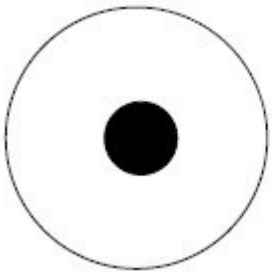
IPSI TIR
JOURNAL

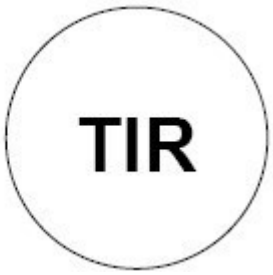
IPSI Transactions on Internet Research

New Trends in Secure Computer Science and Applications
Guest Editors: Bilal Zaka and Munam Ali Shah
and
Innovative Software and Data Services for Modern Business
Guest Editor: Ivan Luković

JULY 2024, VOLUME 20, ISSUE 2

A publication of IPSI Internet Research Society
New York, Frankfurt, Tokyo, Belgrade
ISSN 1820-4503
DOI: 10.58245/ipsi.tir





The IPSI Transactions on Internet Research

Multi-, Inter-, and Trans-disciplinary Issues in Computer Science and Engineering

A publication of IPSI Internet Research Society New York, Frankfurt, Tokyo, Belgrade
July 2024 Volume 20 Number 2 (ISSN 1820-4503)

New Trends in Secure Computer Science and Applications

Guest Editors: Bilal Zaka and Munam Ali Shah

and

Innovative Software and Data Services for Modern Business

Guest Editor: Ivan Luković

Table of Contents:

The IPSI Internet Research Society

The IPSI Internet Research Society is an association of people with professional interests in the field of the Internet. All members will receive IPSI Transaction Journals upon registering at the Society.

Member copies of Transactions are for personal use only
IPSI TRANSACTIONS ON ADVANCED RESEARCH
www.ipsitransactions.org

Editorial Board

Editorial

New Trends in Secure Computer Science and Applications

Guest Editors: Bilal Zaka and Munam Ali Shah

And

Innovative Software and Data Services for Modern Business

Guest Editor Ivan Luković

New Trends in Secure Computer Science and Applications

Guest Editors: Bilal Zaka and Munam Ali Shah

This section explores the latest trends in secure computer science and their real-world applications. Our editorial team has curated six insightful articles that examine the use of advanced machine learning algorithms, knowledge models, and the critical importance of security in today's evolving technological landscape. Each article provides in-depth analysis and valuable insights into the challenges and advancements shaping our digital world.

Paper 1. **Smart Urban Planning: An Intelligent Framework to Predict Traffic Using Stack Ensembling Approach**, by Muhammad Adeel Anjum and Ahmad Alanzi

Forecasting road traffic conditions and creating routes for vehicles is an important aspect of smart cities. Artificial intelligence-based road traffic monitoring systems allow drivers to identify the route that will take them to their destination with the least amount of difficulty and the least amount of time spent in congested regions. This paper presents a unique method for predicting road traffic and air pollution by making use of relevant data. The authors have worked in two major areas: Firstly, they compared ten different regression approaches to determine which technique provides better results and accuracy in accurately identifying road traffic patterns. Secondly, the authors have chosen regression analysis approaches in which they chose those base learners which give better results in Level 1. These predictions are combined as an input to the Level 2 meta-regressor. Their proposed technique lowers the mean square error, the relative absolute error, and root means square error and improves the R-squared value which significantly helps the decision makers to make appropriate decisions for better road traffic management.

Paper 2: **Detecting Malicious Botnet in IoT Networks Using Machine Learning Techniques**, by Muhammad Nabeel Asghar, Zara Murad, and Muhammad Asif Raza

The authors of this paper have investigated the botnet attacks in the context of the Internet of Things (IoT) with the Mirai botnet being a mostly analyzed which is a source of distributed denial of service (DDoS) attacks. According to the authors, the Mirai attack has gained notoriety for its involvement in large-scale attacks that compromised numerous IoT devices through weak authentication credentials. In this paper, they have applied classical machine learning techniques like Random Forest, Support Vector Machine and Logistic Regression to classify the malicious traffic from Mirai and Bashlite botnets. The publicly available N-BaIoT dataset is used to train the selected algorithms to identify the most informative features for detecting botnet attacks on IoT devices. The dataset contains traffic data from nine infected devices against five protocols. The employed machine learning algorithms achieved test validation accuracy above 99%, with Random Forest performing the best. Their analysis has revealed that the devices generating combo floods share common characteristics like weight or variance calculated within a certain time window. These findings can be used by security service providers to prevent botnet attacks in their networks.

Paper 3: **Enhancing Security of Text Using Affine Cipher, and Image Cryptography**, by Mehmoona Jabeen and Carsten Maple

In this paper, the authors have aimed to enhance the security of the well-known cryptographic algorithm i.e., Affine Cipher. The authors have presented an improved version of the said algorithm which can be used for image cryptography. The contribution of the authors in this paper is manyfold. Firstly, they have presented a critical review of symmetric and asymmetric cryptographic algorithms. They have compared several classical algorithms and compared their performance in terms of their resilience against attacks. Secondly, the authors have presented a classification of the existing cryptographic algorithms with a focus on image-based cryptography. Lastly, the authors have presented an enhanced version of the Affine Cipher. The enhancements have been made in Affine Cipher by performing several operations in a sequence such as ASCII to binary conversion, applying XOR with keys and converting the secret text as an image. Their results show that their proposed scheme provides better security and is time efficient.

Paper 4: **Futuristic Blockchain-based Secure and Verifiable Drone Surveillance System: Chain in the Sky**, by Usama Arshad and Yasir Faheem

This paper discusses the Blockchain-based secure solution for drone and unmanned aerial vehicles (UAV). We are aware that drones are increasingly being adopted for multifaceted surveillance missions. This is the reason, data trustworthiness, security vulnerabilities, and privacy encroachments are important questions to be addressed in any research related to drone technology. The authors have presented an avant-garde blueprint, amalgamating blockchain's robust features with drone-centric surveillance to confront and remedy security challenges. Their proposed system offers distinctive drone identification, bolstering their traceability while shielding the acquired data from potential security compromises and unwarranted modifications. The salient feature of this research is that the authors have provided ushering in a new epoch of trustworthy and drone-sourced data.

Paper 5: **Discover and Automate New Adversarial Attack Paths to Reduce Threat Risks for The Security of Organizations**, by Azhar Ghafoor, Munam Ali Shah, Sardar Nawaz and Bilal Zaka

This paper discusses the phishing attack which is a type of fraud and identity theft that involves the use of both social engineering and technological misrepresentation to obtain user credentials, such as login information. The most usual way for a phishing attack to be launched is for a phishing message to arrive in the user's mailbox appearing to be from a bank, security agency or NGO directing the user to a web page and encouraging him to enter his login credentials, but the web page is not actually linked with the bank. In this paper, we focused on discovering different attack paths that savvy attackers leverage to launch various kinds of attacks by masquerading the identity of some well-known celebrity or website, or by spoofing the email addresses. Importantly, we have highlighted some of the previously not reported use cases that are used for malicious purposes, such as attackers using share features provided on various websites to spoof emails, and they need no further efforts to bypass the antivirus or email gateways because they are sending emails from the real mail server of the website's parent organization. We performed various experiments to prove that these attack vectors are important to be taken into consideration and our findings proved to be right.

Paper 6: **Deep Acoustic Modelling for Quranic Recitation - Current Solutions and Future Directions**, by Muhammad Aleem Shakeel, Hassan Ali Khattak and Numan Khurshid

The Holy Quran is the most sacred book for Muslims all around the world. Like other holy scriptures, the Quran contains knowledge and wisdom. According to Muslim beliefs, only reciting the Quran is also beneficial. For this, the Quran should be read according to the Arabic language rules. In the past few years, this field gained a lot of importance in the eyes of researchers who aim to automate the Quranic reading. Moreover, the understanding process of the Quran is widely being investigated with the help of Machine Learning and Deep Learning algorithms. In this paper, the authors have provided a detailed survey on Deep Modeling for Quranic Recitation and have discussed several categories related to speech analysis, including the most advanced feature extraction techniques, mispronunciation detection using Tajweed rules, Reciters and speech dialect classification, and implementation of Automatic Speech Recognition (ASR) on Quranic Recitations. Lastly, the authors have identified the challenges faced by researchers in this domain and identified possible future gaps.

Bilal Zaka, PhD**,** is the Head of the IT Service Directorate at COMSATS University Islamabad, Pakistan. With a professional career spanning over 20 years in academia, he has extensive experience in developing and managing complex information systems, research management, and system/network administration. Bilal's current research focuses on leveraging AI (machine learning) to enhance the capabilities of conventional information systems and unlock the value of both structured and unstructured data. Additionally, he provides technical consultancy services to the Higher Education Commission of Pakistan. Bilal can be reached at zaka@comsats.edu.pk, ORCID: 0000-0001-6176-2541.

Munam Ali Shah, PhD, is serving as an Associate Professor at King Faisal University, Al-Ahsa, Saudi Arabia. He is the author of over 250 research articles published in reputed international conferences and journals. His name appeared thrice in Stanford University's list of the world's top 2% researchers. His research work includes cybersecurity and privacy preservation techniques using modern solutions. Profiles of Dr Shah can be accessed through the following links: [ACM, Scopus, Google Scholar IEEE Xplore ResearchGate ORCID: 0000-0002-4037-3405]

Innovative Software and Data Services for Modern Business

Guest Editor Ivan Luković

In this section, we intend to address open questions and the real potential for various innovative applications aimed at supporting effective software and data services in support of information management in various business systems. We intended to address the interdisciplinary character of a set of theories, methodologies, processes, architectures, and technologies in the development of innovative software and data services. In a rigorous reviewing process, in the July 2024 issue, we selected 4 papers.

Paper 7. **Addressing Class Imbalance in Customer Response Modeling Using Random and Clustering Based Undersampling and SVM**, by Ljiljana Kašćelan and Sunčica Vuković

In the paper, Ljiljana Kašćelan and Sunčica Vuković advocate that one of the main challenges in machine learning-based customer response models is the class imbalance problem, specifically a small number of respondents, compared to non-respondents. Aiming to overcome this issue, they propose an approach of preprocessing training data using a Support Vector Machine (SVM), trained on a balanced sample obtained by random undersampling (B-SVM), as well as on a balanced sample obtained by clustering-based undersampling (CB-SVM). Several classifiers were tested on such a balanced dataset to compare their predictive performances. In the paper, they demonstrate that the approach effectively pre-processes the training data, and, in turn, reduces noise and overcomes the class imbalance problem. Better predictive performance was achieved compared to standard training data balancing techniques such as undersampling and SMOTE. CB-SVM gives a better sensitivity, while B-SVM gives a better ratio of sensitivity and specificity. Organizations can utilize this approach to automatically and simply balance training data and more efficiently select customers that should be targeted in the next direct marketing campaigns.

Paper 8. **The event processing network for systematic reduction of interoperability deviations in a business ecosystem**, by Daliborka Mačinković and Vidan Marković

The authors of this paper propose the model of the event processing network for the systematic reduction of interoperability deviations in a business ecosystem. Complex event processing technology supports real-time events monitoring in collaborative business processes for the behaviors specified as interoperability deviations, generating alerts when such situations occur. As a reaction, alerts on interoperability deviations are delivered as personalized information to the right consumer as a designated collaboration partner. The event processing network enables collaboration partners to be proactive in interoperability deviations and to eliminate the impact of interoperability deviations on the business process objectives in the business ecosystem. In their research, the authors identified characteristic patterns of events. The logic of event processing was specified for the systematic reduction of interoperability deviations in the business ecosystem.

Paper 9. **Enhancing Semantics Learning: A Dynamic Environment for Abstract Language Implementation Education**, by William Steingartner and Igor Sivý

In their paper, William Steingartner and Igor Sivý propose an abstract machine for structural operational semantics as a stack machine with two different model representations of memory. They propose a complex tool enabling compilation from a higher imperative (toy) language into an abstract machine allowing, in addition, the visualization of individual computational steps, interactive memory manipulation, and feedback by compiling back to a higher language. In this paper, the authors present an abstract machine designed primarily for educational purposes, enabling the visualization and interaction with the compilation process of a simple imperative language.

Paper 10. **Innovative Solutions for Tetraplegia: A Smart Hand Orthosis Design**, by Norbert Ferenčík, Veronika Sedláková, Petra Kolembusová, Branko Štefanovič, Radovan Hudák, and William Steingartner

In this paper, the authors advocate that Spinal cord injury (SCI) poses a significant medical challenge, affecting both hand dexterity and locomotor abilities, while ongoing advancements in medical technologies, spanning a spectrum of wearable devices, coupled with concurrent progress in rehabilitation treatments, aim to enhance hand function among individuals affected by SCI. The emergence of three-dimensional (3D) printing provides a cost-effective avenue for crafting personalized devices, fostering a surge of interest in integrating this technology with rehabilitation equipment, thereby complementing advancements in scientific research. Myoelectric control plays a pivotal role in achieving enhanced rehabilitation outcomes. It involves the detection and processing of weak electromyographic signals (EMG) from affected limb muscles to activate orthotic motors. The authors propose in their research a novel 3D-printed hand orthosis, responsive to electromyography signals, to facilitate grasping functionality in cervical SCI patients.

Finally, let us express our great thanks to all the authors for their hard work, great enthusiasm, research efforts, and high-quality submissions. In the reviewing process, more than 25 reviewers were actively involved. we would like also to express our warm thanks to all the reviewers for their great efforts and valuable comments that significantly contributed to raising the overall quality of the selected papers.

Ivan Luković, PhD, Eng. in Informatics, University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia, Orcid: 0000-0003-1319-488X, H-index = 18 (Google Scholar). His research interests are related to Database Systems, Information Systems, Business Intelligence Systems, Data Science, and Software Engineering. Chair of Managing Board of the Computer Science and Information Systems (ComSIS) journal. For several years, a chair of a series of MADEISD workshops at the ADBIS conference, Topical Area 5 – Software, System, and Service Engineering at the FedCSIS conference, as well as IADSP and IADT invited sessions at the KES conference. Chair of M.Sc. study program in Information Engineering, i.e. Data Science, at the Faculty of Organizational Sciences. From 2015 to 2021, created and chaired a new set of B.Sc. and M.Sc. study programs in Information Engineering, at the Faculty of Technical Sciences from the University of Novi Sad. E-mail: ivan.lukovic@fon.bg.ac.rs.

Smart Urban Planning: An Intelligent Framework to Predict Traffic Using Stack Ensembling Approach

Anjum, Muhammad Adeel and Alanzi, Ahmad

Abstract: The intelligent transportation system needs to accurately assess the volume of traffic in the environment in which it operates to ensure that people are moved in a timely and hassle-free manner. Forecasting systems allow drivers to identify the route that will take them to their destination with the slightest difficulty and the least time spent in congested regions. At present, both the corporate sector and government organizations require accurate and timely traffic flow information. There have been no significant efforts to enhance road traffic prediction by utilizing air pollution data. This paper aims to present a new method for predicting road traffic using data related to pollution. Our contribution to this research is twofold. Firstly, we compared ten regression approaches to determine which technique provides better results and accuracy. Secondly, we present a technique based on regression analysis approaches in which we choose those base learners who give better results on Level 1. These predictions are combined as an input to a Level 2 meta regressor. A method is proposed to show that it generates more satisfactory results than any of the regression procedures discussed previously. Compared with the various regression methodologies, the proposed method successfully lowers the mean square error, the relative absolute error, and the root mean square error and improves the R-squared value.

Index Terms: ***Machine learning, ensembling, traffic prediction, air pollution, stack ensembling***

1. INTRODUCTION

The increase in the number of cars on the road is one factor that has contributed to the increase in air pollution, which severely impacts people's health and their standard of living. The number of cars on the roads has increased due to population growth and the economy. Traffic congestion is a major problem, particularly considering the growing number of vehicles on the roadway. The amount of time that is wasted in traffic and the amount of air pollution produced, the amount of fuel consumed, the amount of energy consumed on infrastructure improvements, the amount invested in infrastructure upgrades, and the frequency that transportation infrastructure and roads require maintenance are only a few of the ways the way that traffic can have a significant impact on the daily lives of people. [1]. The issue of traffic is crucial to a city and its residents' well-being because of the vast number of automobiles in use. Vehicular emissions contribute to high urban traffic congestion levels by adding to air pollution. Increased emissions from increased traffic are a significant contributor to air pollution in metropolitan areas [2]. Pollution and lousy air quality are substantial causes of traffic jams in urban areas. It makes it harder for people to get to the hospital, which raises the death rate in big cities.

