and in 2007, there were 17 data mining conferences [10].

This research work initiated the use of Text Mining approach of knowledge repository to mine the empirical literature in MEDLINE database to build scientific evidence that consolidates

publicly known (but not clinically proven) remedies for given diseases e.g., building scientific evidence that for instance garlic reduces the blood pressure.

Text Mining in biomedical literature has grown in the past few years to be a major tool for bioinformatics. For many applications, numerous methods have been developed. To contribute to this still-growing field, it is important to systemize the methods that are already in use. Usually, the Text Mining approach is based on natural language processing, mathematical and statistical backgrounds, it can be a strategic source of evidence for literature-based of discoveries in biomedical sciences. For this reason, researchers and practitioners from various fields are using Text Mining.

Three types of Text Mining approaches are used for new knowledge discovery in the biomedical domain: rule-based or knowledge-based approaches, statistical or machine-learning based approaches, and co-occurrence-based approaches [8]. The most famous work on theory-toresearch using the basic associative relationship as an evidence tool is the research work undertaken by Dr. Swanson [1] who has generated scientific evidence for several literature-based hypotheses that have been corroborated experimentally and clinically. He used an associative relationship between two concepts A and C to retrieve articles that contain the third set of concepts B_1, B_2, B_3...B_n that connect A and C concepts. The major limitation of Swanson approach is that the number of B_i connectors could grow exponentially as the number of articles retrieved from MEDLINE could be millions which will complicate or even make impossible the generation of scientific evidence. Another research work reported by Hui [8] uses the rule-based approach of Text Mining to automatically identify the status of obesity and related co-morbidities based on the patient's clinical discharge summaries. Some of the authors of this paper have undertaken several research works combining Text Mining term co-occurrence and statistical techniques approach to mine social networks and the holy Quran for developing knowledge repository for diabetes and Islamic financial business processes respectively [6,7].

Building on Swanson's successful research work and tackling the limitations of Swanson's basic associative relationship approach and being able to deal with millions of retrieved articles, this research aims at enhancing this research area through the development and implementation of Text Mining approach that combines term co-occurrence and domain Knowledge Base to filter relevant B_i terms retrieved from articles in MEDLINE. This new Text Mining approach will reduce the number of B_i to those semantically relevant to the analysis making it easier to generate scientific evidence for publicly known disease remedies and help in producing more relevant and coherent outcomes.

3. KNOWLEDGE DISCOVERY PROCESS AND IMPLEMENTATION

This research falls within theory-building methodology as it aims at building scientific evidence (theory) that links natural or chemicals existing in a dietary ailment that could be a remedy to a recurrent disease using Text Mining techniques and domain Knowledge Base to mine MEDLINE database. Data Mining and knowledge discovery support the two common strategies of theory building or scientific knowledge discovery [10] either by validating an existing theory known as a theory-to-research strategy or by developing a new theory known as a research-to-theory strategy. Text Mining supports the two strategies common to theory building. The first is of a research-to-theory strategy, whereas the second is of a theory-to-research strategy [11]. The research work fits well with the research-to-theory strategy which aims at deriving the laws of nature from a careful examination of all the available data which in this case is the 25 million medical articles contained in the MEDLINE dataset. As described by Reynold [11], the essences of this research-to-theory strategy are as follows:

1. Select a phenomenon and list all the characteristics of the phenomenon. The phenomenon is to identify all possible links between remedy-disease pairs giving a detailed list of their characteristics and the level of impact of remedy on the diseases.
2. Measure all the characteristics of the phenomenon in a variety of situations which means investigating the effect of the remedy on different persons e.g., female, male, old, and young person.
3. Analyze the resulting data carefully and determine any systematic patterns among the data "worthy" of further attention. This is the core of this research as it enhanced Swanson associative tool [4] with the advanced co-occurrence text mining tool [5,6] to semantically filter retrieved B_i concepts that have a medical relationship to both A and C.
4. Identify significant patterns during the process of mining the MEDLINE database, to formalize the discovered patterns as theoretical statements or scientific evidence that validate the hypothesis linking remedy-disease.

3.1. The Co-occurrence Algorithm and Domain Knowledge Base

First, we developed a Knowledge Base on the components of Dietary aliment for instance "*The major bioactive compounds of garlic are its organosulfur compounds, such as diallyl thiosulfonate (allicin), diallyl sulfide (DAS), diallyl disulfide (DADS), diallyl trisulfide (DATS), E/Z-ajoene, S-allyl-cysteine (SAC), and S-allyl-cysteine sulfoxide (alliin)"* [6] In addition to statements on the impact of garlic on blood pressure for instance: *"Kyolic garlic has also shown promise in improving cardiovascular health by reducing arterial stiffness, elevated cholesterol levels and blood 'stickiness'* [8]. In addition, the prebiotic properties of garlic increase in increasing gut microbial richness and diversity" [9]

Once the Knowledge Base on known pairs of cure-disease was identified, we used R programming language to build a corpus of retrieved B_i terms to undertake the usual Text Mining process of eliminating stop words, and undertaking stemming and Part-of-Speech (PoS) Tagging to tag words according to the grammatical context of the word, hence dividing up the words into nouns, verbs, etc. This was important for the exact analysis of relations between words in the B_i list of terms itself and in relation to the domain Knowledge Base. In the same way, we produced B_i terms, and used a co-occurrence algorithm to filter B_i shared by "Garlic" and "Blood pressure" Knowledge Base as follow:

1. *Prepare the resulting* B_i *list of terms by eliminating stop words, stemming, and tagging special biomedical words,*
2. *Use co-occurrence algorithm to match terms in the Bi list of terms with the A (KB_Garlic) and C (KB_BloodPressure)* Knowledge Base
3. *Use the result of co-occurrence mining process to produce an ordered list of A components (shared by KB_Garlic and KB_BloodPressure) that could be used as evidence of links between "Garlic" and "Blood pressure"*

3.2. The Implementation of the Knowledge Discovery Process

The knowledge discovery process was used to find whether the dietary aliment "Garlic" may help in reducing "Blood pressure". First, we used the Arrowsmith tool to extract from the MEDLINE bio-medical database all medical articles and abstracts that contained the term "Garlic" (Figure 2).

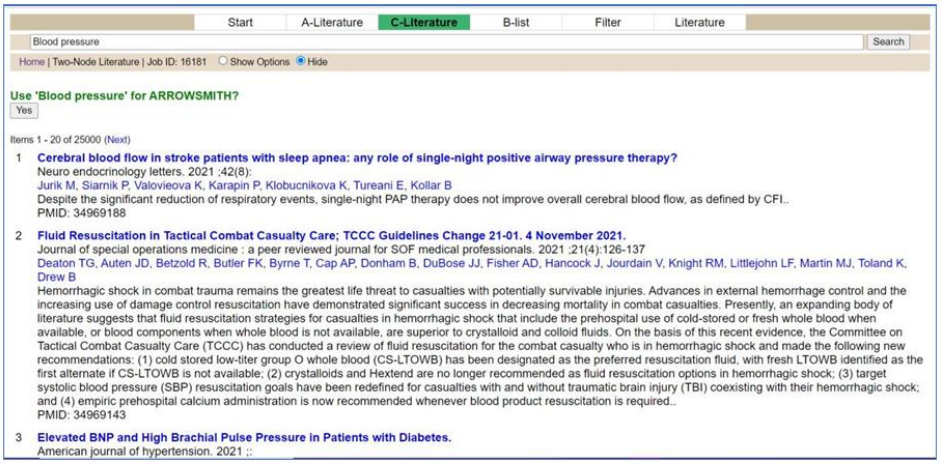


Figure 2. A - Literature Retrieval

Figure 2 shows that 7,103 bio-medical papers and abstracts were retrieved that contained the term "Garlic". The same process was used to retrieve from the literature containing the C term i.e., "Blood pressure" (Figure 3).

Figure 3. C - Literature Retrieval

Figure 3 shows that 25,000 bio-medical papers and abstracts were found that contained the term "Blood pressure". Next, we undertook the process of extracting B lists of terms that are shared by both A-Literature and C-Literature lists giving the results as shown in Figure 4.

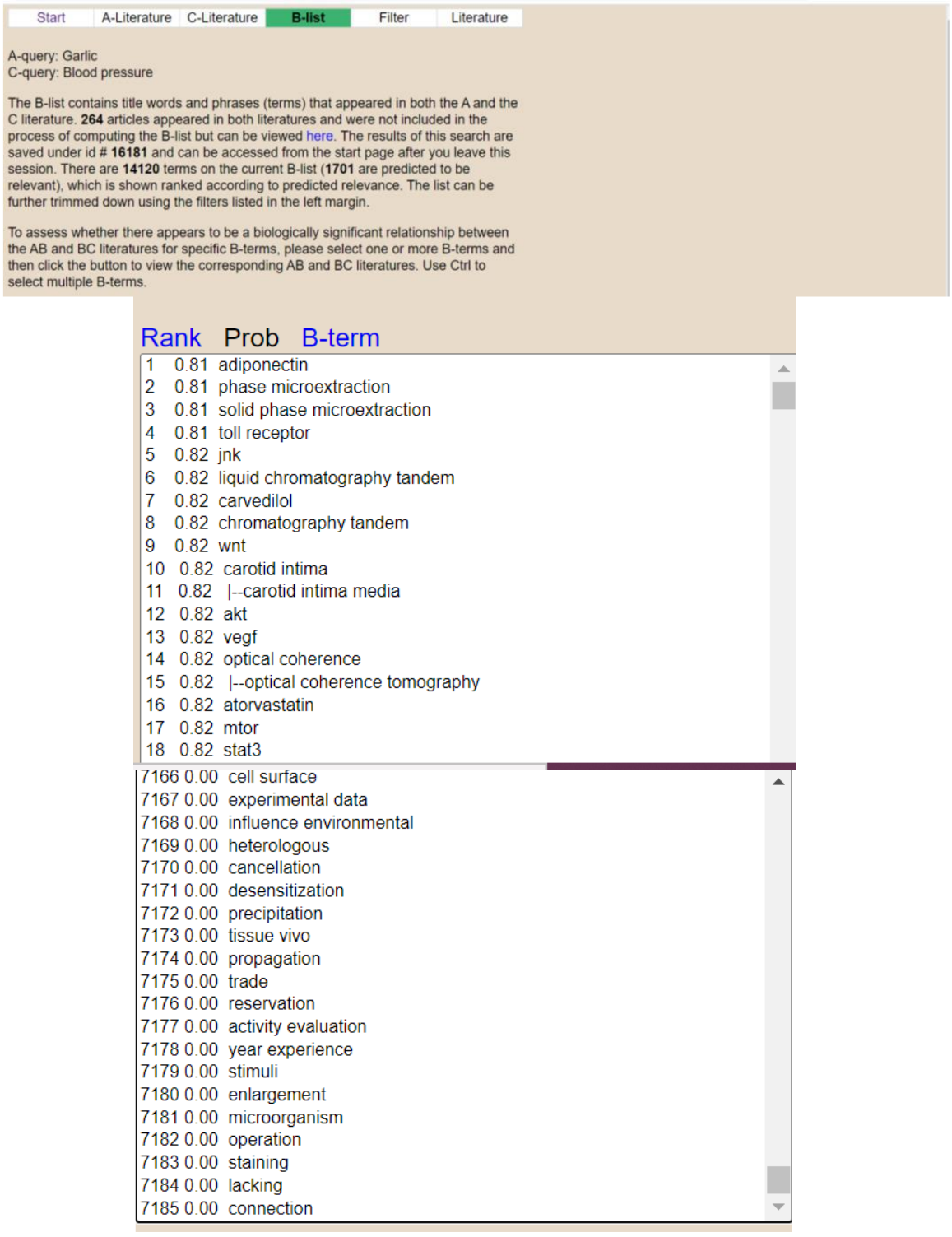


Figure 4. The retrieved B list of terms shared by A-Literature and C-Literature Lists

Figure 4 retrieved the huge number of 7,185 B_i terms and for this reason, there was a need to develop semantic filters to be able to build the evidence that “Garlic” reduces “Blood pressure”.

For this reason, after preparing the B list of terms the third step was to implement the cooccurrence algorithm using Python programming language as shown in Figure 5.

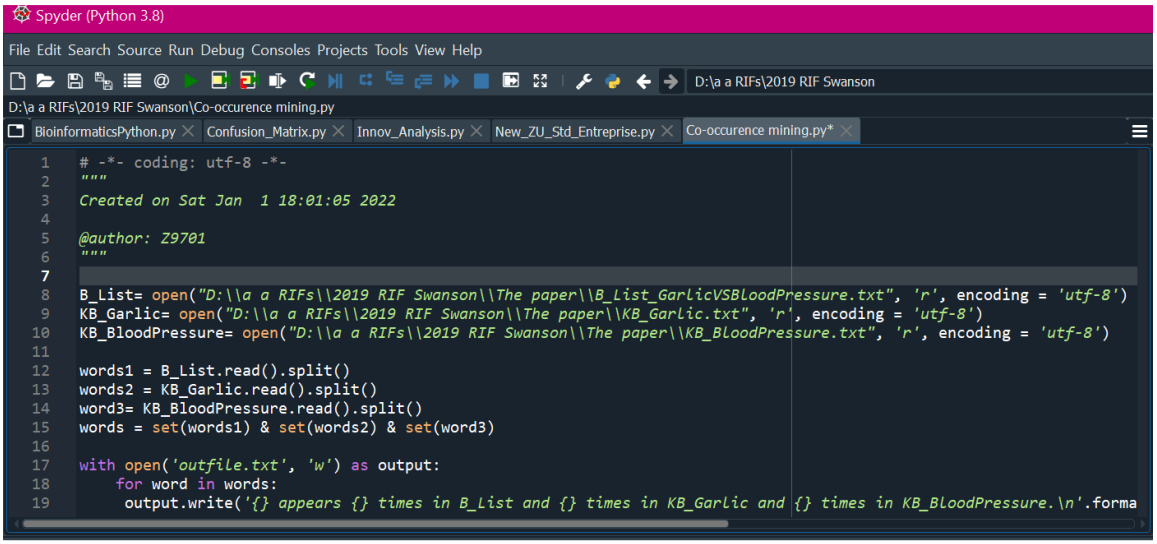


Figure 5. The co-occurrence Python code for filtering B List of terms

Figure 6 shows the filtered list of B terms that resulted from the co-occurrence Python code (Figure 5). The filtering process retrieved terms that occurred simultaneously in B terms lists, "Garlic" Knowledge Base, and "Blood pressure" Knowledge Base (Figure 6).

inhibiting appears 9 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
human appears 85 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
enzyme appears 11 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
reducing appears 5 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
hydrogen appears 4 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
linked appears 2 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
animal appears 8 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
production appears 7 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
in appears 2 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
pressure appears 7 times in B_List and 1 times in KB_Garlic and 5 times in KB_BloodPressure.
blood appears 24 times in B_List and 2 times in KB_Garlic and 6 times in KB_BloodPressure.
content appears 7 times in B_List and 1 times in KB_Garlic and 1 times in KB_BloodPressure.
angiotensin appears 5 times in B_List and 1 times in KB_Garlic and 2 times in KB_BloodPressure.

Figure 6. Filtered List of B terms

Finally, after further analysis we found that the term "*angiotensin*" (Figure 7) could be used as part of the evidence that garlic could reduce the impact of blood pressure as it fits with the following statement:" *Blood pressure reducing properties of garlic have been linked to its hydrogen sulphide production [4,5] and allicin content – liberated from alliin and the enzyme alliinase [6,10] – which has angiotensin II inhibiting and vasodilating effects (the dilatation of blood vessels, which decreases blood pressure), as shown in animal and human*". Other terms like enzymes, hydrogen, and blood could be further analyzed to further consolidate the evidence.

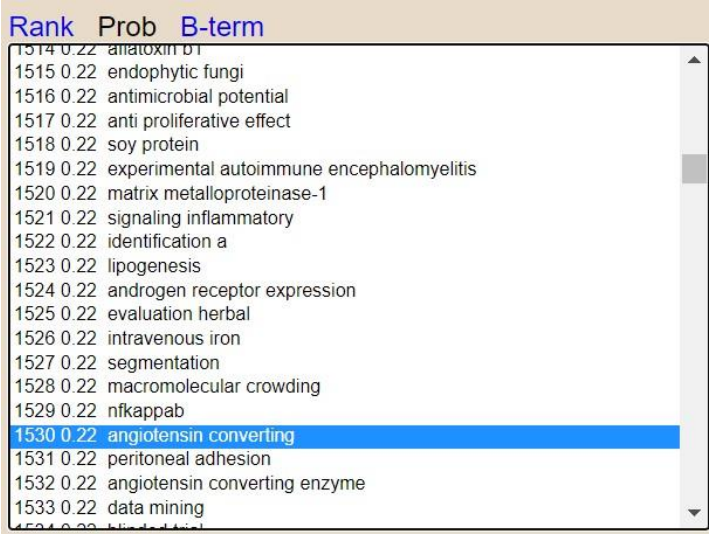


Figure 7. the coccurence ofd the the530th term “angiotensen” in the original B terms

4. CONCLUSIONS AND DISCUSSION

This research aimed at building scientific evidence that linked natural or chemicals existing in a dietary ailment that could be a remedy to a recurrent disease using Text Mining techniques and domain Knowledge Base to mine MEDLINE database knowledge discovery as a research-to theory strategy. It was successfully able to discover and establish an important association on the specific case of “Garlic” and its impact on “Blood pressure”. The results of this research work will be added to the results of other experimental studies to ascertain the efficacy of garlic in reducing blood pressure. As this research project is ongoing, we will be investigating other important associations reported by practicing leading physicians in the region.

Non-pharmacological treatment options for blood pressure or hypertension have the potential to reduce the risk of cardiovascular diseases at a population level. A systematic study on the impact of garlic on blood pressure was undertaken [15] and concurs with the results of our study on mining biomedical literature. They undertook a systematic clinical study on animals and found evidence that garlic reduced blood pressure. [15] Also primary studies were conducted in humans and in a meta-analysis of 12 studies in subjects with uncontrolled hypertension, garlic was found to be effective in reducing the blood pressure. [16] The results stated that garlic lowers blood pressure in hypertensive subjects, improved arterial stiffness and gut microbiota. [16] However, as the number of subjects per trial was too small this warrants further research to be able to ascertain a causal inference between the positive effect of garlic in reducing blood pressure. Randomized controlled clinical trials with adequate sample size in subjects with elevated blood pressure is the gold standard to test the efficacy of this new therapy of the use of garlic in lowering the patients’ blood pressure. However, conducting a large randomized controlled clinical trial is extremely costly and time-consuming. For this reason, using Swanson’s Arrowsmith machine will be a first step to find associations as it is quick and an inexpensive tool where data mining and knowledge discovery can be used to seek any possible connections between garlic and lowering the blood pressure. Furthermore, as interest in complementary medicine for blood pressure is increasing, we conducted this approach of mining biomedical literature to find in the B list the term “angiotensin” shared by “Garlic” literature and by “Blood pressure”. In the literature garlic is a natural angiotensin converting enzyme inhibitor which has an effect on reducing blood pressure. [17] To consolidate our findings, we also ran the mining process on some of a couple of cure-disease discovered by Swanson like for instance Fish oil (A)

and Blood viscosity (C), Magnesium Deficiency (A), Migraine headache (C), and we reached at the same conclusion as reported in the paper [10]. The findings of our literature-based mining on the impact of garlic on blood pressure can be validated only with further research such as in double blind controlled clinical trials which is the golden standard to prove a causal association.

ACKNOWLEDGEMENTS

The authors would like to thank Zayed University for their financial support through the scheme of Research Incentive Funding (RIF code: R19048 and PRFA code: R20092).

COMPETING INTERESTS

The authors declare that there is no conflict of interest

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper. Dr. Farhi Marir, Dr. Hussein Fakhry and Aida J. Azar contributed to the study conception and design, draft manuscript, revising the manuscript critically for important intellectual content, and final approval. All authors reviewed the results and approved the final version of the manuscript.

TRANSPARENCY

The authors affirm that all information submitted in this manuscript is accurate and true. All data has been reported and no data has been omitted.

REFERENCES

[1] Swanson, D. R. (1986). Fish Oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30, 7-18.

[2] Ralph A. Digiacomo, Joel M. Kremer, Dhiraj M. Shah. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study, The American Journal of Medicine, Volume 86, Issue 2,1989,Pages 158-164

[3] Swanson, D. R. and Smalheiser N. R. (1996). Undiscovered Public Knowledge: A Ten-Year Update. KDD-96 Proceedings. AAAI (www.aaai.org).

[4] Arrowsmith tool located in http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html, accessed on the 25th of March 2022

[5] Marc Weeber et al. (2001). Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud–Fish Oil and Migraine–Magnesium Discoveries JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 52(7):548–557, 2001

[6] Marir, F., and Huwida, S. (2016). Text Mining Social Networks to Discover New Symptoms and Treatment for Diabetes Patients, 3rd International Workshop on Machine Learning and Data Mining for Sensor Networks (MLDM-SN), May 23-26, 2016 in Madrid, Spain

[7] Marir, F. et al. (2018). A Recursive Co-Occurrence Text Mining of the Quran to Build Corpora for Islamic Banking Business Processes. The International Conference on Intelligent Human Systems Integration (http://ihsint.net/info.html).

[8] Shang, A., Cao, S. Y., Xu, X. Y., Gan, R. Y., Tang, G. Y., Corke, H., Mavumengwana, V., & Li, H. B. (2019). Bioactive Compounds and Biological Functions of Garlic (Allium sativum L.). Foods (Basel, Switzerland), 8(7), 246. https://doi.org/10.3390/foods8070246)

[7] Ried K. (2020). Garlic lowers blood pressure in hypertensive subjects, improves arterial stiffness and gut microbiota: A review and meta-analysis. Experimental and therapeutic medicine, 19(2), 1472–1478. https://doi.org/10.3892/etm.2019.8374

[8] Ried, K., Travica, N., & Sali, A. (2018). The Effect of Kyolic Aged Garlic Extract on Gut Microbiota, Inflammation, and Cardiovascular Markers in Hypertensives: The GarGIC Trial. Frontiers in nutrition, 5, 122. https://doi.org/10.3389/fnut.2018.00122
[9] Bretonnel K. C. and Lawrence Hunter: Getting Started in Text Mining. PLoS Comput Biol. Vol. 4(1). (2008)
[10] Hui Y., Irena S., John A. K., Goran N.: A Text Mining Approach to the Prediction of Dis-ease Status from Clinical Discharge Summaries. Journal of the American Medical Informatics Association Vol.16, No. 4. (2009)
[11] Kdnuggets, Meetings/Conf. in Data Mining, Knowledge Discovery, and Web Mining (2007). Available online at http://www.kdnuggets.com/meetings/index.html.
[12] Reynolds, P. D. (1971). A primer in theory construction. New York: Macmillan.
[13] Lynham, S. A. (2000b). Theory building in the human resource development profession. Human Resource Development Quarterly, 11(2), 159-178.
[14] Karin Ried, Oliver R Frank, Nigel P Stocks, Peter Fakler & Thomas Sullivan (2008). Effect of garlic on blood pressure: A systematic review and meta-analysis, BMC Cardiovascular Disorders 8, Article number: 13
[15] Ried K. Garlic lowers blood pressure in hypertensive subjects, improves arterial stiffness and gut microbiota: A review and meta-analysis. Exp Ther. Med. 2020; 19(2):1472-1478. doi:10.3892/etm.2019.8374).
[16] Sharifi AM, Darabi R, Akbarloo N. Investigation of antihypertensive mechanism of garlic in 2K1C hypertensive rat. J Ethnopharmacol 2003;86:219-24

AUTHORS

Dr. Farhi Marir is Fellow of British Computer Society. He joined ZU in 2013 and he is director of the CTI Research Centre on Big Data & Analytics. He was Acting Director of the Institution Research Office in ZU for two years (2014-2016) where he initiated ZU project for the development of ZU Data warehouse and Business Intelligent Dashboards. He is currently teaching Data analytics he developed, Knowledge Management, Emerging Technologies, and Data Warehousing. As for research in ZU, he is either a PI or Co-PI in six university RIFs grants worth more than 1 million AED applying data analytics in the domain of Islamic Finance, Health (Diabetics & Autism), Education, and Security. He is also leading an 175K AED Research Project on Big Data Analytics for Dubai Tourism which is funded by the UAE National Research Funding. Prof Marir received his PhD in Deductive Databases in UK in 1993 and worked for around twenty five years (1988-2013) in UK university where he taught a large number of courses related to Data science, artificial intelligence and knowledge management. He was also the Director of the Knowledge Management Research Center (KMRC) at London Metropolitan University (2000-2013). As the director of KMRC, Prof. Marir led around thirty-five UK and EU funded research projects worth more than £7 Million, supervised thirty PhD students to completion and published more than hundreds research papers. Well before studying for PhD in the UK, Prof Marir was head of a computer science department for six years at the University of Batna in Algeria (1982-1988). During this time, he set up one of the first university computer center in Algerian Universities.

Dr Hussein Fakhry received his PhD in October 1994 in Intelligent Control Systems & Robotics from the University of Waterloo, Canada. Since then, he assumed different academic and administrative positions at University of Windsor, Cairo University, University of Dubai and Zayed University.

Dr. Hussein Fakhry has a rich profile working at academia in different roles. Since September 2001 he worked at University of Dubai in different roles as faculty, then Assistant Dean and later as Dean of the College of Information Technology. After moving to Zayed University in September 2014, he assumed different roles as Program Coordinator, Assistant Chair and later as Assistant Dean for Students at the College of Technological Innovation. He is currently an Associate Professor in the College of Technological Innovation at Zayed University. Dr Fakhry has extensive professional experience in teaching and training particularly in Systems Analysis & Design, Database Design, Data warehousing, Expert Systems, Decision Support Systems, Strategic Management

Modelling using System Dynamics, Modelling and Simulation of the Supply Chains, and Operations Research Models. Dr Fakhry's research interests are in the areas Applications of Artificial Intelligence, Information Systems Research using System Dynamics, Information Systems Security, E-Commerce and E-Business, Decision Support Systems, and Assessment of Academic Programs. His research has appeared in several international journals and conferences such as Review of Business Information Systems, Communications of the International Information Management Association CIIMA, An International Journal of Information & Security, The Journal of Mathematical and Computer Modelling, Proceedings of the IEEE International Conference on Information and Communication Technologies, and the IADIS International Conference on Information Systems.

Dr. Aida J. Azar joined Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU) since its inception in 2016. In her current position, she is Associate Professor Epidemiology and is one of the founding members. She is responsible for developing and coordinating the epidemiology, biostatistics, and research curriculum for the undergraduate medical students. In her previous positions, Dr. Aida was part of several collaborative groups on writing cardiovascular guidelines for Dutch cardiologists. Also, she was involved in the conduct and analysis of numerous multicenter clinical trials. Her dissertation on 3404 myocardial infarction patients was used as a reference for the Food and Drug Administration (FDA) to adjust the optimal intensity of long-term anticoagulant therapy in post-myocardial infarction patients, and to define the optimal anticoagulation level. She also took part in the FDA report for the approval of coronary stenting for selective placement in selected patients to prevent restenosis after angioplasty. The results were a milestone in interventional cardiology.

BRAND NAME: AN INTELLIGENT MOBILE-BASED ENVIRONMENTAL PROTECTION RATING AND SUGGESTION PLATFORM USING ARTIFICIAL INTELLIGENCE AND TEXT RECOGNITION

Ximeng Zhang[1] and Yu Sun[2]

[1]Yorba Linda High School, 19900 Bastanchury Rd, Yorba Linda, CA 92886
[2]California State Polytechnic University, Pomona,
CA, 91768, Irvine, CA 92620

ABSTRACT

Recycling is an essential measurement to change waste into reusable material. In the US, only about 30% of solid wastes are properly recycled, compared to other developed countries such as Northern Ireland (50.6%), Japan (50%), Schotland (46.9%), Wales (56.9%), it is much lower. However, in the US, the amount of solid waste disposal has increased in the past decade, which leads to air pollution, water pollution, soil pollution and solid waste is also a cause of many diseases. Specifically, it is noticed that many people have difficulty realizing how well they are doing in the process of recycling. Therefore, an app based on dart language is created to check how well people recycle through a scoring system and collect data from grocery receipts to see if the app can help make the consumers' receipts full of more recyclable items. And it is hypothesized that this app can increase efficiency in recycling and promote people to encourage individuals to use more recyclable items. A clear trend in my data of scores gained from my family grocery receipt each week shows that the amount of recyclables increase as the weeks go by since the number from the app did increase. The number I get from recycling pops up in my head as I do the weekly grocery shopping with my family and reminds me to buy more recyclable items. The app is proven helpful and does increase recycling efficiency to 95%. The product, as an app, will be widely used by smartphones.

KEYWORDS

Environmental protection, Artificial Intelligence, NLP.

1. INTRODUCTION

Recycling is an essential measurement to change waste into reusable material. In the US, only about 30% of solid wastes are properly recycled, compared to other developed countries. Recycling is a major problem in the US compared to many other developed countries. I compared recycling rates in the United States and other developed countries and found that the US is relatively behind. Therefore, I wish to create an app that would tell the user how nice they are doing their recycling directly and in a more efficient manner. This app would help users all around the country.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 59-67, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121105

In 2014, ERNST Worrell and Markus A. Reuter pointed out: With contemporary recycling literature scattered across disparate, unconnected articles, this book is a crucial aid to students and researchers in a range of disciplines, from materials and environmental science to public policy studies [1].

In 2021, Amar Bhutani put forward that plastic wastes should be recycled [9]. All the plastic wastes in the homes, shops and industry should be collected and sent for recycling to plastic making factories. In plastic factories, the waste plastic articles are melted and used, which means the importance of plastic recycling.

Also in 2021, Pradeep Singh thinks suitable plastic will be from Recycle. recycling recycled [10]. In the product, materials plastic harmful This of used The and means factory.

In 2020 , Heather Knowles put forward that turn your classroom into an environmentally friendly learning zone with these three articles [2]. In the same year, Trevor M. M Letcher thinks that this is an essential guide for anyone involved in plastic waste or recycling, including researchers and advanced students across plastics engineering, polymer science, polymer chemistry, environmental science, and sustainable materials [3].

In 2018, Beth Porter pointed out that Reduce, Reuse, Reimagine makes sense of the complex system for any reader who wants to learn how it works, what the problems are, and what they can do to help recycling thrive [6].

In 2016, Paul Bulteel, Nadine Barth [4]. considered the cycle & recycle" project brings a photographic view not only of waste streams but also of the efforts taking place in Europe to recycle waste on an unprecedented scale". Then in 2021, Jennie Romer thinks “If you’ve ever been perplexed by the byzantine rules of recycling, you’re not alone…you’ll want to read Can I Recycle This [5]?

In 2013, Edward Humes pointed out in Garbology, Edward Humes investigates trash—what’s in it; how much we pay for it; how we manage to create so much of it; and how some families, communities, and even nations are finding a way back from waste to discover a new kind [7].

In 2012 Jinger Jarrett recycled articles into other ebooks [8]. JinShanWei LoongAng En-HuaYang proposes a new mobile app-aided RANAS approach for recycling behavioral change in Singapore with the goal to increase household recycling rates while reducing the proportion of contaminants in recyclables [14]. Also Fábio Oliveira da Silva pointed without some type of inspection/incentive/appropriate disposal site or help, many inhabitants choose to dispose of garbage in a simpler and easier way [15].

From our investigations, recycling apps that are currently in the market are largely inefficient. Although the theory is so full, the people cannot easily and conveniently use it. The related method can be realized by using Dart. In the source code construction, it is designed by three components: the history page, the main page, and the login page. Different pages serve different functions. The history page keeps the login information and the login page is a machine-person interface, which is convenient to land on a smartphone. Lastly, the main page connects the whole system.

Two apps that we have specifically investigated are ‘iRecycle’ and ‘Recycle Coach.’ These two apps, along with many others can be great apps; however, they do not have enough features as we believe are helpful enough for people to understand if they themselves are doing enough recycling. ‘iRecycle’, for example, only contains the feature of finding recycle centers to recycle

differently, and 'Recycle Coach' also only contains information about recyclables and no information specific to the user. It does not focus on the user itself and does not contain statistics that the user would be able to tell straightforwardly if they have been doing enough recycling or buying enough recyclable items, which would help the environment straightforwardly.

Other apps we investigated does not focus on the user itself and does not contain statistics that the user would be able to tell straightforwardly if they have been doing enough recycling or buying enough recyclable items, which would help the environment straightforwardly. Our design is much better in how it directly approaches the user's needs.

The main function of our app includes asking the user to scan a grocery receipt and receiving a score that tells them how many recyclables they have purchased. A user can use this over a period of time to see if their levels of recycling have increased. After finishing developing the fully functional app, we collected user feedback from several user's families. The app is proven to be effective and useful.

The rest of the paper is organized as follows: Section 2 gives the details of 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 concluding remarks, as well as points 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. Changing original plan

The first challenge we encountered was that we needed to change from our original plan of scanning barcodes code into scanning a receipt. Barcodes codes sounded like a great plan at first; however, the method is proven largely inefficient because of how little information about the recyclables can be obtained from depicting barcodes. We believe that scanning words directly would give us better data of what exact products the recyclables are.

2.2. Developing new function

A user is now able to take a picture right now using our app and will get the test result instantly. However, we did not have this function at first, and we used an obnoxious system of having the user upload their own picture through many sources. Fixing this problem has proven to give our users better experiences using our app.

2.3. Some functions are divided

Some main functions of the code of our app are divided into the history page, main page, and log-in page. Right now the app is limited to three parts above. Later, we plan to expand the function of the app to other phases.

3. SOLUTION

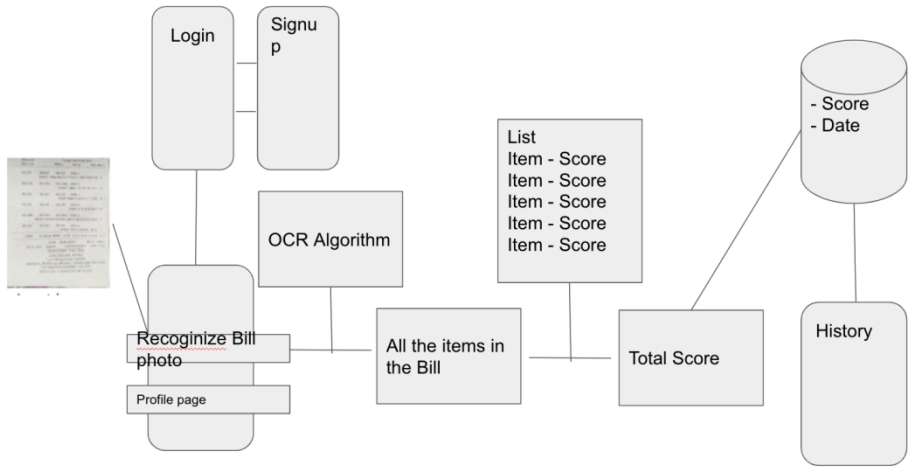


Figure 1. The System Structure

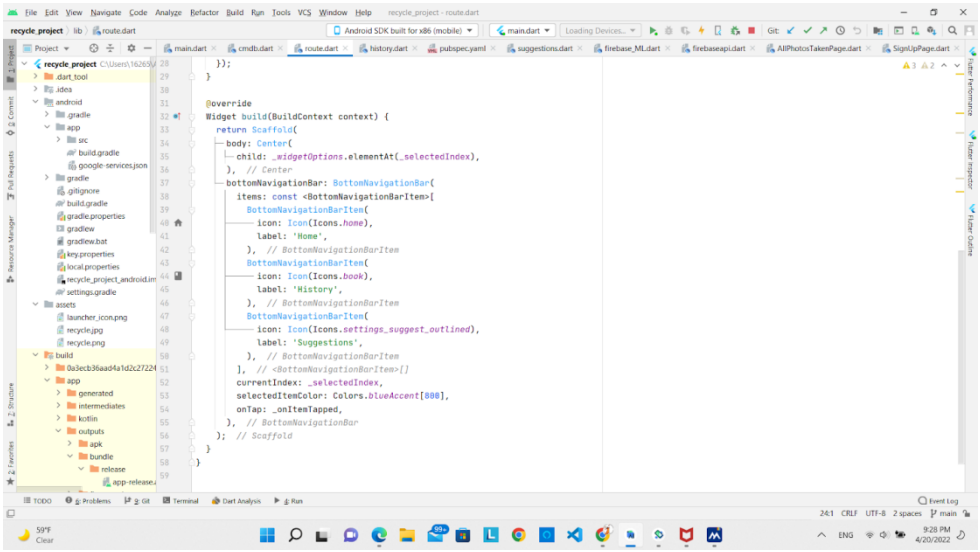


Figure 2. Main Page

We used this page to collect all other pages together. We used a scaffold that had three attributes. The first is the body. The second is the navigation bar, and the third the App Bar. Then, for the body, we used a center class. In the bottom navigation part, we had three different icons representing three different pages that we will use later. For the selected item color, we changed the color of the icon.

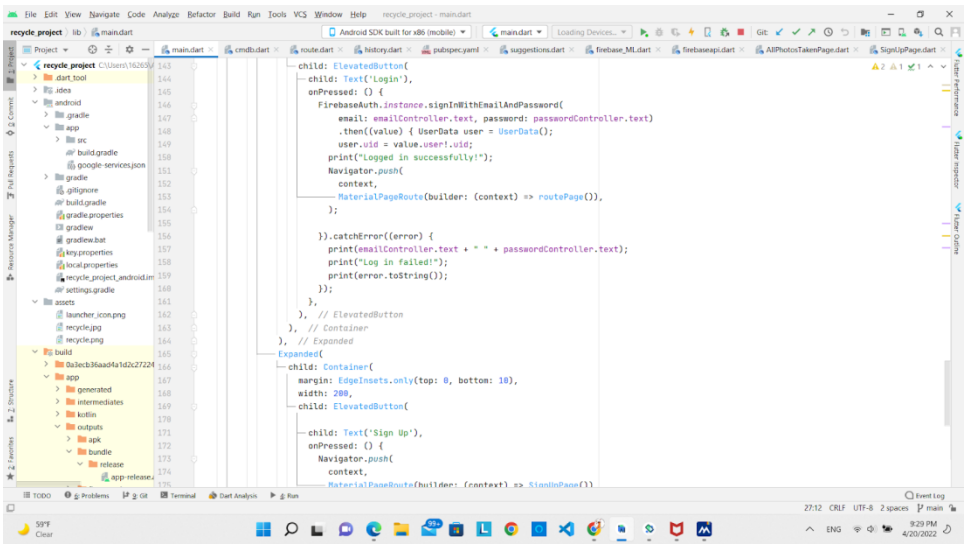


Figure 3. Login page

1. For the Login page, we use the Firebase database as the database to store the username and password.
2. For the UI design, we use an Elevated button to be clicked to send the request to the database to verify the login information.
3. If the data is incorrect, it will show the alert message.
4. We also have an Expanded class to transfer to Sign up page

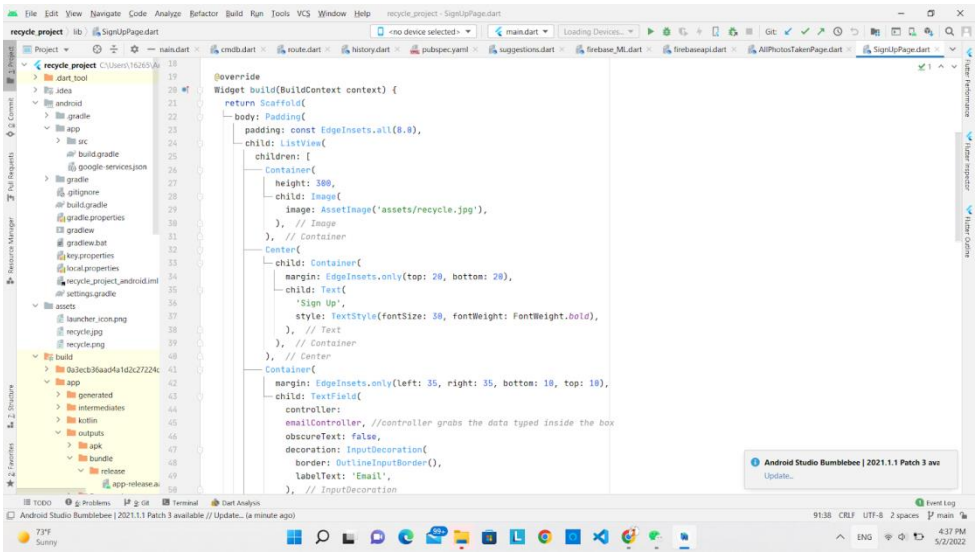


Figure 4. Sign Up Page

For the Signup page, we use the Firebase Auth database as the database to handle the email and password.

For the UI design, we use also an Elevated button to be clicked to send the request to the database to verify the login information.

The difference between this page and Login is that for this page we need user to input the Password information one more time to verify the code.

If the data is incorrect, it will show the alert message.

We also have an Expanded class to transfer to Login page.

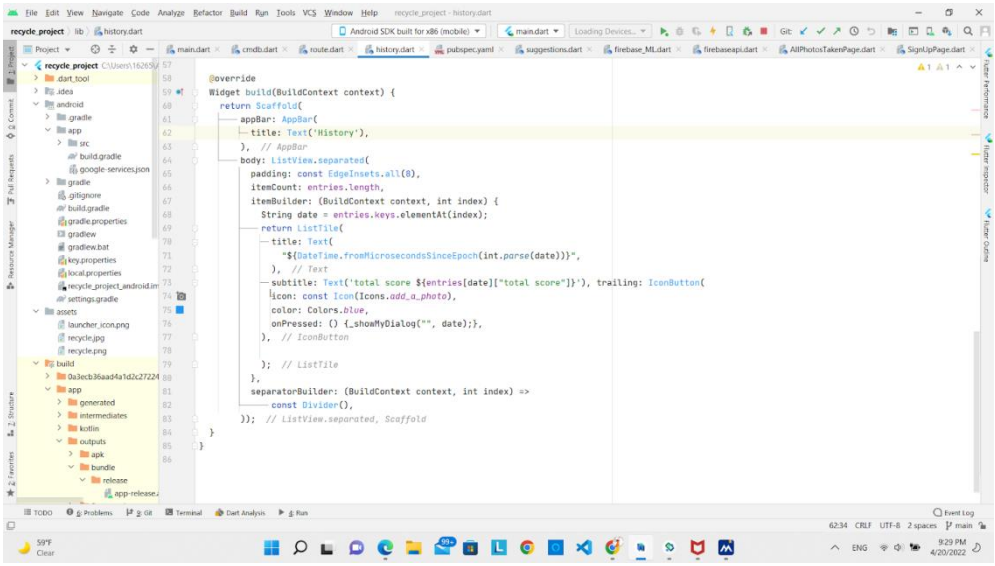


Figure 5. History Page

We used the Dart Language to conduct the algorithms each with different, interconnected pages.

These components are connected by the main page.

4. EXPERIMENT

4.1. Experiment 1

Recycling is a major problem in the US compared to many other developed countries. I compared recycling rates in the United States and other developed countries and found that the US is relatively behind. On the other hand, in the US, the amount of solid waste disposal has increased in the past decade, which leads to air pollution, water pollution, soil pollution and solid waste is also a cause of many diseases. Therefore, I wish to create an app that would tell the user how nice they are doing their recycling directly and in a more efficient manner.

We tried to determine whether that has improved households efficiency of recycling by analyzing the trend in the score the app returns of how environmentally friendly the user is. We collected samples from participating families over a period of time.

4.2. Experiment 2

Results show that the app can improve recycling efficiency, as the scores gained from our weekly receipts have increased. A clear trend in my data of scores gained from family grocery receipts each week shows that the amount of recyclables increase as the weeks go by since the number from the app did increase. We have also found from user feedback that one user says the

recycling number pops up in their head when they do the weekly grocery shopping with their family and reminds them to buy more recyclable items. The app is proven helpful and does increase recycling efficiency.

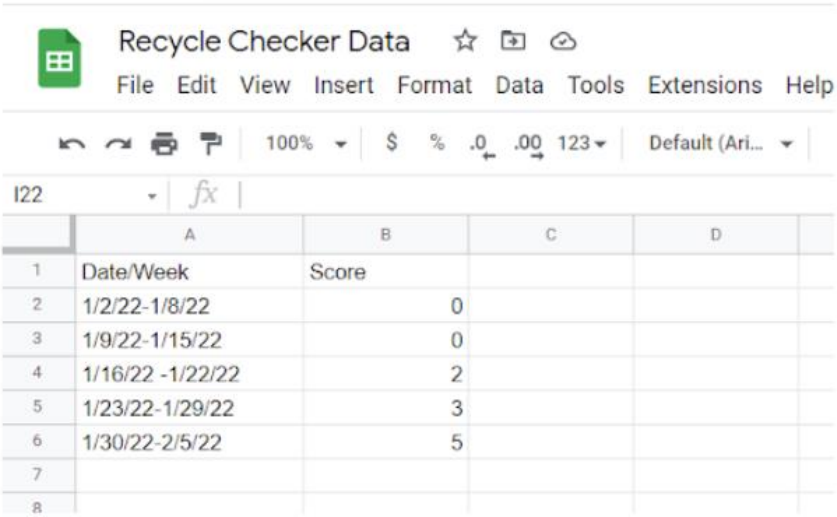


Figure 6. Data set from a specific group of family

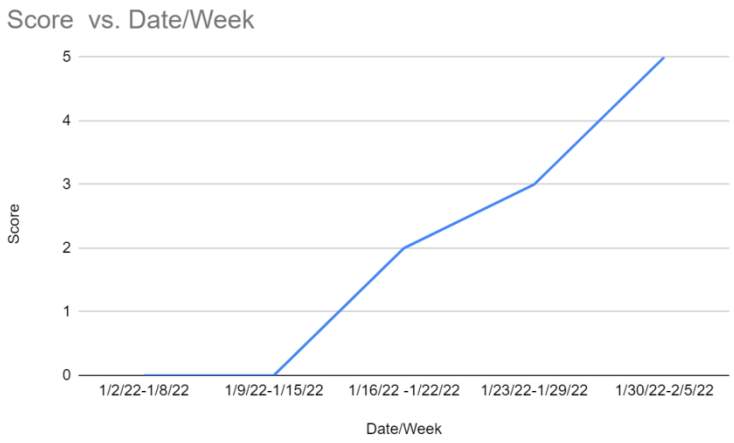


Figure 7. The trend of increase in our recycling score

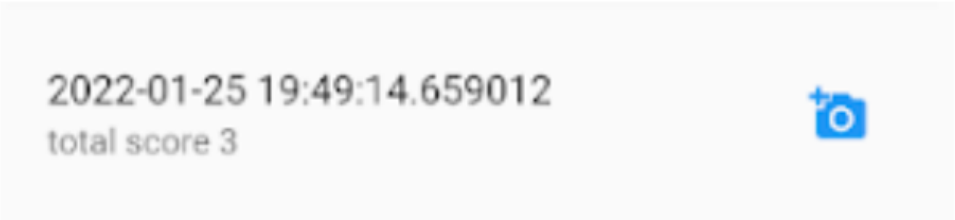


Figure 8. an example of data history that a user can reference anytime after they can the app

The app being proven as effective shows that it can indeed help to solve the current problem with the lack and inefficient recycling in the United States. Although we encountered many problems throughout the development process of our app, our app eventually is proven effective and working.

5. RELATED WORK

From Smart City to Smart Citizen: Rewarding Waste Recycle by Designing a Data-Centric IoT-based Garbage Collection Service [11]. 2020 IEEE International Conference on Smart Computing (SMARTCOMP). In this paper, they proposed a reward-based data-centric solution based on both enabling IoT technologies and cloud architectures to promote waste recycling in urban environments. We extend the consolidated rewarding approaches based on the smart bin model by proposing an incentive system that focuses on door-to-door waste collection. We designed an app based on the app, which will be more convenient for the customer.

Ombretta. Learning how to recycle waste using a game 2020 [12]. This paper teaches citizens how to recycle in real life. Proper waste sorting is crucial for both economic and environmental reasons; however, its effectiveness can be largely limited when citizens do not know how to correctly separate waste, sometimes even due to different regulations depending on their municipality. However, on the other hand, our project is an app, which is easier to use. Their main difference is whether they are installed on smart cell phones.

Filomena Compagno. Recycling 2020- Reduce, Reuse, and Recycle: The case Terracina-Filomena Compagno-Terracina Zero Waste activist, Italy [13]. In order to improve RECYCLE the Municipality of Terracina together with De Vizia, the local sanitation transfer distributed to the families and traders 5 bins for the door-to-door collection. They also created separate waste collection centers. Our project/app is much easier accessible and effective than their system.

6. CONCLUSIONS

In this paper I designed an app that can be easily used by any person, no matter consumer or seller, to check how well they are doing with recycling. The app returns a score to tell the user how well they are doing with the recycling using the scanning and uploading receipt function of our app. The user can see information of all receipts they have scanned. This can help the readers to clearly see what they need to improve in what they have been doing with recycling. After the app was developed, I tested grocery receipts from one set of families over a certain time period to test the functionality of our app, and it was proven to have great results. The app did indeed prove to increase recycling efficiency as seen from the data we collected. Experiment results indicate its effectiveness and solve challenges related to recycling.

Certain parts of the app can include a bit more useful and direct information. For instance, we can add a recycle knowledge button so that as time goes on, the customer can increase some aspects of recycling knowledge. We developed the base of this function in the suggestion feature of our app; however, much more features such as distributing information of recycling centers to the users can be added later. In addition, the practicability of the app is proven to be well functional. On the other hand, the algorithms I currently use need to increase accuracy. We plan to do this after modifications after more experiments. Optimization also needs to be done more accurately, which I will make better along with conducting more experiments.

In the future, I will add more features to make the app more convenient. I will modify the suggestion feature and add more information about recycling to help the user to the best of our

abilities. I will continue to collect feedback from users to make our app even more efficient and convenient for each user individually and to increase accuracy of our individual app.

REFERENCES

[1] Worrell, Ernst, and Markus A. Reuter, eds. Handbook of Recycling: State-of-the-art for Practitioners, Analysts, and Scientists. Newnes, 2014.

[2] Kibona, Deogratias, Gloria Kidulile, and Fredrick Rwabukambara. "Environment, climate warming and water management." Transition Studies Review 16.2 (2009): 484-500.

[3] Letcher, Trevor M., ed. Plastic waste and recycling: environmental impact, societal issues, prevention, and solutions. Academic Press, 2020.

[4] Gaines, Linda. "To recycle, or not to recycle, that is the question: Insights from life-cycle analysis." MRS bulletin 37.4 (2012): 333-338.

[5] Romer, Jennie R., and Leslie Mintz Tamminen. "Plastic Bag Reduction Ordinances: New York City's Proposed Change on All Carryout Bags as a Model for US Cities." Tul. Envtl. LJ 27 (2013): 237.

[6] Porter, Beth. Reduce, reuse, reimagine. Rowman & Littlefield,, 2018.

[7] Humes, Edward. Garbology: Our dirty love affair with trash. Penguin, 2013.

[8] Leonas, Karen K. "The use of recycled fibers in fashion and home products." Textiles and clothing sustainability (2017): 55-77.

[9] Vollmer, Ina, et al. "Beyond mechanical recycling: Giving new life to plastic waste." Angewandte Chemie International Edition 59.36 (2020): 15402-15423.

[10] Joseph, Blessy, et al. "Recycling of medical plastics." Advanced Industrial and Engineering Polymer Research 4.3 (2021): 199-208.

[11] Pelonero, Leonardo, Andrea Fornaia, and Emiliano Tramontana. "From smart city to smart citizen: rewarding waste recycle by designing a data-centric iot based garbage collection service." 2020 IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2020.

[12] Gaggi, Ombretta, et al. "Learning how to recycle waste using a game." Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good. 2020.

[13] Compagno, Filomena. "Recycling 2020-Reduce, Reuse, and Recycle: The case Terracina-FilomenaCompagno-Terracina Zero Waste activist, Italy." Journal of Nuclear Energy & Power Generation Technologies 4.1 (2020): 1-2.

[14] Shan, Xin, Wei Loong Ang, and En-Hua Yang. "Mobile app-aided risks, attitudes, norms, abilities and self-regulation (RANAS) approach for recycling behavioral change in Singapore." Resources, Conservation and Recycling 162 (2020): 105049.

[15] da Silva, Fábio Oliveira, and Paulo Duarte Branco. "Smart Recycle – Development of a waste collection application."

A NOVEL APPROACH TO NETWORK INTRUSION DETECTION SYSTEM USING DEEP LEARNING FOR SDN: FUTURISTIC APPROACH

Mhmood Radhi Hadi[1] and Adnan Saher Mohammed[2]

[1]Department of Computer Engineering, Karabük University, Karabük, Turkey
[2]Adnan Saher Mohammed, Karabük University, Karabük, Turkey

ABSTRACT

Software-Defined Networking (SDN) is the next generation to change the architecture of traditional networks. SDN is one of the promising solutions to change the architecture of internet networks. Attacks become more common due to the centralized nature of SDN architecture. It is vital to provide security for the SDN. In this study, we propose a Network Intrusion Detection System-Deep Learning module (NIDS-DL) approach in the context of SDN. Our suggested method combines Network Intrusion Detection Systems (NIDS) with many types of deep learning algorithms. Our approach employs 12 features extracted from 41 features in the NSL-KDD dataset using a feature selection method. We employed classifiers (CNN, DNN, RNN, LSTM, and GRU). When we compare classifier scores, our technique produced accuracy results of (98.63%, 98.53%, 98.13%, 98.04%, and 97.78%) respectively. The novelty of our new approach (NIDS-DL) uses 5 deep learning classifiers and made pre-processing dataset to harvests the best results. Our proposed approach was successful in binary classification and detecting attacks, implying that our approach (NIDS-DL) might be used with great efficiency in the future.

KEYWORDS

Network Intrusion Detection System, Software Defined Networking, Deep Learning.

1. INTRODUCTION

The architecture of traditional networks has not changed for decades to rum that it suffers from many problems and singled out security problems. Software-defined networking new solution or approach to address these problems, and it is characterized by many features that make it the future structure of the Internet. The most prominent feature of this network is that it is inexpensive, flexible, expandable, and increases the size of its infrastructure without the complexity of the traditional network. All operations in this architecture are controlled by a controller [1]. Instructions are exchanged between the controller and the switches via the OpenFlow protocol. The SDN architecture has many advantages, as it provided many solutions to the problems of the old network infrastructure, which made it the focus of attention and interest of authors [2]. OpenFlow protocol is based on the concept of different IP packets that are exchanged between the controller and the switches. SDN provided a comprehensive overview of the entire network through the controller controlling the entire network. The controller is considered the brain of the network, which is completely isolated from the network, and targeting it from attackers means the fall of the entire network. Accordingly, the controller is the most harmful part and the most affected by attacks. It is necessary to have a network intrusion

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 69-82, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121106

detection system (NIDS) located in the network that protects the SDN, especially the controller that is in the network part from attacks, detecting and reducing their impact. There are several types of NIDS, an approach that uses a signature, that relies on data from previous attack logs that are stored and requires continuous updating, is called the signature-based NIDS approach [3], and a second approach that uses anomaly detection that monitors the traffic pattern is more efficient and effective is called the NIDS approach Based on anomaly detection [4], which compares traffic behavior to normal and abnormal traffic. Machine learning is used with NIDS to identify attacks, but the efficiency is low. Within NIDS, a flow-based approach and anomaly detection are used together. Many factors have led to the lack of success and reliability of using machine learning in intrusion detection techniques in networks, the most prominent of which is the complexity to handle huge amounts of data that are unclassified where the performance and reliability of these systems are inefficient. Deep learning technology is a new and recent technology that predicts the possibility of solving machine learning problems, and it can deal with inconsistent data, find possible correlations, and give good and reliable performance. A reliable NIDS approach can be designed with accuracy and performance using deep learning. With deep learning, various attacks can be identified with high accuracy and with a high detection rate. SDN protection using NIDS based on deep learning is an effective method and a powerful defense mechanism. NIDS focuses on the detection of types of traffic as normal or abnormal behavior. Attacks cannot be completely prevented, but they can be detected early and identified, and their impact reduced if effective methods such as deep learning methods are used [5]. We propose a (NIDS-DL) approach for SDN using deep learning. More than one type of deep learning algorithm has been used to evaluate it based on several Metrics such as (Accuracy, F-score, Recall, Precision, etc.). we applied features selection methods to train our classifiers on high correlations features. The approach was applied to an NSL-KDD [6] dataset.

This paper is organized as follows: Section 1 Introduction. Section 2 is Related work that described some relevant previous work. Section 3 Proposed Methodology that clarified the proposed approach, also explains in brief classifiers model used and summary of architecture. Section 4 discussed the dataset and preprocessing methods applied. Section 5 Experiment results of the approach. Section 6 Study Comparative. Finally, Section 7 explains the conclusion and future work for the approach.

2. RELATED WORK

The application of machine learning systems with SDN has attracted the attention of many authors.

In [7] the author's purpose approach was based on five types of machine learning algorithms (RF, Naïve Bayes, SVM, CART, J84) to obtain an accurate and high-performance approach, this approach was applied to the NSL-KDD dataset with the employs 41 features, this approach achieved good detection accuracy in recognition of attacks and anomaly detection, the RF algorithm achieved the highest accuracy rate of 97%.

After the emergence of deep learning technology, several authors attempted to design several systems that use deep learning in NIDS for SDN in their approach. In [8] the authors built a deep learning-based network intrusion detection approach for the SDN environment, using the DNN algorithm in their approach. Six features from the NSL-KDD dataset used. The authors contrasted the outcomes of his approach with machine learning classifiers. The approach exhibited high detection accuracy and better performance than the machine learning classifier approach, demonstrating the feasibility and potential of using deep learning to construct network intrusion detection systems for SDN. the authors compared the results of the approach he used with machine learning classifiers.

Also, in [9] the same author proposed using a hybrid deep learning approach, the goal was to improve the accuracy and reach a better and more applicable approach, these approaches used two types of deep learning classifiers Gated recurrent unit and Recurrent Neural network to design a hybrid approach called (GRU-RNN), apply these approach was based on NSL-KDD dataset, where the author used in his approach six features in training the classifier. The hybrid approach method achieved 89% better accuracy and proved to be superior to the previous method, as well as its easy and flexible application in the SDN working environment.

Another work in [10] The goal of this approach was to build intrusion detection systems for SDN, the researcher used machine learning and deep learning systems to compare the results. A deep learning algorithm (GRU) was used in the approach, the algorithm achieved better accuracy and performance than machine learning classifiers, more than one type of dataset was used in training and comparison, six types of different attacks were categorized with a benign approach, the approach achieved great success indicating the possibility of applying deep learning in NIDS with great efficiency to SDN.

In this paper, several types of deep learning classifiers (CNN, DNN, RNN, LSTM, GRU) are applied. NSL-KDD dataset was used as the approach was applied to 12 features extracted. Each classifier was evaluated based on a different set of metrics. A broad approach to deep learning and its classifiers has been used to build a robust and effective NIDS system in detection and identifying attacks for future application within the SDN environment, which differs from the rest of the research in that it relies on more than metrics in assessment, not just accuracy and trying to get the best and highest result compared to related work.

3. PROPOSED METHODOLOGY

3.1. System Methodology Description

The adoption of most of the methods applied in the machine learning approach will become less effective with the development of attack and penetration systems and the tools used for them. Machine learning method needs more configured data and it also needs less data to process, moreover performance and accuracy become poor. Most of the methods that use deep learning, discussed by the authors, use classifiers. The classifier is mainly evaluated on the accuracy of the matching metric, and the accuracy is also low, which does not lead to building a reliable and efficient NIDS system to detect attacks.

All of these prompted us to build our methodology shown in Figure 1, this methodology is based on building the NIDS-DL approach for SDN, this approach uses more than one classifier for deep learning with training classifiers on 12 features extracted from 41 features in the NSL-KDD dataset, training the classifier on best correlation features will lead to the possibility of detecting various attacks. Applied feature selection method to select the best features that are effective and get correlations on the result, also the system will be powerful and reliable against attacks. The approach is evaluated on several Metrics and the classifiers are compared with each other.

In our approach, we evaluated CNN, DNN, RNN, LSTM, GRU classifiers are used, Results are compared where the (normalization) mechanism is used on the data to speed up the training process and get the best possible outcomes for generating an efficient NIDS classifier, also using feature selection method to avoid missing in training algorithm and try to reach the best accuracy and performance through selecting the best feature for training.

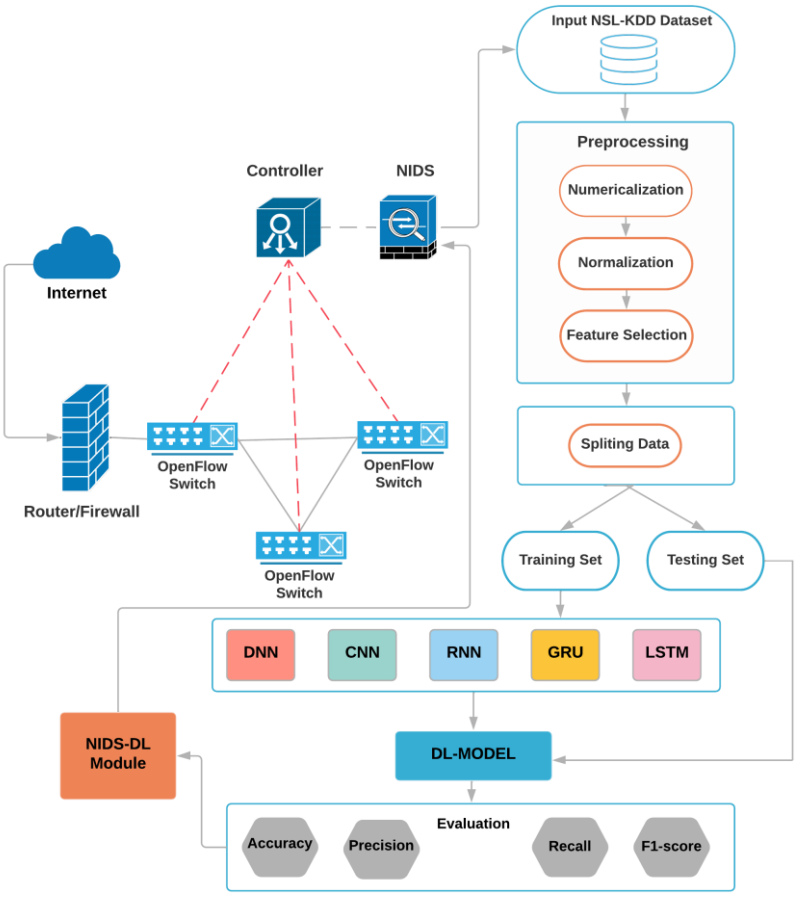


Figure 1. Proposed Methodology for (NIDS-DL) in SDN

3.2. Model Classifiers

In our study we use DNN, CNN, RNN, LSTM, GRU classifiers, architectures summary is given visualization in Figures 2-6.

a) DNN is a deep neural network is a developed class of a simple neural network. a deep neural network is called when it consists of more than three hidden layers. Increase the number of hidden layers, will be led to need for additional computer resources for processing, also that will raise ability and efficiency to process a large amount of data.

b) CNN is a convolutional neural network that processes and classifies input in the form of images. This type of neural network has the property of extracting information and reducing features and this is reason makes it widely used in most applications. CNN uses a feed-forward feature when processing.

c) RNN is Recurrent neural networks are also considered one of the simple neural networks, also considered a powerful type developed in the eighties. The most important thing that distinguishes this type and makes it a strong type is that it contains the internal memory

d) LSTM is Long short-term memory is one of the types of a type of RNN. It came to address several problems that the RNN suffers from. LSTM has the feature of retaining data and information stored for a long period.

e) GRU is Gated Recurrent Unit is also a type of standard recursive network. The specific architecture and interior design are similar to LSTM. Gated Recurrent Unit is designed to address the vanishing gradient problem in RNN.

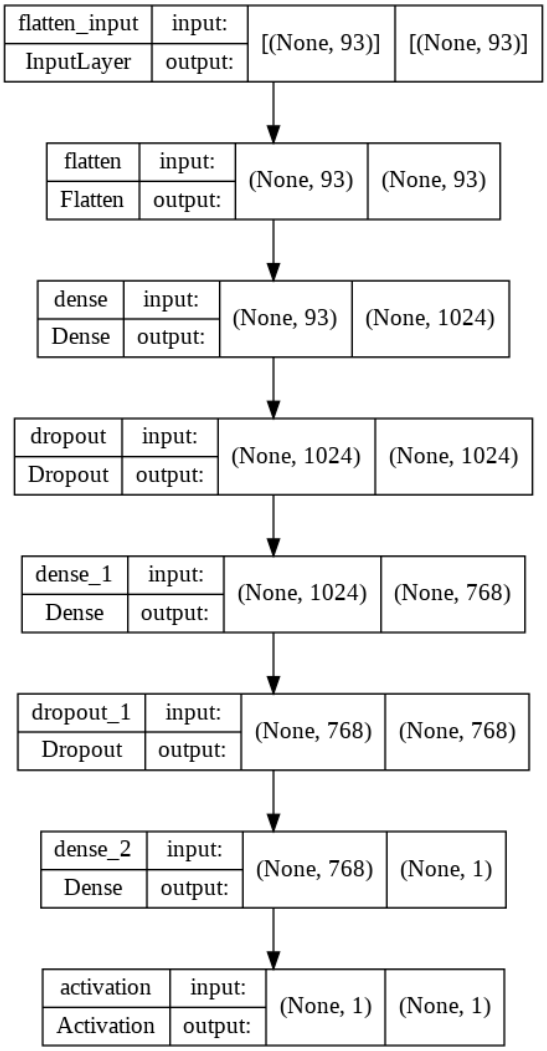


Figure 2. Summary of DNN model.

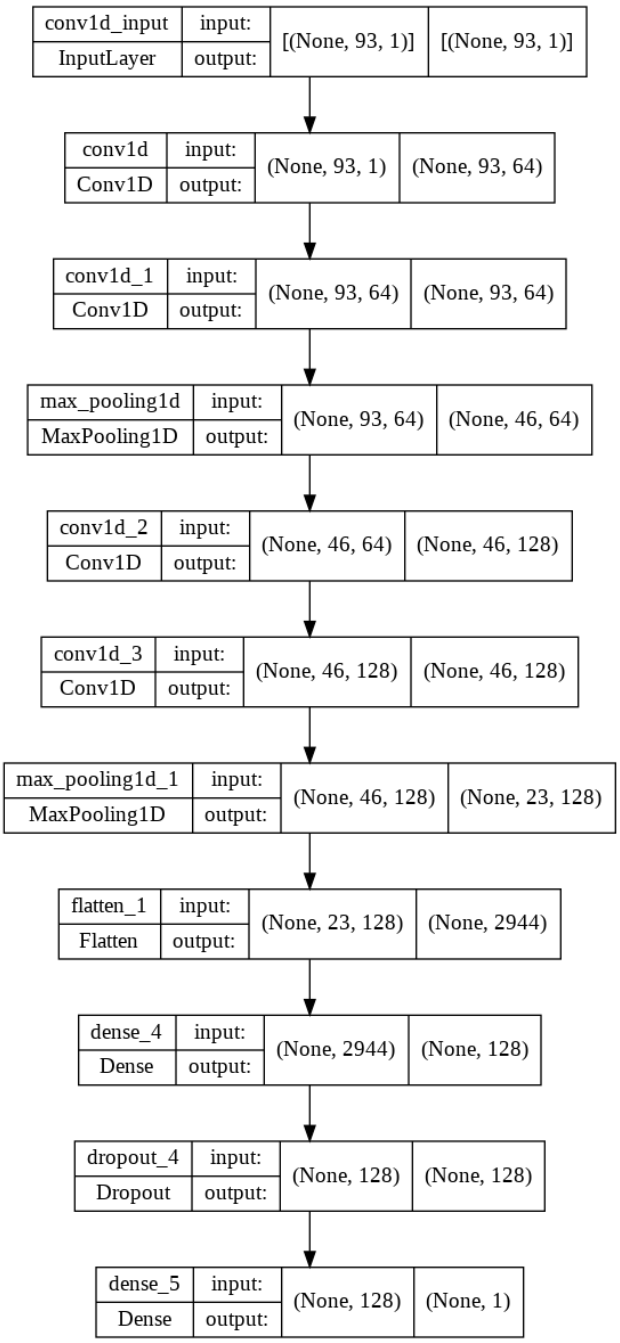


Figure 3. Summary of CNN model.

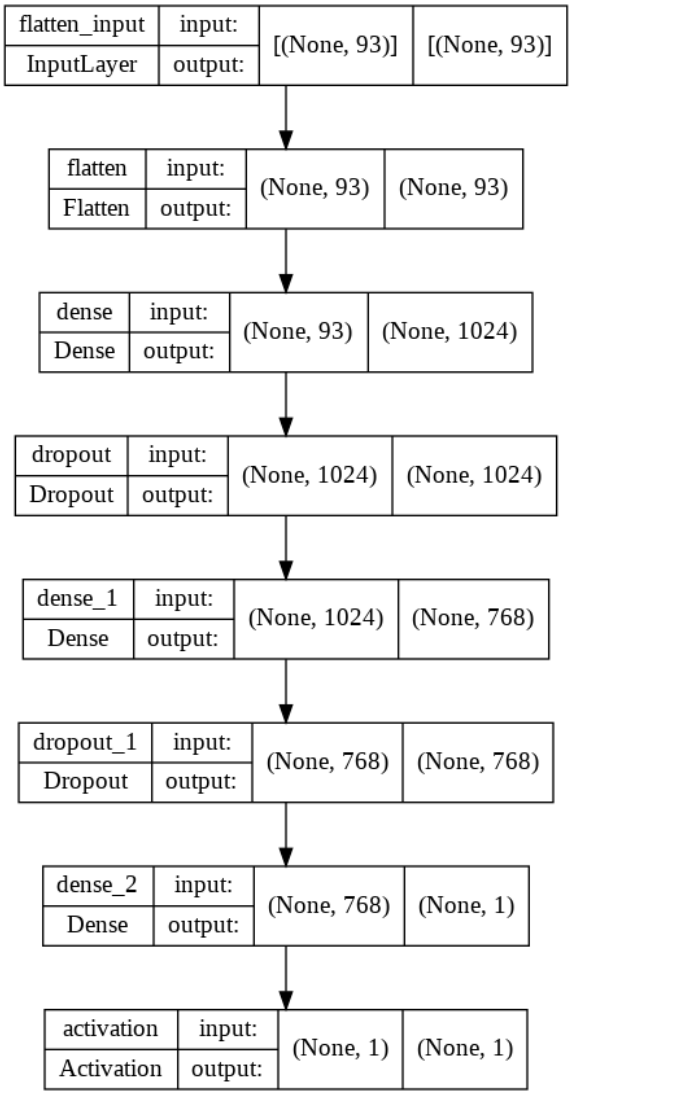


Figure 4. Summary of RNN model.

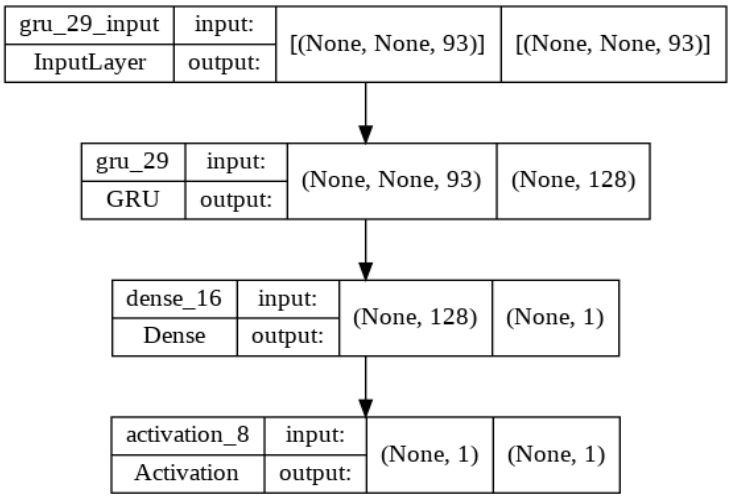


Figure 5. Summary of GRU model.

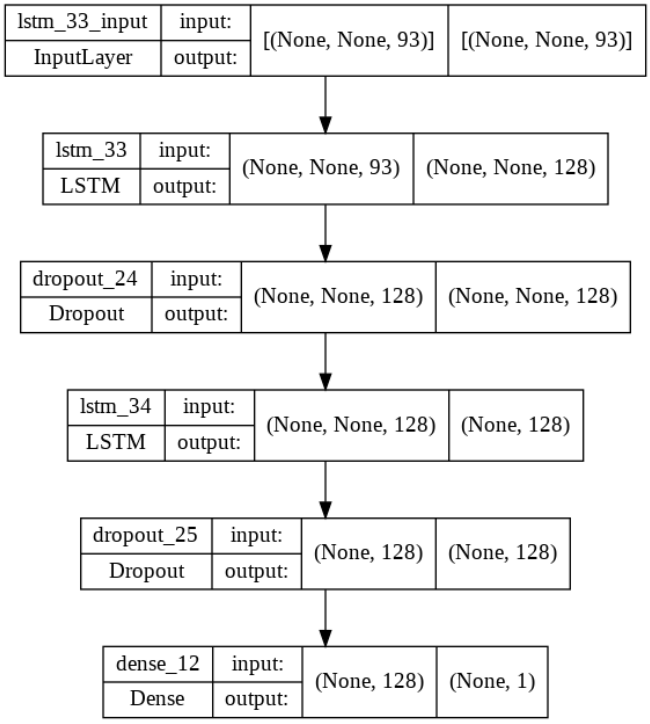


Figure 6. Summary of LSTM model.