According to the World Health Organization [3], the transportation sector contributes significantly to air pollution, resulting in more than seven million people's deaths annually. More than eighty per cent of the urban population resides where air pollution exceeds WHO guidelines [4]. These two ideas are connected, and numerous towns are working to solve this problem by installing sensors that detect the amount of traffic and the quality of the air. Several air pollutants, such as carbon dioxide (CO_2), carbon monoxide (CO), volatile organic compounds (VOC_s), nitrogen dioxide (NO_2), and particulate matter (PM), have been caused mainly by vehicle emissions [5]. It has been discovered that exposure to air pollution brought on by traffic can affect a person's health in the short and long term. These effects include asthma, reduced lung growth in children, cardiovascular disease in adults, and poor academic performance [6]. Pollutants such as ozone and $PM_{2.5}$ induce respiratory severe abnormalities. ($PM_{2.5}$) is 2.5-micrometre-diameter particulate matter. Road networks are the basis of every nation's growth plan. The free flow of

Manuscript received Sep 25, 2023.

Author Muhammad Adeel Anjum is with the Computer Science Department, Comsats University Islamabad, Pakistan (e-mail: adeel.anjum456@gmail.com).

Author Ahmed Alanzi is with the University of Taibah, Saudi Arabia (e-mail: Aenzi@taibahu.edu.sa).

vehicles on the road is crucial for quicker communications and transportation networks. Cameras will be installed on the streets of an intelligent city to monitor traffic flow and transportation. A range of road users, including drivers, private vehicle passengers, and passengers on public transport, require accurate traffic flow data. This information would aid road users in making more intelligent travel decisions, enhance traffic service quality, lower noise, and overcome traffic congestion. Traffic congestion forecasting aims to provide advanced knowledge about traffic congestion.

The use of the traffic congestion forecast has grown due to the rapid growth and application of intelligent transportation systems. If traffic flow can be forecasted in advance, it will be easier for drivers to avoid congestion. It is because they will be able to choose the route to their destination that is both the most convenient and the least congested, or they will be able to adjust their travel plan to account for the time they expect to spend in traffic. Many studies have shown that traffic movement data can help predict air pollution levels. For instance, Batterman et al. [7] estimated ($PM_{2.5}$) and (NO) air pollutants using an emission inventory with the R-LINE dispersion model. The researcher Ly et al. [5] evaluated the amounts of NO_2 and (CO) by combining data from multimodal devices with meteorological data, including absolute humidity, relative humidity, and temperature. Their studies, however, did not consider the amount of traffic density. When travellers are prepared with information regarding the quickest path to their destination, their journey is simplified and enhanced in convenience and ease. Traffic flow administration is among the smart city system's most essential subcomponents.

The primary objective of this paper is to develop a model for predicting road traffic by including data on air pollution levels. To provide predictions, traditional traffic forecasting methods employ multiple variables, including historical traffic patterns and up-to-date data. Nevertheless, including pollution data introduces a novel dimension that improves the precision of these forecasts. So far, there has been a distinct lack of significant initiatives to investigate the relationship between air pollution and traffic flow with the aim of prediction. This paper makes two important contributions to the field of traffic prediction:

1) Comparative Analysis of Regression Approaches: The first contribution extensively evaluates ten different regression techniques to identify the most effective approach for integrating pollution data into traffic prediction models. We aim to determine which regression method yields the highest accuracy and performance metrics through rigorous experimentation and analysis.

2) Novel Regression Ensemble Method: Our second contribution involves creating a novel regression ensemble method explicitly designed for traffic prediction. Based on insights gained through comparative analysis, we propose an ensemble method combining predictions from multiple base learners at Level 1, then transitioning into meta-regression at Level 2. Our goal with this ensemble approach is to leverage individual regression models' strengths for more accurate traffic forecasts.

By providing an in-depth assessment of regression approaches and developing an innovative ensemble method, this research seeks to advance road traffic prediction. The proposed method is expected to perform better than currently employed methods as it lowers error metrics such as percent error of mean square error or root mean square error while increasing the R-squared value. It will make traffic forecasting systems more dependable and useful.

The smart city platform is responsible for providing several services, one of the most significant of which is smart mobility. In many cities, traffic congestion harms health due to increased air pollution. Smart congestion management assists drivers in avoiding congested areas and lowering pollutant concentrations. Due to the traffic flow's unpredictable and nonlinear character, it is difficult to predict the extent to which congestion will be spread. Our research shows that air pollution significantly influences traffic forecasts. The accuracy of traffic forecasts will improve due to lower pollution levels. Road traffic was the primary variable considered in every one of the earlier studies that attempted to forecast levels of air pollution. However, only a few academics have investigated using air quality to improve traffic forecasting. Most of those who have tried to increase traffic forecasting did not take air pollution into account. This is despite air quality being shown to impact traffic significantly. Additionally, they have exclusively utilized standard statistical models in their research. These constraints pose a significant barrier to air pollution-based traffic forecasting. The vast number of automobiles on the road is a major contributor to the problem of air pollution. Consequently, quantifying or using the resulting pollution levels to determine the number of vehicles present might not be a surprise. In the real world, things are more complicated, such as: 1) The levels of air pollution are higher near major roadways, and population density is higher near major roadways. One rationale for the proximity of

pollution sensors to traffic flow sensors is as follows. 2) the automobile set's arrangement is more intricate than in the past.

1.1. Problem Statement

This work investigates the problem of traffic prediction from pollution data in smart cities. Predicting road traffic is one of the important design challenges of a smart city. Different machine learning techniques [8][9][10] have been used to predict road traffic using pollution data; however, the existing methods are unable to improve the R-squared and reduce the error rate. There is a need to propose a new technique that uses stack ensembling to enhance the results and minimize the error rate.

2. BACKGROUND

This section gives an in-depth review of the various deep and machine learning models that are used to model regression to traffic forecasting.

2.1. Deep Learning

Deep Learning (DL) is one of the subfields within Machine Learning (ML). Machine learning includes deep learning. It learns and improves by analyzing computer algorithms. Deep learning imitates how humans think and learn using artificial neural networks. Deep learning helps classify images, translate languages, and recognize the voice. It can solve any pattern recognition problem automatically. Deep learning uses multilayered neural networks. Deep Neural Networks (DNNs) can perform complex tasks such as representation and abstraction to interpret images, audio, and text. Deep learning is the rapidly expanding field of machine learning that many firms use to create innovative business models. The Artificial Intelligence (AI) subfield concentrates on developing systems that can self-learn and enhance performance with increased experience.

2.1.1 Long Short-Term Memory

In the sequence forecasting problem, LSTM trains a model to learn order dependence. Machine translation, speech recognition, etc., are only a few examples of the numerous complicated problem domains in which it is applied. The use of LSTMs is a complex area of DL. Long Short-Term Memory Network is an enhanced RNN that stores data. It solves RNN's vanishing gradient problem. RNNs provide permanent memory. The authors of reference [11] presented a residual graph convolution long short-term memory (RGCLSTM) model. Data with geographical and temporal components were predicted using this method. We ran it at a rate of 10 minutes for each use.

Two sets of information were used to compile this report: traffic statistics from Shanghai, China, and information from Caltrans' Performance Metrics System (PeMS). There is a significant incidence of inaccuracy because fewer characteristics were used, and data on air pollution were ignored. In the paper [12], An artificial neural network that predicted traffic was designed using the Stacked Bidirectional and Unidirectional LSTM (SBU-LSTM) network topology. LSTM is the model's main component. Bidirectional long-term memory (BDLSM) tracks forward and reverse temporal dependencies when working with spatiotemporal data. To avoid missing temporal-spatial data, "data imputation" was proposed. In [13], the graph attention mechanism was used to determine how different road segments depend on each other in space. In addition, an LSTM network was created to extract characteristics from the temporal domain. Pollution data was ignored using the California Department of Transportation's PeMSD7 data.

2.1.2 CNN

Convolution Neural Network (CNN) and DL architecture form the basis of the model reported in [13]. As such, this model is concerned with forecasting near-future traffic patterns. The Spatio-Temporal Feature Selection Algorithm (STFSA) was implemented into the design to determine the input data time, spatial data delays, and volume that were determined to be the most appropriate. The Spatiotemporal traffic flow characteristics were derived from the raw data. The data was then transformed into a two-dimensional matrix. CNN figured out how to use the information it had to construct a model for making predictions. Researchers utilized data from the Washington State Department of Transportation (WSDOT) along the Interstate 5 Freeway in Seattle to show how artificial intelligence helps computers perform activities previously done manually by humans. Fantastic advancements in the area result from the combined efforts of researchers and fans working on various aspects of the subject [14]. Technology that uses a camera to detect objects is one example of this. This field of study aims to teach robots to perceive and understand the world similarly to humans, enabling them to conduct tasks that involve auditory, visual, and contextual information. Over a Convolutional Neural Network, Deep Learning-based computer vision has been developed and refined. A Convolutional Neural Network (ConvNet/CNN) may classify objects using an input image and learnable weights and biases.

Zhang et al. [15] introduced a congestion prediction model utilizing a Convolutional Neural Network (CNN) integrated with a long short-term memory neural system. They examined the unprocessed traffic congestion maps presented as a matrix series to construct their model.

Vehicles or road visual characteristics are not explicitly considered in this process. A possible result might be that it will not have a reliable estimate of traffic congestion. For short-term traffic forecasting, the study's authors [16] created and implemented a modified Spatiotemporal K-Nearest Neighbors (SKNN) called D-STKNN, a dynamic transformation of STKNN. Thus, a spatiotemporal model of traffic that is not static should be used. In this case, we used data sets that tracked how fast cars moved on expressways in California and Beijing, China. In [17, the A3T-GCN model uses the road network's design to learn time-space relationships and gates to identify short-term direction. The tests used Los-loop and SZ-taxi datasets. Insufficient sample size may lead to overgeneralization.

2.2. Machine Learning

Machine learning (ML) is one subfield of artificial intelligence (AI). AI and ML techniques are used to create decisions comparable to those made by people utilizing various models and algorithms capable of adapting to new situations and gaining new knowledge. Machine learning models are trained using data and get smarter and better over time. To avoid misunderstandings, we need to define machine learning. Machine learning enables autonomous learning and improving systems with little human intervention. Machine learning creates algorithms with the ability to assimilate current information. Machine learning inputs things like training data or knowledge graphs, which help it understand domains, entities, and relationships.

2.2.1 Naive Bayes

Naive Bayes is effective for classifiers where class labels are finite and feature instances are vectors of feature values. Bayes' theorem computes the likelihood of event A occurring given event B. Evidence B supports hypothesis A. Assume independent predictors/features. One feature does not impact another. It is naive. Let us see with an example. Consider golf. The day's features determine whether it is good for golf. Columns represent features, and rows represent entries. In the dataset's first row, golf is not ideal if the weather is wet, hot, humid, and windless [18]. First, these predictors are independent, as noted above. If it is hot, that does not indicate it is humid. All predictors have an equal impact on the outcome, another assumption. Windy days do not affect whether you play golf. y is a class variable (play golf) that indicates if conditions are acceptable for golf. X represents parameters/features. x1x2....xn, i.e., outlook, temperature, humidity, and wind. Now, look at the dataset for each value and substitute it into the equation. The denominator remains static for all dataset elements. The denominator was removed; proportionality was introduced.

Class variable(y) can only be yes or no. Multivariate classification is possible. Given the predictors, we must determine class y with the greatest probability. Those Naive Bayes classifiers and Documents on sports, politics, technology, and other fields are classified using Multinomial Naive Bayes [19]. Classifier features/predictors are document word frequency. The yes or no values, such as the presence or absence of a word in the text, are the parameters we use to predict the class variable. Naive Bayes, when predictors are continuous, we suppose they are sampled from a Gaussian distribution since dataset values fluctuate.

In [20], the causal relationship was modelled using a Naive Bayes classifier approach. Python's Scikit-learn module and field survey data were utilized. The data was first split into the test set and the training set. This model has 72.25 percent accuracy for testing and training datasets and 85.03 percent for testing datasets. It has an RMSE of 0.46 and an MAE of 0.28. The naive Bayes classification system shows promise in analyzing weather-related traffic consequences. The strategy was created to create an ATM and an ATIS for Dhaka. Drivers will be free to choose alternate routes that are less congested, thereby easing traffic.

2.2.2 KNN

The KNN approach is a supervised learning-based solution to a classification or regression problem; it assumes that similar objects are adjacent or that the same things are nearby. KNN is an easy-to-understand supervised learning method for regression and classification. Supervised machine learning techniques employ labelled input data to learn a function that yields the correct output with unlabeled data.

The issue of static model designs without spatial, temporal, or dependent link specificity is addressed in reference [21]. We presented adaptiveSTKNN, a K Nearest-Neighbor model incorporating space and temporal factors to forecast traffic at brief intervals. We used data from California expressways and Beijing city streets to evaluate the adaptive-STKNN model's relative vehicle speed prediction. It was proposed in [22] that we use a kernelized KNN technique. It was designed for the varying traffic conditions on the road, as shown by the clock. First, a sample of data representative of the road traffic situation was collected. Reference sequences were used to refine the features of how traffic flows on the road. Finally, the kernel module for the road traffic time series was built. Sequences from both the cited and present data were compared and matched. It has to do with the chaos on the roads. Wang et al. [23] proposed integrating the GMDH and SARIMA

methods to estimate traffic flow in Guiyang, Guizhou province, China. However, an extensive comparison is not made. Only LSTM is considered for comparison. The impact of air pollution is not considered. With an integrated empirical evaluation, P. Fernandes et al. [24] compare several sub-urban roundabouts traffic efficiencies, air pollution, and noise emissions. Exhaust emissions, engine operation, and noise levels are measured in the real world using three different instruments: a portable emission measures system, an onboard diagnostic scan tool, and a sound level meter. The proposed concept, however, will only be used at roundabouts with the same traffic conditions.

2.3. Ensemble Techniques

Ensemble models combine several models or learning methods to create dependable prediction models. This ultimate model drastically outperforms the basic learners. Other applications of ensemble learning include feature selection, data fusion, and so forth. Bagging, boosting, and stacking are the three basic categories of ensemble methods. Ensemble approaches in machine learning combine many learning models to make better conclusions. These approaches work like the air conditioner example. Boosting, bagging, and stacking are the three primary types of ensemble procedures. Also known as "ensemble techniques."

2.3.1. Boosting

The reference authors [25] utilized hierarchical reconciliation and a gradient-boosting method to forecast upcoming traffic volumes. Research has been done on temporal and spatial space models and their interconnections. The dynamics of traffic density in diverse locations are important because of their significance in traffic flow. Three distinct datasets were utilized to examine the proposed framework's performance compared to SARIMA, the Kalman filter model, and RF methods. Gradient boosting is used to learn information from vast datasets and provides automated and extremely flexible learning techniques. This data will be useful if you are trying to predict the traffic volume on a large road system.

A multivariate Gradient Boosting Regression Tree (GBRT) in [13] considers the interdependencies of the outputs to generate numerous outputs, distinguishing it from previous methods. We recorded the traffic conditions every five minutes. The PeMS utilized three-loop detectors on the US101-N roadway to accomplish the mission. We used the Support Vector Regressor (SVR) method as a standard. Three models were evaluated based on these criteria: how well and consistently they predicted and how long it took them to make their predictions. Experiments show that using GBRT or multivariate GBRT directly yields better results than using SVR to make predictions. Using an iterative technique, GBRT achieves good prediction accuracy in the short-step-ahead setting, whereas the precious accuracy reduces dramatically in the long-step-ahead setting. When it comes to stability, multivariate GBRT is unparalleled. Therefore, multi-step-ahead prediction provides greater dependability than the iterative GBRT. The stability offered by the GBRT is the worst possible. The method has a high RMSE and MAE error; the problem is that it is not general.

In [9], the authors employ a boosting ensemble technique to improve the findings of a single regression model before feeding them into a second, even more refined model; a thorough comparison of the various regression models follows this. The dataset utilized comprises publicly available information about Pulse Aarhus City. Standard methods like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were used. Therefore, the method is overly dependent on exceptional cases. Another reason boosting is not scalable beyond a certain point is that each model relies on the accuracy of the preceding model to determine its accuracy. The proposed method was evaluated with a condensed set of model combinations. Further improvements to the error rate are possible.

2.3.2. Bagging

For their traffic forecasts, the authors of [26] used ensemble learning, combining elements of multitask learning with those of traditional ensemble methods. As used in conventional traffic density forecasting methods, a single-task learning model may overlook crucial information embedded in other related tasks. On the other hand, multitask learning uses the overlap between distinct types of work. There have been recent advancements in traffic forecasting using ensemble learning. MTLBag is an approach to traffic flow prediction that combines multitask learning with bagging, a popular ensemble learning method. The benefits of multitask learning over single-task learning were initially highlighted using neural network learners to forecast traffic patterns. Studies showed that MTLBag outperformed a neural network in juggling multiple tasks simultaneously.

Researchers used a hybrid method [27] in which ANN and statistical techniques were used to estimate traffic flow in an urban environment over one hour. Experiments were conducted on three distinct types of actual streets; however, pollution data was not examined. Due to the one-hour duration of the study, the model ignores generic forecasts. The paper [28] presents a novel approach to forecasting the intensity of a traffic

accident using a blend of the balanced bagging classifier and light gradient boosting machines. Using hybrid models to combine the strengths of different classification models is a novel and effective way to improve the accuracy of traffic safety systems. Evaluation of real-world datasets gives credence to the proposed method and demonstrates its practical applications in real-life scenarios. However, this paper could benefit from including more in-depth details regarding the specific dataset used and experiments conducted, as this would provide greater insight into the performance of the proposed method. The author [29] thoroughly compares machine learning and ensemble learning algorithms for predicting highway lane-changing manoeuvres in this paper. Ensemble methods offer a promising method for improving accuracy in predicting highway lane change manoeuvres. Evaluating a real-world dataset gives credibility to a proposed method and suggests it has real-world applications. However, more details regarding specific datasets and experiments would provide more insight into its performance.

2.3.3. Stacking

Researchers have employed RNN-based algorithms to predict traffic movements [12]. A vital component of this method is its proposed stacking architecture, which contains two networks: unidirectional long short-term memory (LSTM) and bidirectional. Spatial-temporal data were analyzed using BDLSM models to establish past and future temporal connections. We added an imputation component to the LSTM model to detect missing values in spatial-temporal data and help anticipate traffic flow. The SBU-LSTM design incorporates the LSTM's bidirectional capabilities. Real-world information was used to create the datasets. Two traffic-based network-wide datasets were used in the research.

It was released for publication and use in traffic-flow forecasting research. We evaluated a variety of different multi-layer LSTM and BDLSTM models. The outcomes demonstrate that the proposed method yielded excellent performance, notably the two-layer BDLSTM network, which could get a respectable performance for predicting traffic flow. The RMSE is too high, and information on pollution was ignored.

The stacked autoencoder was used to predict traffic [36], but this method has flaws. Multiple stacked auto-encoders were trained using a sample replication strategy, and the learned autoencoders were then ensembled using an adaptive boosting technique. Although a comparative examination was not conducted, the strategy significantly enhanced traffic flow forecasts. Table 1 summarizes some of the approaches that have been used in the past. Voting

Voting is a simple ensembling method in which the predictions of multiple machine learning models are combined to make a final prediction. Each model is trained independently on the same training data in a voting ensemble, producing its prediction. The final prediction is then made by aggregating the individual predictions using a majority or weighted vote.

The prediction with the most votes from individual models is chosen as the final one by majority vote. This method works when all models perform similarly, and no single model stands out significantly from its rivals. With weighted voting, however, each model is assigned an appropriate weight that represents its relative performance on validation data; then, their predictions are combined by multiplying each with its weight to create one product and taking the sum. Afterwards, this score determines what prediction is chosen as the final based on scoring; it is useful when some models perform better than others.

Voting ensembles are simple solutions for improving the accuracy and stability of machine learning models. Voting ensembles have become popular in competitions and real-world applications to optimize model performance while decreasing the risks of overfitting.

The paper [37] presents a novel and effective approach for traffic congestion detection in smart city applications using ensemble-based methods combining multiple algorithms. These ensemble methods have proven successful at improving the accuracy of traffic detection systems. Evaluation of real-world data adds credibility to the proposed method and suggests its practical applications in real-world scenarios. However, this paper could benefit from more specific details about the dataset used and experiments conducted to gain greater insight into the effectiveness of its proposed method. In this paper, the author [38] proposes an innovative vehicle detection and traffic density estimation solution within traffic surveillance systems. Employing ensemble-based methods to combine the strengths of two popular object detection models is a creative and effective approach to improving traffic detection systems' accuracy. Evaluation of public datasets and comparison with state-of-the-art methods lend credibility to the proposed method and suggest its practical applications in real-world scenarios. However, more details regarding specific datasets used and experiments performed could provide more insight into their performance.

Table 1: Review Matrix for Literature Review

Authors	Description of Research	Model	Dataset Used	Evaluation Measures	Limitations
[22]	The SZ-taxi and Los-loop datasets showed an improvement above the baselines.	A3T-GCN	The Shenzhen taxi tracker (SZ-taxi) and Los Angeles loop detector datasets	RMSE, MAE, Accuracy, R-2	The error rate is significant, and pollution factors have not been considered.
[30]	A novel deep learning-based network was proposed for traffic prediction.	LSTM + MDC	Traffic data from PeMS	MAE, RMSE, MAPE	Fewer features to use did not include pollution features.
[13]	IoT-driven congestion forecasting for intelligent, eco-friendly urban areas	LSTM	Dataset including traffic flow data in Buxton, UK, gathered by IoT devices at two specific sites.	RMSE, Accuracy	There is no comparison. Factors such as air pollution are not considered.
[31]	The effects of suburban roundabouts on congestion-specific vehicle speed profiles	Predictive discrete choice models	The Performance Measurement System (PMS) and a sound level meter measured the traffic flow.	RMSE, MAE	Due to the site-specific nature of the proposed concept, it will exclusively be implemented at roundabouts that present comparable traffic conditions.
[32]	A model capable of being interpreted to forecast short-term traffic flow	GMDH + SARIMA	Dataset of vehicles in Guiyang district, Guizhou province, China.	RMSE, MRE	Compared with only one model
[18]	A comparative examination of ensemble approaches is performed using road traffic congestion data.	RF, SVM, DT, and LR	Real-life road traffic flow dataset	Accuracy, F1-score, Recall, and Precision	They did not use pollution data.
[33]	Smart city short-term traffic flow prediction using 5g internet of vehicles	EC-DCRFNN	SUMO is an open-source traffic simulation software.	Accuracy	It does not propose a suitable task-scheduling algorithm.
[34]	An evacuation plan for a smart city is focused on adjusting the road network layout for increased resilience.	RnR-SMART	The public GIS data	Comparing the shortest route	It needs to be evaluated on a multipath framework
[11]	Examined traffic flow spatially and temporally.	CNN + STFSA	Transportation Department of Washington dataset.	MAP, MAE	A small dataset with a high error rate.
[35]	Gradient boosting is used to construct a neural network design.	gaNet	Real-time IoT traffic data from a mobile service provider	MAE, MSE, MAD, R2, NRMSD	For gaNet-C, the time needed to train and predict is quite limited.

2.3.4. Blending

Blending is an aggregation method that combines predictions from multiple machine learning models into a final prediction. In blending, some training data is withheld and used instead to train numerous base models, which then make predictions on the data held out; these predictions then serve as input features to train a meta-model.

Held-out data is typically selected from the original training data to train a meta-model without overfitting, ensuring it can generalize effectively to new data sets.

Meta models are straightforward machine learning algorithms, including logistic regression, linear regression, and decision trees, which employ predictions from base models as input features to enhance the precision of variable target predictions. Blending is like stacking, yet it differs by taking an alternative approach to data division for meta-model training. When employing stacking, base models train data in developing meta-models, unlike in blending, where data division occurs from various sources within base models. Blending involves purposely withholding certain data sets for the sole purpose of developing meta-models. Blending is an innovative technique used to increase the precision and reliability of machine learning models, which has gained significant traction both competitively and for practical uses. Blending is more complex and uses more computational resources than simpler approaches such as voting or bagging; however, it leads to improved model performance while decreasing the risk of overfitting.

This section comprehensively analyzes machine learning (ML) and deep learning (DL) models employed in traffic prediction while delineating their advantages and constraints. Ensemble techniques are becoming popular methods for improving prediction accuracy. They achieve this by leveraging various model knowledge to generate more precise forecasts. Additional studies should explore mitigating constraints by incorporating supplementary data modalities, such as pollution data, into current frameworks to develop more comprehensive

traffic management solutions.

3. ARCHITECTURE

3.1 Data Gathering

Data collection and processing are central components of any methodology and integral to its success. Data came from a real-time Aarhus city pulse-open dataset. At our disposal were two sets of information, specifically pollution data and traffic intensity [39], collected through various sensors placed throughout the city. We collected information every five minutes on passing automobiles under these sensing devices. The air dataset includes information regarding (CO), sulfur oxides (SO_2), ozone (O_3), and (PM) that were released into the atmosphere from these vehicles (PM). The pollution and vehicle density data used to create this report spanned over a year's worth of traffic. It includes 120,000 unique instances, each of which is equipped with data on the percentage of oxygen in the air, the number of particles in the air, the percentage of (CO), the percentage of sulfur dioxide (SO_2), the number of nitrogen oxides (NO_x) in the air, the number of vehicles in the area, and the time each car arrived. Vehicle traffic was predicted using pollution data; therefore, both statistics were provided. Pollution and traffic data from Aarhus' Denmark website was integrated using the timestamps provided; additionally, this study utilized an open-source dataset incorporating real-time sensor data that provided insight into Aarhus's life dynamics.

This data set has been referenced in numerous academic papers, including [9][10]. It can be downloaded and used at no cost from their website. The data is offered in a compressed ZIP archive. This file can be easily retrieved and put to effective use. Experiments were conducted using pollution data, including levels of (NO_2), (CO), (SO_2), (O_3), and (PM), location data (including latitude and longitude), and traffic data (including levels of traffic intensity). Information regarding the duration of travel between two locations. Examining the relationship between the dataset's attributes is accomplished through a correlation matrix graph. A positive correlation between the characteristics is evident.

According to the traffic dataset, the total number of cars was calculated, and that number was then added to the pollution dataset, together with the timestamps for each pollutant. Due to the proximity and precision of the sensors that collected the data from the roadsides, the two datasets were combined. Also, the emission of pollutants like (NO_x), (CO), and (SO_2) is proportionate to the number of vehicles on the road; therefore, higher emissions lead to heavier traffic. The model can be used for comprehensive monitoring in urban areas, and the decision to use only pollutant data to analyze traffic flow was made to lower the infrastructure cost. Using the pollution dataset, we can make broad, less-expensive predictions instead of expensive, highly specialized sensors. A correlation matrix is given in Figure 1.

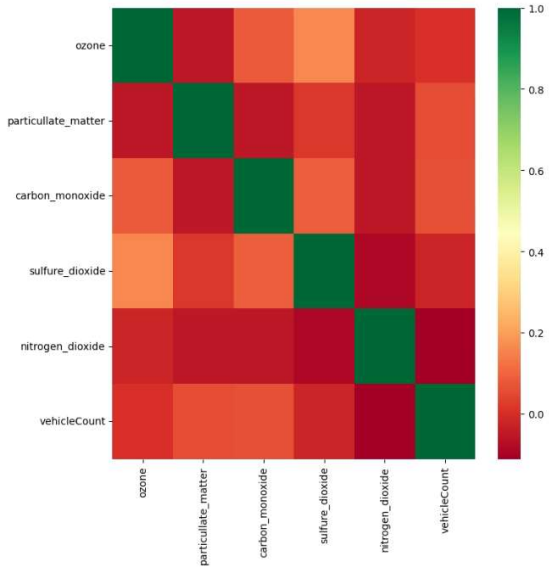


Figure 1: The Correlation heatmap for the used dataset.

The rationale for choosing the given datasets lies in their accessibility, real-time nature, comprehensive coverage of pollution and traffic data, high granularity, relevance to the study context, and previous academic references. These factors collectively support the study's objectives and contribute to the robustness and validity of its findings.

A sample of the dataset is given in Figure 2

	ozone	particullate_matter	carbon_monoxide	sulfure_dioxide	nitrogen_dioxide	longitude	latitude	timestamp	vehicleCount
0	127	38	62	22	39	10.104986	56.231721	8/1/2014 7:50	5
1	122	35	61	17	34	10.104986	56.231721	8/1/2014 7:55	6
2	117	36	65	24	34	10.104986	56.231721	8/1/2014 8:00	4
3	112	36	70	29	35	10.104986	56.231721	8/1/2014 8:05	1
4	115	34	69	34	35	10.104986	56.231721	8/1/2014 8:10	3

Figure 2: A sample of the dataset.

3.2 Data Preprocessing

This stage involved data processing operations. Since the data were raw and contained missing values, outliers, etc. The key factor in establishing the correct relationship between output and input variables was converting the data into a format that the model could analyze. Data was prepared for input into models using a variety of methods.

Many factors, including incomplete surveys and typos in data entry, might lead to missing values. Missing data were wiped out using a median/mean imputation technique. We used a program from Pandas to identify the column with a missing value

and then removed the value. After locating the index of that number, we then substituted the column's mean for it. Outliers are data points that are far from the mean or median. In a dataset, they would be quite unusual.

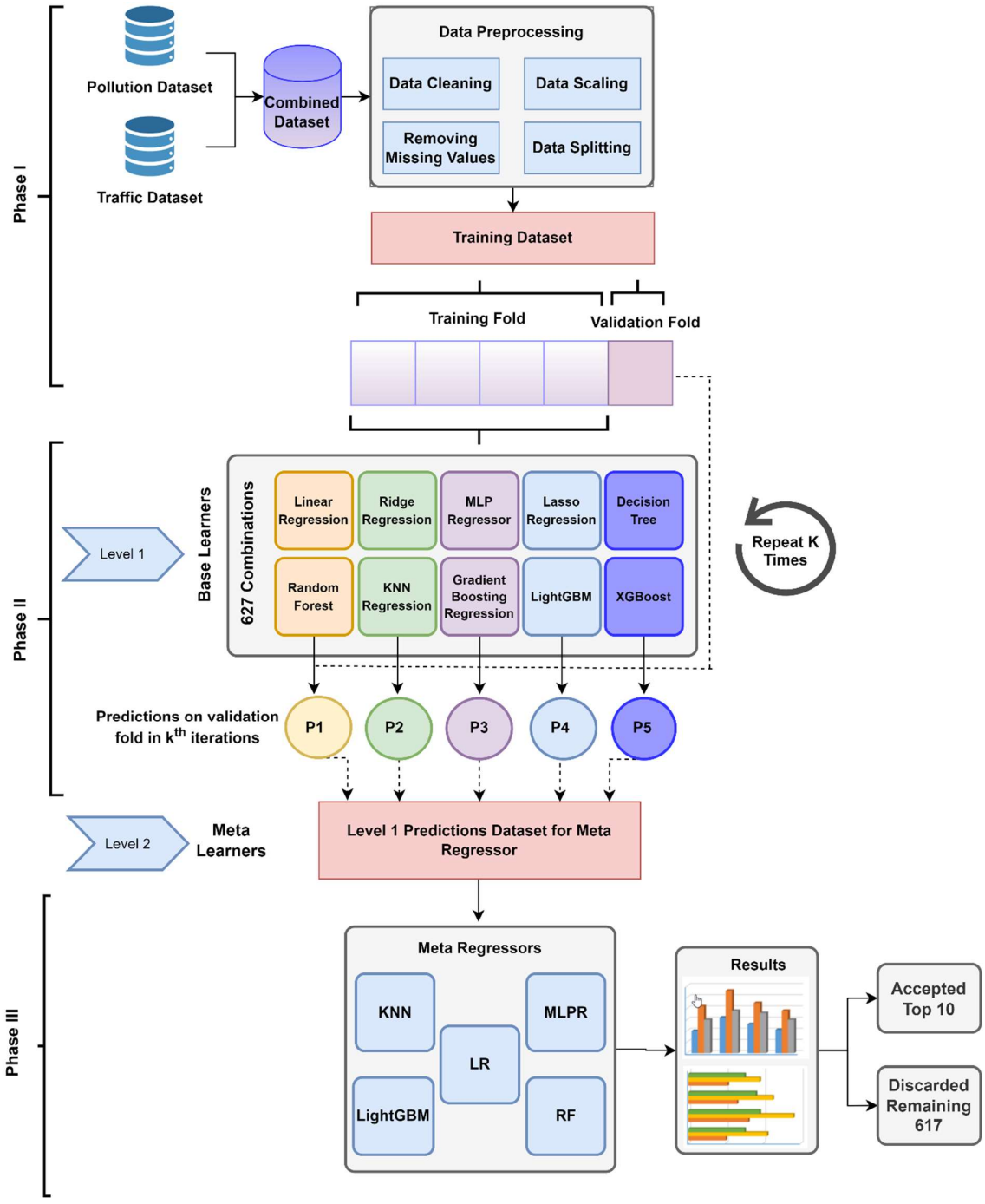


Figure 3: Proposed Stack Ensemble Architecture

Outliers can be a problem in statistical research because they cause tests to overlook significant findings or destroy legitimate results. These details may explain why the tests fail to detect the important findings or the necessary results. Since the best method for eliminating anomalies is problem-specific, there is no universal technique. Z-scores were applied to identify and eliminate anomalies. The distance from the centre of the graph is displayed. It is based on an average of all the numbers. A value with a Z-score greater than zero is significantly out of the ordinary and vice versa. SciPy was employed to apply the Z-score to a dataset. We used SciPy to calculate Z-scores and filter out values below 3 to eliminate outliers. Information was cleaned and normalized so that algorithms like KNN could process it. This is because they measure similarities by comparing the distances between data sets. The min-max scaler and normalization tools from the Scikit-learn module were used for this step. The process of normalizing data is an important part of any pre-processing procedure. If you want your ML or DL model to have reliable input, you need to normalize your data to be used without losing quality during the training process [41]. Data normalization becomes crucial when features in the data have varying values. Consider a data set where the values for traffic density shift from 0 to 50 while the ranges for (CO_2) and (NO_2) change from 0 to 300 and 0 to 4.2, respectively. As a result of the disparity in scales, model performance may be subpar. We can manage these varying scales in the dataset by normalizing the data. It is a boon to us in terms of cutting down on training time. Data normalization approaches include min-max, median, and Z-score decimal scaling. The tests used min-max normalization, a typical data adjustment strategy [42].