4. DATASET

In this part, we will discuss the NSL-KDD [11] dataset that was used in our proposed approach. The NSL-KDD dataset is an update and development of the KDDCup99 dataset [12], which is much older than it was proposed in 1999, as it contained several problems and contained null or it is a recursive dataset which many of its problems have been solved in the NSL-KDD dataset, but this does not mean that it does not contain mistakes. NSL-KDD contains 41 features, we extracted 12 features are more corrections using the feature selection method. NSL-KDD is used as a simulator for network data and internet traffic as it was used in several research and applied by authors in their approach. The main feature of the NSL-KDD dataset that made it preferable to many authors is that its size is almost consistent and contains reasonable several features that help in obtaining the best and most reliable classifiers.

4.1. Data Preprocessing

In this section, we will discuss the methods used in preprocessing datasets.

4.1.1. Numericalization

To handle the NSL-KDD dataset into deep learning classifiers, all data must be in numeric format. The NSL-KDD dataset contains three non-numeric features and 38 numeric features. The features are converted to numeric form so that they can be handled by classifiers after they are converted to array form. The features that are converted are ('flag', 'service', 'protocol_type'). For example, the feature ('protocol_type') contains three types of data ('icmp', 'udp','tcp'), which are encoded into (1,0,1), (1,1,0), (0,0,1). After using this method, all the 12 turns into a map of 122-dimensions.

4.1.2. Normalization

The normalization mechanism is applied for several tasks, it is used to speed up the training process for classifiers as it works to make the data set consistent and make the difference between the data small when we have the difference between the big and small data is large. Among the features in the NSL-KDD data set and contains the difference between its data are dst_bytes [0,9.11×109], duration [0,58329], src_bytes [0,9.11×109]. The formula shown in 1 is applied, which transforms the data range and makes it between [0,1].

$$xi = (xi - Min) / (Max - Min) \quad (1)$$

4.1.3. Feature Selection

In this processing method, we extracted the features that are most correlated to the target feature, and the purpose is to reduce the loss of the classifier during training and try to get the best accurate results and high performance. Table 1. illustrates 12 features extracted from the NSL-KDD dataset.

Table 1. Feature extracted from NSL-KDD dataset.

No.	Features	No.	Features
1	protocol_type	7	srv_serror_rate
2	service	8	same_srv_rate
3	flag	9	dst_host_srv_count
4	count	10	dst_host_same_srv_rate
5	logged_in	11	dst_host_serror_rate
6	serror_rate	12	dst_host_srv_serror_rate

4.1.4. Data Splitting

The features are a selection from NSL-KDD Dataset are splitting by 75% for training and 25% for testing. Table 2. Showing partitioning of training and testing data into the NSL-KDD dataset with 12 features.

Table 2. A distribution instance of the NSL KDD dataset.

	Training set	Test set
Number of instances	107,077	18,896

4.2. Evaluation Metrics

NIDS performance is evaluated by several different metrics, the most prominent of which are Accuracy (AC), Precision (P), recall (R), and F1-score (F). These metrics must be of the highest value, especially the accuracy on which NIDS reliability depends. Another is centered which is the confusion matrix within which several parameters are calculated. One of these parameters is True Positive (TP), which indicates the number of attacks that are successfully categorized as attacks. True Negative (TN) represents the number of ratings of normal records that are correctly categorized as normal. False Positive (FP) refers to the number of normal records that are incorrectly classified as attack records. False Negative (FN) indicates the number of records for attacks that are incorrectly categorized as normal records.

- Accuracy (AC): Calculate the total number of true classifications.

$$AC = (TP+TN) / (TP+TN+FP+FN) \quad (2)$$

- Precision (P): It calculates the true classifications that NIDS can predict.

$$P = TP / (TP+FP) \quad (3)$$

- Recall (R): It calculates the number of correct classifications compared to each intrusion.

$$R= TP / (TP+FN) \quad (4)$$

- F1-score (F1): It is a method for calculating the harmonic mean of precision and recall.

$$F1= 2\times precision\times recall / precision \times recall \quad (5)$$

5. IMPLEMENTATION OF MODELS

Our classifiers are trained in the Google Collab environment using the Keras and Scikit-Learn libraries. The DNN, CNN, RNN, LSTM, GRU models were trained with 100 epochs and using the Adam optimizer with a learning rate of 0.01 for all classifiers. The loss type for all classifiers is also binary cross entropy with a validation distribution of 0.2.

6. EXPERIMENT RESULTS

The goal of our approach is to try to get the best results for several metrics. The approach was made and implemented using the Python 3.5.6 programming language, also using (TensorFlow, Keras) with (NumPy, Pandas) library for preprocessing. The computer Hardware configuration is (Intel i7-2720 QM, 16 GB of RAM, AMD Radeon 2 GB, 256 GB SSD).

The algorithm results are presented for all algorithms in our approach using the metrics in (Accuracy, Precision, Recall, F1 score). The CNN classifier performed better than the other classifiers used in the metric (Accuracy, Precision, F1-score), with results (0.9863, 0.9845, 0.9872), respectively. The DNN classifier showed good results and was ranked after the CNN classifier by metrics (Accuracy, Precision, F1 score) and the results were (0.9853, 0.983, 0.9863) or better than these results. The rest of the classifiers except CNN. The RNN classifier obtained the best result in terms of metric (Recall) with (0.9902), outperforming all classifiers. The results of the LSTM algorithm are metrically similar (Recall) to GRU, in that it also obtains results with the metric (Accuracy, Precision, Recall, F1-score) giving the corresponding results (0.9804, 0.9767, 0.9856, 0.9816). The GRU classifier generated the (accuracy, precision, recall, F1 score) scores (0.9813, 0.98, 0.98, 0.982) respectively, resulting in the lowest score compared to the other classifiers. The GRU classifier gets close results and it looks like a valuable result, but it is low compared to the other classifiers shown in Figure 7 and Table 3.

Table 3. Evaluation Metrics Classifiers.

DL-Algorithm	Accuracy	Precision	Recall	F1-score
DNN	0.9853	0.983	0.9896	0.9863
CNN	0.9863	0.9845	0.9898	0.9872
RNN	0.9813	0.9751	0.9902	0.9826
LSTM	0.9804	0.9767	0.9856	0.9816
GRU	0.9778	0.973	0.9856	0.9793

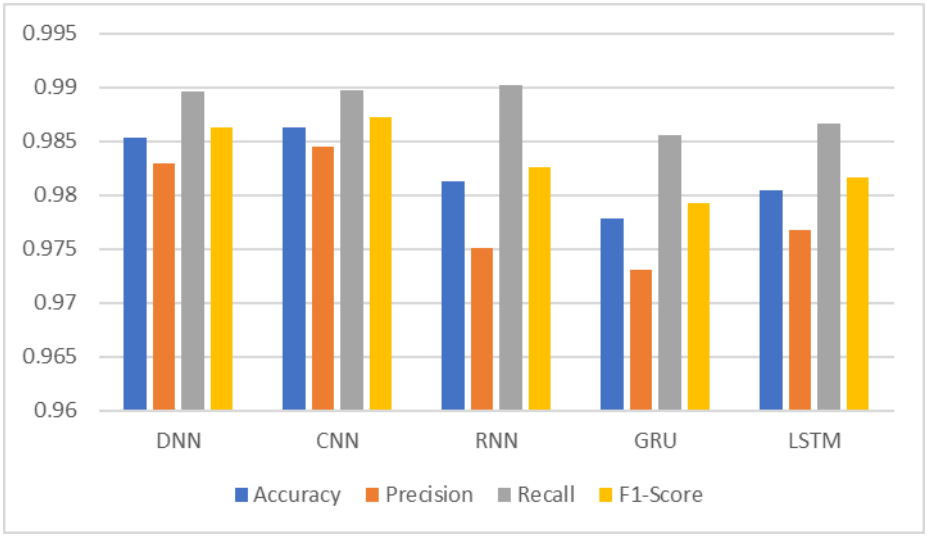


Figure 7. Evaluation Metrics of Deep learning classifiers

The goal of classifiers during their evaluation on the confusion matrix is to obtain the highest value of the measures (TP, TN) and reduce the value of (FP, FN) as much as possible. The CNN classifier has the highest value (TP) and the lowest value (FP) as shown in Table 9. The RNN classifier has a higher value (TN) than all other classifiers. The DNN classifier got results in (TP, TN) higher than the classifier LSTM, GRU and also higher RNN in the parameter (TP). The rest of classifiers like LSTM achieve better results in parameter (TP, TN, FP, FN) than GRU classifier, the results of these algorithms are shown in Table 4.

Table 4. Evaluation Metrics Classifiers.

DL-classifiers	Confusion Matrix-Parameters			
	TP	TN	FP	FN
DNN	14433	16601	287	173
CNN	14460	16604	260	170
RNN	14262	16611	424	163
LSTM	14326	16550	394	224
GRU	14262	16534	224	240

Another important metric, such as ROC (Receiver Operating Curve), by which the results of deep learning classifiers are evaluated, are shown in Table 10. The results of the algorithms DNN, CNN are similar, so the result of the classifier is (0.998). The algorithms RNN and LSTM also obtained the same results (0.997), the GRU algorithm obtained (0.996) as in Table 5.

Table 5. ROC Metrics.

Algorithm	ROC
DNN	0.998
CNN	0.998
RNN	0.997
LSTM	0.997
GRU	0.996

7. STUDY COMPARATIVE

In this section, we will discuss and compare our approach to results with another related study.

In [8], the author had a detection accuracy of 75.75% on the binary classifier. Similarly, the same author in [9] achieved a detection accuracy result of 82.02% using the hybrid approach from the deep learning classifier. The author in [13] achieved a detection accuracy of 93.72% using the LSTM classifier. In [14], more than one machine learning classifier was used and good results were obtained. Compared with previous results and methods, our approach provides an accurate description of the methods used to process the data set, and it uses multiple classifiers to measure the impact of the same method used for the results, in addition, our approach is also based on the extraction of features that affect the results, leading to the performance of the training and high detection process. Our approach to evaluating results also relies on a variety of different metrics. A comparison of the studies is presented in Table 6.

Table 6. Accuracy Result Comparison with another Study Related.

Ref.	Method	Dataset	Accuracy
[8]	DNN	NSL-KDD	75.75 %
[9]	GRU-RNN	NSL-KDD	82.02 %
[13]	LSTM	CSIC 2010	93.72 %
[14]	LR, SVM, DT, RF, ANN	DS2OS traffic traces	98.3 % 98.2 % 99.4 % 99.4 % 99.4 %
Our Method (NIDS-DL)	**CNN** **DNN** **RNN** **LSTM** **GRU**	NSL-KDD	98.63 % 98.53 % 98.13 % 98.04 % 97.78 %

8. CONCLUSION

In this paper, more than one type of deep learning algorithm is used and applied to detect abnormality in NIDS. The approach was evaluated on different metrics and the approach achieved high and reliable results. One of the most contributions of this work is using the feature selection method to train the classifiers on most feature correlations and avoid miss led during training to reach the best result. Our approach focused on binary classification using deep learning algorithms. The results of the algorithms are compared with each other, the results of some classifiers are close, and the CNN classifier achieved the highest results. The use of deep learning demonstrated the possibility and superiority when applied in the binary classification of

network intrusion detection systems. Since the proposed approach harvest high results, future work will be to evaluate the results of classifiers on more than one type of dataset and compare the results. A hybrid approach of deep learning algorithms can also be used as a future work, and its results compared with our approach. These approaches can also be used to detect a specific type of attack, such as (DOS) attacks also we apply this approach inside SDN environment.

REFERENCES

[1] "Software-Defined Networking (SDN) Definition - Open Networking Foundation." https://opennetworking.org/sdn-definition/ (accessed Apr. 25, 2022).

[2] McKeownNick *et al.*, "OpenFlow," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Mar. 2008, doi: 10.1145/1355734.1355746.

[3] S. Sangeetha, B. Gayathri Devi, R. Ramya, M. K. Dharani, and P. Sathya, "Signature Based Semantic Intrusion Detection System on Cloud," *Adv. Intell. Syst. Comput.*, vol. 339, pp. 657–666, 2015, doi: 10.1007/978-81-322-2250-7_66.

[4] S. K. Dey and M. M. Rahman, "Effects of Machine Learning Approach in Flow-Based Anomaly Detection on Software-Defined Networking," *Symmetry 2020, Vol. 12, Page 7*, vol. 12, no. 1, p. 7, Dec. 2019, doi: 10.3390/SYM12010007.

[5] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," *Proc. - IEEE Symp. Secur. Priv.*, pp. 305–316, 2010, doi: 10.1109/SP.2010.25.

[6] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symp. Comput. Intell. Secur. Def. Appl. CISDA 2009*, Dec. 2009, doi: 10.1109/CISDA.2009.5356528.

[7] S. Revathi and A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection," *Int. J. Eng. Res. Technol.*, 2013.

[8] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep learning approach for Network Intrusion Detection in Software Defined Networking," *Proc. - 2016 Int. Conf. Wirel. Networks Mob. Commun. WINCOM 2016 Green Commun. Netw.*, pp. 258–263, Dec. 2016, doi: 10.1109/WINCOM.2016.7777224.

[9] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep Recurrent Neural Network for Intrusion Detection in SDN-based Networks," *2018 4th IEEE Conf. Netw. Softwarization Work. NetSoft 2018*, pp. 462–469, Sep. 2018, doi: 10.1109/NETSOFT.2018.8460090.

[10] I. I. Kurochkin and S. S. Volkov, "Using GRU based deep neural network for intrusion detection in software-defined networks," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 927, no. 1, Sep. 2020, doi: 10.1088/1757-899X/927/1/012035.

[11] "NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/nsl.html (accessed Apr. 25, 2022).

[12] "KDD Cup 1999 Data." http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed Apr. 25, 2022).

[13] S. Althubiti, W. Nick, J. Mason, X. Yuan, and A. Esterline, "Applying Long Short-Term Memory Recurrent Neural Network for Intrusion Detection," *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2018-April, Oct. 2018, doi: 10.1109/SECON.2018.8478898.

[14] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. M. A. Hashem, "Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches," *Internet of Things*, vol. 7, p. 100059, Sep. 2019, doi: 10.1016/J.IOT.2019.100059.

AUTHORS

Mhmood Radhi Hadi is a master's student of Computer engineering at Karabük University, Turkey. Before joining Karabük university in 2020, he was getting BSc degree in 2019 from Networking engineering at Iraqi University, Iraq. Before to university engagement he was working as a graphic designer, and through the BSc journey he was a working as internship and training Network simulation in Cisco academy also through the MSc period he was getting internship in Curve AI company as A Machine Learning. His research interests are Network security using AI/ML/DL, Software-defined Network, wireless and communication, Intelligent system.

Adnan Saher Mohammed received his B.Sc. degree in computer engineering technology in 1999 from Northern technical university, Mosul, Iraq. In 2012 obtained an M.Sc. degree in communication and computer network engineering from UNITEN University, Kuala Lampur, Malaysia, and in 2017 received Ph.D. degree from Ankara Yildirim Beyazit University, Ankara, Turkey. He is currently assistant professor at Karabük university, Turkey. His research interests include computer networks, Algorithms and Artificial Intelligence.

QUALITY INCREASES AS THE ERROR RATE DECREASES

Fabrizio d'Amore

Department of Computer, Control and Management Engineering
Sapienza University of Rome, Italy

ABSTRACT

In this paper we propose an approach to the design of processes and software that aims at decreasing human and software errors, that so frequently happen, making affected people using and wasting a lot of time for the need of fixing the errors. We base our statements on the natural relationship between quality and error rate, increasing the latter as the error rate decreases. We try to classify errors into several types and address techniques to reduce the likelihood of making mistakes, depending on the type of error.

We focus on this approach related to organization, management and software design that will allow to be more effective and efficient in this period where mankind has been affected by a severe pandemic and where we need to be more efficient and effective in all processes, aiming at an industrial renaissance which we know to be not too far and easily reachable once the path to follow has been characterized, also in the light of the experience.

KEYWORDS

Errors, Quality, Processes, Governance, Digitalization, Digital hygiene, Information technology, Computer Science.

1. INTRODUCTION

This paper is not aiming at presenting an innovative computer science's result, rather it wants to make people more sensitive and ready to an approach to organization, governance and design aimed at greater effectiveness and efficiency, as deriving from the increase in the overall quality of the process under consideration, because of the reduction in the quantity of errors, of any nature. The main question we want to answer to is "how can digitalization best help innovation and development?" This isn't exactly research on computer science, rather is research on how computer science, or (better) information technology, should be interfaced to other human activities aiming at innovation and at supporting their development, and trying the use resources at best, and not for fixing errors or solving computer problems. The tools are old at least two decades but we have seen that, because of a strong motivation like the pandemic, solutions technically available twenty years ago have today allowed to increment our efficiency, being however yet far from the best we can obtain.

It is well-known that unexpected (often unintentional) events, can seriously damage any process, and the time spent recovering from the error is far greater than what it would have taken if the process would have gone without surprises. Further, ore we observe that way we recover from the error is frequently improvised due to a typical college education that does not focus on error handling.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 83-89, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121107

Yet we consider important the subject and the recovery process, because an education at such subject, carried out using consolidated and tested effective methodologies, would lead to a better handling of the error situation. Hence the necessity to collect best practices and conceptual frameworks for obtaining a powerful set of instruments for error mitigation.

In this paper we'll often use the terms "user" and "operator." The two words should be meant similar, but we want to address the fact that the former is more generic, and referred to someone not expert of information technology, the latter is referred to a user that must accomplish simple task such as data entry. Also, we'll use the terms "computer science" and "information technology." The former is to describe a wide discipline, also having theoretical problems, and on which much research has being carried out; the latter references to some of the practical repercussions obtained from the former, which impact on human activities.

We wish to point out that we are not aware of any other work like this one, so we must present our model as new research that cannot cite previous literature nor make comparisons. Nevertheless, we feel that new research, addressing well-known problems of the transfer of computer science to information technology, deserves anyway to be considered, and looks extremely modern. In the same way, we cannot yet carry out experimental evaluations of what we propose, because what we are introducing is not yet so well defined as to be able to identify precise tools, although some can be easily imagined. However, part of our statements is referring to well-known issues and close to the common experience of many actors.

The paper has the following structure. Section 2 introduces an initial classification of errors and emphasizes the importance of digital hygiene; Section 3 presents a discussion on the topics introduced; and Section 4 draws some conclusions.

2. TYPES OF ERROR

Here we give a first classification, without claiming to be all-encompassing. The attempt is to take into consideration the most frequent or significant causes of error. We dedicate a subsection to each recognized cause.

2.1. Governance Errors

In many organizations, especially SMEs, the management is not educated at the information security, leaving decisions to be taken by IT people, often a handful of people. Yet, IT people are not educated at governance decisions, such choosing the appropriate policy for some given class of documents and are not completely aware of impact on the organization of the requirements of the information security [1]. For instance, choosing what category of documents should be confidential is a strategic decision and should not be taken by a technician; it is in fact the responsibility of the top management. Technicians will choose how to ensure confidentiality of some documents, like determining [2] the best level which to introduce the encryption in.

Another governance error is associating wrong or ambiguous requirements to information. This is a strategic error, and every organization should define a method for leaving users to let the management know such errors.

2.2. Operator Errors

Operators can apply policies in a wrong way. It is a mistake done during a procedure and (based on the professionalism of the operator) the organization should not define a standard countermeasure.

We believe that typos are a very common type of error; every human can make them. Hence, the need to minimize the writing at operator side, by letting the counter-interested, or the owner of the information, type it (probability of typos is much minor), and then help the data input by means of automated procedures. For instance, (s)he could type and check information, have an automatic tool producing a corresponding QR-code, to be shown to the operator, that should only use an automatic reader for inputting data. To this purpose it is appropriate to mention the huge research done on QR-codes and other types of bidimensional barcodes in the last decade, their improvement and efficientization; see e.g., [3, 4, 5, 6, 7, 8]. In general, for operators, the less they write, the better.

2.3. Omissions

To omit information or suitable details should be considered an error: often it is very expensive to retrieve proper information when one realizes its lack. The repetition of similar problems enables the operator to be able to predict in advance that it will need certain details, so that a retrieval procedure can be provided for time, perhaps by asking it to IT staff. The goal is to abolish every omission.

2.4. Software Issues

We know that various problems can arise from using applications: bugs, functions missing or hidden, etc. Of course, we don't want to tell software engineers how to design and test the software and its user-interface. However, we focus the fact that engineers view could be different from users/operators view. Therefore, we recommend that for the entire cycle of life of the software product some expert user/operator should complete the development team.

A typical source of issues is the duplicate input of information, or the input of information already available, perhaps on other software platform. All computer scientists know that this is increasing the probability of errors, due to possible problems of consistency and maintenance. With respect to this question, we propose an enlargement of the security by-design paradigm [9] by letting the design process include the awareness about other applications/platforms and we call it "awareness by-design." The same arguments that motivate the security by-design approach are at the base of the awareness by-design. This means that it should be a rule to design new software without ignoring pre-existing one and letting new and old platforms able to exchange data and avoiding any duplication. Of course, this is not always possible, especially due to the vendor lock-in policies, that prevent the open approach. To this matter we observe that the open approach should be pursued at any cost, not only for avoiding the vendor lock-in, but to make it simpler the exchange of data between platforms. Vendors will do not love this approach, because apparently in contrast with their profit goals, but this is a myopic vision because in the long run they will benefit from the open approach, the preferred one (or so it should be) by industries, public administration, universities, and all other organizations. Closed platforms should be abandoned, because strongly anti-economic.

Yet about software we mention the need of satisfying the information security requirements, and this can be done in several ways. However, we'd choose the cryptographic mode, especially if information-theoretically secure, rather a traditional approach at an application level, because an

attacker can more easily make an application crashing, for some unexpected input or other application-level detail but cannot break solutions that are information-theoretically secure.

2.5. Inventory not Up to Date

In many cases information about hardware and software are incomplete. This is due to a fast-evolving situation, or to a provisional setting (only temporary), or other causes. A partial description of the inventory is the source of issues when handling critic events, like incident handling, update/upgrade handling, decisions about cloud services and others. At critical moments incomplete/wrong inventories could lead to neglect special or relevant cases, what could greatly increase the inefficiency and the waste of resources.

2.6. Incidents not Well-Managed

Incidents and anomalies should be managed keeping into account the appropriate information security requirements. Accounting and non-repudiation should be guaranteed in such a way that every action can be attributed to one subject, and this cannot repudiate it. Information on incidents/anomalies should be quickly collected and made available. Of course, an outdated inventory would make the handling much more complicated.

2.7. Lack of Digital Hygiene

By "digital hygiene" we mean that part of cybersecurity that intersects the daily life of operators and users. Provocatively, we claim that a correct digital hygiene would make useless an antivirus, because some natural caution would do the job. And this hygiene should be spread as much, by addressing which behaviors could be virtuous and which could be reprehensible. We need real awareness, especially regarding the well-known social engineering, and passwords management. The view of complicated rules for creating a new password (e.g., a smaller-case letter, one capitalized, a figure, a special character, etc.) is at this point obsoleted: the increase in security with stronger passwords is completely (and beyond) balanced with the increase of insecure behaviors. Better to resort to alternative authentication methods, also based on multi-factor authentication, which have existed for some time but are struggling to overcome the traditional user/password scheme. Very instructive the Schneier's intervention [10, 11]. Of course, any education on digital hygiene should be mandatory, addressing caution behaviors and explaining social engineering (and not neglecting what phishing, spoofing, and ransomware are, etc.) and done inside the working time (in order not to incur a bad propensity). We should understand that what is obvious, if not trivial, for a computer scientist or a technologist is totally alien to a final user/operator.

Finally, we want to underline that confidentiality and authentication/integrity are very often requirements of the e-mail, but messages are too often left completely unprotected while stored in some server. An organization could easily implement the OpenPGP protocol [12], at zero cost, while managing only the public keys of local users in a centralized quasi-static manner (e.g., an LDAP) in such a way that all public keys of local users can be trusted, so to have an easy-to-use method of user encryption/signature, at least at a local level. Users should practice, when required, the user encryption, as discussed in [2].

3. DISCUSSION

In this paper two themes strongly interconnected are presented. First, reducing errors or issues, let us gain the real advantage of a full digitalization. Up to date, we haven't been careful enough, and

in addition to errors, that are a component of human beings, we had to spend a lot of time (inefficiency!) in issues coming from the use of computers. Who didn't expend long time in recovering a password, just because it was forgotten and, for some reason, we weren't using a password manager? The traditional rules for creating a new password are too restrictive and we think that the cost of the total time wasted in password recovering is greater than the benefits coming from having a more secure password (that – don't forget – additionally pushes people towards incautious behaviors). In a broader sense, we need to eliminate any inefficiency or issue coming from digitalization. As a further example we saw the pandemic has forced us to make greater use of digital resources, which have already existed for several years. No new technologies but solutions that can be traced back to the last millennium. A gap between research and full technological transfer has always existed, but this isn't a good reason to accept it and make it a rule. Once more, we don't need issues coining from digitalization itself.

As for errors, in addition to adopting mechanisms to reduce them, we should stop dealing with them in an artisan and improvised way. Errors have existed for some long time, so we should predispose, from the very first education, methods that have proved successful in their resolution, as well as other best practices that help to reduce them. Who should take care of this? In our opinion, computer scientists, that are sensitive and qualified, have a mathematical method for designing and testing. We aren't saying that they are the only people prepared enough, but what it should be clear is that we need a multidisciplinary team, that offers a new type of specialization: the horizontal extent is the new vertical depth.

The second theme, strongly interconnected with digitalization, is the digital hygiene. Yes, this is a part of cybersecurity, but all users should be able to understand it. The modern meaning of cybersecurity is "computer security" (this is how Wikipedia re-directs the word "cybersecurity") and a correct hygiene is at its basis. This hygiene should extensively cover subjects like password management, social engineering, inserting USB devices, phishing and spoofing, net-etiquette, privacy, ransomware, etc. Certainly, the list is not complete, but it is sufficient to describe the address of such a formation. We find it impossible to prepare technical solutions that definitively defeat these dangers, and this leads us to focus more on awareness. And whoever does not make it or does not want to, remain completely out of it. We can be sure that widespread digital hygiene will definitively defeat dangers not destined for a specific target, and these are to a much lesser extent, and generally independent of a digitalized scenario. Put simply, there cannot be efficient digitalization without corresponding digital hygiene. Of course, several companies that have special security needs will need more, but in any case, they too will benefit from digital hygiene; they will only have to add other, more specific, and technical measures.

Another point is related to education. We see two types of education, one for the management, another for the final users. The two paths should not be confused, being the former more interested in governance, policies, and other high-level questions. The latter should be oriented to create the digital hygiene, awareness, etc., being much more operational. And the educators, although the ease of the subjects, should be experienced people, because needing to be able to view the subject even with the eyes of a manager/user, what exactly comes from a long experience.

4. CONCLUSIONS

What we have described is not a result, rather it is a meta-result: addressing a new line of research for benefitting at most from digitalization. In Fig. 1 we try to explain how we see the digitalization process, heavily using information technology resources but also deeply entering the several – different – application domains. We believe that this is the task of an information technology expert, assisted by others, including those in the application domain. A computer

scientist's view of the scenario is undoubtedly the most complete and correct. (S)he certainly has the competence and sensitivity necessary for this goal, including the important ones of reducing problems, errors, and making all processes more efficient. Avoiding the vendor lock-in is today not only a clear requirement, but first something coming from the experience/errors. And only an expert can assist in that.

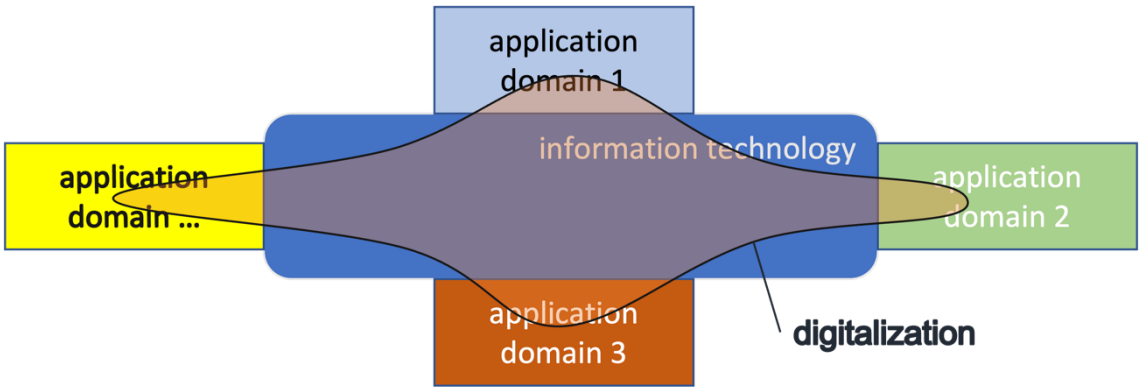


Figure 1. Relationships between information technology, digitalization, and application domains. In blue: information technology; in transparent orange: the digitalization; in light blue, green, red, and yellow: the application domains.

Be careful not to confuse these notes with the traditional computer science research, which will continue to develop results and products; let's think of machine learning, computer vision, robotics, cybersecurity, theoretical computer science, software management models, etc., which will continue to be areas of great interest in computer science. Here we are only proposing a further point on which to discuss and work. Soon we intend to produce precise specifications on tools and procedures to be used to eliminate the errors described and test some existing solutions, perhaps adapting them to our needs.

We don't pretend to include all significant aspects of the question and we have limited ourselves to collecting a handful of obvious requirements, with a view to making the most of the experience of the past in order not to stumble over the same mistakes and to focus on a post-pandemic world that has been able to take advantage of the circumstance to make better use of pre-existing technologies.

ACKNOWLEDGEMENTS

This work has been partially supported by the IoT-STYLE project RG12117A7CE68848.

REFERENCES

[1] ISO 27000 Directory, "An introduction to ISO 27001, ISO 27002... ISO 27008" https://www.27000.org/, accessed: 2022-01-01.

[2] d'Amore, F., Fantozzi, P., Laura, L., Padovan, D. (2020) "On Enterprise Data Encryption: Good, Bad and Ugly", in *Proceedings of MENACIS20*. Springer, to appear in Lecture Notes in Information Systems and Organization (LNISO), 2022.

[3] Chen, C., Huang, W., Zhou, B., Liu, C., Mow, W.H. (2016) "Picode: A new picture-embedding 2D barcode", *IEEE Transactions on Image Processing*, Vol. 25, No. 8, pp3444-3458.

[4] Hung, S.H., Yao, C.Y., Fang, Y.J., Tan, P., Lee, R.R., Sheffer, A., Chu, H.K. (2020) "Micrography QR codes", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 26, No. 9, pp2834-2847. DOI: 10.1109/TVCG.2019.2896895.

[5] Chu, H.K., Chang, C.S., Lee, R.R., Mitra, N.J. (2013) "Halftone QR codes", *ACM Transactions on Graphics (TOG)*, Vol. 32, No. 6, pp1-8.

[6] Garateguy, G.J., Arce, G.R., Lau, D.L., Villarreal, O.P. (2014) "QR images: optimized image embedding in QR codes", *IEEE Transactions on Image Processing*, Vol. 23, No. 7, pp2842-2853.

[7] Lin, S.S., Hu, M.C., Lee, C.H., Lee, T.Y. (2015) "Efficient QR code beautification with high quality visual content", *IEEE Transactions on Multimedia*, Vol. 17, No. 9, pp1515-1524.

[8] Lin, Y.H., Chang, Y.P., Wu, J.L. (2013) "Appearance-based QR code beautifier", *IEEETransactions on Multimedia*, Vol. 15, No. 8, pp2198-2207.

[9] Sveikauskas, D. (2016) "Security by design principles according to OWASP", Online, June, URL: https://patchstack.com/security-design-principles-owasp/, accessed 02-11-2022.

[10] Schneier, B. (2016) "Stop trying to fix the user", *IEEE Secur. Priv*., Vol. 14, No. 5, p96. DOI: 10.1109/MSP.2016.101.

[11] Schneier, B. (2020) "Technologists vs. policy makers" Online, February, URL: https://www.schneier.com/essays/archives/2020/02/technologists_vs_pol.html, accessed 02-11-2022.

[12] Barenghi, A., Mainardi, N., Pelosi, G. (2017) "A security audit of the OpenPGP format", in *14th International Symposium on Pervasive Systems, Algorithms and Networks & 11th International Conference on Frontier of Computer Science and Technology & Third International Symposium of Creative Computing, ISPAN-FCST-ISCC*, Exeter, United Kingdom, June, pp21-23, pp.336-343, *IEEE Computer Society*. DOI: 10.1109/ISPAN-FCST-ISCC.2017.35

AUTHOR

Fabrizio d'Amore is an associate professor at Sapienza University of Rome, Italy. He teaches Theoretical Computer Science and Cybersecurity. He is responsible for the graduated master program "Information Security and Strategic Information". His teaching/scientific interests include Theoretical Computer Science, Digitalization, Applied Cryptography, Privacy, and Algorithms. He also serves for many public Italian bodies.

A NEW DEEP-NET ARCHITECTURE FOR ISCHEMIC STROKE LESION SEGMENTATION

Nesrine Jazzar[1] and Ali Douik[2]

[1]University of Sfax National Enginneering School of Sfax NoCCS Lab, Tunis, Tunisia
[2]National Enginneering School of Sousse, University of Sousse

ABSTRACT

Ischemic stroke, brain cells death due to a lack of oxygen, is a leading cause of long-term disability and death. Accurate diagnosis and timely intervention can effectively improve the blood supply of the ischemic stroke area and minimize brain damage. Recent studies have shown the potential to use magnetic resonance imaging (MRI) to provide contrast imaging to visualize and detect lesions. However, manual segmentation of the stroke lesion produced by MRI is a tedious and time-consuming task. Therefore, the automatic ischemic stroke lesion segmentation method may show excellent advantages. In this paper, we propose a novel deep learning method used to detect and localize brain ischemic stroke, a generalization encoder-decoder by modifying U-Net architecture.

We integrate multi-path architecture into both encoder and decoder blocks to captures different levels of the encoded state, which helps in more robust decision-making for stroke lesion segmentation. In bottleneck of the architecture, we applied dilated blocks to improve the underlying predictive capabilities. The proposed method has been tested on the publicly accessible web platform provided by the MICCAI Ischemic Stroke Lesion Segmentation (ISLES) challenge. The results demonstrate that the proposed method achieves a mean dice coefficient 0.91 of with the training and 0.84 with the testing data respectively.

KEYWORDS

Ischemic stroke segmentation, Convolutional neural network, U-Net, MRI, Dilated blocks.

1. INTRODUCTION

A cerebrovascular accident, i.e., a stroke is a failure of the blood flow that affects a large or small brain area. It occurs when a blood vessel is blocked or ruptured, and it causes the nerve cells to die, which are deprived of oxygen and essential nutrients for their appropriate function. The severity of the stroke depends on the location and the extent of the affected brain areas. According to the World Health Organization (WHO), an ischemic stroke is the leading national cause of acquired physical disability in adults and the second leading cause of death globally [1]. Early diagnosis and timely intervention are very critical for the recovery of stroke patients.

Magnetic resonance images (MRI) are the standard gold examination, they provide essential information for optimized treatment, eliminating a haemorrhagic accident, and detecting the lesion area from the first hour after the onset of clinical signs. MRI is much more accurate than a CT scan detecting multiple or small lesions or assessing the necrotic area's extent. These elements are essential for the patient's prognosis and the treatment decision.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 91-100, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121108

However, the automatic identification and segmentation of ischemic stroke lesions is not a trivial task because of the scarcity of datasets, image complexity, and the high variability of stroke's location contrast shape.

Most of the automatic segmentation methods use hand-designed features. Recently, there has been a growing interest in applying CNNs in image classification and segmentation. Recent studies show that it is more effective and suitable for complex neuroimaging tools such as neurological disorders and psychiatrists. Kaminatas et al. [2] proposed an approach for the segmentation of brain lesions using multimodal brain MRI based on 11-layer-deep, multi-scale 3D convolutional neural networks (CNNs) called Deep Medic. The proposed new training scheme is based on two main components: a 3D CNN, which produces exact flexible segmentation maps, and a fully connected 3D CRF conditional random field, which imposes regularization constraints on the CNN output and produces the segmentation labels. Chen et al. [3] proposed a framework with two CNN modules to segment stroke lesions using DWI in MRI. The first CNN was a combination of two DeconvNets (EDD Net), and the second one was a Multi-Scale Convolutional Label Assessment Network (MUSCLE Net) to focus on lesions detected at a small scale and aim to reduce false potentials detected by the EDD network. The dataset was constructed with clinically acquired DWI scans of 741 patients with acute stroke, exhibiting a high lesion detection rate and high accuracy. Liangliang Liu et al. [4] proposed a new Res-CNN automatic segmentation network that combines a similar U-shaped architecture with residual units. This network could alleviate the problem of the leakage gradient. The architecture of Res-CNN consists of 10 convolutional layers, 4 residual units, 4 concatenations layers, 4 deconvolution layers, and some batch normalization (BN) layers, and Leaky Rectified Linear Units (LReLU) [5]. The dataset was constructed with DWI scans and T2-to-DWI fusion (DWI-T2) as the multi-modality of input to improve lesion segmentation performance. Zhang et al. [6] proposed a fully convolutional and densely connected neural network (3D FC-DenseNet) to segment stroke lesions from DWI diffusion-weighted images. The network could use contextual information and learn end-to-end discriminating characteristics. The network is built based on the idea of densely connected convolutional networks, which allows each layer to take as input all of its previous feature maps, and two layers of a DenseNet network are directly connected. Liu et al. [7] proposed a Residual Structure Fully Convolutional Network (Res-FCN) to segment ischemic stroke lesions from multimodal MRI scans. In Res-FCN, the residual block can capture characteristics of large receptive fields for the network. Havaei et al. [8] proposed a CNN approach to segment subacute and ischemic stroke lesions from DWI, FLAIR, and T2 diffusion-weighted images. Lucas et al. [9] proposed a fully convolutional neural network (FCN) based on 2D-UNet networks with multiscale information propagation to segment acute stroke lesions. Some of the techniques give erroneous segmentation results when the lesions are small especially in the case of ischemic stroke segmentation.

The endeavor of this paper proposes a new deep learning architecture by modifying the U-Net [10] model to perform a fully automated stroke lesion segmentation task. We integrate multi-path architecture into both encoder and decoder blocks to outstanding feature representation ability and preserve low-level information. This multi-path architecture captures different levels of the encoded state, which helps in more robust decision-making [11] for stroke lesion segmentation. Furthermore, we applied dilated blocks in the bottleneck of the architecture to improve the underlying predictive capabilities [12]. We have conducted experiments on the publicly accessible web platform provided by the MICCAI Ischemic Stroke Lesion Segmentation (ISLES) challenge [13] with different images modalities.

This paper is organized as follows. Section II provides an overview of the dataset utilized for stroke segmentation. Section III illustrates in detail the proposed deep method. Then, Section IV includes the evaluation metric used to evaluate the segmentation results, followed by the findings

of the experimental results in section V. Finally, the efficiency of the proposed architecture is summarized in section VI.

2. DATASET DESCRIPTION

2.1. Data Acquisition

The proposed model was assessed on a multi-modal MRI sub-task, the sub-acute ischemic stroke segmentation (SISS) of the MICCAI ISLES 2015 challenge [13] dataset. The SISS dataset contains 64 cases (28 training and 36 test) of patients with sub-acute ischemic stroke. Each patient has four co-registered MRI modalities, namely Fluid-Attenuated Inversion Recovery (Flair), Diffusion-Weighted Imaging (DWI), T1, and T2. The images acquired had a 3D voxel size of 230 ×230 ×153 with an image resolution of 1mm × 1mm × 1mm spacing. The dataset is provided with pixel-accurate ground truth labels. The lesions in this dataset were small and diffuse.

2.2. Data Pre-processing

The training data is one of the essential keys to obtaining a high prediction model's accuracy. That is why pre-processing is mandatory. Hence, we have prepared data for use in the following steps: First, we eliminated any black slices with no data for every patient. Second, all the images of 230×230×3 shape are timed to a new shape of 129×129×3, which is more convenient for the following work steps.

The total number of images in our dataset was 3900 images. Therefore, the data was further augmented using randomly a left-right flip of the images, horizontal flipping, shearing, rotation, and zooming to produce six times as much data, giving 19000 total images. Later, the intensity values of these images are normalized in the range of [0,1] based on the minimal value. The main reason for pre-processing is to increase the robustness and validation accuracy of the network and tackle data insufficiency issues in medical imaging. The final step is data partitioning: 80% for the training and the rest 20% is for the validation.

3. METHODOLOGY

3.1. Network Description

This work proposes a deep learning architecture using a residual CNN inspired by the U-Net [10] architecture multipath network and dilated convolution, illustrated in Figure 1. Our architecture is an encoder-decoder network that takes advantage of a multipath network by integrating M-blocks in a single path as the elementary module [11]. Using a multipath procedure will enhance the possibility to constructs a more extended feature than a single path. Both the encoder and decoder path comprise several M-Blocks, as shown in Figure 2. The M-Block consists of 4 different paths Pi where i in 0 to 3. In each path, we use a different number of convolutional layers. For example, P1, P2, and P3 have one, two, and three convolutional layers, respectively. This technique will make the learning more comprehensive features easier and more precise than a single path by allowing flexibility to the amount of encoding/decoding required for precise segmentation. On the other side, to get advantages of residual connection and minimize information loss along with the depth of the network, P0 does not include any convolutional layer. The convolutional layers consist of Kernel size 3 × 3 number of filter zero-padding followed by batch-normalization and ReLU activation [14]. The outputs of the four paths are concatenated and passed through the activation function. Each encoder block output is processed in two different ways. The output is

max-pooled and sent to the next encoder block as its input, or the same output is concatenated with the transposed output of the lower encoder block and sent to the corresponding decoder block as feedback which enhances the feature. In the encoder block, the max-pooling is done 5 times on the image data, resulting in 6 times smaller than the initial state. The decoder is the same as the encoder, except the max-pooling operation is replaced by a convolutional transpose operation. The output of the last decoder block is passed through a convolutional block with a single filter of kernel size 1 × 1 with sigmoid ($\frac{1}{1+e^{-x}}$) activation function.

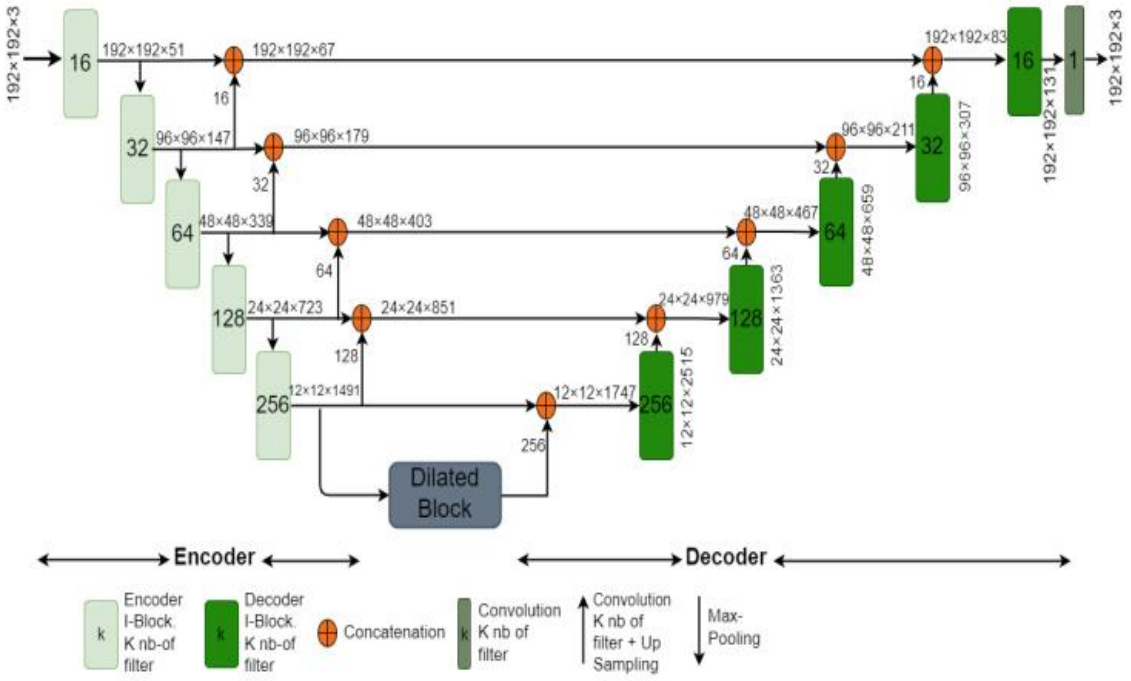


Figure 1. The architecture of the proposed model for ischemic stroke segmentation

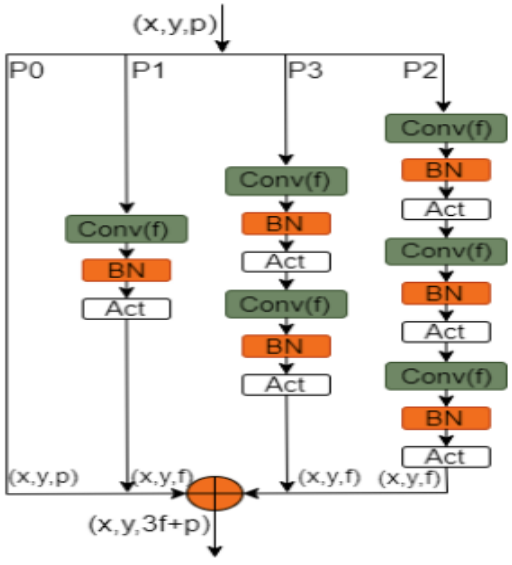


Figure 2. Schematic diagram of the M-Block

In our network, we also integrate the dilated block at the lowest layer of the network to summarize the global information and generate the output of the encoder. Dilated convolution also named Atrous convolution, was originally developed for Wavelet decomposition. The goal of using of a dilation convolution is to insert a hole between the pixels in the convolutional kernel to capture the texture information with different receptive fields. The receptive field is how large the pixels in the high-level feature map are affected by the original image of each layer of the convolutional neural network. We use low dilation rate convolutional to capture the texture information on a small scale and use a high dilation rate convolutional to capture the texture information on a large scale. The dilated block illustrated in Figure 3, consisting of four dilated convolution layers employed in the bottleneck of the network, is configured such that the first layer uses a dilation rate of one. Furthermore, each subsequent layer increases the dilation rate by a multiple of two as proposed in [13]. The dilation rate is 1, 2, 4, and 8. Each convolutional kernel's feature scale is $(2k+1)^2$, where k is the kernel's dilation rate. However, the features extracted from the dilated convolution result produce a different scale of 3 × 3, 5 × 5, 9 × 9, and 17 × 17 as shown in Figure 4. We applied Batch normalization for the output of each dilated convolution layer to enhance the network's stability, followed by ReLU is used as an activation function. Then, a concatenation of the four ReLU outputs, followed by 1 × 1 convolution, to reduce the dimension, Batch normalization, and finally the ReLU activation function. We chose ADAM optimizer with an initial learning rate of Lri = 10^{-4}, decay factor =0.2, step =2.

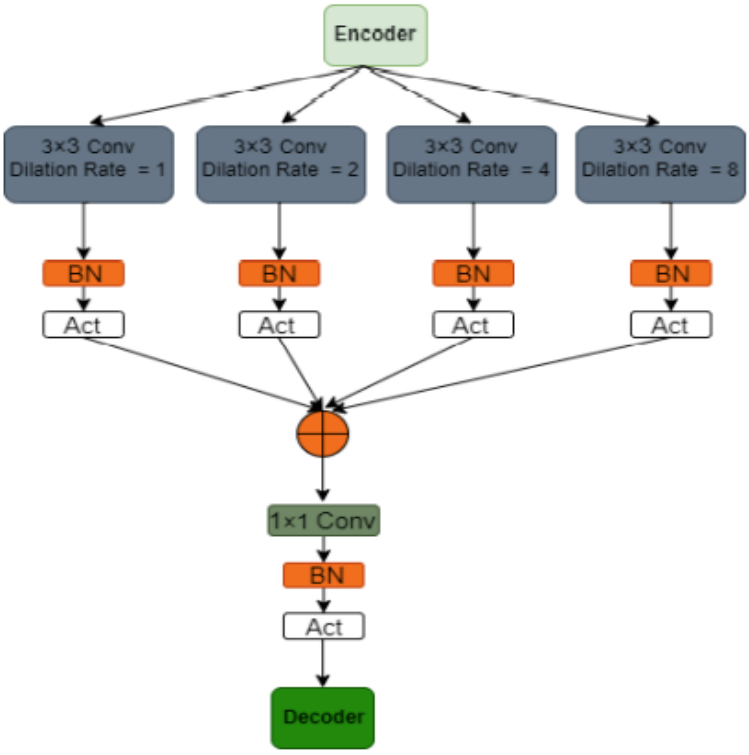


Figure 3. Schematic diagram of the dilated Block

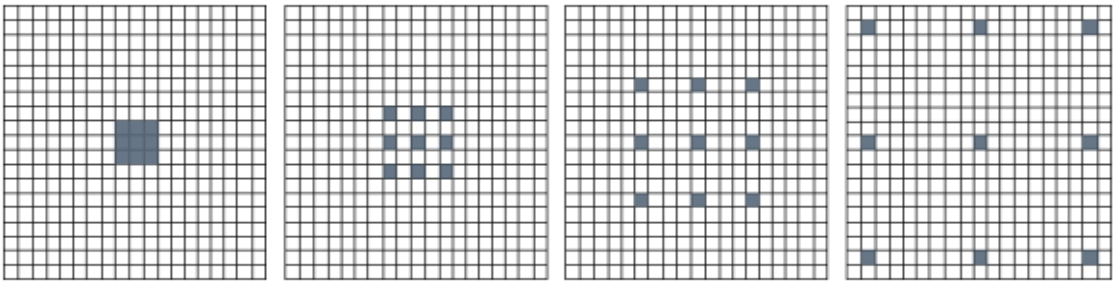


Figure 4. Dilated convolution with 3×3 Kernel with a dilation rate of 1, 2, 4 and 8

3.2. Training Techniques

The input data dimension is 192×192×3, and the segmented result obtained from the network is 192 × 192 × 1. In the training process, the weight initialization plays a vital role in converging the model. It is initialized using he-normal, which draws the ample from a truncated normal distribution centred on 0 with $\sigma = \sqrt{\frac{2}{n}}$, where n is the number of input units in the weight is updated using the Adam optimizer a batch size of 5. The loss function (Binary Cross Entropy BCE, Dice Loss, BCE Dice Loss, Sigmoid BCE) is used to measure the error, as defined in Table 1. It is also the function to be minimized during training, and we use early stopping by monitoring validation loss function with a patient of 10 to bypass the severe class imbalance problem. Our proposed model has trained these four-loss functions separately, as mentioned in Table 2.

Table 1. Loss Function.

Binary Cross Entropy(BCE)	-y $\log(\hat{y}) + (1-y)(\log(1-\hat{y}))$
Dice Loss	1 - 2 $\lvert\frac{y \cap \hat{y}}{y \cap \hat{y}}\rvert$
BCE Dice Loss	BCE-2 $\lvert\frac{y \cap \hat{y}}{y \cap \hat{y}}\rvert$
Sigmoid BCE	-y $\log(f(\hat{y})) + (1-y)(\log(1-f(\hat{y})))$

Table 2. Performance of proposed model in term of Dice coefficient on each modality on SISS Dataset for various loss functions.

Modality	BCE		Dice Loss		BCE Dice Loss		SCE	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
DWI	0.871	0.814	0.892	0.809	0.855	0.714	0.685	0.581
Flair	0.895	0.827	0.916	0.844	0.881	0.759	0.809	0.772
T1	0.784	0.671	0.858	0.699	0.877	0.738	0.815	0.793
T2	0.755	0.674	0.863	0.695	0.869	0.740	0.694	0.618

4. EVALUATION METRICS

Evaluation metrics are essential tools to analyse the performance of segmentation. The value of each criterion increases with the quality of the segmentation result. These values have been standardized to facilitate their comparisons. A criterion value close to 1 reflects an excellent segmentation result. Our model has been evaluated on four widely used quality metrics, which are defined as follows.

4.1. Dice Similarity Coefficient

The Dice score (DSC) [16] measures the similarity; overlap between the manually segmented ground truth and our segmentation results, which will be calculated as follows:

$$DSC = \frac{2TP}{2TP + FP + FN}$$

Where, True Positive (TP) indicates that the method correctly segmented pixels. False Positive (FP) indicates the pixel that the method classifies negative as positive. False Negative (FN) denotes that the positive pixel is incorrectly classified as negative by the segmentation method.

4.2. Accuracy

The accuracy related to the percentage of pixels that were correctly predicted over the total number of pixels segmented in an image [19]. It is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.3. Specificity

The specificity measures the percentage of pixels correctly predicted as belonging to the background region among all the pixels belonging to the background. It is defined as:

$$Specificity = \frac{TP}{TP + FP}$$

4.4. Sensitivity

The sensitivity measures the percentage of pixels correctly segmented that are correctly identified. It is defined as:

$$Sensitivity = \frac{TP}{TP + FN}$$

5. RESULTS AND DISCUSSION

The proposed architecture is trained in ubuntu environment, Dell Predator, inter-core i5, 24 GB RAM installed with NVIDIA GeForce GTX 1050 Ti. Keras and TensorFlow are used as frameworks to implement the model. The detailed statistics of the outcome are shown in Table 2, 3. We have analysed that the Dice Loss function on the SISS dataset gives the optimum results

for all modalities, which is better than other functions on both training and testing datasets. However, the Dice loss function is the most suitable loss function for all modalities. We also noticed that the FLAIR and DWI are the most suitable image modalities while producing consistent outcomes using the Dice loss function. Finally, the average has been calculated to compare the performance of our architecture with some of the well-known methods which are shown in Table 3. It should be noted that our model surpasses the state-of-the-art results marginally. The visual performance of our model on the SISS dataset is shown in Figure 5, which is a curve for DSC and accuracy for training and validation set, and Figure 6.

Figure 5. From left to right: Plot for DSC and accuracy for training and validation set

Table 3. Comparison of Dice for various methods on the SISS training and testing datasets. The average is calculated over the total number of patients. The showcased data has been obtained from the ISLES 2015 Challenge.

Method	DSC	
	Training	Testing
Robben et al. [18]	057±0.28	0.43 ±0.30
Maier et al. [12]	0.58 ±0.29	0.42 ±0.33
Feng et al. [19]	0.63 ±0.28	0.55 ±0.30
Chen et al. [3]	0.55 ±0.29	0.44 ±0.30
proposed	0.91 ±0.11	0.84 ±0.24

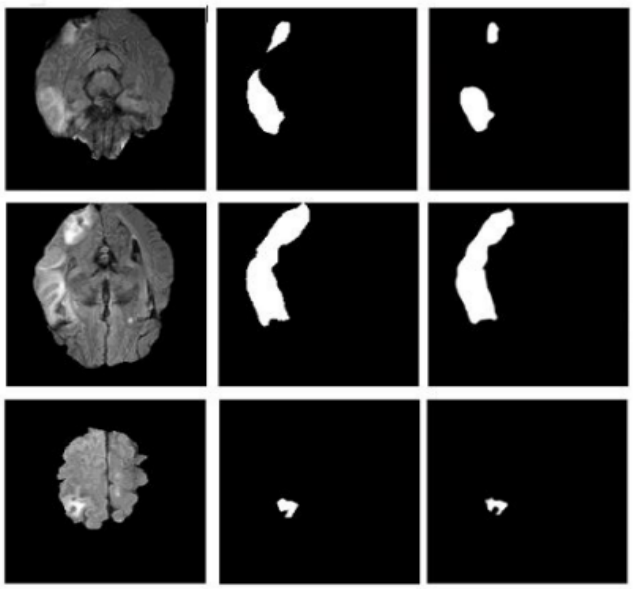


Figure 6. Visual result of our model from the axial view for three different patients on SISS dataset: Original image, ground truth and predicted image

6. CONCLUSIONS

In this paper, we propose a fully automated ischemic stroke segmentation from various modalities inspired by U-Net architecture fused with a multipath network and we integrate a dilated block in the bottleneck. We optimized our model with various loss functions to tackle the severe class imbalance problem. However, we conclude that the Dice loss function gives the optimum results for all modalities. We also evaluate our architecture on a public challenge dataset SISS 2015, where its effectiveness and generalization capability are further demonstrated. However, the proposed architecture presents high segmentation accuracy with different modalities MRI images with a Dice coefficient for subsequent work, we aim to do a fusion of two different modalities such as FLAIR and DWI images such input for our model.

REFERENCES

[1] Global Diffusion of eHealth, (2017), Making Universal Health Coverage Achievable, Report of the Third Global Survey on eHealth, World Health Org., Geneva, Switzerland.

[2] Kamnitsas K, Ledig C, Newcombe VFH, Simpson JP, Kane AD, Menon DK, et al. (2017), Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med Image Anal.

[3] Chen L, Bentley P, Rueckert D, (2017), Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks, NeuroImage.

[4] Liangliang Liu, Shaowu Chen, Fuhao Zhang, Fang-Xiang Wu, Yi Pan, Jianxin Wang, (2019), Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI, part of Springer Nature.

[5] Nair V, Hinton GE, (2010), Rectified linear units improve restricted boltzmann machines, In: International conference on international conference on machine learning, pp 807–814,

[6] Zhang R, Zhao L, Lou W, Abrigo JM, Mok VC, Chu WC, Wang D, Shi L, (2018), Automatic segmentation of acute ischemic stroke from dwi using 3d fully convolutional densenets, IEEE Trans Med Imaging 37(9):2149–2160.

[7] Liu Z, Cao C, Ding S, Han T, Wu H, Liu S, (2018), Towards clinical diagnosis, automated stroke lesion segmentation on multimodal mr image using convolutional neural network.

[8] Havaei M, Dutil F, Pal C, Larochelle H, Jodoin PM, (2015), A convolutional neural network approach to brain tumor segmentation, In: International Workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries, pp 195–208.
[9] Christian Lucas, André Kemmling, Amir Madany Mamlouk, Mattias P. Heinrich, (2018), Multi-Scale neural network for automatic segmentation of ischemic strokes on acute perfusion images, IEEE 15th International Symposium on Biomedical Imaging (ISBI), Washington, D.C., USA.
[10] O. Ronneberger, P. Fischer and T. Brox, (2015), U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer assisted intervention, pp. 234-241.
[11] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollar, (2016), A multipath network for object detection.
[12] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen et al, (2017), Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri, Medical image analysis, vol. 35, pp. 250-269.
[13] Vesal, S., Ravikumar, N., Maier, (2018), A.: Dilated convolutions in neural networks for left atrial segmentation in 3d gadolinium enhanced-mri.
[14] V. Nair and G. E. Hinton, (2010), Rectified linear units improve restricted boltzmann machines, in Proceedings of the 27th international conference on machine learning (ICML-10).
[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, (2017), Rethinking atrous convolution for semantic image segmentation.
[16] L. R. Dice, (1945), Measures of the amount of ecologic association between species, Ecology, vol. 26, no. 3, pp. 297–302.
[17] Chang, P.D, (2016), Fully convolutional neural networks with hyperlocal features for brain tumor segmentation, MICCAI- BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark.
[18] D. Robben, D. Christiaens, J. R. Rangarajan, J. Gelderblom, P. Joris, F. Maes, and P. Suetens, (2015), A voxel-wise, cascaded classification approach to ischemic stroke lesion segmentation, in BrainLes 2015. Springer, pp. 254–265.
[19] C. Feng, D. Zhao, and M. Huang, (2015), Segmentation of Stroke Lesions in Multi-spectral MR Images Using Bias Correction Embedded FCM and Three Phase Level Set.

AUTHORS

Nesrine Jazzar received the B.A. degree in mathematical from Ali Bourguiba College, Monastir, Tunisia, in 2012, and obtained licence degree in computer science in 2015 and in 2017 she had the Master degree Industrial computer science. Since 2019, She is a Ph.D. student at the National Engineering School of Sfax, Tunisia. She is a member of the Networked Objects, Control, and Communication Systems (NOCCS) laboratory. Her research focuses on computer sciences, in particular, medical image segmentation.

ALi Douik Director at National Engineering School of Sousse. Having a doctorate in Electrical Engineering and a qualification to direct research. He was able to provide lessons in Automatics, system control, Image processing. He supervised several doctoral theses. He published numerous scientific articles in peer-reviewed international conferences and journals.

EVALUATION OF SEMANTIC ANSWER SIMILARITY METRICS

[1]Farida Mustafazade and [2]Peter F. Ebbinghaus

[1]GAM Systematic

[2] Teufel Audio

ABSTRACT

There are several issues with the existing general machine translation or natural language generation evaluation metrics, and question-answering (QA) systems are indifferent in that context. To build robust QA systems, we need the ability to have equivalently robust evaluation systems to verify whether model predictions to questions are similar to ground-truth annotations. The ability to compare similarity based on semantics as opposed to pure string overlap is important to compare models fairly and to indicate more realistic acceptance criteria in real-life applications. We build upon the first to our knowledge paper that uses transformer-based model metrics to assess semantic answer similarity and achieve higher correlations to human judgement in the case of no lexical overlap. We propose cross-encoder augmented bi-encoder and BERTScore models for semantic answer similarity, trained on a new dataset consisting of name pairs of US-American public figures. As far as we are concerned, we provide the first dataset of co-referent name string pairs along with their similarities, which can be used both for training and as a benchmark.

KEYWORDS

Question-answering, semantic answer similarity, exact match, pre-trained language models, cross-encoder, bi-encoder, semantic textual similarity, automated data labelling

1. INTRODUCTION

Having reliable metrics for evaluation of language models in general, and models solving difficult question answering (QA) problems, is crucial in this rapidly developing field. These metrics are not only useful to identify issues with the current models, but they also influence the development of a new generation of models. In addition, it is preferable to have an automatic, simple metric as opposed to expensive, manual annotation or a highly configurable and parameterisable metric so that the development and the hyperparameter tuning do not add more layers of complexity. SAS, a cross-encoder-based metric for the estimation of semantic answer similarity [1], provides one such metric to compare answers based on semantic similarity.

The central objective of this research project is to analyse pairs of answers similar to the one in Figure 1 and to evaluate evaluation errors across datasets and evaluation metrics.

The main hypotheses that we will aim to test thoroughly through experiments are twofold. Firstly, lexical-based metrics are not well suited for automated QA model evaluation as they lack a notion of context and semantics. Secondly, most metrics, specifically SAS and BERTScore, as described in [1], find some data types more difficult to assess for similarity than others.

After familiarising ourselves with the current state of research in the field in Section 2, we describe the datasets provided in [1] and the new dataset of names that we purposefully tailor to our model in Section 3. This is followed by Section 4, introducing the four new semantic answer similarity approaches described in [1], our fine-tuned model as well as three lexical n-gram-based automated metrics. Then in Section 5, we thoroughly analyse the evaluation datasets described in

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 101-115, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121109

Figure 1: Representative example of a question and all semantic answer similarity measurement results.

Question: Who makes more money: NFL or Premier League?
Ground-truth answer: National Football League
Predicted Answer: the NFL
EM: 0.00
$\mathbf{F_1}$: 0.00
Top-1-Accuracy: 0.00
SAS: 0.9008
Human Judgment: 2 (definitely correct prediction)
$\mathbf{f_{BERT}}$: 0.4317
$\mathbf{f'_{BERT}}$: 0.4446
Bi-Encoder: 0.5019

the previous section and conduct an in-depth qualitative analysis of the errors. Finally, in Section 6, we summarise our contributions.

2. Related work

We define semantic similarity as different descriptions for something that has the same meaning in a given context, following largely [2]'s definition of semantic and contextual synonyms. [3] noted that open-domain QA is inherently ambiguous because of the uncertainties in the language itself. The human annotators attach a label `2` to all predictions that are "definitely correct", `1` - "possibly correct", and `0` - "definitely incorrect". Automatic evaluation based on exact match (EM) fails to capture semantic similarity for definitely correct answers, where 60% of the predictions are semantically equivalent to the ground-truth answer. Just under a third of the predictions that do not match the ground-truth labels were nonetheless correct. They also mention other reasons for failure to spot equivalence, such as time-dependence of the answers or underlying ambiguity in the questions.

QA evaluation metrics in the context of SQuAD v1.0 [4] dataset are analysed in [5]. They thoroughly discuss the limitations of EM and F1 score from n-gram based metrics, as well as the importance of context including the relevance of questions to the interpretation of answers. A BERT matching metric (Bert Match) is proposed for answer equivalence prediction, which performs better when the questions are included alongside the two answers, but appending contexts didn't improve results. Additionally, authors demonstrate better suitability of Bert Match in constructing top-*k* model's predictions. In contrast, we will cover multilingual datasets, as well as more token-level equivalence measures, but limit our focus on similarity of answer pairs without accompanying questions or contexts.

Two out of four semantic textual similarity (STS) metrics that we analyse and the model that we eventually train depend on bi-encoder and BERTScore [6]. The bi-encoder approach model is based on the Sentence Transformer structure [7], which is a faster adaptation of BERT for the semantic search and clustering type of problems. BERTScore uses BERT to generate contextual embeddings, then match the tokens of the ground-truth answer and prediction, followed by creating a score from the maximum cosine similarity of the matched tokens. This metric is not one-size-fits-all. On top of choosing a suitable contextual embedding and model, there is an optional feature of importance weighting using inverse document frequency (idf). The idea is to limit the influence of common words. One of the findings is that most automated evaluation metrics demonstrate significantly better results on datasets without adversarial examples, even when these are introduced within the training dataset, while the performance of BERTScore suffers only slightly. [6] uses machine

translation (MT) and image captioning tasks in experiments and not QA. [8] apply BERT-based evaluation metrics for the first time in the context of QA. Even though they find that METEOR as an n-gram based evaluation metric proved to perform better than the BERT-based approaches, they encourage more research in the area of semantic text analysis for QA. Moreover, [5] uses only BERTScore base as one of the benchmarks, while we explore the larger model, as well as a finetuned variation of it.

Authors in [1] expand on this idea and further address the issues with existing general MT, natural language generation (NLG), which entails as well generative QA and extractive QA evaluation metrics. These include reliance on string-based methods, such as EM, F1-score, and top-n-accuracy. The problem is even more substantial for multi-way annotations. Here, multiple ground-truth answers exist in the document for the same question, but only one of them is annotated. The major contribution of the authors is the formulation and analysis of four semantic answer similarity approaches that aim to resolve to a large extent the issues mentioned above. They also release two three-way annotated datasets: a subset of the English SQuAD dataset [9], German GermanQuAD dataset [10], and NQ-open [3].

Looking into error categories (see Table 1 and Section 5) revealed problematic data types, where entities, particularly those involving names of any kind, turned out to be the leading category. [11] analyse Natural Questions (NQ) [12], TriviaQA [13] as well as SQuAD and address the issue that current QA benchmarks neglect the possibility of multiple correct answers. They focus on the variations of names, e.g. nicknames, and improve the evaluation of Open-domain QA models based on a higher EM score by augmenting ground-truth answers with aliases from Wikipedia and Freebase. In our work, we focus solely on the evaluations of answer evaluation metrics and generate a standalone names dataset from another dataset, described in greater detail in Section 3.

Our main assumption is that better metrics will have a higher correlation with human judgement, but the choice of a correlation metric is important. Pearson correlation is a commonly used metric in evaluating semantic text similarity (STS) for comparing the system output to human evaluation. [14] show that Pearson power-moment correlation can be misleading when it comes to intrinsic evaluation. They further go on to demonstrate that no single evaluation metric is well suited for all STS tasks, hence evaluation metrics should be chosen based on the specific task. In our case, most of the assumptions, such as normality of data and continuity of the variables behind Pearson correlation do not hold. Kendall's rank correlations are meant to be more robust and slightly more efficient in comparison to Spearman as demonstrated in [15].

Soon after Transformers took over the field, adversarial tests resulted in significantly lower performance figures, which increased the importance of adversarial attacks [16]. General shortcomings of language models and their benchmarks led to new approaches such as Dynabench [17]. Adversarial GLUE (AdvGLUE) [18] focuses on the added difficulty of maintaining the semantic meaning when applying a general attack framework for generating adversarial texts. There are other shortcomings of large language models, including environmental and financial costs [19]. Hence, analysing existing benchmarks is crucial in effectively supporting the pursuit of more robust models. We therefore carefully analyse state of the art benchmarking for semantic answer similarity metrics while keeping in mind the more general underlying shortcomings of large pre-trained language models.

3. DATA

We perform our analysis on three subsets of larger datasets annotated by three human raters and provided by [1]. Unless specified otherwise, these will be referred to by their associated dataset names.

Table 1: Category definitions and examples from annotated NQ-open dataset.

Category	**Definition**	**Question**	**Gold label**	**Prediction**
Acronym	An abbreviation formed from the initial letters of other words and pronounced as a word	what channel does the haves and have nots come on on directv	OWN	Oprah Winfrey Network
Alias	Indicate an additional name that a person sometimes uses	who is the man in black the dark tower	Randall Flagg	Walter Padick
Co-reference	Requires resolution of a relationship between two distinct words referring to the same entity	who is marconi in we built this city	the father of the radio	Italian inventor Guglielmo Marconi
Different levels of precision	When both answers are correct, but one is more precise	when does the sympathetic nervous system be activated	constantly	fight-or-flight response
Imprecise question	There can be more than one correct answers	b-25 bomber accidentally flew into the empire state building	Old John Feather Merchant	1945
Medical term	Language used to describe components and processes of the human body	what is the scientific name for the shoulder bone	shoulder blade	scapula
Multiple correct answers	There is no single definite answer	city belonging to mid west of united states	Des Moines	kansas city
Spatial	Requires an understanding of the concept of space, location, or proximity	where was the tv series pie in the sky filmed	Marlow in Buckinghamshire	bray studios
Synonyms	Gold label and prediction are synonymous	what is the purpose of a chip in a debit card	control access to a resource	security
Biological term	Of or relating to biology or life and living processes	where is the ground tissue located in plants	in regions of new growth	cortex
Wrong gold label	The ground-truth label is incorrect	how do you call a person who cannot speak	sign language	mute
Wrong label	The human judgement is incorrect	who wrote the words to the original pledge of allegiance	Captain George Thatcher Balch	Francis Julius Bellamy
Incomplete answer	The gold label answer contains only a subset of the full answer	what are your rights in the first amendment	religion	freedom of the press

3.1. Original datasets

SQuAD is an English-language dataset containing multi-way annotated questions with 4.8 answers per question on average. **GermanQuAD** is a three-way annotated German-language question/answer pairs dataset created by the deepset team which also wrote [1]. Based on the German counterpart of the English Wikipedia articles used in SQuAD, GermanQuAD is the SOTA dataset for German question answering models. To address a shortcoming of SQuAD that was mentioned in [20], GermanQuAD was created with the goal of preventing strong lexical overlap between questions and answers. Hence, more complex questions were encouraged, and questions were rephrased with synonyms and altered syntax. SQuaD and GermanQuAD contain a pair of answers and a hand-labelled annotation of `0` if answers are completely dissimilar, `1` if answers have a somewhat similar meaning, and `2` if the two answers express the same meaning. **NQ-open** is a five-way annotated open-domain adaption of [20]'s Natural Questions dataset. NQ-open is based on actual Google search engine queries. In case of NQ-open, the labels follow a different methodology as described in [3]. The assumption is that we only leave questions with a non-vague interpretation (see Table 1). Questions like *Who won the last FIFA World Cup?* received the label `1` because they have different correct answers without a precise answer at a point in time later than when the question was retrieved. There is yet another ambiguity with this question, which is whether it is discussing FIFA Women's World Cup or FIFA Men's World Cup. This way, the two answers can be correct without semantic similarity even though only one correct answer is expected.

The annotation of NQ-open indicates truthfulness of the predicted answer, whereas for SQuAD and GermanQuAD the annotation relates to the semantic similarity of both answers which can lead to differences in interpretation as well as evaluation. To keep the methodology consistent and improve NQ-open subset, vague questions with more than one ground-truth labels have been filtered out. We also manually re-label incorrect labels as well as filter out vague questions.

Table 2 describes the size and some lexical features for each of the three datasets. There were 2, 3 and 23 duplicates in each dataset respectively. Dropping these duplicates led to slight changes in the metric scores.

Table 2: Percentage distribution of the labels and statistics on the subsets of datasets used in the analyses. The average answer size column refers to the average of both the first and second answers as well as ground-truth answer and predicted answer (NQ-open only). $F_1 = 0$ indicates no string similarity, $F_1 \neq 0$ indicates some string similarity. Label distribution is given in percentages.

	SQuAD	**GermanQuAD**	**NQ-open**
Label 0	56.7	27.3	71.7
Label 1	30.7	51.5	16.6
Label 2	12.7	21.1	11.7
$\mathbf{F_1 = 0}$	565	124	3030
$\mathbf{F_1 \neq 0}$	374	299	529
Size	939	423	3559
Avg answer size	23	68	13

3.2. Augmented dataset

For NQ-open, the largest of the three datasets, names was the most challenging category to predict similarity. While names includes city and country names as well, we focus on the names of public figures in our work. To resolve this issue, we provide a new dataset that consists of ∼40,000 (39,593) name pairs and employ the Augmented SBERT approach [21]: we use the cross-encoder

model to label a new dataset consisting of name pairs and then train a bi-encoder model on the resulting dataset. We discuss the deployed models in more detail in Section 4.

The underlying dataset is created from an open dbpedia-data dataset [22] which includes the names of more than a million public figures that have a page on Wikipedia and DBpedia, including actors, politicians, scientists, sportsmen, and writers. Out of these we only use those with a U.S. nationality as the questions in NQ-open are on predominantly U.S. related topics. We then shuffle the list of 25,462 names and pair them randomly to get the name pairs that are then labelled by the cross-encoder model.