3.3 Proposed Model

This paragraph reviews how we calculate traffic patterns based on pollution levels. As part of the first stage, information regarding pollution levels and traffic congestion was collected from the city of Aarhus's website. In the second stage, we performed preprocessing by eliminating outliers and missing values. Then, the time stamps from both sets were used to create a single dataset. The resulting dataset was combined and normalized. In the third stage, many regression models were compared to identify the one that produced the best R-squared value and lowest error rate. The best ensemble approach was compared to several regression models in the fourth stage. The approach diagram shows that this way has the lowest error rate and is the most effective overall.

3.4 Building Model by using Stack Ensembling

A stacking model is an ensemble learning approach that combines multiple regression models into a single model via a meta-regressor. The standard stacking algorithm (implemented as the StackingRegressor) gets the data ready for the level-2 regressor by using out-of-fold predictions when used with the StackingCVRegressor.

To avoid overfitting, it is usual to fit first-level regressors to the same training set as second-level regressors. On the other hand, the StackingCVRegressor utilizes out-of-fold predictions: the dataset is separated into k folds. In k consecutive rounds, the first level regressor is fitted using the first k-1 folds from the dataset. It allows for more accurate predictions. First-level regressors are applied to the one subset not used for model fitting in a previous iteration after each model fitting is complete. The resulting predictions are passed into the second-level regressor as input data. After the StackingCVRegressor has been trained, the first-level regressors are fitted to the entire dataset to provide the most accurate predictions. In the proposed stack ensemble regression model in Figure 3, five regression models relate to the meta-regressor using the stack ensemble technique. Traditional and boosting models are used to enhance the results and predictions of the models. Firstly, base classifiers are applied to the datasets, and then the five classifiers that perform well on that dataset are chosen. After that, the prediction is performed on this and then used as input for meta-learners. A meta-regressor is applied to these prediction datasets to perform the final predictions.

3.5 Putting Model into Work

Our proposed model framework workflow is as follows: First, we select the two publicly available datasets. We used the pollution dataset and the traffic dataset. We combined these datasets based on timestamps. After that, we applied data preprocessing to clean the data and make it suitable for working with machine learning models. In data pre-processing, we first deal with missing values, replacing or deleting the missing records, and then with duplicate values. After duplicate values, we do data standardization to make data in one format and standard to make the model work efficiently with the data. Then, we remove outliers from the data to make it more optimal for the best fit with our proposed model. Then, we select the features that are more relevant to us and remove the extra ones that are irrelevant to us. Then, we apply base machine learning classifiers at the first level and get the results. Then, in the second level, we use these results as input for meta-classifiers. We apply some meta-classifiers and make the final predictions.

4. EXPERIMENTS

In this section, an analysis of the various tests that were carried out on the dataset is presented. Using the pollution dataset, straightforward regression models were utilized to make predictions about the level of traffic. After that, various other combinations of the ensemble model were used. A comparison of the outcomes of the different methods was conducted to determine which combination successfully fulfilled the tasks at hand. The criteria used to select the classifiers were based on their prevalence in the research and previous successful applications to the dataset. Specifically, the classifiers chosen Linear Regression, Ridge Regression, MLP Regression, Lasso Regression, Decision Tree, Random Forest, KNN, Gradient Boosting Regression, LightGBM, and XGBoost were selected because they are widely utilized in the academic and research community. Additionally, past researchers have demonstrated their effectiveness when applied to the same dataset.

4.1. Experimental Setup

Visual studio code or Jupiter Notebook with Python 3.10.7 was used. The dataset was split 70/30 using the sci-kit-learn test train split function, and various sci-kit-learn predefined models were utilized. Using sci-kit-learn, we performed standardization and normalization. Data reading and processing were conducted using NumPy, SciPy, and pandas. Data visualization was accomplished using Seaborn and Matplotlib. We used MAE, RMSE, R-square, and RAE to evaluate the proposed model.

4.2. Evaluation

The four most used metrics for measuring the accuracy of continuous variables are *RMSE*, *MAE*, *RAE*, and R-squared (R2).

4.2.1. Mean Absolute Error (MAE)Building Model by using Stack Ensembling

The *MAE* is a metric for calculating the average number of errors in a series of data values (predictions) without considering direction [43]. A sample's *MAE* is the sum of all the absolute discrepancies between the actual and expected data. It is interpreted as follows:

$$MeanAbsoluteError = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

where,

$y_i =$ prediction, $x_i =$ True Value,

$n =$ Total Number of Data Points

Suppose we do not take the absolute value. Then, the average error resulting from including both positive and negative signs in the mistakes is called the Mean Bias Error (MBE), representing the average bias in the model. It is important to consider MBE because positive and negative faults might cancel one other out, but it is still valuable.

4.2.2. Root Mean Square Error (RMSE)

The *RMSE* is a widely used metric for quantifying the accuracy of a model in predicting numerical outcomes. It is recognized as an exceptional all-purpose error measurement for predictive forecasting. A smaller *RMSE* indicates a more accurate fit to the data. The calculation involves finding the square root of the average of the squared differences between actual and anticipated values [43]. It is estimated as follows:

$$\text{RootMeanSqaureError} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

To compare datasets or modes of different scales, the following procedure [44] is used to normalize the *RMSE*:

$$\text{NormalizedRootMeanSquareError} = \frac{RMSE}{y_{max} - y_{min}}$$

4.2.3. Relative Absolute Error (RAE)

RAE is used to gauge predictive models' accuracy. Common applications of RAE include machine learning, data mining, and operations management; specifically, its application in machine learning is its most frequent form. RAE can be calculated by dividing the mean error (also called residual) against the pre-error produced by oversimplified or naive models and finding their ratio with one another. It is calculated as follows:

$$\text{RelativeAbsoluteError} = \frac{\sum_{j=1}^{n}|P_{ij} - T_j|}{\sum_{j=1}^{n}|T_j - \hat{T}|}$$

Where P_{i} does the individual program estimate the value *i* for sample case *j* (out of *n* sample cases) and T_{j} is the target value for sample case *j* and is given by the formula [45]:

$$\hat{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

4.2.4. R-Squared (R^2)

R-squared is a linear regression model metric measuring how much variance each independent variable accounts for within a dependent variable. R squared can be seen as an easy and direct way of indicating whether our model and interaction are significant, using its scale from 0-100% as its measure. R-squared statistics can often be understood by seeing how well a regression model matches up against real-world findings - this being one method. For instance, if the n r-squared is 60%, this suggests that the regression model can explain 60% of the data. The model is more accurate if the r-squared value is higher [46]. It is calculated using this equation:

$$\text{R} - \text{squared} = \frac{SS_{regression}}{SS_{total}}$$

where $SS_{\mathrm{regression}}$ is the sum of squares due to regression and SS_{total} is the total sum of squares.

4.3. Dataset

It is a well-known dataset that is utilized by many researchers [9] [10] [47] [48]. This study used extensive, live Internet of Things data from a publicly accessible dataset, including details on Aarhus urban activity[20]. It has already been published and is available without charge. The [20] was compiled using information from various sources within Aarhus. The dataset comprises 449 unique sensors located throughout the city. Sensors were positioned at multiple locations throughout the city. Two datasets are utilized for experimental assessment.

Pollution Dataset: This pollution dataset consists of (NO_2), (CO), (SO_2), ozone, longitude, latitude, and (PM) of the various locations.

Traffic Dataset: This traffic dataset consists of different traffic flows. This dataset consists of vehicles, timestamps, longitude, and latitude.

Our experiment combined the number of vehicles with pollution data by time stamp. We accomplished this since the pollution and traffic sensors happened to be in the same area. Air pollution, including (CO), (NO_2), (SO_2), (PM), and ozone increases as the number of vehicles on the road increases. Utilizing air pollution data for forecasting can reduce the number of traffic sensors, cut repair expenses, and establish a more comprehensive environmental monitoring system by transitioning from controlled surveillance to broader sensing in urban areas. This means the model may make predictions based on air pollution data alone, without the need for sensors, which can be a significant cost saving when estimating traffic.

5. PERFORMANCE

5.1. Model Comparison on Different Dataset Size:

With a dataset including 120,000 instances, the best model, Stack Ensembling with MLPR Meta Regressor, was trained and evaluated for accuracy. The dataset was then randomized; between 5 and 30 thousand instances were taken, and models were trained. This was done so that the performance of the model could be evaluated regardless of the size of the dataset on which it was trained. Following that, model outcomes were assessed, each presented on its own in Table 2. As can be observed, the findings vary depending on the sample. The distinction is that as the amount of data increases, the model will acquire more patterns, leading to improved performance. Some combinations perform well with a large dataset. At the same time, the same combination did not perform well in a small or medium dataset, as shown in Table 2. Greater datasets lead to improved results because the increased amount of data allows the model to learn more patterns and perform better.

Linear Regression, Ridge Regression, Lasso Regression, MLP Regression, and Decision Tree do not perform well on our dataset of 120k instances. The performance of different classifiers depends on numerous factors, such as the complexity of the data, the presence of nonlinear relationships or interactions, the size of the dataset, and the effectiveness of regularization techniques. The dataset that we used was nonlinear.

Enhancing predictive precision, the stack ensemble design exhibits favorable results in finance, healthcare, and marketing. Within the area of finance, it serves the purpose of predicting stock prices and optimizing investment portfolios, with the ability to maximize profits. Similarly, in the field of healthcare, it aids in the identification of diseases and the selection of appropriate treatments, resulting in enhanced patient outcomes. Furthermore, marketing contributes to customer segmentation and personalized recommendations, improving sales and customer satisfaction. However, implementation complexity, computational demands, and model interpretability remain. These complexities may hinder real-time application in resource-constrained environments and challenge understanding the ensemble's decision-making process. Thus, while stack ensemble architectures offer significant benefits, their practical deployment requires careful consideration of these limitations.

5.2. Comparative Analysis of Different Techniques

The model's performance was evaluated by comparing it with other baseline models using four popular error measures for regression models. The comparison findings are displayed in Table 3. The RMSE, MAE, RAE, and R-squared values were considered indicators of the model's error rate and fitness, representing the variance between actual values and model predictions. A high RMSE, MAE, and RAE indicate a high error rate in the model, while a low error rate suggests better learning. If the R-squared value belongs to one, then its mean model fitness is good, but if the value of the R-squared is higher than 1 or lower than 1, then its mean model does not fit well. R-squared near 1 shows more model accuracy and should not be negative.

Table 2: Comparison with different datasets

Dataset	Model Combinations	Root Mean Square Error	Mean Absolute Error	Relative Absolute Error	R Squared
Dataset with 5K Instances	LR, RR, RF, KNN, LightGBM	0.5477	0.3519	0.4647	0.6997
	LR, LASSOR, RF, KNN, LightGBM	0.5478	0.3520	0.4648	0.6995
	RR, LASSOR, RF, KNN, LightGBM	0.5479	0.3520	0.4650	0.6994
	RR, MLPR, LASSOR, RF, KNN	0.5478	0.3541	0.4673	0.6995
	LR, RR, MLPR, RF, KNN	0.5474	0.3540	0.4678	0.7000
	LR, MLPR, LASSOR, RF, KNN	0.5479	0.3546	0.4678	0.6993
	RR, RF, KNN, LightGBM, XGBoost	0.5482	0.3548	0.4692	0.6991
	MLPR, LASSOR, RF, KNN, XGBoost	0.5477	0.3550	0.4679	0.6996
	MLPR, RF, KNN, LightGBM, XGBoost	**0.5473**	**0.3551**	**0.4682**	**0.7000**
	LR, RF, KNN, LightGBM, XGBoost	0.5482	0.3548	0.4690	0.6991
Dataset with 30K Instances	**RF, KNN, XGBoost**	**0.5220**	**0.3483**	**0.4491**	**0.7274**
	RF, KNN, GBR	0.5221	0.3483	0.4497	0.7274
	MLPR, RF, KNN	0.5222	0.3472	0.4501	0.7272
	RR, RF, KNN	0.5224	0.3476	0.4504	0.7271
	RF, KNN, LightGBM	0.5224	0.3482	0.4500	0.7271
	LR, RF, KNN	0.5224	0.3473	0.4504	0.7270
	LASSOR, RF, KNN	0.5228	0.3475	0.4509	0.7267
	RR, MLPR, RF, KNN, GBR	0.5227	0.3565	0.4548	0.7268
	LR, MLPR, RF, KNN, GBR	0.5228	0.3571	0.4551	0.7267
	DT, RF, KNN	0.5234	0.3497	0.4505	0.7260
Dataset with 120K Instances	LR, LASSOR, RF, KNN, LightGBM	0.3701	0.2155	0.3083	0.8613
	RR, RF, KNN, GBR	0.3701	0.2160	0.3100	0.8613
	DT, RF, KNN, GBR	0.3701	0.2162	0.3093	0.8613
	RF, KNN, LightGBM, XGBoost	0.3701	0.2162	0.3097	0.8614
	LR, RR, RF, KNN, LightGBM	0.3700	0.2159	0.3087	0.8614
	RR, MLPR, RF, KNN, LightGBM	0.3700	0.2156	0.3087	0.8614
	RR, LASSOR, RF, KNN	0.3701	0.2165	0.3099	0.8613
	RF, KNN, GBR, LightGBM	**0.3698**	**0.2178**	**0.3105**	**0.8615**
	LR, RR, LASSOR, RF, KNN	0.3700	0.2164	0.3090	0.8614
	LR, RR, RF, KNN, GBR	0.3700	0.2162	0.3090	0.8614

Table 3:Performance Comparison of Different Techniques

Model Name	Root Mean Square Error	Mean Absolute Error	Relative Absolute Error	R Squared
Boosting [9]	1.54	1.07	0.59	0.58
Bagging [10]	1.04	0.51	0.38	0.38
Proposed	**0.37**	**0.22**	**0.31**	**0.86**

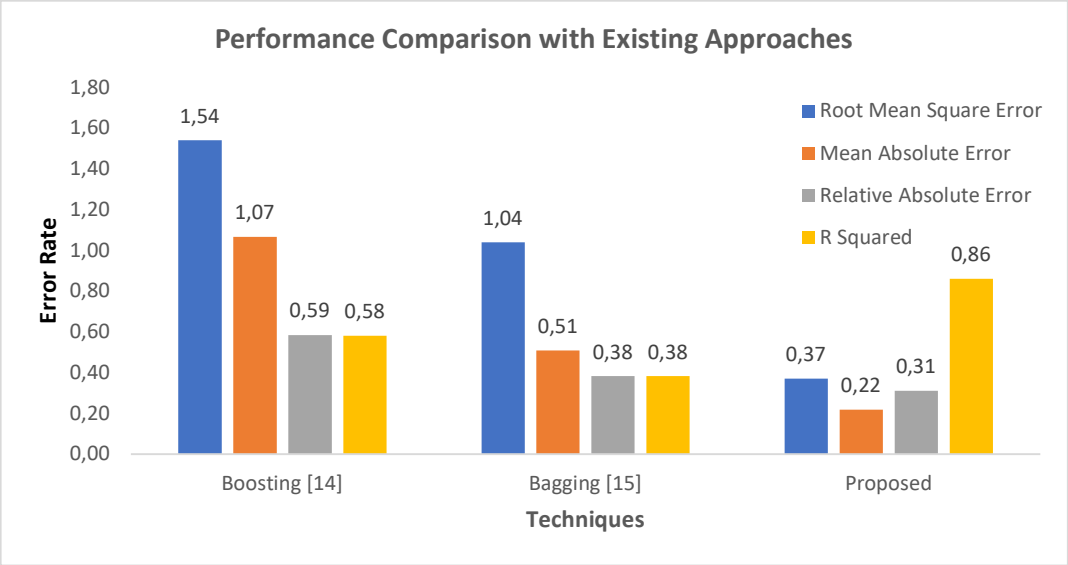


Figure 4: Performance Comparison with Existing Approaches

As can be observed in Table 3, the fact that this model has the lowest error rate of any of the baseline models indicates that it has trained more effectively than others. The base paper is compared with this study [9] [10]. The tests showed that the error rate was reduced while utilizing a stacked ensemble compared to the boosting ensemble in reference [9] and the bagging ensemble in reference [10]. The two research studies are comparable in using the same dataset and investigating the same issue. The initial stage of the machine learning process, known as data pre-processing, is always the same across the board. The Lasso Regression, LightGBM, and XGBoost models were used, while [9] [10] have not used these techniques. The core paper for this study evaluated the suggested system's performance to others, as shown in Figure 4.