The dataset includes different ways of writing a person's name including aliases. For example, *Gary A Labranche* and *Labranche Gary*, or aliases like *Lisa Marie Abato*'s stage name *Holly Ryder* as well as e.g. Chinese ways of writing such as *Rulan Chao Pian* and 卞趙如蘭. We filter out all examples where more than three different ways of writing a person's name exist because in these cases these names don't refer to the same person but were mistakenly included in the dataset. For example, names of various members of Tampa Bay Rays minor league who have one page for all members. Since most public figures in the dataset have a maximum of one variation of their name, we only leave out close to 800 other variations this way, and can add 14,131 additional pairs. These are labelled as 1 because they refer to the same person.

4. MODELS / METRICS

The focus of our research lies on different semantic similarity metrics and their underlying models. As a human baseline, [1] reports correlations between the labels by the first and the second annotator for subsets of SQuAD and GermanQuAD and omits these for the NQ-open subset since they are not publicly available. Maximum Kendall's tau-b rank correlations are 0.64 for SQuAD and 0.57 for GermanQuAD. The baseline semantic similarity models considered are bi-encoder, BERTScore vanilla, and BERTScore trained, whereas the focus will be on cross-encoder (SAS) performance. Table 3 outlines the exact configurations used for each model.

Table 3: Configuration details of each of the models used in evaluations. The architectures for the first two models and our model follow corresponding sequence classification. T-systems-onsite model, as well as our trained model, follow `XLMRobertaModel`, and the other two - `BertForMaskedLM` & `ElectraForPreTraining` architectures respectively. Most of the models use absolute position embedding.

	deepset/ gbert-large-sts	**cross-encoder/ stsb-roberta-large**	**T-Systems-onsite/ cross-en-de-roberta -sentence-transformer**	**bert-base-uncased**	**deepset/ gelectra-base**	**Augmented cross-en-de-roberta -sentence-transformer**
hidden_size	1,024	1,024	768	768	768	768
intermediate_size	4,096	4,096	3,072	3,072	3,072	3,072
max_position_embeddings	512	514	514	512	512	514
model_type	bert	roberta	xlm-roberta	bert	electra	xlm-roberta
num_attention_heads	16	16	12	12	12	12
num_hidden_layers	24	24	12	12	12	12
vocab_size	31,102	50,265	250,002	30,522	31,102	250,002
transformers_version	4.9.2	-	-	4.6.0.dev0	-	4.12.2

A cross-encoder architecture [23] concatenates two sentences with a special separator token and passes them to a network to apply multi-head attention over all input tokens in one pass. Pre-computation is not possible with the cross-encoder approach because it takes both input texts into account at the same time to calculate embeddings. A well-known language model that makes use of the cross-encoder architecture is BERT [24]. The resulting improved performance in terms of more accurate similarity scores for text pairs comes with the cost of higher time complexity, i.e. lower speed, of cross-encoders in comparison to bi-encoders. A bi-encoder calculates the

embeddings of the two input texts separately by mapping independently encoded sentences for comparison to a dense vector space which can then be compared using cosine similarity. The separate embeddings result in higher speed but result in reduced scoring quality due to treating the text pairs completely separate [25]. In our work, both cross- and bi-encoder architectures are based on Sentence Transformers [26].

The original bi-encoder applied in [1] uses the multi-lingual T-Systems-onsite/cross-en-de-roberta-sentence-transformer [27] that is based on xlm-roberta-base which was further trained on an unreleased multi-lingual paraphrase dataset resulting in the model paraphrase-xlm-r-multilingual-v1. The latter then in turn was fine-tuned on an English-language STS benchmark dataset [28] and a machine-translated German STS benchmark.

[1] used a separate English and German model for the cross-encoder because there is no multi-lingual cross-encoder implementation available yet. Similar to the bi-encoder approach, the English SAS cross-encoder model relies on cross-encoder/stsb-roberta-large which was trained on the same English STS benchmark. For German, a new cross-encoder model had to be trained, as there were no German cross-encoder models available. It is based on deepset's gbert-large [29] and trained on the same machine-translated German STS benchmark as the bi-encoder model, resulting in gbert-large-sts.

BERTScore implementation from [6] is used for our evaluation, with minor changes to accommodate for missing key-value pairs for the [27] model type. For BERTScore trained, the last layer representations were used, while for vanilla type BERTScore, only the second layer. **BERTScore vanilla** is based on bert-base-uncased for English (SQuAD and NQ-open) and deepset's gelectra-base [29] for German (GermanQuAD), whereas **BERTScore trained** is based on the *multi-lingual* model that is used by the bi-encoder [27]. BERTScore trained outperforms SAS for answer-prediction pairs without lexical overlap, the largest group in NQ-open, but neither of the models perform well on names. We use our new name pairs dataset to train the Sentence Transformer with the same hyperparameters as were used to train paraphrase-xlm-r-multilingual-v1 on the English-language STS benchmark dataset.

We did an automatic hyperparameter search Table 6 for 5 trials with Optuna [30]. Note that cross-validation is an approximation of Bayesian optimization, so it is not necessary to use it with Optuna. The following set of hyperparameters was found to be the best: 'batch': 64, 'epochs': 2, 'warm': 0.45.

We have scanned all metrics from Table 5 for time complexity on NQ-open as it is the largest evaluation dataset. Note that we haven't profiled training times as those are not defined for lexical-based metrics, but only measured CPU time for predicting answer pairs in NQ-open. N-gram based metrics are much faster as they don't have any encoding or decoding steps involved, and they take ∼10s to generate similarity scores. The slowest is the cross-encoder as it requires concatenating answers first, followed by encoding, and it takes ∼10 minutes. Concatenation grows on a quadratic scale with the input length due to self-attention mechanism. For the same dataset, bi-encoder takes ∼2 minutes. BERTScore trained takes ∼3 minutes, hence computational costs of BERTScore and bi-encoders are comparable. Additional complexity for all methods mentioned above except for SAS would be marginal when used during training on the validation set. Please note the following system description:

```
System name='Darwin', Release='20.6.0', Machine='x86_64',
Total Memory=8.00GB, Total cores=4, Frequency=2700.00Mhz
```

5. ANALYSIS

To evaluate the shortcomings of lexical-based metrics in the context of QA, we compare BLEU, ROUGE-L, METEOR, F_1 and the semantic answer similarity metrics, i.e. Bi-Encoder, BERTScore vanilla, BERTScore trained, and Cross-Encoder (SAS) scores on evaluation datasets. To address the second hypothesis, we delve deeply into every single dataset and find differences between different types of answers.

5.1. Quantitative Analysis

As can be observed from Table 4 and Table 5, lexical-based metrics show considerably lower results than any of the semantic similarity approaches. BLEU lags behind all other metrics, followed by METEOR. Similarly, we found that ROUGE-L and F1 achieve close results. In the absence of lexical overlap, METEOR gives superior results than the other n-gram-based metrics in the case of SQUAD, but ROUGE-L is closer to human judgement for the rest. The highest correlations are achieved in the case of BERTScore based trained models, followed closely by bi- and cross-encoder models. We found some inconsistencies regarding the performance of the cross-encoder based SAS metric. The superior performance of SAS doesn't hold up for the correlation metrics other than Pearson. We observed that SAS score underperformed when $F_1 = 0$ compared to all other semantic answer similarity metrics and overperformed when there is some lexical similarity.

NQ-open is not only by far the largest of the three datasets but also the most skewed one. We observe that the vast majority of answer-prediction pairs have a label 0 (see Table 2). In the majority of cases, the underlying QA model predicted the wrong answer.

All four semantic similarity metrics perform considerably worse on NQ-open than on SQuAD and GermanQuAD. In particular, answer-prediction pairs that have no lexical overlap ($F_1 = 0$) amount to 95 per cent of all pairs with the label 0 indicating incorrect predictions. Additionally, they perform only marginally better than METEOR or ROUGE-L.

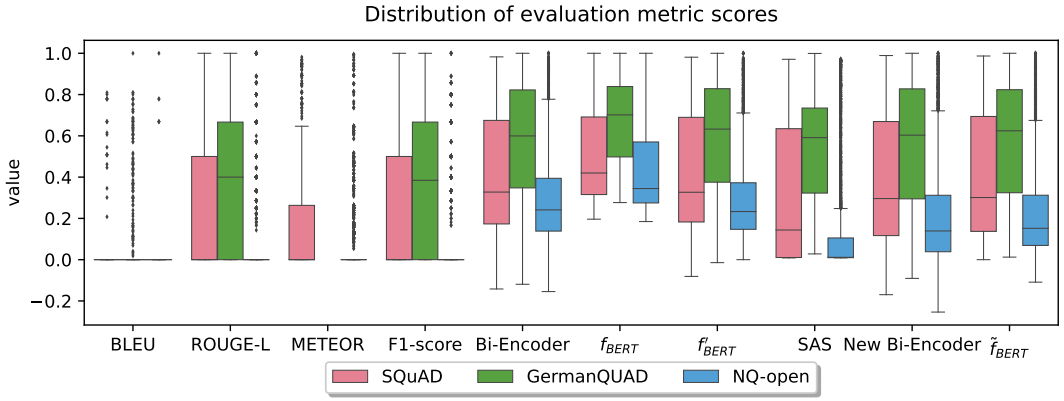


Figure 2: Comparison of all (similarity) scores for the pairs in evaluation datasets. METEOR computations for GermanQuAD are omitted since it is not available for German.

Score distribution for SAS and BERTScore trained shows that SAS scores are heavily tilted towards 0 Figure 2.

In Figure 3, we analyse SQuAD subset dataset of answers and we observe a similar phenomenon as in [1] when there is no lexical overlap between the answer pairs: the higher in layers we go in case of BERTScore trained, the higher the correlation values with human labels are. Quite the opposite is observed in the case of BERTScore vanilla, where it is either not as sensitive to embedding representations in case of no lexical overlap or correlations decrease with higher embedding layers.

Table 4: Pearson, Spearman's, and Kendall's rank correlations of annotator labels and automated metrics on subsets of GermanQuAD. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

	GermanQuAD					
	$F_1 = 0$			$F_1 \neq 0$		
Metrics	r	ρ	τ	r	ρ	τ
BLEU	0.000	0.000	0.000	0.153	0.095	0.089
ROUGE-L	0.172	0.106	0.100	0.579	0.554	0.460
F_1-score	0.000	0.000	0.000	0.560	0.534	0.443
Bi-Encoder	0.392	0.337	0.273	0.596	0.595	0.491
f_{BERT}	0.149	0.008	0.006	0.599	0.554	0.457
f'_{BERT}	0.410	0.349	0.284	0.606	0.592	0.489
SAS	**0.488**	**0.432**	**0.349**	**0.713**	**0.690**	**0.574**

Table 5: Pearson, Spearman's, and Kendall's rank correlations of annotator labels and automated metrics on subsets of SQuAD and NQ-open. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained, and $\tilde{f}_{BERT}$ is the new BERTScore trained on names.

	SQuad						**NQ-open**					
	$F_1 = 0$			$F_1 \neq 0$			$F_1 = 0$			$F_1 \neq 0$		
Metrics	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
BLEU	0.000	0.000	0.000	0.182	0.168	0.159	0.000	0.000	0.000	0.052	0.054	0.051
ROUGE-L	0.100	0.043	0.041	0.556	0.537	0.455	0.220	0.163	0.159	0.450	0.458	0.377
METEOR	0.398	0.207	0.200	0.450	0.464	0.378	0.233	0.152	0.148	0.188	0.179	0.139
F1-score	0.000	0.000	0.000	0.594	0.579	0.497	0.000	0.000	0.000	0.394	0.407	0.337
Bi-Encoder	0.487	0.372	0.303	0.684	0.684	0.566	0.294	0.212	0.170	0.454	0.446	0.351
f_{BERT}	0.249	0.132	0.108	0.612	0.601	0.492	0.156	0.169	0.135	0.165	0.142	0.112
f'_{BERT}	0.516	0.391	0.318	0.698	0.688	0.571	0.319	0.225	0.181	0.452	0.449	0.354
SAS	**0.561**	0.359	0.291	**0.743**	**0.735**	**0.613**	**0.422**	0.196	0.158	**0.662**	**0.647**	**0.512**
New Bi-Encoder	0.501	0.391	0.318	0.694	0.690	0.572	0.338	0.252	0.203	0.501	0.501	0.392
$\tilde{f}_{BERT}$	0.519	**0.399**	**0.324**	0.707	0.698	0.581	0.351	**0.257**	**0.208**	0.498	0.507	0.398

5.2. Qualitative Analysis

This section is entirely dedicated to highlighting the major categories of problematic samples in each of the datasets.

5.2.1. SQuAD

In **SQuAD** there are only 16 cases where SAS completely diverges from human labels. In all seven cases where SAS score is above 0.5 and label is 0, we notice that the two answers have either **a common substring** or could be used often in the same context. In the other 9 extreme cases when the label is indicative of semantic similarity and SAS is giving scores below 0.25, there are three

Table 6: Experimental setup for hyperparameter tuning of cross-encoder augmented BERTScore.

Batch Size	{16, 32, 64, 128, 256}
Epochs	{1, 2, 3, 4}
warm	uniform(0.0, 0.5)

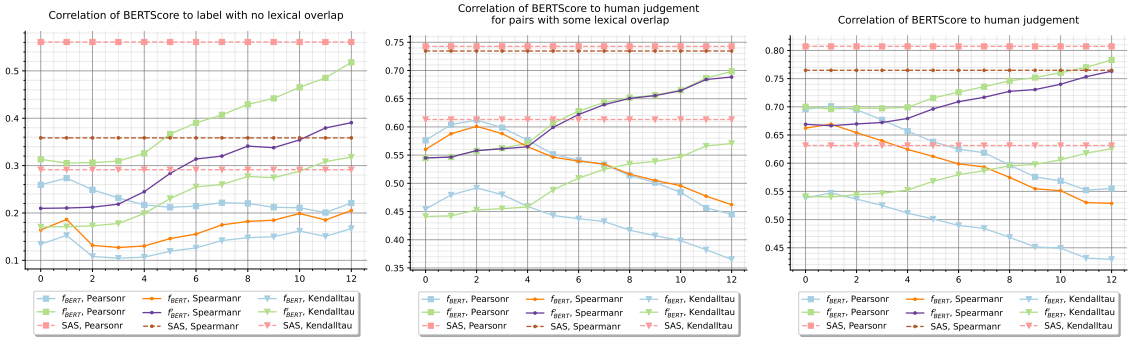


Figure 3: Pearson, Spearman's, and Kendall's rank correlations for different embedding extractions for when there is no lexical overlap ($F_1 = 0$), when there is some overlap ($F_1 \neq 0$) and aggregated for the SQuAD subset. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

spatial translations. There is an encoding-related example with 12 and 10 special characters each which seems to be a mislabelled example.

5.2.2. GermanQuAD

Overall, error analysis for GermanQuAD is limited to a few cases because it is the smallest dataset of the three and all language model based metrics perform comparably well - SAS in particular. SAS fails to identify semantic similarity in cases where the answers are **synonyms or translations** which also include technical terms that rely on Latin (e.g. *vis viva* and *living forces* (translated) (SAS score: 0.5), *Anorthotiko Komma Ergazomenou Laou* and *Progressive Party of the Working People* (translation) (0.04), *Nährgebiet* and *Akkumulationsgebiet* (0.45), *Zehrgebiet* and *Ablationsgebiet* (0.43)). This is likely the case because SAS does not use a multilingual model. Since multilingual models have not been implemented for cross-encoders yet, this remains an area for future research. Text-based **calculations and numbers** are also problematic: (translated) *46th day before Easter Sunday* and *Wednesday after the 7th Sunday before Easter* (0.41).

SAS also fails to recognise **aliases or descriptions of relations** that point to the same person or object: *Thayendanegea* and *Joseph Brant* (0.028) are the same people. BERTScore vanilla and BERTScore trained both find some similarity (0.36, 0.22). *Goring House* and *Buckinghams Haus* (0.29) refer to the same object but one is the official name, the other one a description of the same, again BERTScore vanilla and BERTScore trained identify more similarity (0.44, 0.37).

5.2.3. NQ-open

We also observe that similarity scores for answer-prediction pairs which include numbers, e.g. an amount, a date or a year, SAS, as well as BERTScore trained, diverge from labels. The only semantically similar entities to answers expected to contain a numeric value should be the exact value, not a unit more or less. Also, the position within the pairs seems to matter for digits and their string representation. For SAS that the pair of *11* and *eleven* has a score of 0.09 whereas the pair of *eleven* and *11* has a score of 0.89.

Figure 4 depicts the major error categories for when SAS scores range below 0.25 while human annotations indicate a label of 2. We observe that entities related to names, which includes spatial names as well as co-references and synonyms, form the largest group of scoring errors. After correcting for encoding errors and fixing the labels manually in the NQ-open subset, totalling 70 samples, the correlations have already improved by about a per cent for SAS. Correcting wrong labels in extreme cases where SAS score is below 0.25 and the label is 2 or when SAS is above

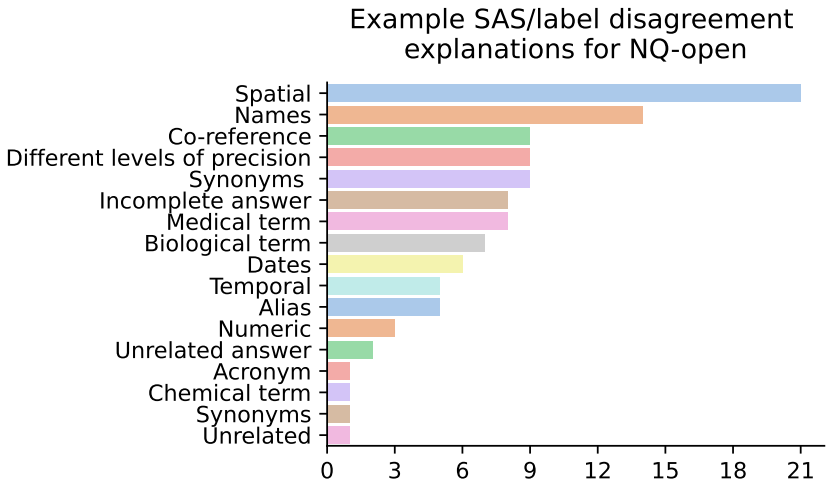


Figure 4: Subset of NQ-open test set, where SAS score < 0.01 and human label is 2, manually annotated for an explanation of discrepancies. Original questions and Google search has been used to assess the correctness of the gold labels.

0.5 and label is 0 improves results almost across the board for all models, but more so for SAS. After removal of duplicates, sample with imprecise questions, wrong gold label or multiple correct answers, we are left with 3559 ground-truth answer/prediction pairs compared to 3658 we started with.

An example for the better performance on names when applying our new bi-encoder and SBERT trained models can be seen in Figure 5, where both models perform well in comparison to SAS and human judgement.

6. Conclusion

Existing evaluation metrics for QA models have various limitations. N-gram based metrics suffer from asymmetry, strictness, failure to capture multi-hop dependencies and penalise semantically-critical ordering, failure to account for relevant context or question, to name a few. We have found patterns in the mistakes that SAS was making. These include **spatial awareness**, **names**, **numbers**, **dates**, **context awareness**, **translations**, **acronyms**, **scientific terminology**, **historical events**, **conversions**, **encodings**.

The comparison to annotator labels is performed on answer pairs taken from subsets of SQuAD and GermanQuAD datasets, and for NQ-open we have a prediction and ground-truth answer pair. For cases with lexical overlap, ROUGE-L achieves comparative results to pre-trained semantic similarity evaluation models at a fraction of computation costs that the other models require. This holds for all GermanQuAD, SQuAD and NQ-open alike. We conclude that to further improve the semantic answer similarity evaluation in German, future work should focus on providing a larger dataset of answer pairs, since we could find only a few examples where SAS and the other metrics based on pre-trained language models didn't match with human annotator labels. Dataset size was one of the reasons why we focused more heavily on NQ-open dataset. In addition, focusing on the other two would mean less strong evidence on how the metric will perform when applied to model predictions behind a real-world application. Furthermore, all semantic similarity metrics failed to have a high correlation to human labels when there was no token-level overlap, which is arguably the most important use-case for a semantic answer similarity metric as opposed to, say, ROUGE-L. NQ-open happened to have the largest number of samples that satisfied this requirement. Removing duplicates and re-labelling led to significant improvements across the board. We have generated a

Figure 5: Representative example of a question and all semantic answer similarity measurement results.

Question: Who killed Natalie and Ann in Sharp Objects?
Ground-truth answer: Amma
Predicted Answer: Luke
EM: 0.00
$\mathbf{F_1}$: 0.00
Top-1-Accuracy: 0.00
SAS: 0.0096
Human Judgment: 0
$\mathbf{f_{BERT}}$: 0.226
$\mathbf{f'_{BERT}}$: 0.145
Bi-Encoder: 0.208
$\mathbf{\tilde{f}_{BERT}}$: 0.00
Bi-Encoder (new model): -0.034

names dataset, which was then used to fine-tune the bi-encoder and BERTScore model. The latter achieves and beats SOTA rank correlation figures when there is no lexical overlap for datasets with English as the core language. Bi-encoders outperformed cross-encoders on answer-prediction pairs without lexical overlap both in terms of correlation to human judgement and speed, which makes them more applicable in real-world scenarios. This, plus support for multilingual setups, could be essential for companies as well because models most probably won't understand the relationships between different employees and stakeholders mentioned in internal documents. A reason to have a preference towards BERTScore would be the ability to use BERTScore as a training objective to generate soft predictions, allowing the network to remain differentiable end-to-end.

An element of future research would be further improving the performance on names of public figures as well as spatial names like cities and countries. Knowledge-bases, such as Freebase or Wikipedia, as explored in [11], could be used to find an equivalent answer to named geographical entities. Numbers and dates which is the problematic data type in multi-lingual, as well as monolingual contexts, would be another dimension.

7. ACKNOWLEDGEMENTS

We would like to thank Ardhendu Singh, Julian Risch, Malte Pietsch and XCS224U course facilitators, Ankit Chadha in particular, as well as Christopher Potts for their constant support.

8. REFERENCES

[1] J. Risch, T. Möller, J. Gutsch, and M. Pietsch, "Semantic answer similarity for evaluating question answering models," *arXiv preprint arXiv:2108.06130*, 2021.

[2] X.-M. Zeng, "Semantic relationships between contextual synonyms," *US-China education review*, vol. 4, pp. 33–37, 2007.

[3] S. Min, J. Boyd-Graber, C. Alberti, D. Chen, E. Choi, M. Collins, K. Guu, H. Hajishirzi, K. Lee, J. Palomaki, C. Raffel, A. Roberts, T. Kwiatkowski, P. Lewis, Y. Wu, H. Küttler, L. Liu, P. Minervini, P. Stenetorp, S. Riedel, S. Yang, M. Seo, G. Izacard, F. Petroni, L. Hosseini, N. D. Cao, E. Grave, I. Yamada, S. Shimaoka, M. Suzuki, S. Miyawaki, S. Sato, R. Takahashi, J. Suzuki, M. Fajcik, M. Docekal, K. Ondrej, P. Smrz, H. Cheng, Y. Shen, X. Liu, P. He, W. Chen, J. Gao, B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko,

M. Schlichtkrull, S. Gupta, Y. Mehdad, and W.-t. Yih, "Neurips 2020 efficientqa competition: Systems, analyses and lessons learned," in *Proceedings of the NeurIPS 2020 Competition and Demonstration Track* (H. J. Escalante and K. Hofmann, eds.), vol. 133 of *Proceedings of Machine Learning Research*, pp. 86–111, PMLR, 06–12 Dec 2021.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv:1606.05250*, 2016.

[5] J. Bulian, C. Buck, W. Gajewski, B. Boerschinger, and T. Schuster, "Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation," *arXiv preprint arXiv:2202.07654*, 2022.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, Nov. 2019.

[8] A. Chen, G. Stanovsky, S. Singh, and M. Gardner, "Evaluating question answering evaluation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, (Hong Kong, China), pp. 119–124, Association for Computational Linguistics, Nov. 2019.

[9] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

[10] T. Möller, J. Risch, and M. Pietsch, "Germanquad and germandpr: Improving non-english question answering and passage retrieval," *arXiv:2104.12741*, 2021.

[11] C. Si, C. Zhao, and J. Boyd-Graber, "What's in a name? answer equivalence for open-domain question answering," *arXiv preprint arXiv:2109.05289*, 2021.

[12] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: a benchmark for question answering research," *Transactions of the Association of Computational Linguistics*, 2019.

[13] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *CoRR*, vol. abs/1705.03551, 2017.

[14] N. Reimers, P. Beyer, and I. Gurevych, "Task-oriented intrinsic evaluation of semantic textual similarity," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (Osaka, Japan), pp. 87–96, The COLING 2016 Organizing Committee, Dec. 2016.

[15] C. Croux and C. Dehon, "Influence functions of the spearman and kendall correlation measures," *Stat Methods Appl (2010) 19:497–515*, 2010.

[16] T. Niven and H.-Y. Kao, "Probing neural network comprehension of natural language arguments," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4658–4664, Association for Computational Linguistics, July 2019.

[17] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, *et al.*, "Dynabench: Rethinking benchmarking in nlp," *arXiv preprint arXiv:2104.14337*, 2021.

[18] B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li, "Adversarial

GLUE: A multi-task benchmark for robustness evaluation of language models," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[19] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 610–623, Association for Computing Machinery, 2021.

[20] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, Mar. 2019.

[21] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks," *arXiv:2010.08240v2 [cs.CL]*, 2021.

[22] C. Wagner, "Politicians on wikipedia and dbpedia," 2017.

[23] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring," 2020.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[25] J. Chen, L. Yang, K. Raman, M. Bendersky, J.-J. Yeh, Y. Zhou, M. Najork, D. Cai, and E. Emadzadeh, "DiPair: Fast and accurate distillation for trillion-scale text matching and pair modeling," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 2925–2937, Association for Computational Linguistics, Nov. 2020.

[26] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019.

[27] P. May, "T-systems-onsite/cross-en-de-roberta-sentence-transformer," *Hugging Face*, 2020.

[28] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 1–14, Association for Computational Linguistics, Aug. 2017.

[29] B. Chan, S. Schweter, and T. Möller, "German's next language model," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 6788–6796, International Committee on Computational Linguistics, Dec. 2020.

[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

9. AUTHORS

Peter F. Ebbinghaus is currently SEO Team Lead at Teufel. With a background in econometrics and copywriting, his work focuses on the quantitative analysis of e-commerce content. Based in Germany and Mexico, his research includes applied multi-lingual NLP in particular.

Farida Mustafazade is a London-based quantitative researcher in the GAM Systematic Cambridge investment team where she focuses on macro and sustainable macro trading strategies. Her area of expertise expands to applying machine learning, including NLP techniques, to financial as well as alternative data.

A DISTRIBUTED ENERGY-EFFICIENT UNEQUAL CLUSTERING BASED KRUSKAL HEURISTIC FOR IOT NETWORKS

Mohamed Sofiane BATTA[1,2], Zibouda ALIOUAT[2], Hakim MABED[1], and Malha MERAH[2]

[1] FEMTO-ST Institute/DISC, University of Bourgogne Franche-Comte, Montbeliard, France

[2] LRSD Laboratory, Computer Science Dept., Ferhat Abbas University Setif 1, Setif, Algeria

ABSTRACT

Energy efficiency is a major concern and a critical issue for energy constrained wireless networks. In this context, clustering is commonly used for topology management and maximizing the network lifetime. Clustering approaches typically use a multi-hopping mechanism where Cluster Heads (CHs) near the Base Station (BS) consume higher energy since they relay data of farther CHs. Therefore, nodes close to the BS are strangled with an overloaded routing task and tend to die earlier than their intended lifetime, which affects the network performance. This situation is known as the hot spot problem that induces unbalanced energy consumption among CHs. The concern in this work is to address the intra-clustering structure in large scale environments to tolerate the network scaling and reasonably balance the energy consumption among CHs. In this regard, we propose a new Unequal Clustering algorithm based on Kruskal heuristic (UCKA) to optimize the network lifetime. UCKA applies the Kruskal heuristic in a distributed fashion to perform a minimum spanning tree within large cluster which strengthen the intra-cluster routing structure and reduce the energy devoted to wireless communications. To the best of our knowledge, this is the first solution that combines the Kruskal heuristic and the unequal clustering to extend the devices durability and alleviate the hot spot problem. Simulation results indicate that UCKA can effectively reduce the energy consumption and lengthen the network lifetime.

KEYWORDS

IoT, WSN, Energy-aware protocols, Unequal Clustering, Hot spot energy problem, Kruskal Heuristic.

1. INTRODUCTION

The efficient use of the Internet of Things (IoT) and Wireless Sensor Networks (WSNs) enables a practical solution for various applications and impacts several daily life aspects. These networks have become a point of interest for many researchers due to their implementation in various real-world scenarios from environmental monitoring and traffic surveillance to smart city and healthcare systems. These networks are made of a large number of intelligent devices, known as connected devices, distributed over a large geographical area and cooperatively interact to perform a specific task.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 117-129, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121110

The communication with the external world is made using a gateway called Base Station (BS). Each device represents a limited on-board processing, limited memory and low radio capacity. Besides, these devices are usually powered by non-rechargeable batteries. Therefore, due to the extensive tasks handled by network devices, energy awareness is a major concern and a critical design issue for wireless networks. To strengthen the energy efficiency and increase the network scalability, the IoT network is commonly organized into clusters. Several clustering protocols are proposed in the literature [5,6,16]. Network devices are grouped into clusters, with a representative Cluster Head (CH) in each cluster responsible for collecting and aggregating data of Cluster Members (CMs) as shown in Figure 1.

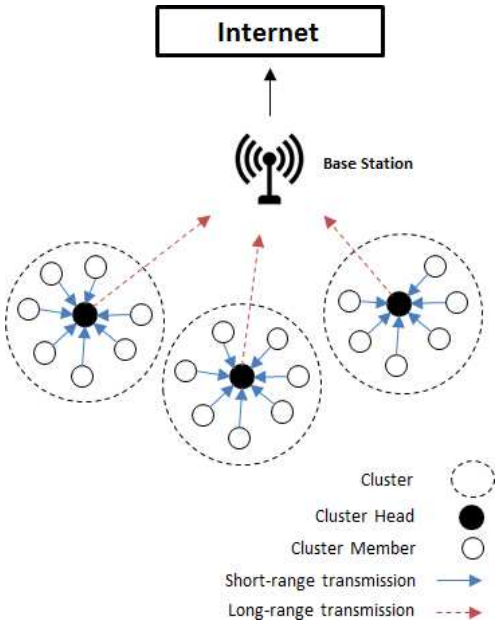


Figure 1. Network clustering topology.

The network clustering is a popular energy efficient technique to improve the energy consumption [9]. In this technique, multi-hop communications between the data sources (CHs) and the gateway (BS) are used to preserve the energy. However, with multi-hop routing, CHs cooperate with each other to forward the data to the base station (BS). Consequently, CHs closer to the BS are burdened with heavy relay traffic of other CHs and tend to die earlier [18]. This scheme reduces the network coverage (sensing area for example) and may cause a network partitioning, especially in large scale networks. Proposed clustering approaches usually generate clusters of even size [21]. Therefore, nodes closer to the BS waste much more energy during the inter-cluster routing because they have a higher load of relaying remote CHs traffic. Thereby, they deplete their energy quickly. This issue is known as the hot spot problem [8]. One solution to mitigate this problem is the use of unequal clustering method, where the network is partitioned into uneven clusters as in Energy-efficient Multi-hop routing with Unequal Clustering (EMUC) [11] and Multi-hop Low Energy Adaptive Clustering Hierarchy (MH-LEACH) [1].

Several unequal clustering proposals have been proposed to extend the network lifetime [19, 22]. However, the majority of these approaches [2, 8] consider small networks and focus on controlling the size of each cluster to balance and reduce energy dissipation, whereas, the intra-clustering structure of the generated clusters is not considered. Moreover, in most cases the connections between the CMs and their CH are assumed to be direct, which may obstruct the communication performances when the network scales (due to the collisions and interference problems [4]). In this work we propose an unequal clustering approach to alleviate the hot spot problem in large scale networks. In the proposed scheme, clusters size varies proportionally to the distance toward the BS. Clustering topology is generated using the Kruskal spanning tree heuristic [14] to reduce the intra-cluster communication cost and expend the cluster lifespan. The objective of integrates the Kruskal heuristic in the unequal clustering is to alleviate the hot spot problem and optimize the network lifetime. Both residual energy and the nodes density are considered in the clustering process to strengthen the energy efficiency.

The proposed approach aims to ensure the load balancing of the energy consumption among CHs by varying clusters size. Clusters, close to the BS, represent small and equal size, where cluster members are directly linked to the corresponding CH. In this case, the CH communicates directly with the BS. Whereas, remote clusters are larger and cluster members are organized into a spanning tree where the CH is the root. The spanning tree is generated using the Kruskal heuristic and defines the intra-cluster communication links between the cluster members. Distant CHs relay their cluster data to the BS through multi-hop transmissions between CHs. With this scheme, smaller clusters are assigned for CHs nearby the BS since they act as a router node for other far CHs. On the other hand, distant nodes spent their energy for multi-hop intra-cluster transmissions.

The rest of this paper is organized as follows: Section 2 describes some related works. Section 3 presents the contribution. Simulation settings and results are discussed in Section 4. We conclude our work in Section 5.

2. RELATED WORKS

The lifespan of sensor nodes is restricted by the lifetime of their batteries. To preserve the energy and increase the lifespan of the wireless devices, many clustering algorithms have been proposed [2,7,8,21]. In this section, some of the relevant works in this field are reviewed.

Low Energy Adaptive Clustering Hierarchy (LEACH) algorithm [13] is one of the most prevalent clustering techniques. To reduce the energy consumption of nodes and enhances their lifetime, each node computes a probability value of subsequently becoming a CH during the next round. The main disadvantage of LEACH is that some CHs may be close to each other. Besides, LEACH uses one-hop communication architecture, i.e. CHs are directly connected to the BS. Several LEACH versions [1,15] attempt to overcome LEACH drawbacks and improve this scheme. Nie et al. [8] proposed a Lifetime-Aware Clustering approach based on a Directed Acyclic Graph (DAG) algorithm. The authors used a linear programming approach to construct a connected routing tree within the network. Xia et al. [23] presented a distributed energy-efficient approach based on Unequal Clustering and Connected Graph theory (UCCGRA). A voting mechanism is used to select the final set of CHs. The clustering takes into count the residual energy of nodes and the distance between the neighbours to elect the cluster heads. A Connected Graph Theory (CGT) is used for inter-cluster communication in order to reduce energy consumption. CHs form a connected routing tree with the BS as root. Therefore, CHs maintain many paths to reach the BS which improves the reliability of transition.

An Unequal Multi-hop Balanced Immune Clustering protocol for wireless networks (UMBIC) is presented in [20]. It uses an Unequal Clustering Mechanism (UCM) and a Multi-Objective Immune Algorithm (MOIA) to avoid the hot spot problem. UMBIC aims to strengthen the network coverage, by producing an optimized routing and ensures low communication cost among nodes. The CH replacement occurs when the residual energy of the CH becomes lower than a particular threshold value. Authors in [12] presented a new clustering approach called Energy Degree Distance Unequal Clustering Algorithm (EDDUCA) to approximate the equalization of energy consumption in a wireless network and eliminate the hot spot problem. The network is divided into unequal clusters by using the Sierpinski triangle method [12]. The clustering is based on the residual energy, node degree, and distance toward the BS. A weight is associated with each node, which is calculated based on the above three criteria. In each cluster, the node with the maximal weight is selected as CH. The objective of EDDUCA is to extend the lifetime of CHs closer to the BS and maintain the clusters' connectivity.

Baranidharan et al. [3] introduced a new Distributed Unequal Clustering algorithm using Fuzzy logic (DUCF). A Fuzzy Inference System (FIS) is adapted for the CHs election and the energy load balancing. The FIS system considers the residual energy, node degree, and distance to the BS as input and assigns a maximum limit of CMs to each cluster. DUCF uses a Centroid method for defuzzification. The FIS specifies the probability of each node to become CH and determines the size of each cluster according to the input parameters. DUCF assigns a maximum limit of CMs for a CH based on its residual energy and number of neighbours in order to bypass the hot spot problem. To extend clusters lifetime, authors in [10] proposed an Energy Efficient intra-clustering technique (EE3C) where multiple high energy nodes act as CHs within each cluster instead of a single cluster head. The BS collects nodes' information to determine their location and lifetime. Then it divides the network into rectangular sectors (clusters) to efficiently distribute the energy in the monitored area. The CH election is centralized at the BS. Only one cluster head acts as master CH for a given cluster to send the collected information to the BS. The master CH changes periodically among CHs after a particular number of rounds. EE3C improves the intra-clustering efficiency, however, the clustering process is fully centralized. The authors also presented a k-hop Energy Constrained intra-clustering technique based on the Dominating Sets theory called K-ECDS. The proposed algorithm takes into account the energy limitation. Besides, K-ECDS models the problem of choosing cluster heads using the quality of communication channels and neighbours cardinality.