6. CONCLUSION

Traffic forecasting is a crucial activity for major cities. Accurate traffic flow estimates can assist drivers in planning their travel routes more efficiently. This paper integrates air quality data and traffic intensity information to enhance the precision of traffic flow calculations and improve forecast accuracy. Various machine learning techniques and combinations of models were used in the dataset to determine the most precise combination. We proposed an ensemble regression technique approach using the stack ensembling approach. In the first phase, we combine and apply two datasets to preprocess. In the second phase, we apply ten base classifiers to our dataset and then predict some values. In the third phase, we use five different meta-regressors to the predicted data from phase two, and then we get the final predictions. MLPR performs well compared to the other four meta-regressors. Results show that RF, KNN, GBR, and LightGBM model combinations provide us with lower error rates and a high R-square. The stacking ensemble technique decreased the error rate by 64% in predicting traffic flow in smart cities, surpassing previous studies utilizing boosting and bagging, as indicated by the experimental data. Because of the considerable number of outliers in the dataset, boosting ensemble models overfit and bagging ensemble models underfit. The trial results indicated that the proposed stacking ensemble technique was effective in reducing the influence of both overfitting and underfitting, which led to a lower error rate.

One significant limitation is the potential for overfitting when training several base models on the same dataset and merging their predictions using a meta-model. This might result in storing irrelevant details and oddities in the training data, leading to a lack of generalization to new data. Additionally, stack ensemble architectures rely heavily on the quality and diversity of the base models and their predictions; if the base models suffer from similar biases or are trained on similar subsets of the data, the ensemble's performance may be compromised. Addressing these limitations requires careful design, diverse base models, and thorough evaluation of varied datasets to ensure effective generalization.

To begin with, it increases storage space and processing time. Second, because so many base classifiers were utilized, stacking could be the root of the model's lack of interpretability. The model's performance may be affected when applied to regions with different seasonal patterns and traffic conditions than Aarhus, Germany, where the dataset was collected. We plan to incorporate more extensive experimental elements into future

studies. Future research will also concentrate on determining time complexity. Furthermore, we can consider weather conditions and air pollution factors to forecast traffic flow.

REFERENCES

[1] Y. Kuang, B. T. H. Yen, E. Suprun, and O. Sahin, "A soft traffic management approach for achieving environmentally sustainable and economically viable outcomes: An Australian case study," *J Environ Manage*, vol. 237, pp. 379–386, May 2019, doi: 10.1016/J.JENVMAN.2019.02.087.

[2] L. Lazić, M. A. Urošević, Z. Mijić, G. Vuković, and L. Ilić, "Traffic contribution to air pollution in urban street canyons: Integrated application of the OSPM, moss biomonitoring and spectral analysis," *Atmos Environ*, vol. 141, pp. 347–360, Sep. 2016, doi: 10.1016/J.ATMOSENV.2016.07.008.

[3] "WHO/Europe | Home." Accessed: Sep. 25, 2023. [Online]. Available: https://www.who.int/europe/home?v=welcome

[4] P. S. Maciąg, N. Kasabov, M. Kryszkiewicz, and R. Bembenik, "Air pollution prediction with clustering-based ensemble of evolving spiking neural networks and a case study for London area," *Environmental Modelling & Software*, vol. 118, pp. 262–280, Aug. 2019, doi: 10.1016/J.ENVSOFT.2019.04.012.

[5] National Research Council. Committee for the Evaluation of the Congestion Mitigation Air Quality Improvement Program, *The Congestion Mitigation and Air Quality Improvement Program Assessing 10 Years of Experience*. Transportation Research Board, 2002. Accessed: Nov. 30, 2022. [Online]. Available: www.TRB.org

[6] B. Brunekreef, N. A. Janssen, J. de Hartog, H. Harssema, M. Knape, and P. van Vliet, "Air Pollution from Truck Traffic and Lung Function in Children Living near Motorways on JSTOR," *Epidemiology*, 1997, Accessed: Nov. 30, 2022. [Online]. Available: https://www.jstor.org/stable/3702257

[7] S. Batterman, R. Ganguly, and P. Harbin, "High Resolution Spatial and Temporal Mapping of Traffic-Related Air Pollutants," *International Journal of Environmental Research and Public Health 2015, Vol. 12, Pages 3646-3666*, vol. 12, no. 4, pp. 3646–3666, Apr. 2015, doi: 10.3390/IJERPH120403646.

[8] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Neural network architecture based on gradient boosting for IoT traffic prediction," *Future Generation Computer Systems*, vol. 100, pp. 656–673, Nov. 2019, doi: 10.1016/J.FUTURE.2019.05.060.

[9] N. Shahid, M. A. Shah, A. Khan, C. Maple, and G. Jeon, "Towards greener smart cities and road traffic forecasting using air pollution data," *Sustain Cities Soc*, vol. 72, p. 103062, Sep. 2021, doi: 10.1016/J.SCS.2021.103062.

[10] N. U. Khan, M. A. Shah, C. Maple, E. Ahmed, and N. Asghar, "Traffic Flow Prediction: An Intelligent Scheme for Forecasting Traffic Flow Using Air Pollution Data in Smart Cities with Bagging Ensemble," *Sustainability 2022, Vol. 14, Page 4164*, vol. 14, no. 7, p. 4164, Mar. 2022, doi: 10.3390/SU14074164.

[11] C. Zhang, J. J. Q. Yu, and Y. Liu, "Spatial-Temporal Graph Attention Networks: A Deep Learning Approach for Traffic Forecasting," *IEEE Access*, vol. 7, pp. 166246–166256, 2019, doi: 10.1109/ACCESS.2019.2953888.

[12] Y. Ma, Z. Zhang, and A. Ihler, "Multi-Lane Short-Term Traffic Forecasting with Convolutional LSTM Network," *IEEE Access*, vol. 8, pp. 34629–34643, 2020, doi: 10.1109/ACCESS.2020.2974575.

[13] S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, and R. Zhu, "Congestion prediction for smart sustainable cities using IoT and machine learning approaches," *Sustain Cities Soc*, vol. 64, p. 102500, Jan. 2021, doi: 10.1016/J.SCS.2020.102500.

[14] I. O. Olayode, A. Severino, L. Tartibu, T. Campisi, and L. K. Tartibu, "Prediction of Vehicular Traffic Flow Using Levenberg-Marquardt Artificial Neural Network Model: Italy Road Transportation System," doi: 10.26552/com.C.2022.2. E74-E86.

[15] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values," *Transp Res Part C Emerg Technol*, vol. 118, p. 102674, Sep. 2020, doi: 10.1016/J.TRC.2020.102674.

[16] J. Guo, Y. Liu, Q. Yang, Y. Wang, and S. Fang, "GPS-based citywide traffic congestion forecasting using CNN-RNN and C3D hybrid model," *https://doi.org/10.1080/23249935.2020.1745927*, vol. 17, no. 2, pp. 190–211, 2020, doi: 10.1080/23249935.2020.1745927.

[17] T. Senthil Kumar, "Video-based Traffic Forecasting using Convolution Neural Network Model and Transfer Learning Techniques," *Journal of Innovative Image Processing*, vol. 02, no. 03, 2020, doi: 10.36548/jiip.2020.3.002.

[18] T. Bokaba, W. Doorsamy, and B. S. Paul, "Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents," *Applied Sciences 2022, Vol. 12, Page 828*, vol. 12, no. 2, p. 828, Jan. 2022, doi: 10.3390/APP12020828.

[19] M. Rasol *et al.*, "GPR monitoring for road transport infrastructure: A systematic review and machine learning insights," *Constr Build Mater*, vol. 324, p. 126686, Mar. 2022, doi: 10.1016/J.CONBUILDMAT.2022.126686.

[20] S. Cheng, F. Lu, and P. Peng, "Short-term traffic forecasting by mining the non-stationarity of spatiotemporal patterns," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 10, pp. 6365–6383, Oct. 2021, doi: 10.1109/TITS.2020.2991781.

[21] J. Bai *et al.*, "A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting," *ISPRS International Journal of Geo-Information 2021, Vol. 10, Page 485*, vol. 10, no. 7, p. 485, Jul. 2021, doi: 10.3390/IJGI10070485.

[22] S. Barua and M. Ahmed, "A naïve Bayes classifier approach to incorporate weather to predict congestion at intersections," *World*

Acad. J. Eng. Sci, vol. 7, no. 2, pp. 72–76, 2020, Accessed: Nov. 29, 2022. [Online]. Available: www.isroset.org

[23] S. Cheng, F. Lu, P. Peng, and S. Wu, "Short-term traffic forecasting: An adaptive ST-KNN model that considers spatial heterogeneity," *Comput Environ Urban Syst*, vol. 71, pp. 186–198, Sep. 2018, doi: 10.1016/J.COMPENVURBSYS.2018.05.009.

[24] Y. Xu *et al.*, "MAF-GNN: Multi-adaptive Spatiotemporal-flow Graph Neural Network for Traffic Speed Forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, p. 036119812211166, Aug. 2021, doi: 10.48550/arxiv.2108.03594.

[25] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with arima time series models to forecast traffic flow," *Transp Res Part C Emerg Technol*, vol. 4, no. 5, pp. 307–318, Oct. 1996, doi: 10.1016/S0968-090X (97)82903-8.

[26] H. W. Kim, J. H. Lee, Y. H. Choi, Y. U. Chung, and H. Lee, "Dynamic bandwidth provisioning using ARIMA-based traffic forecasting for Mobile WiMAX," *Comput Commun*, vol. 34, no. 1, pp. 99–106, Jan. 2011, doi: 10.1016/J.COMCOM.2010.08.008.

[27] Z. Li, Z. Zheng, and S. Washington, "Short-Term Traffic Flow Forecasting: A Component-Wise Gradient Boosting Approach with Hierarchical Reconciliation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 12, pp. 5060–5072, Dec. 2020, doi: 10.1109/TITS.2019.2948381.

[28] J. Niyogisubizo *et al.*, "Predicting traffic crash severity using hybrid of balanced bagging classification and light gradient boosting machine," *Intelligent Data Analysis*, vol. 27, no. 1, pp. 79–101, Jan. 2023, doi: 10.3233/IDA-216398.

[29] B. Khelfa, I. Ba, and A. Tordeux, "Predicting highway lane-changing maneuvers: A benchmark analysis of machine and ensemble learning algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 612, p. 128471, Feb. 2023, doi: 10.1016/J.PHYSA.2023.128471.

[30] H. Lu, D. Huang, Y. Song, D. Jiang, T. Zhou, and J. Qin, "ST-TrafficNet: A Spatial-Temporal Deep Learning Network for Traffic Forecasting," *Electronics 2020, Vol. 9, Page 1474*, vol. 9, no. 9, p. 1474, Sep. 2020, doi: 10.3390/ELECTRONICS9091474.

[31] P. Fernandes *et al.*, "Impacts of roundabouts in suburban areas on congestion-specific vehicle speed profiles, pollutant, and noise emissions: An empirical analysis," *Sustain Cities Soc*, vol. 62, p. 102386, Nov. 2020, doi: 10.1016/J.SCS.2020.102386.

[32] W. Wang *et al.*, "An interpretable model for short-term traffic flow prediction," *Math Comput Simul*, vol. 171, pp. 264–278, May 2020, doi: 10.1016/J.MATCOM.2019.12.013.

[33] S. Zhou, C. Wei, C. Song, X. Pan, W. Chang, and L. Yang, "Short-Term Traffic Flow Prediction of the Smart City Using 5G Internet of Vehicles Based on Edge Computing," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2022, doi: 10.1109/TITS.2022.3147845.

[34] J. Kim, J. Park, K. Kim, and M. Kim, "RnR-SMART: Resilient smart city evacuation plan based on road network reconfiguration in outbreak response," *Sustain Cities Soc*, vol. 75, Dec. 2021, doi: 10.1016/j.scs.2021.103386.

[35] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Neural network architecture based on gradient boosting for IoT traffic prediction," *Future Generation Computer Systems*, vol. 100, pp. 656–673, Nov. 2019, doi: 10.1016/j.future.2019.05.060.

[36] T. Zhou *et al.*, "δ-agree AdaBoost stacked autoencoder for short-term traffic flow forecasting," *Neurocomputing*, vol. 247, pp. 31–38, Jul. 2017, doi: 10.1016/J.NEUCOM.2017.03.049.

[37] M. Bawaneh and V. Simon, "Novel traffic congestion detection algorithms for smart city applications," *Concurr Comput*, vol. 35, no. 5, p. e7563, Feb. 2023, doi: 10.1002/CPE.7563.

[38] U. Mittal, P. Chawla, and R. Tiwari, "EnsembleNet: a hybrid approach for vehicle detection and estimation of traffic density based on faster R-CNN and YOLO models," *Neural Computing and Applications 2022 35:6*, vol. 35, no. 6, pp. 4755–4774, Oct. 2022, doi: 10.1007/S00521-022-07940-9.

[39] "CityPulse Smart City Datasets - Datasets." Accessed: Feb. 27, 2022. [Online]. Available: http://iot.ee.surrey.ac.uk:8080/datasets.html

[40] J. Zenkert, M. Dornhofer, C. Weber, C. Ngoukam, and M. Fathi, "Big data analytics in smart mobility: Modeling and analysis of the Aarhus smart city dataset," *Proceedings - 2018 IEEE Industrial Cyber-Physical Systems, ICPS 2018*, pp. 363–368, Jun. 2018, doi: 10.1109/ICPHYS.2018.8387685.

[41] A. R. Honarvar and A. Sami, "Multi-source dataset for urban computing in a Smart City," *Data Brief*, vol. 22, pp. 222–226, Feb. 2019, doi: 10.1016/J.DIB.2018.09.113.

[42] T. Chai and R. Draxler, "Arguments against avoiding RMSE in the literature," *Geoscientific model*, 2014.

[43] "Root mean square - Wikipedia." Accessed: Sep. 25, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Root_mean_square

[44] "Relative Absolute Error." Accessed: Sep. 25, 2023. [Online]. Available: https://www.gepsoft.com/gxpt4kb/Chapter10/Section2/SS15.htm

[45] "R-Squared - Definition, Interpretation, Formula, How to Calculate." Accessed: Sep. 25, 2023. [Online]. Available: https://corporatefinanceinstitute.com/resources/data-science/r-squared/

[46] S. Ameer *et al.*, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.

[47] M. I. Ali, F. Gao, and A. Mileo, "CityBench: A configurable benchmark to evaluate RSP engines using smart city datasets," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics), vol. 9367, pp. 374–389, 2015, doi: 10.1007/978-3-319-25010-6_25/COVER.

Muhammad Adeel Anjum completed his B.Sc. degree in Computer Science from the University of the Punjab, Pakistan, in 2017. He secured an M.Sc. degree in Computer Science from the University of Agriculture Faisalabad, Pakistan, in 2019. Adeel further completed his MS degree in Computer Science from COMSATS University Islamabad, in 2023. His research interests include Machine Learning and Artificial Intelligence. He authored his research thesis on “Smart Urban Planning: An Intelligent Framework to Predict Traffic Using Stack Ensembling Approach”.

Ahmed Alenezi is a lecturer at Taibah University, Madinah. He earned his MSc degree in Computer Security and Forensics from the University of Bedfordshire, England, in 2011. Following that, he completed his Doctor of Philosophy (PhD) in Computer Science in 2016 at the University of Bedfordshire, England.

Detecting Malicious Botnets in IoT Networks Using Machine Learning Techniques

Asghar, Muhammad Nabeel; Raza, Muhammad Asif; Murad, Zara; and Alyahya, Ahmed

Abstract: The widespread use of the Internet of Things (IoT) has led to a rise in botnet attacks, with the Mirai botnet being a major source of Distributed Denial of Service (DDOS) attacks. Mirai gained notoriety for its involvement in large-scale attacks that compromised numerous IoT devices through weak authentication credentials. Similarly, Bashlite, also known as Gafgyt or Lizkebab, targets vulnerable IoT devices by exploiting the Shellshock vulnerability in Linux-based systems. These botnets leverage compromised devices to carry out malicious activities and the propagation of malware. While Machine Learning (ML) based approaches have been proposed to identify botnets, however, detecting both Mirai and Bashlite botnets simultaneously is challenging as their attack characteristics are not very similar.

In this study, we apply ML techniques like Logistic Regression, Support Vector Machine and Random Forest to classify the malicious traffic from Mirai and Bashlite botnets. The publicly available N-BaIoT dataset is used for the training of algorithms to identify the most informative features to detect botnet traffic targeting IoT devices. The dataset contains traffic data from nine infected devices against five protocols. The employed machine learning algorithms achieved test validation accuracy above 99%, with Random Forest performing the best. Our analysis shows that devices generating combo floods share common characteristics like weight or variance calculated within a certain time window.

Index Terms: ***Internet of Things, Machine Learning, LR, SVM, RF, Botnet, TCP, UDP, Bashlite, Mirai***

1. INTRODUCTION

A botnet entails a collection of compromised devices, commonly known as bots or zombies, which are controlled by a central command-and-control(C&C) server. Botnets are commonly formed by attackers who exploit vulnerabilities in computers, servers, or IoT devices. After being compromised, these devices are enlisted into the botnet network and can be utilized to execute a range of nefarious actions, such as distributed denial of service (DDoS) attacks, spam distribution, data theft, and other malicious activities.

IoT security is crucial across various application areas due to the increase in usage of these devices in everyday life. Consider the prevalence of smart home technologies such as security cameras, door locks, and thermostats are now commonly used. Securing these devices is necessary to prevent the leakage of personal information of residents and improve their security. Similarly, IoT devices are used in healthcare systems for patient and medical staff assistance, in industries for automation processes, and smart cities for different management systems, environmental monitoring, public safety, and more. IoT devices are becoming a part of human life, and it is important to secure these devices from unauthorized access. Our research's major role is to improve the security of these devices through lightweight botnet detection mechanisms, which can detect and prevent IoT devices from becoming part of botnets.

IoT devices act as the primary origin of botnets for launching Denial of Services attacks which comprises thousands of IoT devices. The weaknesses of the Internet of Things are a core risk factor which makes these devices prone to malicious attacks. Botnet attacks pose a substantial threat to the security and integrity of IoT networks. These attacks especially the Mirai botnet, can manifest in various forms, including HTTP flooding, TCP flooding, GRE, ACK flooding, UDP flooding, DNS and UDPPLAIN etc. However, the detection of specific botnet attacks remains a challenge, as there is a lack of research exploring the network traffic characteristics that are most effective in identifying them. In a recent study [1], three classifiers were employed to classify botnet assaults on nine different devices, utilizing the comprehensive N-BaIoT dataset containing 115 features. Nevertheless, the management and monitoring of this high-dimensional dataset for network traffic classification pose inefficiencies due to capacity restrictions in storage and

increased computational costs [2] on low-powered IoT devices.

While existing literature has primarily focused on the detection of botnet attacks, there is no noticeable research proposing effective methods to detect multiple botnets. Consequently, a significant need arises for novel approaches and methodologies. Our study presents an effective strategy to identify IoT devices against multiple botnet attacks, focusing on the identification and prevention of malicious network traffic from Bashlite and Mirai attacks. The primary contributions of our work are as follows:

1. We train ML algorithms using the publicly available N-BaIoT dataset for detecting and classifying botnet attacks aimed at IoT devices. By employing Logistic Regression, Support Vector Machine and Random Forest techniques, we achieve a test validation accuracy above 99%, with Random Forest outperforming the other algorithms.
2. Our analysis reveals that devices generating combo floods exhibit common characteristics, such as weight or variance calculated within a specific time window. This finding aids in the detection and classification of malicious traffic associated with botnet attacks.

The rest of the paper is arranged in the following manner. Section 2 offers a comprehensive literature review on the detection and mitigation approaches for IoT-based botnets. Our proposed methodology for traffic analysis and training of machine learning algorithms is explained in Section 3 while Section 4 demonstrates the performance of the trained machine learning models in identifying malicious traffic. Finally, in Section 5, we conclude our work by emphasizing the significance of our research in developing effective strategies to ensure the security of IoT devices from botnet attacks.

2. *LITERATURE REVIEW*

the following section presents a thorough exploration of the existing relevant literature comprising detection and mitigation approaches targeting IoT devices. For botnet detection, there are mainly two approaches: Network-based detection and Host-based detection. The primary approach is deployed on a central server responsible for processing all network traffic, for instance gateway or router, while the secondary approach is employed using individual or singular devices such as PCs and smartphones. Host-based detection approaches are commonly utilized on devices with sufficient storage capacity to accommodate these detection algorithms. However, taking into account the limited resources of the Internet of Things, storing such resource-intensive detection algorithms becomes challenging. As a result, a majority of the previous botnet detection methodologies in the literature focus on network-based detection algorithms. There are specific signature-based approaches, such as [3], that aim to identify IoT botnets. Although these approaches are primarily network-based, they incorporate Mirai signatures into the detection algorithm to enhance its effectiveness. Additionally, approaches like [4] focus on detecting malicious domains within an IoT network to prevent malicious activities originating from specific domains. In the paper [5], a novel graph-based approach is proposed to identify botnets which can operate across diverse device architectures. The methodology involves analyzing elf files, containing both benign and malicious traffic, to identify the botnet lifecycle. The proposed approach involves creating function call graphs and printable string information (PSI) graphs from botnet lifecycle-related functions and using a convolutional neural network to classify benign and malicious samples with an accuracy of over 95%.

Paper [6] proposes a technique to detect Mirai botnet and its variants on IoT devices using power consumption patterns (PCP). The methodology involves collecting PCP of each device during different stages and training a CNN model on the pre-processed dataset. The model achieved 90% accuracy, but the proposed methodology faces several implementation challenges, such as expensive power consumption tools and a lack of standard datasets of (PCP). In the research paper [3], a network-centric methodology is proposed by authors to detect bots within IoT networks during the scanning phase. They examined the signatures of the Mirai malware and selected the port scanning signature to detect bots. After detection, the authors suggested blocking the traffic from bots or limiting the communication of detected bots to mitigate botnet attacks. The authors in [7] present a method for detecting and isolating vulnerable devices in an IoT network to prevent malware infection and their subsequent participation in botnets. The technique involves examining open ports associated with Telnet, SSH, and HTTP on IoT devices, followed by configuring firewall rules to disconnect vulnerable devices from the internet. In [8], the authors propose a policy descriptor approach to detect abnormal behaviour of IoT devices by comparing their current policies with their previously stored original policies related to access, usage, and communication. The proposed approach provides an effective method for detecting and preventing the escalation of the Mirai malware in IoT-based networks. Similarly, in [9], a deep learning approach is proposed for detecting botnets using network flow data. The method involves capturing network traffic flows and converting them to connection records, which are then used to train classifiers to differentiate between malicious and

legitimate traffic. The proposed technique is demonstrated to be effective in accurately detecting botnet traffic.

The authors of paper [10 propose a method employing deep learning to detect botnets, the methodology uses LAE and BLSTM techniques where LAE is utilized for reducing the number of attributes, while BLSTM analyses the enduring time-frame interrelated changes in the set of attributes created by LAE for attack categorization. The dataset used by the authors is BOT-IoT and achieved almost 92% data size reduction rate using LAE. In the paper [11], the authors propose an RNN-based technique that uses BLSTM at the packet level to detect botnets. The authors collected their data by performing a Mirai attack on IoT devices in a sandbox environment. The data was then classified using the BLSTM-RNN neural network. The authors of [12] propose a machine learning method that uses an artificial neural network to determine data accuracy. They utilized the N-BaIoT dataset collected from nine different devices but only used data from one device in their study. The model's accuracy was determined to be 92%. In [13], the authors presented a novel feature selection methodology called corrauc, which evaluates correlation characteristics and calculates the area under the ROC curve. The approach comprises four stages: selecting features of satisfactory information, applying feature selection algorithms, validating selected features, and evaluating the employing four distinct ML algorithms using the BOT-IoT dataset, which led to achieving an accuracy of 96%.

In the paper [14], the authors proposed a binary and multiclass classification approach to distinguish between normal and malicious IoT traffic. Two feature selection methods were used, and three machine learning algorithms were employed for classification. The results showed that all classifiers performed well, achieving high accuracy in binary and multiclass classification. Finally, in [15], the methodology proposed by the authors is for detecting outliers in green IoT devices using the density-based clustering algorithm DBSCAN to cluster network traffic based on density. Low-density clusters were labelled as malicious traffic and high-density clusters as normal traffic. Three machine learning algorithms were employed to further classify the labelled clusters, achieving over 90% accuracy for each attack. This approach can aid in identifying and mitigating malicious traffic in green IoT devices.

In the article [16], a combination of BO-GP and DT machine learning algorithms was proposed to detect malicious network traffic using the BOT-IoT dataset. The authors used the min-max method for normalization and the SMOTE algorithm to address the unbalanced dataset problem. However, they did not use any feature selection technique and optimal parametric refinement of the model to maximize detection performance using the Bayesian Optimization Gaussian Process. The accuracy of the model was 99%, but the model might be complex for IoT due to the absence of a feature selection approach. The authors did not evaluate the model using the entire dataset. The study in the article [4] suggests an unsupervised machine learning algorithm-based domain name detection technique to classify normal and malicious domains. The authors collected normal and malicious domains and extracted 204 variables. However, they reduced the features to 20 during preprocessing. They evaluated nine different algorithms and found that four of them achieved 99% accuracy, including ANN, DBSCAN, GMM, Hierarchical Clustering, LAC, Mini Batch K-Means, AP, and K-medians. The study concludes that unsupervised machine-learning techniques can effectively classify normal and malicious domains.

The research paper in [17] compares the performance of supervised machine learning and deep learning algorithms on small and large unbalanced datasets. The evaluation criteria used were probability of detection, accuracy, and likelihood of false alarms. The study utilized Multi-Class Decision Forest and also Multi-Class Neural Network algorithms on a dataset of small size, which did not perform well. However, on the larger dataset, the performance improved except for the rate of false positives at 0.3%. The research findings conclude that the accuracy of algorithms is correlated with the scale of a dataset and class imbalance, and further exploration is required to improve their efficacy when dealing with small datasets. The research paper is primarily concerned with the use of supervised ML models for classifying botnet traffic. The authors used the BOT-IoT dataset and applied chi2 as a feature selection method. They used three supervised ML models, Multi-Layer Perceptron Artificial Neural Network (MLP ANN), K-Nearest Neighbours (KNN) and Gaussian Naive Bayes (GNB). The study shows that KNN outperformed the other two algorithms regarding accuracy. The research paper in [18] presented a framework for automatically detecting and mitigating anomalies in IoT networks based on MUD policies. The approach uses the MUD maker to create a MUD profile for each IoT device, and the MUD controller grants network access based on the policies specified in the MUD profile. The approach detects anomalies and notifies device owners in real time via an SMTP server. The study emphasizes the importance of using MUD policies for securing IoT networks and provides a useful tool for real-time network traffic monitoring via a graphical user interface.

Table 1. N_BIoT Categorization

IoT Devices	Malware	Attacks
Device 1: Damini_Doorbell,	Bashlite	Scan, TCP combo, UDP, Junk
Device 2: Ennio_Doorbell,		
Device 3: Provision_PT_737E_Security_Camera,		
Device 4: Provision_PT_838_Security_Camera,		
Device 5: Samsung_SNH_1011_N_Webcam,		
Device 6: SimpleHome_XCS7_1002_WHT_Security _Camera,	Mirai	Scan, Ack, UDP plain, Syn, UDP
Device 7: SimpleHome_XCS7_1003_WHT_Secur-ity_Camera,		
Device 8: Philips_B120N10_Baby_Monitor,		
Device 9: Ecobee_Thermostat		

The paper [19] demonstrates a botnet vs botnet approach using the BDS Botnet Defense System consisting of four stages: first is the monitor module second is the strategy planner third is the worm launcher, and last C&Cs. The authors proposed the Few-Elite launch strategy, which deploys white-hat worms based on the worm's life cycle and network structure density, using the Petri Net 2 simulator. In [20], the focus is on proposing a vulnerability scanner based on Mirai to secure IoT devices by identifying vulnerabilities and generating reports for network administrators or home users. The tool executes attacks similar to Mirai infection against IoT devices to enhance the security of IoT networks. The paper [22] introduces the Hybrid Strawberry African Buffalo Optimizer (HSABO) algorithm to detect botnet attacks which target IoT devices by leveraging the collective intelligence of African buffalos and strawberry plants to enhance the security of these devices. In [21], a whitelisting-based strategy is proposed to mitigate Mirai botnet threats in IoT networks by computing hash codes of each device in the network and enforcing whitelisting using the profiling module and application monitor. The approach ensures that only trusted applications run on the devices in the network, mitigating the risks posed by Mirai botnet attacks.

On the other hand, limited work has been done in the direction of preventing IoT devices from becoming part of botnet. Also, previous research has shown limited solutions for removing malicious activity from compromised IoT devices, with one method being botnet-vs-botnet, which requires significant time, battery power, and storage. The alternative solution of installing IDS on each device presents a drawback in terms of storage utilization proving unsuitable for IoT devices due to their restricted storage capacity and battery life. Another approach is to prevent devices from becoming bots by addressing three primary vulnerabilities: all-time online availability, open ports, and weak credentials. However, these measures are not entirely effective due to user awareness issues and Mirai's ability to compromise devices through brute force attacks.

In the following sections, we employ ML models to identify malicious traffic in the network and subsequently remove the bot devices generating such traffic. Botnet Identification using ML Techniques.

3. METHODOLOGY

The following section presents a methodology for the detection and prevention of botnets in IoT networks using a machine-learning approach. The general machine-learning process is shown in Figure 1. This proposed approach employs ML techniques to analyse network traffic and detect patterns that are indicative of botnet activity. The detected botnets are then removed from the network. The machine learning algorithms are trained on publicly available N-BaIoT dataset. The dataset was introduced in [23] and is publicly accessible through the ML Repository of UCI. The main reason for selecting the N-BaIoT dataset is its detailed labelled classification of attacks. In this dataset, both Mirai and Bashlite affected instances are further labelled with their specific attacks, aiding in the identification of patterns of traffic source attacks The dataset originated from the collection of real-time traffic data obtained from nine commercial Internet of Things (IoT) devices, consisting of four security cameras, one webcam, two doorbells, and one thermostat. Each device was deliberately infected with two types of malware, namely Mirai and BASHLITE, and subjected to various types of attacks, as mentioned in Table 1.

The regular network traffic was recorded after the installation of each IoT device to ensure the exclusion of the malicious data samples from the training part of the dataset.

The dataset comprises 23 distinct features, each sampled across five temporal windows: 1 min, 10 sec, 1.5 sec, 500 ms and 100 ms, resulting in generating a dataset with 115 features. Moreover, several statistics for each packet, such as the

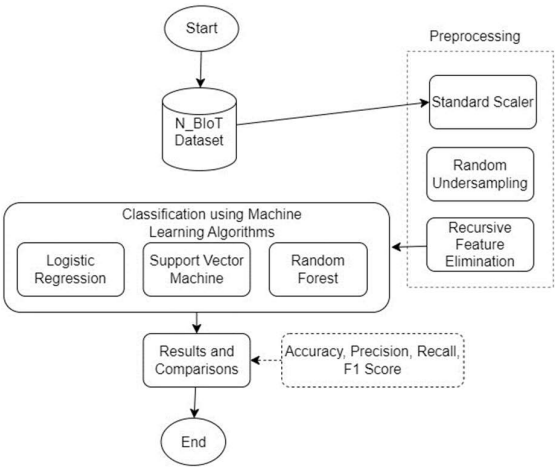


Figure 1: Methodology for Botnet Detection

variance, standard deviation and mean of the packet size, as well as the covariance, Pearson correlation coefficient, radius and magnitude of two stream means. are also computed.

3.1 Data Preprocessing

In the context of machine learning the preprocessing of data is a pivotal stage, as it enables the preparation of data before being fed to the machine learning algorithm to ensure optimal performance Data preparation consists of various phases such as including data transformation to numerical format, elimination of null and duplicate samples, balancing dataset and standardization of dataset values. For instance, within the context of the N-BaIoT dataset, standardization of the dataset as well as class balancing is required due to the high rate of attack samples compared to their corresponding benign data samples. Therefore, it is imperative to select and execute appropriate preprocessing procedures to ensure the dataset is correctly prepared for the machine learning algorithm, leading to accurate and reliable results.