Unequal cluster-based protocols can mitigate the hot spot problem. However, existing schemes focus on managing the clusters size and do not consider the intra-clustering structure of the generated clusters, which may restrict the scalability in case of a large scale network. Thereby, in our proposed approach, the intra-clustering topology is considered and the proposed scheme is evaluated under a large network with different nodes density to cover several use-case scenarios.

3. THE PROPOSED SCHEME

The wireless network is mapped into a graph G(V, E), where V represents the set of nodes (sensor devices) and the edges E constitutes the communication links. The set of neighbouring nodes of a node i is described by $N(i) = \{j \in V \mid (i, j) \in E\}$. The distance (number of hops) between two nodes i and j is represented by $D_{i,j}$. The transmitting range is denoted by T_x. The proposed scheme uses a weight-based technique that selects the node with the maximum weight in its neighbourhood as a CH. Indeed, in addition to designing an adequate cluster, the CHs selection criteria is crucial for balancing the energy usage among the entire network in a way that all devices operations finish at approximately the same time, i.e. this criteria ensures that the network lifespan is optimized [18].

The nodes weight is calculated based on the following two parameters.

1. **Remaining energy (RE_i)**: As the CH performs more tasks than ordinary nodes, it requires more energy. Therefore, the residual energy RE_i of each node i is considered during the CHs election. To estimate the battery charge diminution, we used the energy consumption model proposed in [7].

2. **Node density (δ_i)**: Due to the aggregation and interference problems, the energy consumed by a node increases according to its degree. Thus, this parameter is taken into consideration. The neighbourhood density is computed using the received signal strength [24] of adjacent nodes.

Based on the previous parameters, the weight of a node i is computed as in Equation 1.

$$W_i = \{\alpha \times RE_i + \beta \times \delta_i\} \wedge \alpha + \beta = 1 \quad (1)$$

δ_i represents the degree of the node i. α and β represent the weighting coefficient of the two criteria: residual energy and node degree respectively. α and β are chosen regarding the target application and the surrounding environment.

A particular weight coefficient may be adjusted relatively to the others to acquire an optimal result for a particular network configuration. For example, in a low density environment, the residual energy should be favoured. Whereas, in case of a dense network, the connectivity should be considered. The proposed approach is designed to operate under a typical network with different configurations to cover various use-case scenarios. Hence, in this experiment, the weight coefficients are considered equal ($\alpha = \beta = 0.5$).

In the proposed scheme, the network is divided into three types of unequal clusters: small, medium, and large clusters. The size of a cluster depends on the distance between the CH and the BS. Clusters close to the BS are assigned a small size. Hence, CHs of these clusters regroup a reduced number of members, which reserves their energy for routing remote CHs data. In small clusters, intra-cluster communication is done by direct transmission between the cluster member node (CM) and its CH. Medium and large clusters are two and three times larger than small clusters. As the size increases, clusters tend to regroup more members, and direct transmission between the CMs and their CHs may not be feasible. That is why medium and large clusters employ the two-hop and three-hop intra cluster communications to route CMs data to the corresponding CH. Figure 2 shows an overview of the UCKA clustering topology. Network nodes determine their Cluster Range $CR_i \in \{small, medium, large\}$ using Equation 2.

$$\Omega i = \frac{D_{MAX} - D_{i,BS}}{D_{MAX} - D_{MIN}}$$

$$CR_i = \begin{cases} Small & \Omega i \leq 1/3 \\ Medium & 1/3 < \Omega i \leq 2/3 \\ Large & \text{otherwise} \end{cases} \quad (2)$$

D_{MIN} and D_{MAX} are estimated by the BS, they represent the distance between the nearest and the furthest node with respect to the BS. In order to improve the energy efficiency of routing within clusters, medium and large clusters use the Kruskal spanning tree to design the intra-cluster topology.

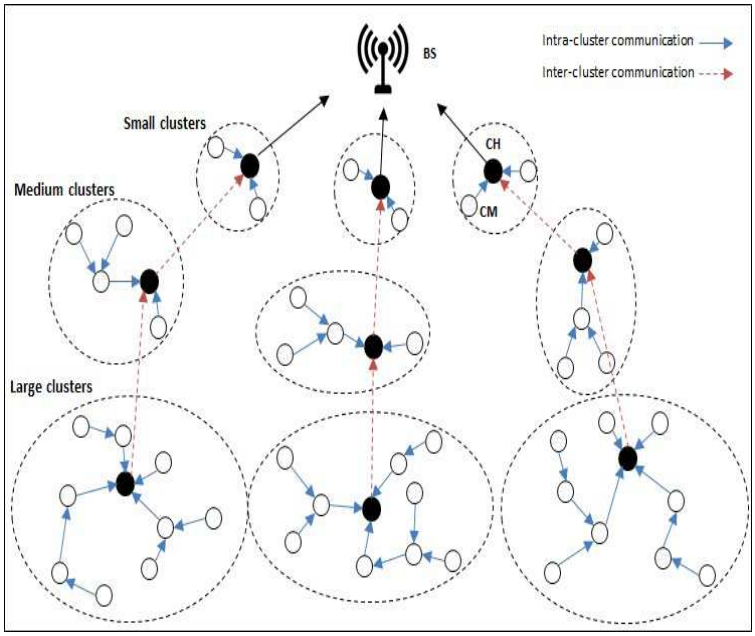


Figure 2. UCKA clustering topology.

The Kruskal heuristic computes the minimum spanning tree within each cluster. For this purpose, each CM constitutes initially a singleton tree. Then at each stage, separated trees come together in pairs. The connection between the isolated trees is made using the shortest path that relays these trees (without forming a cycle). Successively, the process is repeated until connecting all the trees in the cluster. The heuristic result is a single minimum spanning tree (MST) obtained using successive connections between isolated trees. The MST relays each node to its CH using the shortest routing path. The clustering procedure is described in algorithm 1.

Algorithm 1 UCKA clustering process

Begin
STEP 1: Compute the node weight (W_i) using equation 1.
STEP 2: Determine the cluster range CR_i using equation 2.
STEP 3: If the current node has the greatest weight among the neighbourhood **then**
$Wi = Max(Wj \mid \forall j \in N(i))$
Broadcast a *CH_Announcement* message
Else
Upon receiving *CH_announcement* **Do**
Send a join request *CM_join* to the CH with the highest weight.
Upon receiving a reject message (*CM_reject*) **Do**
Repeat STEP 3
STEP 4: Upon receiving *CM_join* from node j **Do**
Update list of cluster members *CM_list*
$CM\_list = CM\_list \cup j$
End

The procedure of the MST formation is centralized at the CH and is illustrated in algorithm 2. This later may require an additional time to form the clusters. However, the generated structure optimizes the energy consumption and allows the tolerance of the dynamic network topology. Figure 3 shows an execution example of the MST formation.

Algorithm 2 Intra-cluster MST formation

Begin
$T = \emptyset$ //contains the set of sub-trees in the cluster.
$S = \emptyset$ //contains the cluster edges.

STEP 1: each node in the cluster constitutes a singleton tree.
$\forall i \in CM\_list$ **Do**
$Tree_i = Create\_tree(i)$
Add $Tree_i$ to T:
$T = T \cup Tree_i$
STEP 2: Create a set S containing all the edges in the cluster
$(\forall (i, j) \in E) \wedge (i, j \subset CM\_list)$: $S = S \cup (i, j)$
STEP 3: Connect all the sub-trees in T
While $S \neq \emptyset$ **Do**
$Tree_i = Find\_Tree(i, T)$
//Find_Tree(i, T): return the tree in T to which node i belong.
$Tree_j = Find\_Tree(j, T)$
If $Tree_i \neq Tree_j$ **then**
If (i,j) connects $Tree_i$, $Tree_j$ without forming a loop **then**
$Tree_z = Merge(Tree_i, Tree_j)$
Remove $Tree_i$, $Tree_j$ from T
Add $Tree_z$ to T
endIf
endIf
Remove *(i,j)* From S
Endwhile
End

4. PERFORMANCE ANALYSIS

The simulation experiment is performed using Java Universal Network/Graph (Jung) [17] in Eclipse platform. Jung is a Java-based library that allows modelling and displaying a wireless network as a graph. The performance of UCKA is evaluated against two similar (in terms of objectives) clustering protocols, namely, EMUC [11] and MH-LEACH [1]. Sensor nodes are randomly deployed over a large area (1000×1000 m^2). Communication links are symmetric and the transmitting range of all nodes $T_x = 90$ m. The initial energy of devices is set to 1 joule. For the performance evaluation of the proposed protocol, three metrics are used, notably, the number of clusters generated, the average energy consumed, and the network lifetime. Simulation results are discussed in the following subsections.

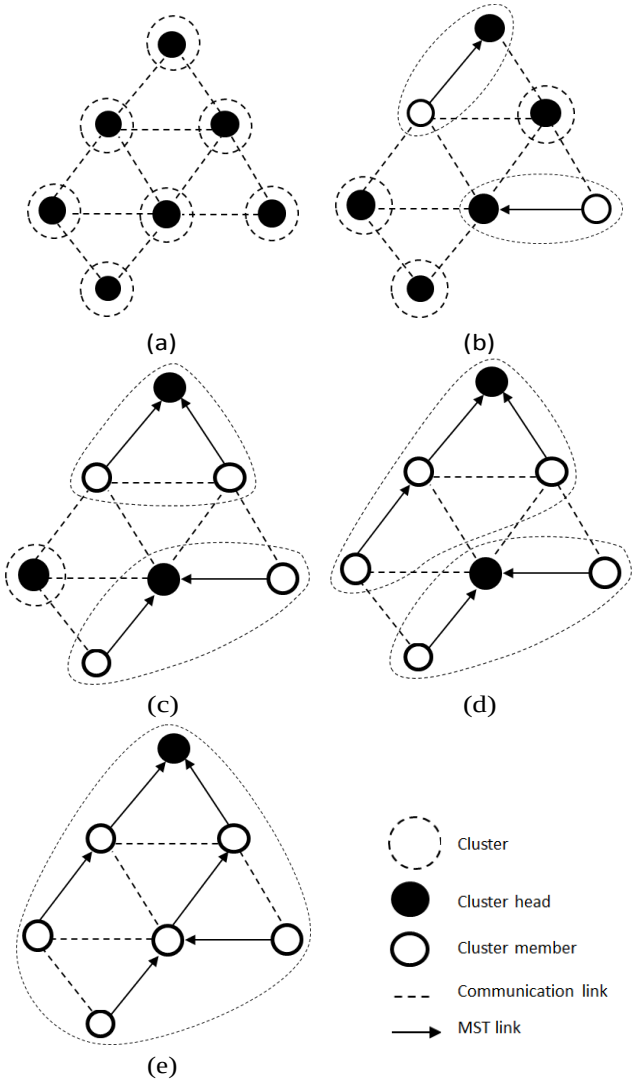


Figure 3. MST execution example.

4.1. Average number of clusters

Figure 4 compares our UCKA algorithm with EMUC and MH-LEACH methods in terms of generated clusters number. We observe that the topology generated by UCKA presents fewer clusters compared to MH-LEACH and EMUC. This is due to the fact that UCKA considers the cardinality of nodes during the clustering process and chooses the cluster size according to the distance toward the BS. Whereas, MH-LEACH does not consider nodes density. EMUC takes into account the cardinality of nodes during the CH election, but the size of the cluster only depends on the transmission range of the CH. Therefore, when the network density increases, the number of clusters generated by EMUC tends to increase. Globally, the proposed scheme shows an average improvement of 43% and 34% compared to EMUC and MH-LEACH, respectively.

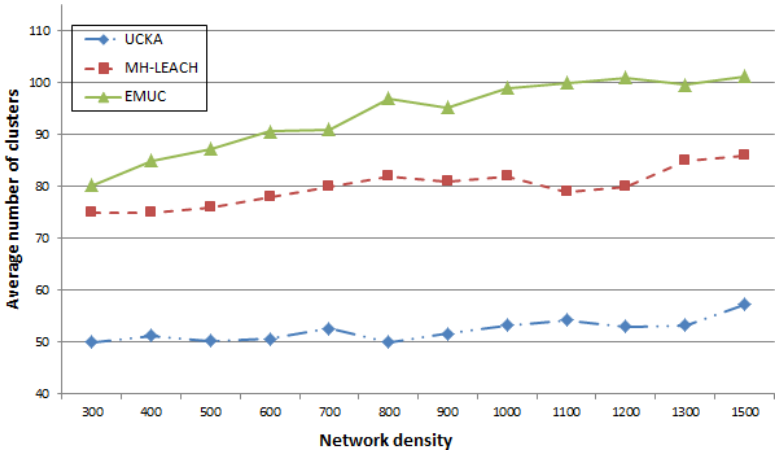


Figure 4. Average number of clusters versus network density.

4.2. Average consumed energy

Figure 5 shows the average energy consumed according to the network density. It shows that the proposed scheme has better performances compared to MH-LEACH and EMUC. The energy efficiency expresses the benefit of the clustering process. As CHs consume more energy compared to other nodes, reducing the set of selected CHs and the use of unequal clustering have balanced and reduced the energy consumption. Indeed, the proposed approach reduces the energy consumption by an average of 24.5% and 8.7% compared to EMUC and MH-LEACH respectively.

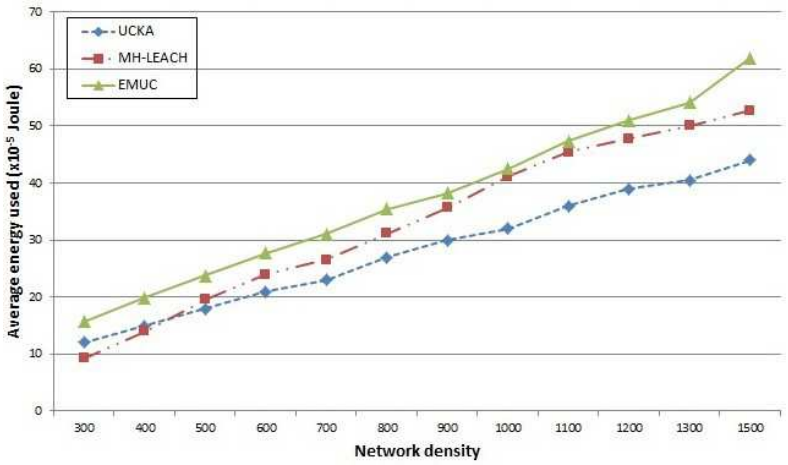


Figure 5. Average energy consumption.

4.2. Network lifetime

In this experiment, the network lifetime is defined as the time duration until the last node in the network died (measured in rounds). It can be observed in figure 6 that nodes using the proposed approach stayed alive for a longer period compared to other approaches. This is due to the usage of the unequal clustering model which enables the load balance of tasks among nodes and improves the energy solicitation. In addition, the usage of Kruskal spanning-tree method has improved the energy consumption inside the clusters. Consequently, UCKA has improved the lifespan of nodes and extended the network durability by an average of 41.8% and 12.8% compared to MH-LEACH and EMUC respectively.

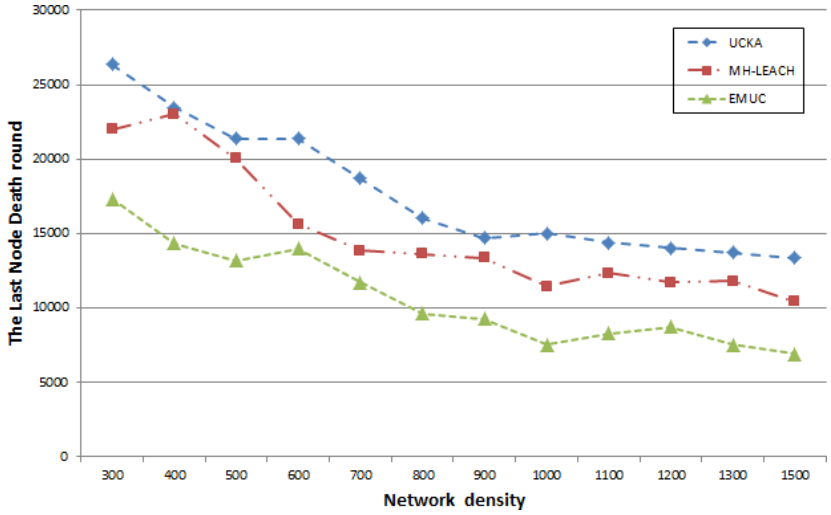


Figure 6. Average network lifetime.

5. CONCLUSION AND FUTURE WORK

The hot spot energy problem represents one of the main limits of wireless networks. Indeed, devices close to the base station experience a higher data relay load, which exhaust their energy and leads to an earlier dead comparing to other devices. In this paper, we presented an energy efficient unequal clustering approach based on the Kruskal heuristic to minimize the energy dissipation and mitigate the hot spot problem in large scale IoT networks. The proposed scheme considers the residual energy and the network density in the clustering process. Moreover, it reasonably balances the energy consumption over network clusters which in turn optimize the network lifetime. Simulation results show that the proposed scheme is effective in terms of energy and network durability. It reduces the energy consumption by an average of 16% and extends the network lifetime by an average of 27% compared to similar approaches in the literature. In future works, we aim to integrate a deep learning technique for the CHs election to further extend the network durability in the long-term. We aim to investigate more factors, such as the state of health of devices batteries to further improve the performance of UCKA.

REFERENCES

[1] Ammar, A., Dziri, A., Terre, M., Youssef, H.: Multi-hop leach based cross-layer design for large scale wireless sensor networks. In: International wireless communications and mobile computing conference (IWCMC). pp. 763–768. IEEE (2016)

[2] Arjunan, S., Pothula, S.: A survey on unequal clustering protocols in wireless sensor networks. Journal of King Saud University Computer and Information Sciences 31(3), 304–317 (2019)

[3] Baranidharan, B., Santhi, B.: Ducf: Distributed load balancing unequal clustering in wireless sensor networks using fuzzy approach. Applied Soft Computing 40, 495–506 (2016)

[4] Batta, M.S., Aliouat, Z., Harous, S.: A distributed weight-based tdma scheduling algorithm for latency improvement in iot. In: IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). pp. 0768–0774 (2019)

[5] Batta, M.S., Aliouat, Z., Mabed, H., Harous, S.: Lteoc: Long term energy optimization clustering for dynamic IoT networks. In: IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (2020)

[6] Batta, M.S., Mabed, H., Aliouat, Z.: Dynamic clustering based energy optimization for iot network. In: International Symposium on Modelling and Implementation of Complex Systems. pp. 77–91. Springer (2020)

[7] Batta, M.S., Mabed, H., Aliouat, Z., Harous, S.: A distributed multi-hop intra-clustering approach based on neighbors two-hop connectivity for iot networks. Sensors 21(3), 873 (2021)

[8] Biswas, T., Kumar, S., Singh, T., Gupta, K., Saxena, D.: A comparative analysis of unequal clustering-based routing protocol in wsns. In: Soft Computing and Signal Processing, pp. 53–62. Springer (2019)

[9] Dehkordi, S., Farajzadeh, K., Rezazadeh, J., Farahbakhsh, R., Sandrasegaran, K., Dehkordi, M.: A survey on data aggregation techniques in IoT sensor networks. Wireless Networks 26(2), 1243–1263 (2020)

[10] Deshpande, V., Patil, A.: Energy efficient clustering in wireless sensor network using cluster of cluster heads. In: tenth international conference on wireless and optical communications networks (WOCN). pp. 1–5. IEEE (2013)

[11] El Assari, Y.: Energy-efficient multi-hop routing with unequal clustering approach for wireless sensor networks. International Journal of Computer Networks & Communications (IJCNC) Vol. 12 (2020)

[12] Guiloufi, A., Nasri, N., Kachouri, A.: An energy-efficient unequal clustering algorithm using ‘sierpinski triangle’for wsns. Wireless Personal Communications 88(3), 449–465 (2016)

[13] Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless micro-sensor networks. IEEE Transactions on wireless communications 1(4), 660–670 (2002)

[14] Kruskal, J.: On the shortest spanning sub tree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society 7(1), 48–50 (1956)

[15] Mabed, H., Batta, M.S., Aliouat, Z.: Optimization of rechargeable battery lifespan in wireless networking protocols. In: 17th EAI International Conference on Mobile and Ubiquitous Systems (MobiQuitous) (December 2020)

[16] Merah, M., Aliouat, Z., Kara-mohamed, C.: An energy efficient self-organizing map based clustering protocol for iot networks. In: IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) (2022)

[17] Omadadhain, J., Fisher, D., Smyth, P., White, S., Boey, Y.: Analysis and visualization of network data using jung. Journal of Statistical Software 10(2), 1–35 (2005)

[18] Phoemphon, S., So-In, C., Aimtongkham, P., Nguyen, T.: An energy-efficient fuzzy-based scheme for unequal multi-hop clustering in wireless sensor networks. Journal of Ambient Intelligence and Humanized Computing (2020)

[19] Ranganathan, R., Somanathan, B., Kannan, K.: Fuzzy-based cluster head amendment (fcha) approach to prolong the lifetime of sensor networks. Wireless Personal Communications 110(3), 1533–1549 (2020)

[20] Sabor, N., Abo-Zahhad, M., Sasaki, S., Ahmed, S.: An unequal multi-hop balanced immune clustering protocol for wireless sensor networks. Applied Soft Computing 43, 372–389 (2016)

[21] Shahraki, A., Taherkordi, A., Haugen, o., Eliassen, F.: Clustering objectives in wireless sensor networks: A survey and research direction analysis. Computer Networks 180, 107376 (2020)

[22] Singh, S., Kumar, P., Singh, J., Alryalat, M.: An energy efficient routing using multi-hop intra clustering technique in wsns. In: IEEE Region 10 Conference (TENCON). pp. 381–386. IEEE (2017)

[23] Xia, H., Zhang, R., Yu, J., Pan, Z.: Energy-efficient routing algorithm based on unequal clustering and connected graph in wireless sensor networks. International Journal of Wireless Information Networks 23(2), 141–150 (2016)

Publishers.

Mohamed Sofiane BATTA received the Master 2 degree in networks and distributed systems from the University of Ferhat Abbas Setif-1, Algeria, in 2018. He is currently pursuing the Ph.D. degree with the LRSD laboratory, Setif, Algeria and the Femto-ST/DISC laboratory, Montbeliard, France. He is working on the field of networks and distributed systems. His research interests include wireless sensor networks, wireless communication management in the IoT networks, dynamic Ad-Hoc networks.

Zibouda ALIOUAT obtained her engineer diploma in 1984 and M.Sc. in 1993 from Constantine University. She received her Ph.D. from Setif University of Algeria. She was an assistant at Constantine University from 1985 to 1994. She is currently a Professor in Computer Engineering Departmentat Setif University of Algeria. Her research interests are in the areas of wireless mobile networks modelling and simulation, wireless sensor networks and fault tolerance of embedded systems.

Hakim MABED is an associate professor at the University of Franche-Comte (UFC), France. He is part of the FEMTO-ST institute (UMR CNRS 6174) and the complex networks team where he does his research. He obtained the Ph.D. degree from the University of Angers, France in 2003; he received the M.S. degree from the USTHB, Algeria in 2000. His research interests are in distributed intelligent MEMS, optimization, distributed algorithms, self-reconfiguration, and mobility.

Malha MERAH received the bachelor's degree (2018), and the master's degree (2020) in Computer Science from Ferhat Abbas University of Setif-1, Algeria. She is currently a Ph.D. student at Ferhat Abbas University of Setif-1, Algeria. Her research interests are concerned with the Internet of things (IoT), wireless communications, routing and energy efficiency using machine learning techniques.

SEPARATION DISTANCE REDUCTION BETWEEN 5G NR BASE STATION AND SATELLITE EARTH STATION AT C-BAND

Mohamed Ahmed M. Khalifa[1], Hebat-Allah M. Mourad[1]
and Mahmoud Abdelaziz[2]

[1]Electronics and Electrical Communications Dept, Faculty of Engineering,
Cairo University, Giza, Egypt
[2]Communications and Information Engineering,
University of Science and Technology in Zewail City, Giza, Egypt

ABSTRACT

Increasing global data traffic made 5G (IMT-2020) a good solution, especially in the mid-band spectrum (the 3.5 GHz range), because it balances coverage and capacity. Thus, some countries have identified it as one of the candidate bands for 5G. C-band (from 3400 to 4200 MHz) is a mainstay of satellite communications and provides broadband connectivity in remote areas today. This paper discusses the feasibility of 5G (IMT-2020) and Fixed Satellite Service (FSS) system to coexist in the C-band range by analyzing the impact of the interference from 5G (IMT-2020) base stations towards the FSS earth station. This analysis is based on the most recent unwanted emissions limits used from 3GPP TS 38.104. The results show that in the adjacent channel scenario and by employing an elevation angle of 48º and a guard band from 41-100 MHz, 5G (IMT-2020) base station needs to be separated by at least 0.295 Km away from the FSS earth station. The protection distance increases by 26.35 Km when decreasing the guard band to 30 MHz. Thus, a new unwanted emission limit is proposed to reduce the protection distance in the 30 MHz guard band scenario.

KEYWORDS

FSS, interference, 5G, C-band, IMT-2020 and satellite communication.

1. INTRODUCTION

There has been a significant increase in mobile data traffic per smartphone worldwide from 2014 to 2027. In 2027, global monthly smartphone data traffic is projected to become 40.64 gigabytes per active smartphone worldwide [1]. As long as mobile data traffic increases, a new technology with suitable bands needs to take place to fulfill the required needs. 105 new 5G commercial networks have used the 3.5 GHz, making it the most popular mid-band 5G spectrum worldwide [2]. 5G (IMT-2020) technology will provide a variety of services and new experiences for the user, industries, and content provider, and also higher data rate, flexibility, and reliability compared to previous technologies like IMT-Advanced. This 5G technology allows a peak data rate of 20 Gbps for the downlink and 10 Gbps for the uplink while providing user data rates up to 100 Mbps for downlink and 50 Mbps for uplink [3]. The availability of at least 80-100 MHz of contiguous spectrum per 5G network operator is required to achieve the previous data rates.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 131-138, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121111

Among the frequency ranges allocated for the satellite systems, the C-band frequency range is the most critical frequency employed in providing the FSS services. Figure 1. shows the satellite services in C-band including broadcasting, mobile backhaul, ATM-networks, E-government, Oil and gas, maritime, etc. The extensive uses are due to its robustness to high rainfall and its technical capability to reach distant areas.

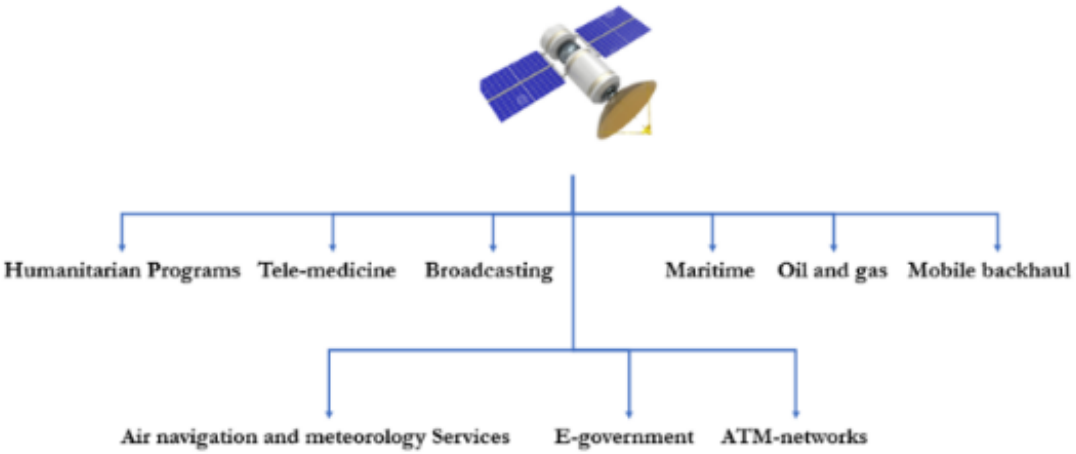


Figure 1. satellite services in C-band

Considering the urgency of the 3.5 GHz frequency band, which is within the C-band range, in providing communication services through satellite, then the coexistence between FSS and 5G (IMT-2020) needs to be considered carefully to avoid harmful effects from 5G (IMT-2020) to FSS and vice versa. Three possible types of interferences result from 5G New Radio and may affect the downlink earth station. The first type is due to in-band 5G emission as the incoming FSS signal's power flux density at the earth station location is very low while 5G base station which is closer to the earth station can produce a higher power level at the input to the FSS receiver than the desired satellite signal. The second type is unwanted emissions (out of band emission (OOBE) and spurious emission) generated by 5G operating in an adjacent band which may create interference to FSS. The third type is due to LNA/LNB overdrive. There are several studies related to the coexistence of FSS and IMT systems within the C-band range when the two systems are adjacent to each other. According to [4] coexistence between FSS and 5G system in TDD mode is feasible in the adjacent bands if separation distance is applied between the base station and FSS earth station. Meanwhile, sharing studies were conducted during the ITU WRC-07 cycle and concluded that across all studies, a common feature is that co-channel FSS and IMT operation requires a separation distance greater than the separation distance for the adjacent frequency operation [5]. Moreover, at WRC-15 studies concluded that adjacent channel in case of a macrocell requires apre-determined separation distance [6].

Hence, this paper aims to investigate the feasibility of coexistence between 5G (IMT-2020) and FSS in the adjacent band scenario through the interference analysis between those two systems at the C-band leading to determine a suitable separation distance. Furthermore, a new unwanted emission limit is proposed to reduce the separation distance resulting from the current unwanted emission limits, while improving the spectrum efficiency by reducing the guard band.

2. SYSTEM MODEL

This study on the coexistence and interference analysis between FSS and 5G (IMT-2020) is focused on the 3600-3700 MHz frequencies that lie between the 5G NR which operate on 3400-3600 MHz and FSS DL at 3700-4200 MHz where each transponder operates with 36 MHz bandwidth. The 5G operating bandwidth is assumed to be 100 MHz, with a 41-100 MHz guard band as stated in 3GPP TS 38.104 [7]. Figure 2 shows that the interference from the adjacent band is due to the very low power level of the incoming FSS signal. Thus, unwanted emissions generated by the IMT system operating in an adjacent band can create interference to FSS.

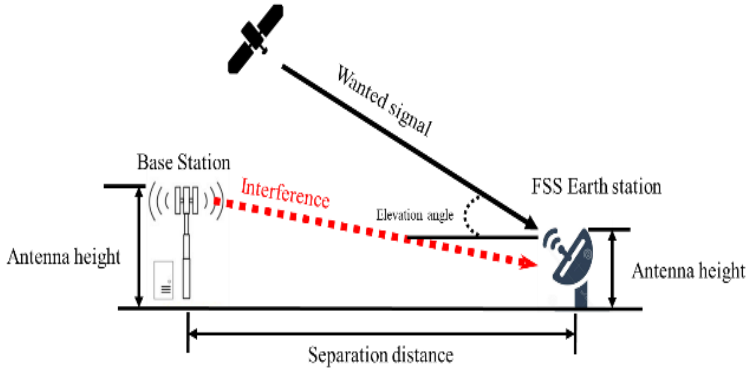


Figure 2. interference from adjacent band 5G emission

The analytical model used in this study to calculate the minimum path loss between 5G BS and FSS earth station is based on the FSS protection criteria specified in ITU-R S.1432 [8] and ITU-R M.2109 [5], and on 3GPP 5G BS out of band emission (OOBE) and spurious emission mask specified in 3GPP TS 38.104. The technical parameters of the FSS and 5G (IMT-2020) used in this paper are shown in Tables 1 and 2.

The propagation model used is based on the ITU model in ITU-R P.452-11 [9]. Path loss between the FSS earth station and 5G (IMT-2020) station is stated as follows

$$path\ loss(dB)=92.5+20*log(d)+20*log(f) \qquad (1)$$

where d is the distance between the 5G base station and FSS earth station (in kilometers), and f is the carrier frequency in MHz.

Table 1. FSS Technical Parameters.

parameters	Value/Reference
Antenna diameter	1.8 - 3 m
Antenna radiation pattern	ITU-R Recommendation S.465 [10]
C-band transponder BW	36 MHz
Planned DL frequencies	3700-4200 MHz
Elevation angle	0-48 degree
Temperature	100°K [6]

Table 2. 5g (IMT-2020) Technical Parameters.

Parameters	Value/Reference
Bandwidth (MHz)	100
Max output power (dBm/100 MHz)	50
Antenna gain (dBi)	24
Spectrum Emission Mask (SEM)	3GPP TS 38.104 [7]
Frequency (MHz)	3400-3600

The FSS earth station protection criteria is based on interference to noise ratio (I/N) = -12.2 dB [6]. The maximum permissible interference power at the input of the receiver [11]

$$I_{max} = 10\ log(kTrB) + (I/N) - w \quad (2)$$

k is Boltzman's constant 1.38*10-23(J/K), Tr is the noise temperature of the receiving system (earth station under clear sky conditions) in (K), B is the reference bandwidth of FSS transponder in (Hz), I/N is the ratio in (dB) of the permissible long-term interfering power from anyone interfering source to the thermal noise power in the FSS system and *w* is a thermal noise equivalent factor in (dB) for interfering emissions in the reference bandwidth (0 dB for digital systems), The interference power from IMT stations to FSS earth stations can be calculated using the following formula [6]:

$$I = EIRP_{5G} - path\ loss\ (dB) + G\ (\alpha) \quad (3)$$

Where $EIRP_{5G}$ is EIRP of the 5G base station, whereas G(α) is FSS earth station off-axis antenna gain toward the local horizon [10] given by:

$$G = 32 - 25log\ (\alpha)\ for\ \alpha \leq \alpha < 48^{o} \quad (4)$$
$$= -10\ dBi \qquad for\ 48^{o} \leq \alpha < 48^{o}$$

Where α is the angle between the interference signal direction and axis main beam from the FSS earth station in the frequency range 2-30 GHz and αmin = 1º. The required separation distance between the IMT station and FSS earth station can be found when $I = I_{max}$ by combining equations (1), (3), and (4) and comparing the path loss in the equations (1) and (3).

3. RESULTS AND ANALYSIS

Calculations were used to obtain the protection distance for the adjacent channel interference case. we assume that the EIRP is 75 dBm. Interference analysis is also carried out for different guard bands and elevation angles. Moreover, in this paper, we perform only the single interferer scenario. In addition, a new unwanted emissions mask is proposed to be used to add more efficiency for the spectrum regulators while planning margins or guard bands in new 5G frequency bands allocations to telecommunication operators.

3.1. Protection distance in case of adjacent channel interference in case of a single interferer

Table 3 shows the protection distance of the 5G (IMT-2020) base station and FSS earth station while considering different elevation angles and guard bands in the adjacent channel interference scenario in the case of a single interferer. The higher the angle between the interferer (5G BS) and the main axis beam from the FSS earth station (α), the smaller the protection distance needed between the interferer (5G BS) and the FSS earth station. Protection distance is constant for 48º

≤α≤180º. The minimum protection distance (295 m) occurs when the elevation angle is higher than or equal 48, while there is a guard band between 41-100 MHz, and it increases to 166 Km in case of zero guard band.

The EIRP limit is one of the essential parameters in IMT- 2020 technology deployment. In the U.S, the Federal Communications Commission has defined very high effective isotropic radiated power (EIRP) limits for the 28 and 39 GHz bands to reach 75 dBm/100 MHz [12].

Table 3. Protection Distances.

parameters	**Protection distance (Km)**			
Guard band (MHz)	0	5-10	10-40	41-100
ES elevation angle (5 º)	2792.1	1247.1	442.5	4.96
ES elevation angle (10º)	1177.4	525.9	186.6	2.09
ES elevation angle (20º)	496.5	221.78	78.69	0.8829
ES elevation angle (48º)	166.31	74.2	26.35	0.295

3.2. New mask proposal

Unwanted emissions consist of out-of-band emissions and spurious emissions [13]. Out of band emissions are unwanted emissions immediately outside the BS channel bandwidth resulting from the modulation process and non-linearity in the transmitter. Spurious emissions are emissions that are caused by unwanted transmitter effects such as harmonics emission, parasitic emission, intermodulation products, and frequency conversion products. Figure 3 shows the wide-area BS operating band unwanted emission limits. For the spurious emissions at a frequency offset higher than 40 MHz, it is assumed that 5G NR base stations shall conform to the limit of -52 dBm/MHz or -62 dBm / 100 kHz to facilitate coexistence with other legacy mobile systems operating in different frequency bands.