3.2 Data Standardization

To enhance the functionality of the model, the Standard Scaler technique was employed to standardize the dataset samples. This technique involves subtracting the mean value from each feature and subsequently standardizing them to unit variance through division by their respective standard deviation. This results in a transformed dataset with a standard deviation of 1 and a mean of 0 [24]. defined as;

$$z = \frac{x-u}{\sigma} \qquad (1)$$

where x denotes the original feature value, while μ and σ signify the mean value and standard deviation of the samples in the dataset. the standardization process was performed using the StandardScaler module available in the sklearn library.

3.3 Class Balancing

The dataset used for binary classification exhibited a significant class imbalance, featuring a significantly large number of samples representing attack class as compared to the benign class data samples. Such an imbalance can lead to bias towards a particular class and consequently, overfitting. Additionally, feature selection methods can also favor features exhibiting a strong correlation with the predominant class. All of these issues are known to stem from unbalanced data. To mitigate these concerns, we balanced the dataset by ensuring that an equal number of samples were present for both the attack and benign classes for each device. This was achieved through the use of the RandomUnderSampler module from the imblearn library, which randomly under-samples the majority class to match the minority class size.

3.4 Feature Selection

Dimensionality reduction is an effective preprocessing step in machine learning that helps eliminate irrelevant and redundant data, which, in turn, enhances the accuracy of the learning process and improves the comprehensibility of results [25]. In our study, we performed feature selection by selecting the top 10 features from the combined attack and benign datasets. There are three main methods for feature selection, including wrapper, filter, and embedded methods. These methods facilitate the identification of important features by evaluating their effects on model efficacy.

3.5 Recursive Feature Elimination

Our strategy incorporated a Wrapper method, Recursive Feature Elimination, and Logistic Regression to select the best feature subset for binary classification. This method, as illustrated in Figure 2, takes all dataset features as input, where a logistic regression model is used as an estimator for feature importance ranking with each iteration, a fixed number (k = 20) of the least important features are removed from the dataset, and the logistic regression model is again trained on the reduced feature set. The process was repeated until the model reached the desired number of features K which is 10. We reduced the dimensionality of our dataset from 115 features to 10 optimal features.

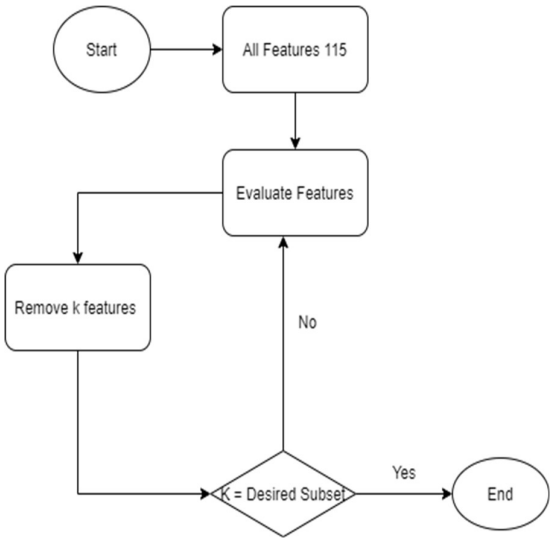


Figure 2 Recursive Feature Examination

3.6 Classification

To detect malicious traffic, we employed binary classification methodology using classifiers: using classifiers: LR, SVM and RF. These classifiers were trained on the selected features to classify network traffic as benign or malicious. By using LR, we aimed to leverage its ability to model non-linear relationships and make accurate predictions based on the selected features. We also used SVM aiming to accurately classify benign and malicious traffic by finding the best decision boundary in high-dimensional data. Finally, Random Forest (RF) is selected for regression and classification tasks when in-depth analysis of data is needed. It consists of multiple decision trees, and the final output is obtained by prediction of all these decision trees combined to obtain the final output. The accuracy and performance of RF are influenced by the number of decision trees.

3.7 Evaluation Metrics

In our study, we have employed a total of five evaluation metrics to assess the performance of trained models. At the forefront of evaluation methods lies the confusion matrix which offers four parameters (TP, TN, FP, FN) that are used for the evaluation of the model. For instance, when the predicted outcome denotes malicious traffic and the actual target sample is indeed malicious then it's called True Positive or TP conversely, it is False Negative (FN) if the actual target sample indicates benign class. Similar to TP if predicted and target results show a benign class then it is True Negative (TN) since both results show it's not malicious but if the target class is malicious in reality, then it is False Positive (FP). All the other remaining metrics are derived based on the calculations of these four parameters. Although accuracy measures the classifiers' accuracy, its assessment alone is insufficient. Thus, we complement it with f1-score, recall and precision as well.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \qquad (2)$$

precision serves as a metric to assess the reliability of true positives by determining the proportion of accurately classified positive results. it is beneficial in contexts where the frequency of false positives escalates.

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

Recall is used for assessing the true positive rate by representing the frequency with which the model correctly identifies instances of a given class when they are indeed true.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

F1-score serves as a comprehensive metric to measure the accuracy of the model by combining both precision and recall.

$$f1 - score = \frac{2(precision)(recall)}{precision+recall} \qquad (5)$$

4. Selected Machine Learning Models for Identifying Malicious Traffic

The methodology of this study focuses on utilizing machine learning models, specifically Logistic Regression, Support Vector Machine and Random Forest., to classify malicious traffic in IoT networks. The dataset, containing traffic data from nine infected devices across five protocols, is employed for training and evaluating these models. The aim is to identify informative features that can effectively identify botnet attacks on IoT devices.

4.1. Random Forest

Random Forest is an ensemble learning technique, whereby numerous decision trees are generated during the training phase and the final prediction is determined by the most frequently occurring class among the predictions made by individual trees. it is acknowledged for its adeptness to handle large datasets, and to capture complex relationships between features.

4.2. Support Vector Machine

As a supervised learning technique, the support vector machine (SVM) is a model that analyzes data and builds a hyperplane to separate different classes. it maps data into feature space of high dimensionality and uses a decision boundary to classify new instances. Linearly separable and non-linearly separable, SVM is efficient in managing data of both forms.

4.3. Logistic Regression

Logistic regression stands as a statistical model applied for estimating the likelihood of a binary outcome using input variables. it uses a logistic function by estimating the probabilities to model the relationship between a dependent variable or one or multiple independent variables. LR is commonly utilized for binary classification problems. The dataset is analyzed to identify

features that capture common characteristics of malicious traffic, such as weight or variance calculated within a specific time window. these attributes are utilized for the training process of the selected ML models with labelled data, where instances of malicious traffic are appropriately labelled. Model performance is assessed based on measures like test validation accuracy. The goal is to assess the effectiveness of Random Forest, SVM, and Logistic Regression in accurately classifying malicious traffic in IoT devices. By applying these machine-learning techniques and analyzing the results, the primary goal of this study is to contribute to the advancement of strategies for the detection and mitigation of botnet attacks. The methodology provides a framework for training models on malicious traffic and evaluating their classification performance.

5. *Results and Evaluation*

the following section is dedicated to discussing the results derived from the application of three classifiers, namely SVM, LR, and RF, for the classification of benign and botnet traffic. From the dataset consisting of 115 features, we carefully selected the ten most informative features for each attack type. All selected features concerning attacks are represented with their serial number 1 to 115. Within this section, there is an extensive analysis of these selected features and the results achieved by the applied classifiers, evaluating their performance for each attack type on individual devices is also discussed. the evaluation metrics used for assessing the classifiers include Recall score, Precision, Confusion metrics, Accuracy and F1-score. The file of the dataset containing benign data was merged and balanced alongside files containing malicious data for each attack in a distinct file and afterwards, feature selection was performed with the help of wrapper methods. in the last binary classification was performed using the features obtained from wrapper methods. Tables 2 to 10 illustrate the results derived from classification models, encompassing data from nine distinct IoT devices, categorizing results based on individual attack vectors. Metrics evaluation is conducted independently for every distinct attack traffic. For instance, in the context of combo attack analysis, three ML algorithms were used for classification, and the resultant evaluation metrics for each algorithm concerning the combo attack are presented for each distinct IoT device result as shown in figures 3 to 11.

5.1 Results of Device 1 Dataset

Table 2 Device 1 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
g_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_scan	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
g_tcp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%
g_udp	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_ack	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_scan	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_syn	LR	99%	99%	100%	99%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp plain	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%

Figure 3 Device 1 Dataset Classification Accuracy

Figure 3 shows the comparison between the accuracy of all selected algorithms against all types of attacks. The accuracy ranges between 99% and 100%, with Random Forest (RF) achieving the highest accuracy and Logistic Regression (LR) achieving the lowest.

5.2 Results of Device 2 Dataset

Table 3 Device 2 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%

	SVM	99%	100%	99%	99%
g_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_scan	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_tcp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_udp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%

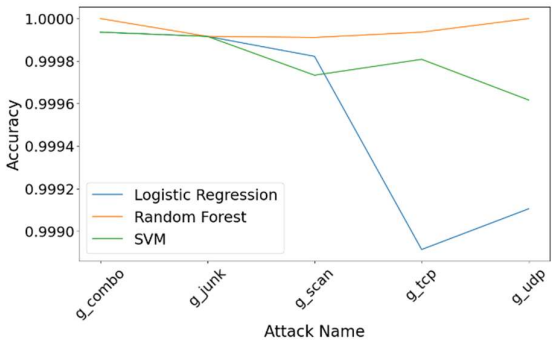


Figure 4 Device 2 Dataset Classification Accuracy

5.3 Results of Device 3 Dataset

Table 4 Device 3 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%
g_scan	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
g_tcp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%
g_udp	LR	99%	99%	100%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
m_ack	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_scan	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_syn	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_udp	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_udp plain	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%

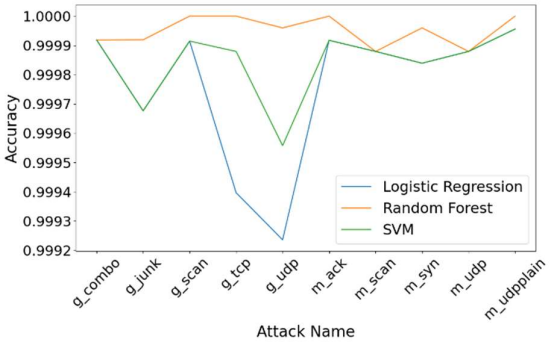


Figure 5 Device 3 Dataset Classification Accuracy

5.4 Results of Device 4 Dataset

Table 5 Device 4 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_scan	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%`
g_tcp	LR	99%	99%	99%	99%
	RF	99%	99%	99%	99%
	SVM	99%	99%	99%	99%
g_udp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%
m_ack	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_scan	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_syn	LR	100%	100%	100%	100%
	RF	99%	100%	99%	99%
	SVM	100%	100%	100%	100%
m_udp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%
m_udp plain	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%

	SVM	100%	100%	100%	100%

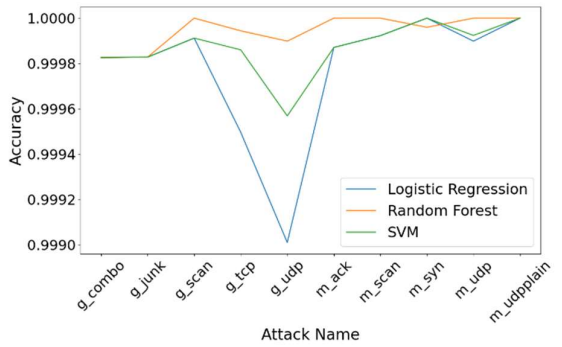


Figure 6 Device 4 Dataset Classification Accuracy

5.5 Results of Device 5 Dataset

Table 6 Device 5 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	99%	99%	100%	99%
	SVM	99%	100%	99%	99%
m_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
m_scan	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_tcp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%
m_udp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%

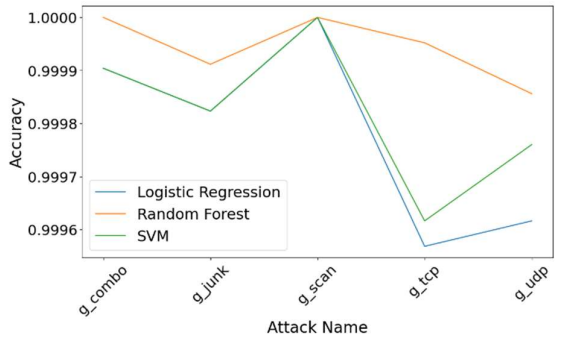


Figure 7 Device 5 Dataset Classification Accuracy

5.6 Results of Device 6 Dataset

Table 7 Device 6 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
g_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_scan	LR	99%	100%	99%	99%
	RF	99%	100%	98%	99%
	SVM	99%	100%	99%	99%
g_tcp	LR	99%	99%	99%	99%
	RF	99%	99%	100%	99%
	SVM	99%	99%	99%	99%
g_udp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%
m_ack	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_scan	LR	99%	99%	100%	99%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_syn	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp plain	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%

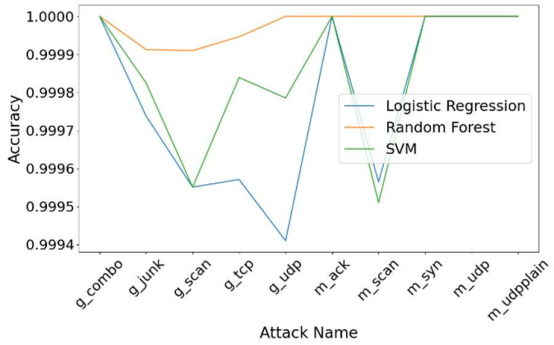


Figure 8 Device 6 Dataset Classification Accuracy

5.7 Results of Device 7 Dataset

Table 8 Device 7 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
g_junk	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
g_scan	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_tcp	LR	99%	99%	99%	99%

	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%
g_udp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%
g_ack	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_scan	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	99%	99%	100%	99%
m_syn	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
m_udp	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp plain	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%

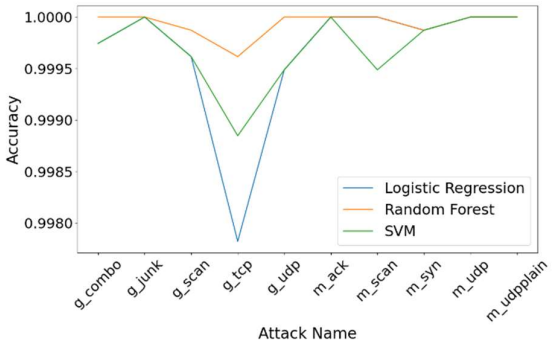


Figure 9 Device 7 Dataset Classification Accuracy

5.8 Results of Device 8 Dataset

Table 9 Device 8 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
g_junk	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_scan	LR	99%	99%	99%	99%
	RF	99%	99%	99%	99%
	SVM	99%	100%	99%	99%
g_tcp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	100%	99%
g_udp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%
m_ack	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_scan	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
m_syn	LR	99%	100%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	100%	99%	99%
m_udp	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
m_udp plain	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%

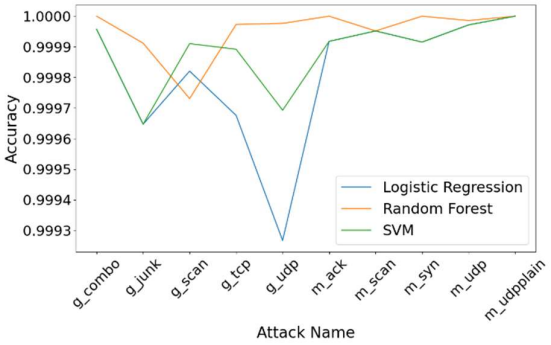


Figure 10 Device 8 Dataset Classification Accuracy

5.9 Results of Device 9 Dataset

Table 10 Device 9 Classification Results

Attack	Algorithm	Accuracy	Precision	Recall	F1 Score
g_combo	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%
g_junk	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
g_scan	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
g_Tcp	LR	99%	99%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	99%	99%	99%
g_udp	LR	99%	99%	99%	99%
	RF	100%	100%	100%	100%
	SVM	99%	99%	99%	99%
m_ack	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_scan	LR	99%	100%	99%	99%
	RF	99%	99%	99%	99%
	SVM	99%	100%	99%	99%
m_syn	LR	100%	100%	100%	100%

	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp	LR	100%	100%	100%	100%
	RF	100%	100%	100%	100%
	SVM	100%	100%	100%	100%
m_udp plain	LR	99%	100%	99%	99%
	RF	99%	100%	99%	99%
	SVM	99%	100%	99%	99%

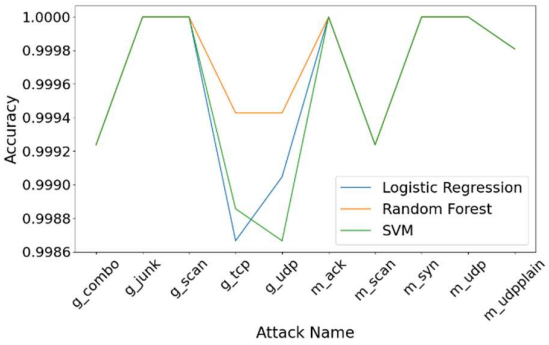


Figure 11 Device 9 Dataset Classification Accuracy

6. CONCLUSION

In this research, we aimed to achieve two objectives, namely dimensionality reduction of the N-BaIoT dataset for quick yet efficient malicious traffic identification and evaluation of the performance of various ML algorithms for malicious traffic using selected features from multiple IoT device-to-device communications. The N-BaIoT dataset contained network traffic originating from nine distinct IoT devices infected by both Mirai and Bashlite's ten attack types and normal traffic data. We reduced the dataset's feature set from 115 to 10 and conducted binary classification using Logistic Regression, Support Vector Machine, and Random Forest algorithms. Our results showed that the selected subset of features achieved above 99% accuracy in detecting botnet traffic.