While the spectrum efficiency is a need for all the telecommunication systems, and as shown in Figure 3, the maximum limit to facilitate co-existence is -52 dBm/MHz or -62 dBm / 100 kHz from a frequency offset higher than 40 MHz (41-100 MHz) will result in a separation distance of 0.295 Km by applying the equations in part A. However, when the guard band decreases to 10-40 MHz range, the maximum limit will be -13 dBm/MHz or -23 dBm / 100 kHz and in this case, the separation distance will increase to 26.35 Km as shown in Table 3. So, to maintain the same separation distance as in the 41-100 MHz case, the maximum limit -62 dBm / 100 kHz should be shifted from a frequency offset higher than 40 MHz to lower parts than 40 MHz. Furthermore, Figure 4. shows a proposed mask with a frequency offset higher than 30 MHz in order to apply a 30 MHz guard band. Substituting the new value (-62 dBm / 100 kHz) at 30 MHz guard band instead of (-23 dBm / 100 kHz) in equations (3) and (4) will result in separation distances equal to the case of 41-100 MHz guard band.

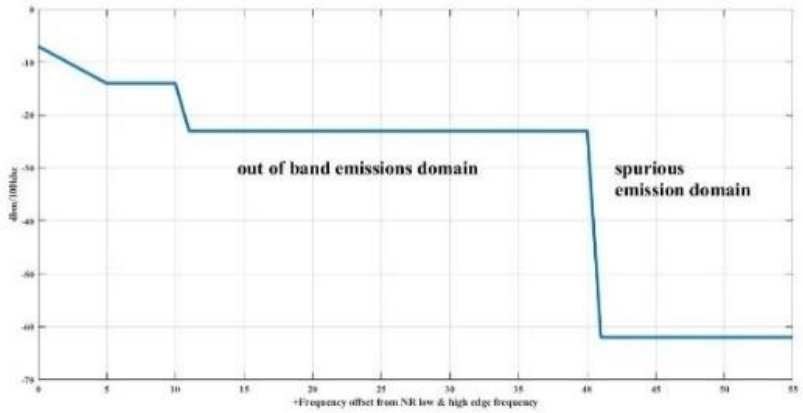


Figure 3. Unwanted emissions limits [7] (dBm/100 kHz)

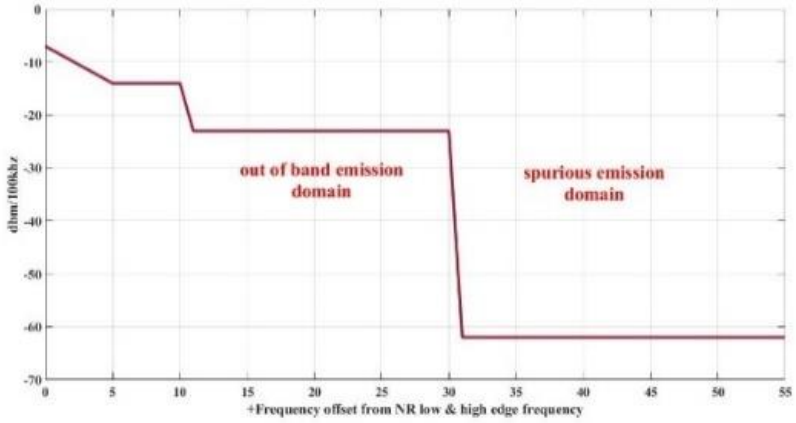


Figure 4. proposed new unwanted emissions limits for the spurious emissions at a frequency offset higher than 30 MHz (dBm/100 kHz)

4. CONCLUSIONS

This paper considered the interference analysis between the FSS and 5G (IMT-2020) system in the C-band, one of the 5G (IMT-2020) spectrum's most important candidates. The results of the adjacent-channel analysis show that the required separation distance is 0.295 km in the case of a guard band from 41 to 100 MHz at a 48º elevation angle.

When the guard band decreases the separation distance increases while increasing the elevation angle will also decrease the separation distance. A new mask is proposed which will lead to the reduction of the required separation distance in the case of the 30 MHz guard band from 26.35 Km to 0.295 Km. thus, efficient use of the scarce spectrum resources. We plan to use signal processing techniques to meet the new proposed mask in the future.

REFERENCES

[1] Mobile data traffic per smartphone worldwide from 2014 to 2027 (https://www.statista.com/statistics/738977/worldwide-monthly-data-traffic-per-smartphone/).

[2] 67 markets worldwide have commercial 5G services (https://www.spglobal.com/marketintelligence/en/news-insights/research/67-markets-worldwide-have-commercial-5g-services).

[3] ITU -R Workshop on IMT -2020 terrestrial radio interfaces, Minimum Technical Performance Requirements for IMT-2020 radio interface(s).

[4] E. Lagunas, C. G. Tsinos, S. K. Sharma, and S. Chatzinotas, ''5G cellular and fixed satellite service spectrum coexistence in C-Band,'' IEEE Access, vol. 8, pp. 72078–72094, 2020.

[5] Report ITU-R M.2109 Sharing studies between IMT-Advanced systems and geostationary satellite networks in the FSS in the 3 400-4 200 and 4 500-4 800 MHz frequency bands.

[6] ITU-R S.2368"Sharing studies between International Mobile Telecommunication-Advanced systems and geostationary satellite networks in the fixed-satellite service in the 3 400-4 200 MHz and 4 500-4 800 MHz frequency bands in the WRC study cycle leading to WRC-15 S Serie," vol. 0, 2015.

[7] 3GPP, 3GPP TS 38.104: Technical Specification Group Radio Access Network; NR; Base Station (BS) Radio Transmission and Reception (Release 17). 2021.

[8] ITU-R S.1432, Apportionment of the allowable error performance degradations to fixed-satellite service (FSS) hypothetical reference digital paths arising from time-invariant interference for systems operating below 30 GHz.

[9] Recommendation ITU-R P.452-11, Prediction procedure for the evaluation of microwave interference between stations on the surface of the Earth at frequencies above about 0.7 GHz.

[10] ITU-R S.465, Reference radiation pattern of earth station antennas in the fixed-satellite service for use in coordination and interference assessment in the frequency range from 2 to 31 GHz .

[11] Recommendation ITU-R SF.1006, Determination of the interference potential between earth stations of the fixed-satellite service and stations in the fixed service.

[12] Connecting the World with 5G: Qorvo® Highlights the Essentials. A collection of the most compelling 5G articles, blogs, videos, e-books, and infographics from Qorvo.com.

[13] ITU-R SM.329, Unwanted emissions in the spurious domain.

AUTHORS

Mohamed Khalifa received a Bachelor's degree (very good) in Electronics and Communications Engineering from Cairo University Faculty of Engineering, Egypt, in 2016. He is currently worked at the National Telecom Regulatory Authority since 2016. He participated in the launching of the 4G in Egypt and the TIBASAT Egyptian satellite. He also represents the Egyptian administration at the world radiocommunication conference in 2019 (WRC-2019). His research interest is in mobile and satellite communication.

Hebat-Allah M. Mourad received her B.Sc., M. Sc. and Ph.D. degrees in electrical communication engineering from Cairo University, Egypt, in 1983, 1987, and 1994 respectively. Since 1983, she has been with the Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, and is currently a professor there. She was the director of the high technology centre, faculty of engineering, Cairo University from 2008 till 2012. She was a member of the technical office of the post graduate and research sector of Cairo university from 2010 till 2012. She is named in Marquis Who's Who in Science and engineering in the year 2004. Her research interests include wireless and mobile communications.

Dr. Mahmoud Abdelaziz received the Ph.D degree (with honors) in Electronics and Communications Engineering from Tampere University of Technology, Finland, in 2017. He then worked one year as a Postdoctoral Researcher at the same University. He received the B.Sc. (with honors) and M.Sc. degrees in Electronics and Communications Engineering from Cairo University, Egypt, in 2006 and 2011, respectively. He is currently an Assistant Professor at the University of Science and Technology in Zewail City. From 2007 to 2012 he has worked as a communications and signal processing engineer as well as an embedded system engineer at Newport Media Inc., Etisalat Egypt, and Axxcelera Broadband Wireless. His research interests include statistical and adaptive signal processing in flexible radio transceivers, in particular, behavioral modelling and digital pre-distortion of power amplifiers in single and multiple antenna transmitters.

Verifying Outsourced Computation in an Edge Computing Marketplace

Christopher Harth-Kitzerow and Gonzalo Munilla Garrido

Department of Informatics, Technical University of Munich,
Garching, Germany

Abstract. An edge computing marketplace could enable IoT devices (Outsourcers) to outsource computation to any participating node (Contractors) in their proximity. In return, these nodes receive a reward for providing computation resources. In this work, we propose a scheme that verifies the integrity of arbitrary deterministic functions in the presence of both dishonest Outsourcers and Contractors who try to maximize their expected payoff. We compile a comprehensive set of threats for this adversary model and show that not all of these threats are addressed when combining verification techniques of related work. Our verification scheme fills the gap by detecting or preventing each identified threat. We tested our verification scheme with state-of-the-art pre-trained Convolutional Neural Network models designed for object detection. On all devices, our verification scheme causes less than 1ms computational overhead and a negligible network bandwidth overhead of at most 84 bytes per frame. Our implementation can also perform our verification scheme's tasks parallel to the object detection to eliminate any latency overhead.

Keywords: Edge Computing, Internet of Things, Function Verification, Computing Marketplaces

1 Introduction

Offloading computational tasks from IoT devices to computation resources at the network edge can improve the responsiveness of existing applications and enable novel latency-sensitive use cases [1]. In an edge computing marketplace, we assume that the Outsourcer is a computationally weak IoT device that outsources real-time data to a Contractor to process. The Contractor can be an edge server or any device in proximity to the Outsourcer with enough unutilized computational resources to execute the assigned function reliably and with low latency.

Compared to fixed client-server assignments, an edge computing marketplace may overcome the challenges of limited availability of servers, insufficient quality of service, and idle server resources. Dynamic assignments in an open marketplace could increase availability and competition among edge servers. This allows IoT applications to profit from increased connectivity and responsiveness due to better matching. Edge servers, on the other hand, profit from higher resource utilization due to increased matching rates.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 139-157, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121112

As IoT devices are often computationally weak, it might be difficult for them to verify whether responses returned by a third-party Contractor are valid. In fact, Contractors have an incentive to return a computationally less expensive probabilistic result to save resources while still collecting the reward. Likewise, an Outsourcer has no incentive to pay an honest Contractor right after receiving all computational results, and expensive micro-transactions prohibit real-time payment.

In this work, we propose a verification scheme for computation marketplaces that can verify the integrity of arbitrary deterministic functions. The Outsourcer verifies the integrity of returned responses by sending some inputs to another Contractor in proximity called the Verifier. We refer to this approach as sampling-based re-execution. We evaluate our scheme's performance with outsourced object detection based on a real-time image stream sent by an IoT device to an edge server. We provide the following contributions:

1. We compile a comprehensive list of potential threats that a computing marketplace might be vulnerable to in the presence of dishonest participants.
2. We combine existing verification techniques proposed by related work and introduce two novel ones to address all identified threats.
3. Our resulting verification scheme requires little interaction with a trusted third party (TTP) and is resistant to dishonest Outsourcers, Contractors, and Verifiers. The TTP does not have to be located at the edge and can act with arbitrary latency.
4. Our implementation demonstrates that our verification scheme causes negligible communication and latency overhead.

2 Related Work

In previous work, authors have identified different components that an edge computing marketplace should provide. These components include a matching and price-finding algorithm [2], a payment scheme [3] [4], in some cases privacy preservation [5] [6] [7], and a verification scheme [8] [9]. We focus on designing a verification scheme in this work and assume that the other components are present.

Compared to schemes based on cryptographic techniques such as Secure Multiparty Computation, or Fully Homomorphic Encryption, re-execution adds only a negligible computational and network overhead to the computation. Re-execution can be implemented in several ways. In [10], the authors propose outsourcing computation to multiple Contractors in a multi-round approach. The Outsourcer sends different inputs to each Contractor in every round. A trusted master node compares results if the same input is sent to more than one Contractor. The disadvantage of this scheme is that it requires many available Contractors in proximity to the Outsourcer.

In [11], the authors propose a scheme based on sampling-based re-execution. Within specified intervals, the Contractor commits to signed Merkle root hashes based on the responses it sent to the Outsourcer. The Outsourcer can instruct a third-party Verifier to verify these results by checking if the signed Merkle root's signature is correct and if random samples were computed and returned correctly. The scheme assumes that the Verifier and the Outsourcer are honest.

In [12], the authors propose a scheme based on complete re-execution. The Outsourcer sends inputs to $n = 2$ Contractors and accepts the result if they are identical. If responses do not match, the job gets outsourced again to $n_{new} = n^2$ Contractors until no conflicts arise. This approach is similar yet less efficient than our Contestation protocol introduced in section 3. As their scheme does not distinguish between Contractors and Verifiers, the resulting overhead is higher. Also, their scheme relies on a trusted time-stamping server that monitors communication between edge devices. This TTP can end up being the bottleneck of the ecosystem.

In [13], the authors propose a scheme based on sampling-based re-execution with third-party Verifiers. They utilize smart contracts running on Ethereum to set incentives for Outsourcers and Contractors. The incentives discourage dishonest behavior. Verifiers are assumed to be partially trusted.

In [14], the authors propose a scheme based on sampling-based re-execution. The Contractor commits to a Merkle Tree root hash every few intervals and sends it to the Outsourcer. The Outsourcer then randomly selects a few samples to re-compute them and sends a proof of membership challenge to the Contractor. It aborts the contract if the verified output does not match the Contractor's response or if the proof of membership challenge is unsuccessful. The Outsourcer is assumed to be fully trusted.

3 Design of our Verification Scheme

As our scheme uses re-execution as a verification approach, it can be applied to any deterministic function. We assume the following setting for outsourced computation in an edge computing marketplace.

1. Edge servers are stationary (reappearing actors) and offer outsourced computation for a fee. They can either act as Contractors or Verifiers.
2. Outsourcers are mobile (reappearing and adhoc actors).
3. Outsourcers owe a reward to Contractors and Verifiers for each processed input.
4. Each edge participant may act dishonestly but tries to maximize its expected payoff.
5. A TTP or Blockchain is present that provides a public key infrastructure, a reputation system, and handles payments. We refer to this party as the payment settlement entity (PSE). The PSE does not have to be located at the edge.

Before outsourcing of computation starts, we assume an Outsourcer and a Contractor have agreed to a contract. The contract contains a unique ID, the participants' public keys, a reward per processed input, and the function/model to be used is specified. Additionally, the Contractor and Outsourcer may agree on fines, deposits, and bounties if a participant is caught cheating to increase the protocol's robustness. We assume multiple Verifiers are available and willing to agree on a contract with the Outsourcer to process random sample inputs.

3.1 Preparation Phase

The preparation phase is responsible for assigning a Verifier to a contract while preventing collusion. We refer to planned collusion if two participants know each other beforehand and try to collude. We refer to ad-hoc collusion if two participants do not initially know each other but still try to communicate and collude.

Randomization Randomization ensures that the Outsourcer and the Contractor commit to a random Verifier. Additionally, the Verifier and the Contractor do not learn each other's identities. The protocol consists of the following steps: The Outsourcer signs the hash $h(x)$ of a large random number x and the contract hash ch. It sends $h(x)$ with a signature to the Contractor. The Contractor signs the received hash of the Outsourcer along with a large random number y, the contract hash, and a list of available Verifiers sorted by their public keys. Along with this signed hash, the Contractor sends the value y and the list of available Verifiers to the Outsourcer. By signing the initially sent hash of the Outsourcer, the Contractor commits to x and y without knowing x. Figure 1 illustrates this protocol.

If the list of available Verifiers matches Outsourcer's local list, it contacts the Verifier at $(x+y) \mod n$, where n is the total number of Verifiers. If the Outsourcer contacts a different Verifer, it will not be able to present the necessary Contractor signatures during Contestation. Thus, Randomization prevents planned collusion.

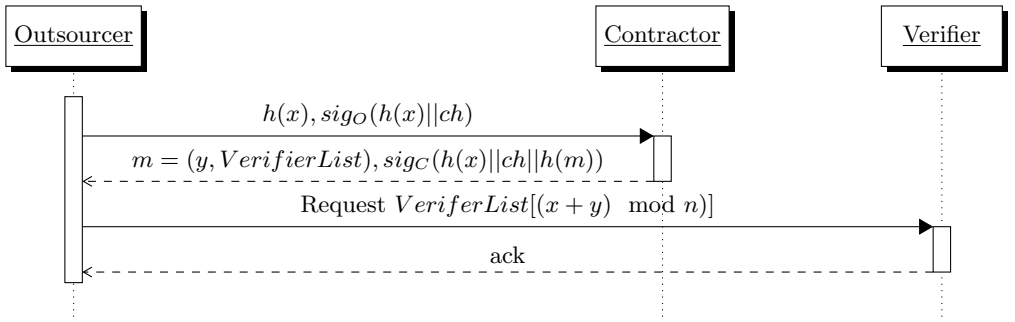


Fig. 1: Randomization

Table 1: Payoff matrix with honesty-promoting incentives

Contractor \ Verifier	Diligent	Dishonest
Diligent	$r - c_h$	$r - c_h + b$
Dishonest	$rq - (f + b)(1 - q) - c_d$	$r - c_d$

Game-theoretic Incentives Even if the Verifier and the Contractor do not know each other, there is a risk of ad-hoc collusion. Suppose there exists a q-algorithm that is computationally inexpensive but provides a correct result with a certain probability q. For object detection, this might be the naive response that no object is detected and, therefore, no bounding boxes at specific coordinates have to be estimated. There is a reward r for returning a valid result and computation costs when computing the desired function honestly c_h or dishonestly c_d. From a game-theoretic perspective [15], there are two nash equilibria [16]. One nash equilibrium exists when both players act honestly, but the other when both players act dishonestly [17] [12].

It is crucial to design incentives that eliminate the Nash equilibrium of both players acting dishonestly. In [12] and [17], the authors have identified a relationship of incentives by adding fees for cheating players and bounties for dishonest players such that being honest is a dominant strategy from a game-theoretic point of view. Table 1 illustrates the payoff matrix of the Contractor with the use of a bounty b and a fee f. The payoff matrix for the Verifier looks identical. Our scheme uses an initial deposit to enforce that a cheating participant pays the fine after detection.

3.2 Execution Phase

After agreeing on a contract with the random Verifier, the Outsourcer starts sending inputs to the Contractor and the Verifier to process.

Sampling Sampling refers to picking one random input out of a collection of inputs. In our verification scheme, the Outsourcer sends samples to the Verifier to check whether its response matches the Contractor's response belonging to the same input. We call this process sampling-based re-execution. Sampling-based re-execution has a significant advantage over complete re-execution. Just with a few samples, a dishonest Contractor with a cheating rate c can be detected with nearly 100% confidence. Thus, we can significantly improve the efficiency of the verification process at a negligible security drawdown.

During sampling, the Outsourcer chooses an interval length l and sends only one random sample per interval to the Verifier. The chance p of detecting a cheating attempt within n intervals is $p = 1 - (1 - c)^n$. Even if the Contractor has a low

cheating rate, e.g., $c = 0.1$, the Outsourcer needs less than $n = 50$ intervals to detect cheating with 99% confidence.

Digital Signatures Since participants communicate in an unmonitored, peer-to-peer fashion in our verification scheme, we need a way to securely record payment promises and dishonest behavior. Otherwise, an Outsourcer could claim never to have received any responses from the other participants. Likewise, the Contractor and Verifier could deny that a dishonest result originated from them and could claim to have processed more responses than they did. The PSE can only solve a dispute and hold entities accountable with tamper-proof records.

Figure 2 shows a high-level overview of sampling in combination with digital signatures. "$i = r$" indicates that the current index i matches the random number r generated in the current interval.

When an Outsourcer sends an input, it always attaches a digital signature signed over the current input index, the contract hash, and the input itself. This ensures that each signature can be traced back to one unique input. Also, the Outsourcer includes a number of currently acknowledged outputs n to the message and signature. Thus, the Contractor and the Verifier receive a signed commitment of redeeming n times the specified reward per response. The unique contract hash ensures that each participant can only redeem payment once per contract.

When the Contractor and the Verifier send a result to the Outsourcer, they attach the associated input index, along with a digital signature forged over the contract hash, input index, input signature, and the input itself. This signature serves as proof of being the originator of a fraudulent message when detected cheating. If the signature verification fails, the participant aborts the contract.

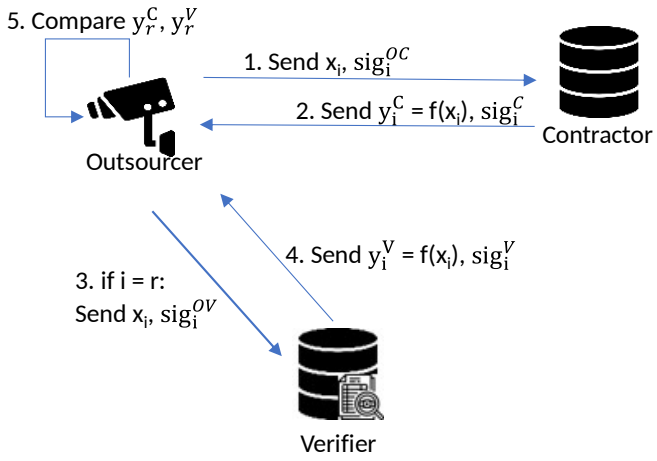


Fig. 2: Execution phase

Commitment to Messages Using Merkle Trees Merkle Trees can be used as a data structure to efficiently verify whether data is contained in a large collection. The root hash of a Merkle tree can be used to verify if data is included in the whole tree with $log_2(n)$ steps. Some verification schemes use this attribute of Merkle trees in combination with signatures to commit to large amounts of inputs or outputs by sending the signed Merkle tree root hash [18].

In our scheme, a Contractor utilizing Merkle Trees commits to a collection of responses and receives a proof-of-membership challenge in even intervals. This way, instead of signing all responses, it is sufficient for the Contractor only to sign one Merkle root hash and challenge-response per interval, thus improving efficiency.

3.3 Closing Phase

A participant can terminate a contract at any time. If the contract is terminated according to custom, the Verifier and the Contractor store their last signed input of the Outsourcer, containing the latest number of acknowledged outputs and the signed contract hash. They send the signature and all values to verify it to the PSE. To prevent insufficient quality of service (QoS), a global reputation system and local blocklists per node ensure that participants providing reliable service can be identified. Thus, all participants can submit a review for the other participants at the end of a contract.

The PSE verifies the signatures and deducts the reward per input specified in the contract times the number of acknowledged output contained in the last input on behalf of the Outsourcer after a deadline. Within that deadline, the Outsourcer can report dishonest behavior if it holds two responses that do not match. In this case, the Outsourcer sends the input, both responses, their signatures, and the contracts to the PSE. For scalability reasons, the PSE does not re-execute the computation. It only checks whether all values match their signatures and verifies if responses are indeed unequal. Provisionally, the Contractor is accused of cheating. Within the specified deadline, the Contractor can decide to engage in a protocol we call Contestation to prove that the Verifier's response was incorrect instead.

Contestation Contestation ensures that a falsely accused Contractor or Verifier can prove its innocence. A participant accused of cheating may decide to re-outsource the original input to two additional random Verifiers within a deadline. If both random Verifiers return a response that matches the participant's response, it presents their responses and signatures to the PSE. The participant having the minority of random Verifier support at the end of Contestation is convicted of cheating.

If a Verifier is accused of cheating, it can use the identical protocol to contact two additional random Verifiers and flip the majority of random Verifier support. This protocol might be repeated until no available Verifiers are left. If more than

50% of available Verifiers are non-colluding, Contestation serves as a guarantee that the participant who returned a fraudulent response is found guilty. In combination with a nearly 100% detection rate of cheating using sampling-based re-execution, any cheating parting will be eventually found guilty with a high probability. The participant found guilty at the end of the protocol has to pay the specified fee in its contract, and all additionally consulted Verifiers. In the first round of Contestation, the Outsourcer must prove to the PSE that it contacted the correct Verifier during Randomization by presenting the received Contractor signatures.

Figure 3 illustrates a message sequence chart of Contestation. Notice that the TTP is involved in minimal computation to ensure scalability. It only needs to verify anything if the convicted participant contests conviction. Also, a convicted participant failing to get a majority of Verifier support has no incentive to send the last message and occupy the TTP.

Note that for a dishonest participant, it is irrational to perform Contestation as additional random Verifiers have to be paid for their service. The computational overhead of Contestation is low as only one input has to be recomputed. However, it requires finding multiple available Verifiers in the system. As latency is not critical in this scenario, those random Verifiers do not have to be located at the edge and can be computationally weak devices. We expect that Contestation is usually not performed as its existence alone ensures that a dishonest participant decreases its expected payoff when cheating.

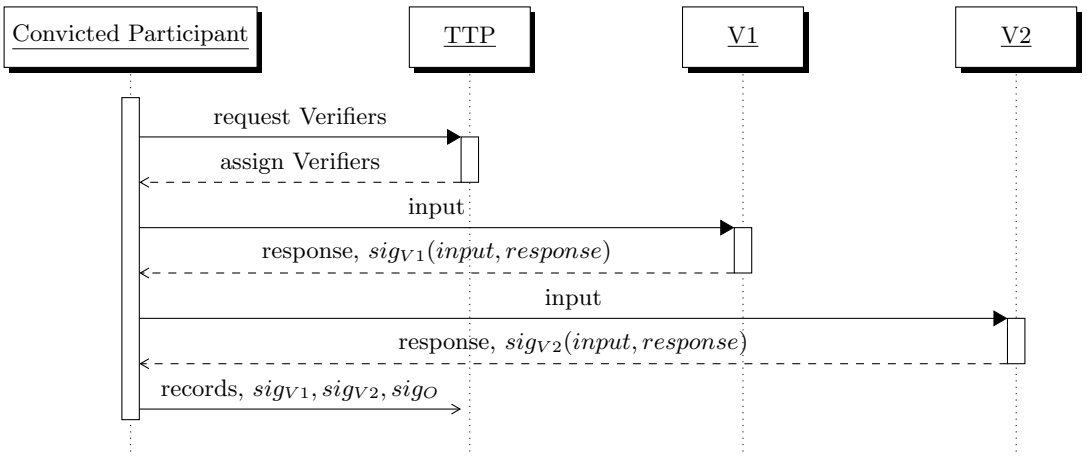


Fig. 3: Contestation

4 Threat Model

In our verification scheme, the Outsourcer, the Contractor, and the Verifier are untrusted and may behave dishonestly. However, we assume that they try to maximize their expected payoff. This type of threat model was first introduced by the authors of [17]. We noticed that existing verification schemes such as [12] assuming payoff-maximizing adversaries address only a subset of possible threats that can result from these assumptions.

Due to the lack of an existing collection of threats in this setting, we compile a comprehensive list of possible threats. We consider internal threats of dishonest behavior by one participant and by collusion. Also, we consider threats of external attackers and quality of service (QoS) violations such as timeouts and low response rates. We assume a distant TTP conducts payments and delegates handling of disputes.

We compiled nine threats in total. These are described in the following paragraphs. Our verification scheme resists all identified threats with high probability. Table 2 summarizes the techniques used by our verification scheme to prevent or detect each possible protocol violation we identified. The confidence column indicates the probability that the protocol violation can be prevented or detected by our techniques. Note that Contestation provides a 100% detection rate of associated protocol violations only if more than 50% of available Verifiers in the ecosystem are non-colluding by not agreeing on an identical incorrect response.

4.1 Contractor sends back false responses to save resources

If a Contractor sends back incorrect responses, the Outsourcer detects this with high probability by sending random samples to the Verifier and comparing if responses belonging to the same input from both participants are equal. In section 3, we show that even with a low number of samples, a cheating Contractor is caught with nearly 100% confidence.

4.2 Verifier sends back false responses to save resources

When the Outsourcer detects two unequal responses from the Verifier and the Contractor, our verification scheme provisionally accuses the Contractor of cheating. However, the Contractor can perform Contestation to prove that the Verifier sent the incorrect response instead.

4.3 Outsourcer sends back different inputs to Contractor and Verifier to refuse payment

The Outsourcer may send two different inputs to the Contractor and the Verifier two receive different responses. It can report these responses to the PSE to refuse

payment. This behavior can only be detected if proof is available that the responses resulted from two different inputs. Thus, the Outsourcer must sign each sent raw input with contract-related information. By concatenating the Outsourcer's signature to the Contractor's and Verifier's responses before signing, Contestation can verify if the Outsourcer sent identical raw inputs to both parties.

4.4 Contractor or Verifier tries to avoid global penalties when convicted of cheating or QoS violations

Even when a Verifier or Contractor is detected cheating by an Outsourcer, they may claim never to have sent the reported response. The use of digital signatures prevents this threat.

4.5 Participant refuses to pay even if obliged to by the protocol

Even if the Outsourcer is obliged to reward an honest Contractor or Verifier, there needs to be a way to enforce the payment. Likewise, the Verifier or the Contractor might try to reject paying a penalty fee when detected cheating.

Microtransactions sent for each response are not an option. Usually, the payment scheme can become a latency bottleneck, and each transaction comes with transaction costs. Therefore, payment has to be handled after a contract ends. We assume that the PSE supports deposits and payments on other participants' behalf. It does not need to recompute any values or execute contract-specific functions.

4.6 Outsourcer and Verifier collude to refuse payment and save resources

The Outsourcer and the Verifier may collude to report the Contractor for cheating. Contestation detects this dishonest behavior. If the Verifier is detected cheating by the Contractor through Contestation, it has to pay a fine. If the incentives are set correctly, acting honestly maximizes the expected payoff. Randomization also prevents planned collusion with high probability.

.

4.7 Contractor and Verifier collude to save resources

The Contractor and the Verifier may collude to save computational resources by agreeing on an incorrect response. The Outsourcer checks if both results match and assumes the responses to be correct. Randomization prevents this behavior with a high probability in case of planned collusion.

Additionally, a contract with honesty-promoting incentives maximizes the expected payoff when acting honestly. This measurement makes ad-hoc collusion between Contractor and Verifier unlikely as well. Beyond our verification scheme,

an Outsourcer may decide to utilize more than one Verifier or perform additional verification techniques at a lower frequency.

4.8 Timeouts, Low Response Rate, High Response time

In our verification scheme, dishonest Contractors and Verifiers do not receive payments and must pay a fine. In contrast, QoS violations such as timeouts, a low response rate, or a high response time come without monetary consequences. Nevertheless, participating in an ecosystem with insufficient processing or networking capabilities should be discouraged.

Whenever a participant receives a message from another participant that exceeds the QoS thresholds specified in its internal parameters, it may abort the contract. It may also blacklist a participant and submit a negative review. Thus, an unreliable participant misses out on the current contract's ongoing payments and may receive fewer assignments or less payoff from future contracts.

4.9 Message Tampering

An external attacker may attempt to tamper with messages sent between participants to harm a participant. Digital signatures prevent this behavior.

Table 2: Utilized Techniques to Prevent Protocol Violations

Type of Violation	Description	Techniques	Confidence
Dishonest Behavior by Individual	1. Contractor sends back false response to save resources	Sampling-based re-execution, utilization of a third party Verifier	Up to 100%
	2. Verifier sends back false response to save resources	Contestation	100%
	3. Outsourcer sends different input to Contractor and Verifier to refuse payment	Digital Signatures (signature chain), Contestation	100%
	4. Contractor or Verifier tries to avoid global penalties	Digital Signatures	100%
	5. Participant refuses to pay even if obliged to by the protocol	PSE authorized to conduct payment on behalf of another entity	100%
Dishonest Behavior via Collusion	6. Outsourcer and Verifier collude to refuse payment and save resources	Randomization, Game-theoretic incentives, Contestation	100%
	7. Contractor and Verifier collude to save resources	Randomization, Game-theoretic incentives	High confidence
QoS Violation	8. Timeout, Low Response Rate, High Response Time	Blacklisting, Review system, Contract abortion	100%
External Threat	9. Message Tampering	Digital Signatures	100%

4.10 Other Threats

Outsourcers and Contractors can join the ecosystem with multiple identities without hurting the system's security. However, the PSE should issue identity checks of new Verifiers in the ecosystem to prevent Sybil attacks [18]. Otherwise, a participant can increase its probability of matching with colluding participants.

5 Discussion

Out of the different re-execution approaches introduced in section 2, sampling based re-execution is the most practical. We further borrow the following other techniques from related work to improve the outsourcing process.

1. Digitial dignatures to hold verifiable proofs of received messages [9] [19] [11] [18].
2. Merkle Trees to reduce the required number of digital signatures that need to be sent [9] [14] [11] [18].
3. TTPs or Blockchains to resolve payments or record communication [9] [13] [19] [18] [12].
4. A global reputation system to promote honest behavior [10].

Other proposed verification schemes often require a TTP at the network edge or assume that one of the participants is a trusted party (TP). A comparison of the trust assumptions of different schemes is shown in Table 3. In our verification scheme, all participants at the edge may act dishonestly. During outsourcing, they collect publicly verifiable proofs from other participants to later present to a distant trusted PSE. This TTP conducts payments and delegates handling of disputes.

Table 3: Trusted parties required in different schemes

Scheme	TPs involved	TPs involved during execution phase	Assumptions
Ours	1	0	Payment settlement entity is fully trusted
[14]	1	1	Outsourcer is fully trusted
[11]	2	2	Outsourcer and Verifier are fully trusted
[20]	1	1	Outsourcer is fully trusted
[18]	> 3	1	Outsourcer is semi-trusted
[10]	> 2	> 2	Multiple Contractors available for one Contract, Pool of trusted nodes available
[13]	1	1	Verifiers are semi-trusted
[12]	2	1	Trusted timestamp server is available

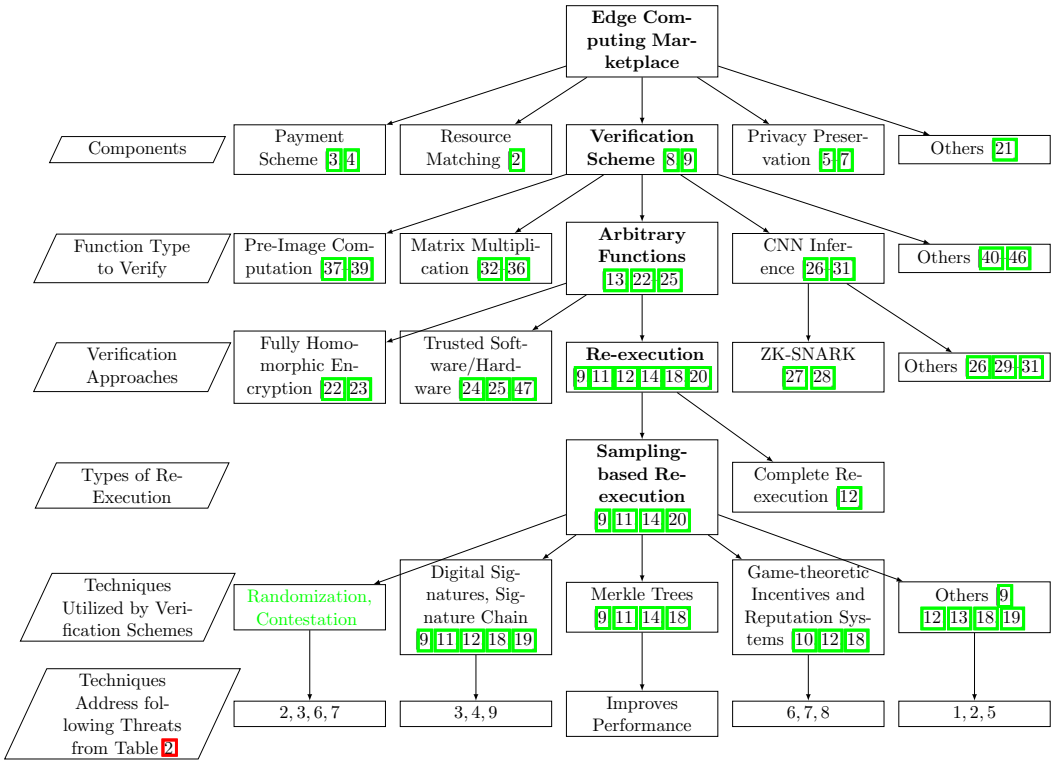


Fig. 4: Overview: Designed verification scheme

Figure 4 shows an overview of our scheme in the context of our literature analysis. The last two comparisons show different verification techniques utilized by existing verification schemes based on re-execution. Even by combining these existing techniques, we can only solve 7 out of our 9 identified possible threats. We address the remaining two by our Randomization and Contestation protocols (marked green in figure 4):

Threat 6 supposes that the Outsourcer and the Verifier collude to refuse payment to the Contractor. Randomization prevents this behavior with a high probability in case of planned collusion. Contestation prevents this behavior in both cases of adhoc and planned collusion. Threat 3 supposes that the Outsourcer sends different inputs to the Contractor and the Verifier. This behavior is detected by Contestation. We discuss both threats in more detail in section 4.

6 Performance

In our test setup, a Raspberry Pi (Outsourcer) sends a real-time webcam stream to two different machines (Verifier and Contractor) in the local network. The Con-

tractor and the Verifier send back bounding boxes of detected objects in a frame. One test setup uses a regular GPU/CPU for inference, and one uses a Coral USB Accelerator. The Coral USB Accelerator is an entry-level Tensor Processing Unit (TPU) that is specifically designed to perform neural network inference [48]. We use weights that were pre-trained on the Microsoft Coco dataset [49]. We use Yolov4 [50] and MobileNet SSD V2 [51] as models for object detection.

Our implementation uses the NaCl ED25519 signature scheme and can perform the verification scheme's task in parallel to the object detection. This eliminates the latency overhead of our scheme as the verification scheme's task utilizes one CPU thread while the GPU is the bottleneck when performing inference. The source code of our implementation is publicly available here. Our implementation supports multithreading, Merkle Trees, and non-blocking message pattern to improve the efficiency of our scheme. An overview of the test setup is provided in the appendix. Table 4 shows the key results of our test implementation.

Our results show that our verification scheme causes less than 1ms of latency overhead per frame. The Contractor's GPU is the system's bottleneck in our test setup, limiting overall performance to 68.06 fps. As we only used a mid-range GPU and an entry-level edge accelerator in our tests, more potent Contractor hardware could increase overall system performance to match the Outsourcer's performance of more than 200fps. The average 416x416 frame has a size of 120 KB in our tests. The Network bandwidth overhead per participant is negligible at a maximum of 84 bytes per frame. It consists of a 512-bit large signature and, at most, five 32 bit integers such as frame index, acknowledged responses, and other contract-related information. When Merkle Trees are utilized, the Contractor and the Verifier only send signatures when the Outsourcer requests a proof-of-membership challenge.

Table 4: Key Results

Participant	Device	CPU	GPU	Model	$\frac{Frames}{Second}$	Time spent on application (%)	Time spent on verification (%)	Time spent on verification (ms)
Outsourcer	Raspberry Pi Model 4B			MobileNet SSD V2 300×300	236.00	78.70	21.30	0.90
Outsourcer	Raspberry Pi Model 4B			Yolov4 tiny 416×416	146.90	85.10	14.90	1.01
Contractor	Desktop PC	Core i7 3770K	GTX 970	Yolov4 tiny 416×416	68.06	100.00	0.00	0.00
Contractor	Desktop PC	Core i7 3770K	Coral USB Accelerator	MobileNet SSD V2 300×300	63.59	100.00	0.00	0.00
Contractor	Notebook	Core i5 4300U	Coral USB Accelerator	MobileNet SSD V2 300×300	49.30	100.00	0.00	0.00
Verifier	Notebook	Core i5 4300U	Coral USB Accelerator	MobileNet SSD V2 300×300	28.75	-	0.00	0.00

7 Conclusion

In this work, we proposed a scheme for verifying arbitrary outsourced functions in an edge computation marketplace. Our verification scheme is resistant to a comprehensive set of protocol violations that might occur in a computation marketplace with untrusted participants. We benchmarked our verification scheme's performance on consumer hardware and TPUs. Our verification scheme achieves less than 1ms of latency overhead per frame on all tested machines. By utilizing concurrency, the overhead can be reduced to 0 by running the verification scheme's tasks in a parallel thread to the outsourced computation. The network bandwidth overhead of our scheme caused mainly by digital signatures is negligible (at most 84 bytes per frame).

In comparison with verification schemes proposed by the current academic literature, our verification scheme provides additional security by preventing or detecting all threats we identified. At the same time, it only requires third-party involvement outside the network edge. These performance and security characteristics make our scheme ideal for use within an edge computing marketplace that matches computationally weak untrusted IoT devices with untrusted third-party resources to outsource latency-sensitive tasks.

Our verification scheme implements one essential component of a fully functioning edge computing marketplace. As illustrated in figure 4, the remaining components to make an edge computing marketplace viable are: Payment, Matching and price-finding, and Privacy preservation. Future work may optimize and aggregate all components to build an end-to-end system serving as a standalone edge computing marketplace for arbitrary functions.

References

1. W. Tang, X. Zhao, W. Rafique, L. Qi, W. Dou, and Q. Ni, "An offloading method using decentralized p2p-enabled mobile edge servers in edge computing," *Journal of Systems Architecture*, vol. 94, pp. 1–13, 2019.
2. A. Zavodovski, S. Bayhan, N. Mohan, P. Zhou, W. Wong, and J. Kangasharju, "Decloud: Truthful decentralized double auction for edge clouds," 05 2019.
3. R. Rahmani, Y. Li, and T. Kanter, "A scalable distriubuted ledger for internet of things based on edge computing," in *Seventh International Conference on Advances in Computing, Communication and Information Technology-CCIT 2018, Rome, Italy, 27-28 October, 2018*. Institute of Research Engineers and Doctors (IRED), 2018, pp. 41–45.
4. L. Zhao, Q. Wang, C. Wang, Q. Li, C. Shen, X. Lin, S. Hu, and M. Du, "Veriml: Enabling integrity assurances and fair payments for machine learning as a service," *arXiv preprint arXiv:1909.06961*, 2019.
5. J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: Survey and open issues," *IEEE Access*, vol. 6, pp. 18 209–18 237, 2018.
6. Y. Wang, Z. Tian, S. Su, Y. Sun, and C. Zhu, "Preserving location privacy in mobile edge computing," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 05 2019, pp. 1–6.

7. M. Gheisari, Q.-V. Pham, M. Alazab, X. Zhang, C. Fernández-Campusano, and G. Srivastava, "Eca: An edge computing architecture for privacy-preserving in iot-based smart city," *IEEE Access*, vol. PP, pp. 1–1, 08 2019.
8. W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE Access*, vol. 6, pp. 6900–6919, 2018.
9. H. Huang, X. Chen, Q. Wu, X. Huang, and J. Shen, "Bitcoin-based fair payments for outsourcing computations of fog devices," *Future Generation Computer Systems*, vol. 78, 12 2016.
10. R. Di Pietro, F. Lombardi, F. Martinelli, and D. Sgandurra, "Anticheetah: Trustworthy computing in an outsourced (cheating) environment," *Future Generation Computer Systems*, vol. 48, pp. 28–38, 2015.
11. L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, and A. V. Vasilakos, "Security and privacy for storage and computation in cloud computing," *Information sciences*, vol. 258, pp. 371–386, 2014.
12. A. Küpçü, "Incentivized outsourced computation resistant to malicious contractors," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 6, pp. 633–649, 2015.
13. S. Eisele, T. Eghtesad, N. Troutman, A. Laszka, and A. Dubey, "Mechanisms for outsourcing computation via a decentralized market," *arXiv preprint arXiv:2005.11429*, 2020.
14. W. Du, J. Jia, M. Mangal, and M. Murugesan, "Uncheatable grid computing," in *24th International Conference on Distributed Computing Systems, 2004. Proceedings.* IEEE, 2004, pp. 4–11.
15. M. J. Osborne *et al.*, *An introduction to game theory.* Oxford university press New York, 2004, vol. 3, no. 3.
16. M. Aghassi and D. Bertsimas, "Robust game theory," *Mathematical Programming*, vol. 107, no. 1-2, pp. 231–273, 2006.
17. M. Belenkiy, M. Chase, C. C. Erway, J. Jannotti, A. Küpçü, and A. Lysyanskaya, "Incentivizing outsourced computation," in *Proceedings of the 3rd international workshop on Economics of networked systems*, 2008, pp. 85–90.
18. M. Nabi, S. Avizheh, M. V. Kumaramangalam, and R. Safavi-Naini, "Game-theoretic analysis of an incentivized verifiable computation system," in *International Conference on Financial Cryptography and Data Security.* Springer, 2019, pp. 50–66.
19. X. Chen, J. Li, and W. Susilo, "Efficient fair conditional payments for outsourcing computations," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1687–1694, 2012.
20. H. Wang, "Integrity verification of cloud-hosted data analytics computations," in *Proceedings of the 1st International Workshop on Cloud Intelligence*, 2012, pp. 1–4.
21. I. Psaras, "Decentralised edge-computing and iot through distributed trust," 06 2018, pp. 505–507.
22. Y. C. Chunming Tang, "Efficient non-interactive verifiable outsourced computation for arbitrary functions," Cryptology ePrint Archive, Report 2014/439, 2014, https://eprint.iacr.org/2014/439.
23. C. Xiang and C. Tang, "New verifiable outsourced computation scheme for an arbitrary function," *International Journal of Grid and Utility Computing*, vol. 7, p. 190, 01 2016.
24. T. Combe, A. Martin, and R. Di Pietro, "To docker or not to docker: A security perspective," *IEEE Cloud Computing*, vol. 3, no. 5, pp. 54–62, 2016.
25. S. van Schaik, A. Kwong, D. Genkin, and Y. Yarom, "Sgaxe: How sgx fails in practice," 2020.
26. H. Chabanne, J. Keuffer, and R. Molva, "Embedded proofs for verifiable neural networks," *IACR Cryptol. ePrint Arch.*, vol. 2017, p. 1038, 2017.
27. S. Lee, H. Ko, J. Kim, and H. Oh, "vcnn: Verifiable convolutional neural network," *IACR Cryptol. ePrint Arch.*, vol. 2020, p. 584, 2020.
28. J. Groth, "On the size of pairing-based non-interactive arguments," 05 2016, pp. 305–326.
29. X. Chen, J. Ji, L. Yu, C. Luo, and P. Li, "Securenets: Secure inference of deep neural networks on an untrusted cloud," in *ACML*, 2018.

30. Z. Ghodsi, T. Gu, and S. Garg, "Safetynets: Verifiable execution of deep neural networks on an untrusted cloud," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4675–4684.
31. A. A. Badawi, J. Chao, J. Lin, C. F. Mun, J. J. Sim, B. H. M. Tan, X. Nan, K. M. M. Aung, and V. R. Chandrasekhar, "Towards the alexnet moment for homomorphic encryption: Hcnn, thefirst homomorphic cnn on encrypted data with gpus," 2018.
32. R. Freivalds, "Probabilistic machines can use less running time." in *IFIP congress*, vol. 839, 1977, p. 842.
33. X. Lei, X. Liao, T. Huang, and F. H. Rabevohitra, "Achieving security, robust cheating resistance, and high-efficiency for outsourcing large matrix multiplication computation to a malicious cloud," *Inf. Sci.*, vol. 280, pp. 205–217, 2014.
34. Z. Cao and L. Liu, "A note on "achieving security, robust cheating resistance, and high-efficiency for outsourcing large matrix multiplication computation to a malicious cloud"," 03 2016.
35. D. Benjamin and M. J. Atallah, "Private and cheating-free outsourcing of algebraic computations," in *2008 Sixth Annual Conference on Privacy, Security and Trust*, 2008, pp. 240–245.
36. M. J. Atallah and K. B. Frikken, "Securely outsourcing linear algebra computations," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 48–59. [Online]. Available: https://doi.org/10.1145/1755688.1755695
37. B. Carbunar and M. Tripunitara, "Fair payments for outsourced computations," in *2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, 2010, pp. 1–9.
38. B. Carbunar and M. V. Tripunitara, "Payments for outsourced computations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 2, pp. 313–320, 2012.
39. P. Golle and I. Mironov, "Uncheatable distributed computations," vol. 2020, 04 2001, pp. 425–440.
40. G. Xu, G. T. Amariucai, and Y. Guan, "Delegation of computation with verification outsourcing: Curious verifiers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 3, pp. 717–730, 2017.
41. X. Hu and C. Tang, "Secure outsourced computation of the characteristic polynomial and eigenvalues of matrix," *Journal of Cloud Computing*, vol. 4, 12 2015.
42. C. Wang, K. Ren, J. Wang, and Q. Wang, "Harnessing the cloud for securely outsourcing large-scale systems of linear equations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1172–1181, 2013.
43. X. Lei, X. Liao, T. Huang, H. Li, and C. Hu, "Outsourcing large matrix inversion computation to a public cloud," *IEEE TRANSACTIONS ON CLOUD COMPUTING*, vol. 1, pp. 78–87, 07 2013.
44. W. Song, B. Wang, Q. Wang, C. Shi, W. Lou, and Z. Peng, "Publicly verifiable computation of polynomials over outsourced data with multiple sources," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 10, pp. 2334–2347, 2017.
45. X. Wang, K.-K. R. Choo, J. Weng, and J. Ma, "Comments on "publicly verifiable computation of polynomials over outsourced data with multiple sources"," *IEEE Transactions on Information Forensics and Security*, vol. PP, pp. 1–1, 08 2019.
46. J. Meena, S. Tiwari, and M. Vardhan, "Privacy preserving, verifiable and efficient outsourcing algorithm for regression analysis to a malicious cloud," *Journal of Intelligent & Fuzzy Systems*, vol. 32, pp. 3413–3427, 04 2017.
47. V. Costan and S. Devadas, "Intel sgx explained," *IACR Cryptol. ePrint Arch.*, vol. 2016, p. 86, 2016.
48. A. Ghosh, S. A. Al Mahmud, T. I. R. Uday, and D. M. Farid, "Assistive technology for visually impaired using tensor flow object detection in raspberry pi and coral usb accelerator," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 186–189.

49. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
50. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
51. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

8 Appendix

Codebase: `https://github.com/chart21/Verification-of-Outsourced-Object-Detection`

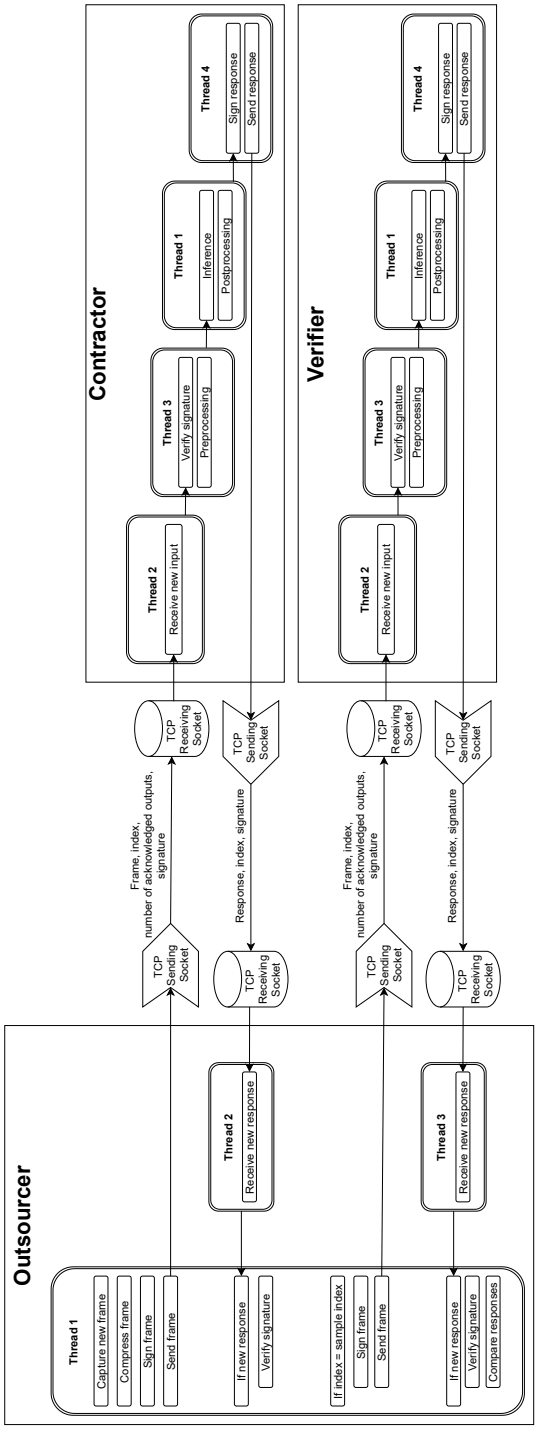


Fig. 5: Test Setup

AN INTELLIGENT MOBILE PLATFORM TO ASSIST CUSTOMIZED COSMETIC SELECTION USING ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING

Jenny Sun[1] and Yu Sun[2]

[1]JSerra Catholic High school, 26351 Junipero Serra Rd,
San Juan Capistrano, CA 92675
[2]California State Polytechnic University, Pomona,
CA, 91768, Irvine, CA 92620

ABSTRACT

Most teenagers have some types of skin blemishes and allergies that often go undiagnosed [4][5]. This can lead to many unnecessary skin ailments and treatments that do not address the root of the problems [6]. This is especially true for young women who are beginning to use makeup. This can lead to embarrassments and even isolation and depression. But with this Intelligent Mobile Platform app, teenagers, parents, and dermatologists can all be assisted in identifying specific skin conditions such as oily or dry skin, skin prone to breakouts and rashes, acne, or even more severe conditions such as eczema and psoriasis [7]. Once this app has made a proper diagnosis, it will recommend and prescribe proper skin treatments intended to avoid or solve potentially major physical and emotional problems. This paper is designed to discuss the ways that this app evolved into a productive and reputable device capable of improving the lives of millions of young teenagers as well as those of older ages who continue to struggle with skin problems.

KEYWORDS

Natural Language Processing, Artificial Intelligence, Mobile Platform.

1. INTRODUCTION

Teenage years are some of the most difficult years of a person's life as they adjust from childhood, to adolescence, to adulthood [8][9]. It is during these challenging times that puberty begins to develop and most are faced with some type of skin difficulties. During these sensitive years, teens, especially young females, are also introduced to many skin care products. But this can be quite dangerous as most are not compatible with the particular problems and therefore can irritate the skin and even exacerbate the problems. This can cause damage that takes months to clear up. In the meantime, young teens, who are already overly self-conscious, are afraid to be seen in public. Imagine the unnecessary embarrassment, all because the skin irritations were not properly diagnosed and treated. Researchers, however, began working on an intelligent mobile platform app to assist in diagnosing and then selecting the perfect customized cosmetic solution. Now, with the invention of this app designed to mitigate such problems, teens will at least have a better chance to anticipate potential difficulties and apply the safest products for skin care. Now,

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 159-167, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121113

teens will have more enjoyable days and evenings as they no longer worry about their skin problems.

Currently, teenagers engage in “trial and error” methods for determining if products are safe for the skin. They also rely on the “word of mouth” system, whereby friends or family members suggest different skin-care products based on commercials they hear or advice they might receive from neighbors and associates. These methods are dangerous, however, because they do not take into consideration the fact that all skin types are slightly different and each skin has unique sensitivities that react differently to similar products. Dermatologists are also fairly new to the subject of skin ailments and treatments as the medical profession continues to evolve. Their medical training is also limited and they often overlook the nuanced differences that are now detected by this simple app.

There are some products being developed and found online or in medical device centers that diagnose skin disorders, but most are either costly or unreliable. The Skin Analysis System with its corresponding laptop, for instance, retails for $1,490. The Skin Analysis Scope created by The Garfield Company retails for $970. Other, less-expensive diagnostic machines such as the Bio-Therapeutic bt-analyze Skin Identification Device or the Fantexy Portable Skin Analyzer retail for less than $150, but they are notoriously unreliable [10]. Perhaps the best device on the market today is the Dermatoscope Diagnostic Skin Analyzer Handheld Skin Tester that sells for just under $2,000 [11]. Dermatologists often use this machine, however, the new app we are describing today has far more advantages. It is portable, more reliable in terms of diagnostics, results are supplied far more quickly, and all of these benefits are much less expensive than the Dermatoscope Diagnostic Skin Analyzer.

Because this is a mobile app device, the benefits become apparent instantly when compared to the archaic alternatives. This device does not require indoor settings with heavy machine equipment plugged into electrical outlets. Instead, to properly use this app, the user need only take several screenshots of the facial skin from a variety of angles [12]. This allows the hand-held app to properly diagnose all details of the skin ranging from dryness or moisture, to potential skin allergies or diseases, to possible issues involving overexposure to the sun or from wet or bitterly cold weather. Then, the app will make appropriate diagnoses and recommendations regarding appropriate and inappropriate skin care products. In instances where skin care products are new to the market or are virtually untested, this app will determine the chemical compositions of the components and analyze them according to the possible applications for the skin in question. In this way, the app is revolutionary. It is able to anticipate potential problems and cures based on its sophisticated bank of information that is growing daily.

This paper is structured to first provide an overview of the issue and the necessity for this more modern, easier to use, more effective device. Next, we discuss the difficulties that occurred throughout the developing process. Finally, we explain the mechanics involved with this app to give confidence to those interested in supporting its use either professionally or personally. As the pertinent details are explained, every member of the audience can see the work and the benefits for all involved.

2. CHALLENGES

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

2.1. Laptop Issue

Because this was the first time coding and working in an android studio, I made numerous mistakes. This made the entire experience nearly overwhelming. Additionally, the original laptop did not adapt properly, so I purchased a new laptop, which, after much experimentation, was finally resolved by going to a particular Starbucks in Irvine to do most of the coding on their premises.

2.2. App recognize issue

Another challenge we faced was that our app did not recognize some of the skin conditions such as burns to the skin and uncommon types of rashes. Therefore, we needed to conduct far more research on unusual and rare types of skin conditions and store that data into our information bank. Additionally, we secured previously unpublished data from numerous dermatological facilities as well as from the CDC in Atlanta. Georgia [13]. As additional data was collected, our app became more sophisticated and adept in identifying, diagnosing, and determining proper care for each patient. This is and will be a continuing process.

2.3. Prepare App in a presentable way

A third challenge was to prepare our app in a presentable way that attracted investors. We did not need a perfectly functioning system in place, but we needed a prototype available, complete with accompanying presentations by our engineering team and the artificial intelligence designers who were qualified to showcase our product and explain exactly how it functioned. We also needed to have many of the skin diagnoses functioning as test examples so that investors could see exactly how it worked and how our future plans were realistically calculated.

3. SOLUTION

Information is accumulated and stored in two ways. Initially, we measure the subject's facial features by rubbing the test module across the cheeks and forehead. This provides the app the skin cells to react with the components buried inside the app to deliver an initial diagnosis regarding skin types and possible future issues recognized by the app. Then an in-depth interview to determine physical characteristics including diet, activity schedule, physical locations throughout the day, atmospheric conditions, genetic details, and other data are accumulated and compiled and input in the mobile app. All data is then downloaded and stored in the central storage unit and interpreted through artificial intelligence where determinations are made and sent back to the mobile app where each patient is given a unique diagnosis and an individualized treatment plan. Once the personalized data is initially input into the system, weekly or bi-weekly tests and diagnoses are given by the app in the same way to monitor progress and prognosis. It is further recommended that a qualified dermatologist give intermittent reviews of the patient's progress as a way to ensure that the treatment plan is the very best possible. Continual updates to the information bank in addition to the accumulation of patient results will make this app increasingly accurate with every new day, because of its artificial component.

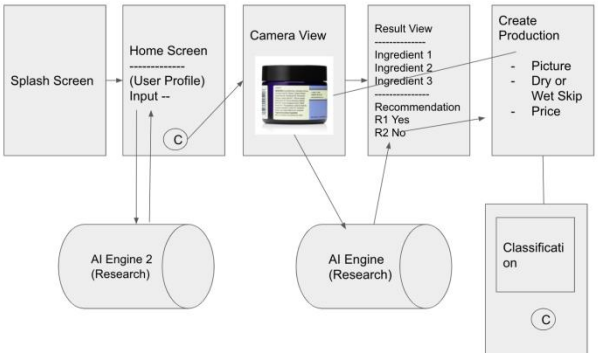


Figure 1. Overview of the solution

1. Login/Create new Account

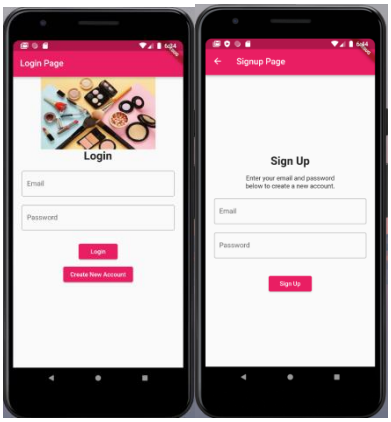


Figure 2. Screenshot of Login page

2. Take a picture and upload to Firebase

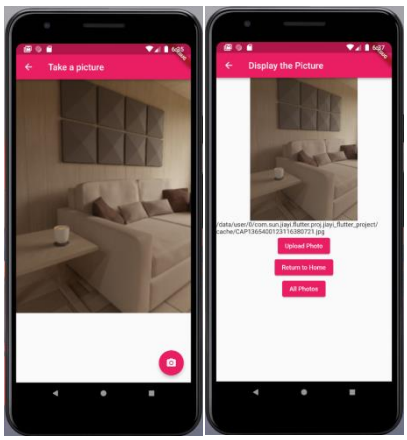


Figure 3. Screenshot of upload page

3. Scan a photo, read the text, and recommend or not recommend

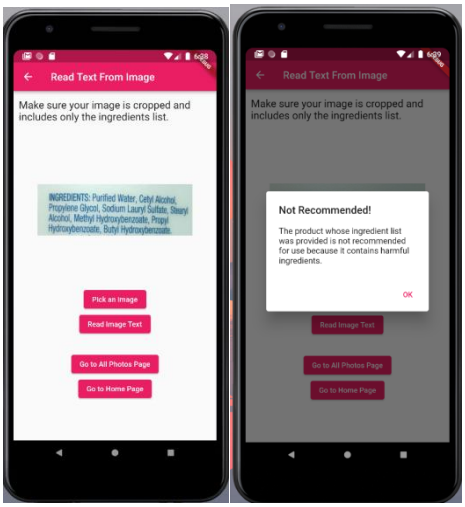


Figure 4. Screenshot of reading text

4. Skin type detector

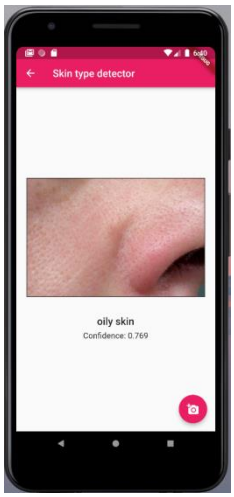


Figure 5. Screenshot of skin

4. EXPERIMENT

4.1. Experiment 1

In order to verify that our solution can effectively solve problems at different levels and have good user feedback, we decided to select multiple experimental groups and comparison groups for several experiments. For the first experiment, we want to prove that our solution works stable and continuously, so we choose a group size of 40 different trials in 2 different kinds of skin. The 2 different types of skin are dry and wet. The goal of the first experiment is to verify if the AI algorithm works good for different types of skin. Through sampling 2 groups of skin problems. Result is collected by statistics if the app tests the skin type correctly. Experiments have shown that all skin in different types tested the right result. Dry Skin has the most correct rates, which means our user are works more better in dry skin. This experiment could explain that the skin

types do have a obvious impact on the arrange results. The average correct rate of 2 different types of the skins shows below:

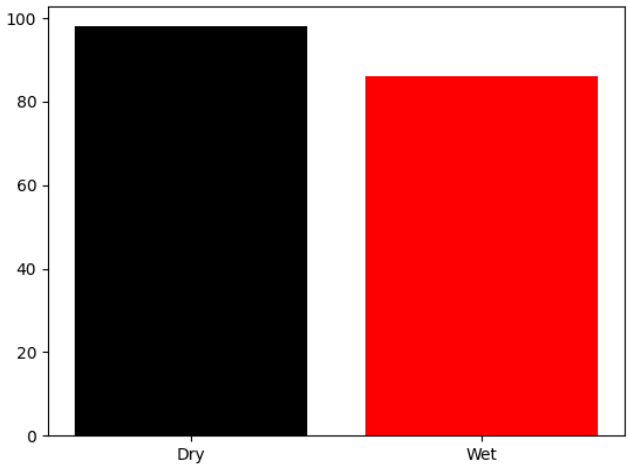


Figure 6. Data of dry and wet

A good user experience is as important as a good product. So a perfect solution should have excellent user experience feedback.In order to prove that our solution has the best user feedback, we specially designed a user experience questionnaire base on the US system usability questionnaire rules。We statistics the feedback result from 100 users, Track the user's data for 5 days, let them explore freely on the functionality. We divide those users into Five different groups. The first group of users ages from 10 - 20, the second group of users ages from 20 - 30, the third group of users ages from 30 - 40, the fourth group of users ages from 40 - 50, the fifth group of users ages from 50 - 60The goal of the first experiment is to verify high feedback scores shows high performance We collect the feedback scores form these 5 different group of users and analyze it.Experiments have shown that users who ages from 30 - 40 give the highest result feedback to our app. Which may because of the age between those range are more likely to use the makeup and need know their skin type more The experiment graph shows below:

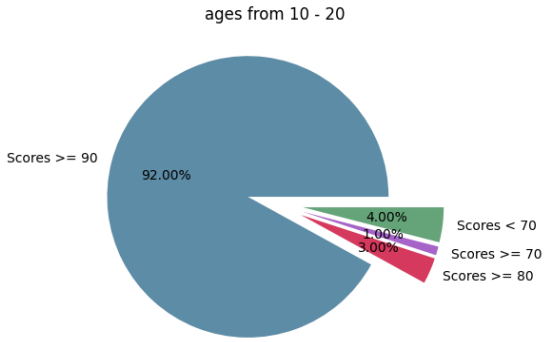


Figure 7. Result of Age 10-20

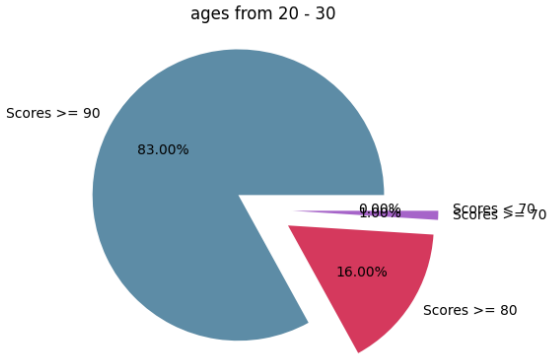


Figure 8. Result of Age 20-30

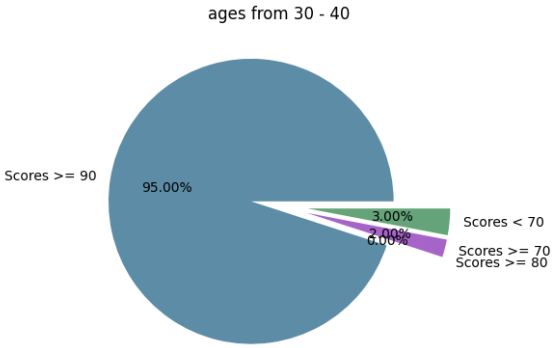


Figure 9. Result of Age 30-40

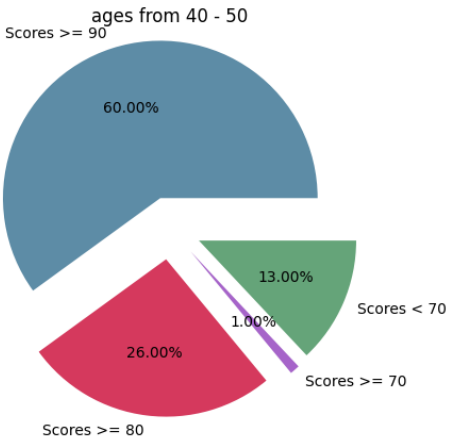


Figure 10. Result of Age 40-50

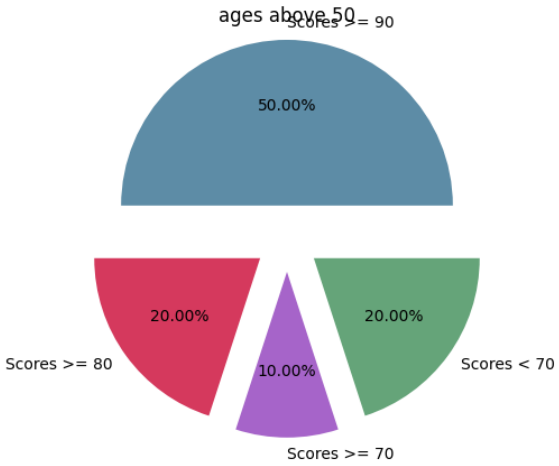


Figure 11. Result of Age above 50

5. RELATED WORK

This article by doctors from the University of Hull in the UK explores the viability and the practicality of collecting medical data and using it to diagnose and treat patients [1]. This is important because it shows that my research does have merit. This, however lays the background for my work which validates what they are saying. My research goes beyond what these doctors discovered about artificial intelligence in the medical field.

This short article adds to an earlier study by these doctors [2]. They state that there are some types of skin disorders that require less severe treatments than originally thought. This is useful to this project, because it shows that diagnosing skin disorders is part of the usual process of a dermatologist. Our research will be able to bypass some of the work of dermatologists because the app can do this much easier.

This article by a Turkish doctor from Akdeniz University in Antalya, Turkey endorses the use of smartphone apps for diagnosing skin disorders [3]. Age and distance limitations can interfere with diagnostic and treatment plans, so this doctor sees apps as a possible solution. This is useful for our research and app, because it gives our app credibility. Our app, however, is more sophisticated than what this doctor is recommending, but this is a good start. Our app not only gives diagnoses but also offers treatment plans.

6. CONCLUSIONS

The Intelligent Mobile Platform is the most versatile and accurate app available for teenagers, parents, and even dermatologists to identify specific skin conditions that range from oily or dry skin, to blemishes and acne, to severe conditions such as eczema and psoriasis. It is a hand-held device that requires just a quick rub across the cheeks and forehead to accumulate cells and forward the information to a central database to determine chemical compositions of the skin followed by a tentative diagnosis and treatment plan [14]. An in-depth interview with the subject is then conducted to learn the daily activities as well as diet and other necessary information on the subject. This is necessary for the Platform to gather and present a comprehensive diagnosis, treatment plan, and prognosis. The Intelligent Mobile Platform is inexpensive and mobile which

allows its use in almost any setting. This is especially important for young teenagers who are shy about confronting skin problems such as acne but desire a quick, reliable diagnosis and treatment plan before the acne worsens [15]. Parents will appreciate this device, not just for its privacy and accuracy, but because they want to know definitively if there are any other skin disorders their teens are suffering from. The Intelligent Mobile Platform can deliver on all of these demands, quickly, accurately, and inexpensively.

This app will only be as good as the latest information on skin disorders is available. As the dermatological field advances with improved diagnoses and treatments of various skin disorders, this app will often be in need of updates. Additionally, as the AI feature in the database is updated, the app will need to be recalibrated periodically.

Future work should focus on the size and versatility of this app. Ideally, it should be smaller and easier to use, with a smoother surface. The results should also be downloaded easier and results returned more quickly.

REFERENCES

[1] Ramesh, A. N., Kambhampati, C., Monson, J. R., & Drew, P. J. (2004). Artificial intelligence in medicine. Annals of The Royal College of Surgeons of England, 86(5), 334–338. https://doi.org/10.1308/147870804290

[2] Ebbert, Jon, et al. “In Reply–Prevalence of Skin Disorders in Patients Seeking Health Care.” Mayo Clinic Proceedings, vol. 88, no. 7, 10 June 2013, pp. 776–777., https://doi.org/https://doi.org/10.1016/j.mayocp.2013.05.007.

[3] Göçeri, Evgin. “2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA).” Impact of Deep Learning and Smartphone Technologies in Dermatology: Automated Diagnosis, 2020. IEEE, 10.1109/IPTA50016.2020.9286706. Accessed 2022.

[4] Jaeger, Bastian, et al. "Effects of facial skin smoothness and blemishes on trait impressions." Perception 47.6 (2018): 608-625.

[5] Angold, Adrian, et al. "Impaired but undiagnosed." Journal of the American Academy of Child & Adolescent Psychiatry 38.2 (1999): 129-137.

[6] Saikia, Abinash Pratim, et al. "Ethnobotany of medicinal plants used by Assamese people for various skin ailments and cosmetics." Journal of Ethnopharmacology 106.2 (2006): 149-157.

[7] Wilmer, Erin N., et al. "Most common dermatologic conditions encountered by dermatologists and nondermatologists." Cutis 94.6 (2014): 285-292.

[8] Sawyer, Susan M., et al. "Adolescence: a foundation for future health." The lancet 379.9826 (2012): 1630-1640.

[9] Hogan, Dennis P., and Nan Marie Astone. "The transition to adulthood." Annual review of sociology 12 (1986): 109-130.

[10] Shah, Lipa, Sunita Yadav, and Mansoor Amiji. "Nanotechnology for CNS delivery of bio-therapeutic agents." Drug delivery and translational research 3.4 (2013): 336-351.

[11] MacKie, R. M., et al. "The use of the dermatoscope to identify early melanoma using the three-colour test." British Journal of Dermatology 146.3 (2002): 481-484.

[12] Arda, Oktay, Nadir Göksügür, and Yalçın Tüzün. "Basic histological structure and functions of facial skin." Clinics in dermatology 32.1 (2014): 3-13.

[13] Hidalgo, Linda García. "Dermatological complications of obesity." American journal of clinical dermatology 3.7 (2002): 497-506.

[14] Liu, Ming-Tzen, et al. "Identification of chemical compositions of skin calcified deposit by vibrational microspectroscopies." Archives of dermatological research 297.5 (2005): 231-234.

[15] Williams, Hywel C., Robert P. Dellavalle, and Sarah Garner. "Acne vulgaris." The Lancet 379.9813 (2012): 361-372.

AUTHOR INDEX