Furthermore, we analyzed the characteristics of selected features for each attack protocol from all 9 devices to determine if the botnets share similar characteristics. Our findings revealed that botnets with the same attack protocol generated by multiple IoT devices infected with the same botnet had nearly identical characteristics. However, if certain devices are infected with different botnets, the characteristics will not be the same.

REFERENCES

[1] S. Bagui, X. Wang, and S. Bagui, "Machine Learning Based Intrusion Detection for IoT Botnet," International Journal of Machine Learning and Computing, vol. 11, no. 6, pp. 399–406, Nov. 2021, doi: 10.18178/ijmlc.2021.11.6.1068.

[2] T. Huang, H. Sethu, and N. Kandasamy, "A New Approach to Dimensionality Reduction for Anomaly Detection in Data Traffic," IEEE Transactions on Network and Service Management 13.3 (2016): 651-665.

[3] Kumar A and TJ Lim, "Early detection of Mirai-like IoT bots in large-scale networks through sub-sampled packet traffic analysis," Conference, - Future of Information and Communication, 2019.

[4] Rezaei and Amirhossein, "Detecting botnet on IoT by using unsupervised learning techniques," International Journal of Computer Science and Information Security (IJCSIS) 18.4 (2020).

[5] H. T. Nguyen, Q. D. Ngo, and V. H. Le, "A novel graph-based approach for IoT botnet detection," International Journal of Information Security, vol. 19, no. 5, pp. 567–577, Oct. 2020, doi: 10.1007/S10207-019-00475-6.

[6] W. Jung, H. Zhao, M. Sun, G. Z.-S. Health, and undefined 2020, "IoT botnet detection via power consumption modeling," Smart Health 15 (2020): 100103.

[7] Dietz et al., "IoT-botnet detection and isolation by access routers," 2018 9th International Conference on the Network of the Future (NOF). IEEE, 2018.

[8] Sajjad, Syed Muhammad, and Muhammad Yousaf, "UCAM: usage, communication and access monitoring based detection system for IoT botnets," 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2018.

[9] Sriram, S.Vinayakumar R., Alazab M., and Soman K. P, "Network flow based IoT botnet attack detection using deep learning," 2020.

[10] S. I. Popoola, Bamidele Adebisi, Mohammad Hammoudeh, Guan Gui, and Haris Gacanin, "Hybrid Deep Learning for Botnet Attack Detection in the Internet-of-Things Networks," IEEE Internet of Things Journal 8, 2020.

[11] Christopher D. McDermott, Farzan Majdani, and Andrei V. Petrovski, "Botnet detection in the Internet of things using deep learning approaches," ieeexplore.ieee.org, 2018.

[12] T. G. Palla and S. Tayeb, "Intelligent Mirai Malware Detection in IoT Devices," in 2021 IEEE World AI IoT Congress (AIIoT), Jun. 2021, pp. 0420–0426. doi: 10.1109/aiiot52608.2021.9454215.

[13] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "CorrAUC: A Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine-Learning Techniques," IEEE Internet of Things Journal, vol. 8, no. 5, pp. 3242–3254, Mar. 2021, doi: 10.1109/JIOT.2020.3002255.

[14] Z. Alothman, M. Alkasassbeh, and S. Al-Haj Baddar, "An efficient approach to detect IoT botnet attacks using machine learning," Journal of High Speed Networks, vol. 26, no. 3, pp. 241–254, 2020, doi: 10.3233/JHS-200641.

[15] M. Shobana and S. Poonkuzhali, "An Efficient botnet Detection Approach for Green IoT Devices using Machine learning Techniques," 2020.

[16] M. Injadat, A. Moubayed, and A. Shami, "Detecting Botnet Attacks in IoT Environments: An Optimized Machine Learning Approach," arXiv, Dec. 2020.

[17] M. Hegde, G. Kepnang, M. al Mazroei, J. S. Chavis, and L. Watkins, "Identification of Botnet Activity in IoT Network Traffic Using Machine Learning," in 2020 International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2020, Oct. 2020, pp. 21–27. doi: 10.1109/IDSTA50958.2020.9264143.

[18] H. J. Hadi, S. M. Sajjad, and K. Un Nisa, "BoDMitM: Botnet detection and mitigation system for home router base on the MUD," in 2019 International Conference on Frontiers of Information Technology, FIT 2019, Dec. 2019, pp. 139–143. doi: 10.1109/FIT47737.2019.00035.

[19] S. Yamaguchi, "Botnet defence system: Concept, design, and basic strategy," Information, vol. 11, no. 11, pp. 1–15, Nov. 2020, doi: 10.3390/info11110516.

[20] S. Vysakh and P. K. Binu, "IoT-based Mirai vulnerability scanner prototype," in Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Aug. 2020, pp. 97–101. doi: 10.1109/ICSSIT48917.2020.9214099.

[21] Tatikayala Sai Gopal, Mallesh Meerolla, Jyostna G, Reddy Lakshmi Eswari P, and E Magesh, "2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)," 2018.

[22] M. G. Karthik and M. B. M. Krishnan, “Securing an Internet of Things from Distributed Denial of Service and Mirai Botnet Attacks Using a Novel Hybrid Detection and Mitigation Mechanism,” International Journal of Intelligent Engineering and Systems, vol. 14, no. 1, pp. 113–123, 2021, doi: 10.22266/IJIES202100%228.12.
[23] Y. Meidan et al., “N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders,” IEEE Pervasive Computing, vol. 17, no. 3, pp. 12–22, Jul. 2018, doi: 10.1109/MPRV.2018.03367731.
[24] “Scale, Standardize, or Normalize with Scikit-Learn | by Jeff Hale | Towards Data Science.” https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02 (accessed Sep. 18, 2021).
[25] “Cloudflare says it stopped the largest DDoS attack ever reported | ZDNet.” https://www.zdnet.com/article/cloudflare-says-it-stopped-the-largest-ddos-attack-ever-rep

Muhammad Nabeel Asghar received the Ph.D. degree from the University of Bedfordshire, U.K. He is currently an Associate Professor in the College of Computer Science and Information Technology, King Faisal University, in Hofuf, Saudi Arabia. He had worked as an Assistant Professor with the Department of Computer Science, Bahauddin Zakariya University, Pakistan. His recent work is concerned with multimedia, incorporating text, audio, and visual processing into one dynamic novel frame work. Email: Nabeel.Asghar@bzu.edu.pk, masghar@kfu.edu.sa

Muhammad Asif earned his PhD in Computer Science and Electrical Engineering from Gwangju Institute of Science and Technology in South Korea. His research focuses on machine learning, deep learning, natural language processing, and the Internet of Things (IoT). Dr. Asif is dedicated to pushing the boundaries of technology and contributing to advancements in these critical areas of technology. Email: asif79@bzu.edu.pk

Zara Murad holds an MS in Computer Science from Bahauddin Zakariya University Multan. Having research interests in Machine learning and IoT security, Zara Murad is dedicated to advancing research in these critical areas of technology. Email: ms.zmurad@gmail.com

Ahmed Alyahya, PhD, is an assistant professor in the College of Computer Science and Information Technology, King Faisal University, in Hofuf, Saudi Arabia. His areas of interest include information security and privacy, penetration testing, web and network security, information security risk management and social engineering. Email: aaalyahya@kfu.edu.sa

Enhancing Security of Text Using Affine Cipher and Image Cryptography

Mehmoona, Jabeen[1] and Carsten, Maple[2]

Air University, Islamabad Pakistan[1]

University of Warwick, UK[2f]

mehmoona.jabeen@au.edu.pk, cm@warwick.ac.uk

Abstract: In the contemporary digital landscape, the escalating reliance on diverse social media platforms for textual communication necessitates the establishment of secure and trustworthy channels to thwart the threats of theft or hacking. Most of these messages contain highly confidential data, underscoring the critical need for robust security measures, primarily through the deployment of encryption techniques. While existing algorithms predominantly employ text-to-text encryption (TOTET) methods, this paper introduces an innovative hybrid approach that amalgamates TOTET with text-to-image encryption (TOIET), thereby fortifying the privacy of transmitted messages. This novel method undergoes rigorous testing using various parameters, including privacy levels and encryption time, to evaluate its effectiveness. Comparative analyses are conducted against established techniques such as DES, 3DES, and AES. The experimental results conclusively demonstrate the superior privacy capabilities of the proposed scheme in comparison to its predecessors. This advancement in encryption technology not only bolsters the security of confidential messages but also positions itself as a noteworthy improvement over existing methods, marking a pivotal step towards ensuring the integrity of digital communications.

Index Terms: cryptography, encryption, decryption, confidentiality

1. *Introduction*

Information security has become a serious issue due to the dependence on digital devices. Because it makes it easy for an intruder to get illicit access to confidential data. Different techniques, which include computational or arithmetical, have been developed by an attacker to get access to the present security system. To minimize this issue, computer security systems use the science of cryptography. Cryptography is a field that transforms plain text into cipher text [1]. The main objective of cryptography is to make it problematic for an attacker to get unauthorized access. The security of algorithms depends on these parameters: length of keys used, complexity of an algorithm and level of coding. These parameters can impact the throughput and the capability of algorithms to encrypt small text. Most of the algorithms are divided into two categories: symmetric and asymmetric algorithms [3].

In symmetric algorithms, only one key is used to encrypt and decrypt the data. Examples of symmetric algorithms are DES, 3DES, and AES. In the asymmetric algorithms, two keys are used [4]. One is public and the other one is private. In an asymmetric algorithm, the sender uses the public key for encryption and the receiver uses his own private key for decryption. These two keys are interlinked.[8]. Examples of asymmetric algorithms are ecliptic curve cryptography (ECC), and RSA.

In this research paper, symmetric cryptography will be used for encryption and decryption of data. The TOTET and TOIET will be used to enhance the security of the text. The organization of this research paper is as follows: In this research paper, Section 2 will describe the related work that has been done so far. In Section 3, we will describe the proposed methodology for text encryption and decryption to make it secure. Section 4 will evaluate the implementation strategy by performing various experiments. Section 5 will conclude the research paper.

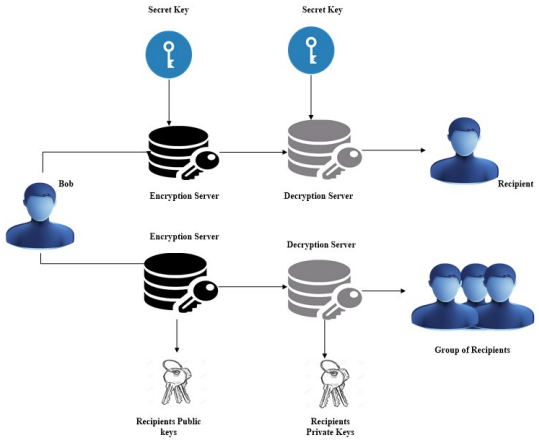


Fig. 1. A Scenario of Transferring a Message from Bob to his Friends

2. *Literature Review*

Data security is described as a group of events or actions that are used to protect sensitive information

from exterior threats such as damage or robbery. Data invasion has become a main danger to the security of internet users, and the devices that relateto it. To mitigate this risk, increase the security of the data, and protect it from surveillance. Some business owners prefer to depend on a secure network that is made secure by using special kinds of algorithms or coding. Some users do not want to transfer sensitive information over the internet. Consequently, they use storage devices or cards to store or transfer data [11]. Some of them are seeking to attain anti-virus software from authenticated sources to make information secure. Different kinds of information are transmitted through several social media platforms. This movement implies a duty on us to safeguard this information from stealing and viewing [26]. To attain this purpose, different kinds of algorithms are used to protect sensitive data. These algorithms convert simple text into cipher text. Some of them use key pairs for the encryption and decryption of data. All these techniques have some unique features. On the other hand, it is difficult to identify the best encryption technique which offers high security, takes minimum time to generate a key, and for the processes of encryption and decryption. They have different kinds of advantages, and disadvantages, and are appropriate for different applications [17]. According to the literature, different kinds of cryptographic algorithms can be the best solution to a particular problem. Some of the most prominent cryptographic techniques are as follows [8]:

2.1 Process of Encryption Using Symmetric Algorithms

The Data Encryption Standard (DES) comes under the umbrella of symmetric cryptographic algorithms. It uses one key in both for the encryption and, also, for the decryption. It has various drawbacks. It can be possible to produce the same output by selecting a taken input to an S box. Furthermore, its initial and final permutation is not clear. The only advantage is the length of the key which makes it tough to launch the brute force to decrypt the data [20]. Another example of a symmetric key algorithm is triple DES. The data is moved through the original DES algorithm three times during the encryption process. It can still be used in fiscal services and other industries. 3DES is slower 3 times than DES, low performance in terms of power consumption and throughput, but it is safer. In 3DES, the first key is used to encrypt the plaintext, the second key is used to decrypt the cipher text, and the third key is used to encrypt. Decryption

is the reverse process of the encryption procedure [18]. The Advanced Encryption Standard (AES) is made to update the original DES. Some of the most prominent uses of the AES include texting platforms such as WhatsApp. AES is a good encryption algorithm, because of its performance and the level of security it offers [14].

2.1.1 Comparison of Symmetric Cipher

In [6], in this paper, they have drawn a comparison between the symmetric and asymmetric algorithms. In the research paper, they have used different parameters: keys that have been used, throughput, encryption ratio, efficiency, length of the key, and security against attacks. However, they do not discuss many parameters to increase the encryption ratio. The performance of different algorithms can be checked by using different file sizes. Various cryptographic algorithms: DES, 3DES, Blowfish, Two-fish, and Three-fish have been checked by giving different inputs. The result showed that the Blowfish performed well as compared to other tested algorithms [26]. In [14], they have described how anyone can choose a better algorithm for encryption. For this purpose, it is mandatory to have deep knowledge of these algorithms. Someone can also check the different parameters and their ratio. The most important parameters can be encryption time, throughput, and the utilization of the CPU.

2.2 Process of Encryption Using Asymmetric Algorithms

RSA was developed in 1977. It is a block cipher, and extensively used for secure communication of data. In this algorithm, three steps are involved: Key generation, encryption, and decryption. The main drawback of RSA is that it is vulnerable to attacks such as timing attacks, and brute force attacks [18].

2.2.1 Comparison of Asymmetric Algorithms

In this paper, they have drawn a comparison among various asymmetric algorithms. It uses complex and geometric equations to produce public keys. It uses two different keys: one for encryption and one for decryption. The ElGamal algorithm was developed by Taher ElGamal. It can also be used in the generation of digital signatures which is called the ElGamal signature scheme. It is an asymmetric algorithm that is built on the Diffie Hellman key exchange algorithm. In [27], to increase the security**.** In this research paper, symmetric cryptography will be of the data, information hiding technique and encryption have been used. In this paper, the least significant algorithm (LSB) is used for information hiding, and the RSAasymmetric algorithm is used for encryption. It provides a robust level of security that

is difficult to breach.

Table 1. Comparative Analysis of the Pre-existing Studies

Ref	Algorithm name	Encryption time	Key size	Type of Cipher	Protection against attack
[1]	AES is used for designing the file security system with a key of 128 bits.	low	fixed	Block	Brute force attack
[2]	Data is encrypted using the XOR with an image of a key.	low	fixed	Block	Unbreakable
[3]	A combination of Vigenère and Hill cipher is used to mitigate the weakness of Vigenère cipher	Low	Not fixed	Block	Statistical attack
[4]	A secure algorithm has been used which consists of Xor operation and bite shifting for encryption.	Low	Not fixed	Block	Avalanche attack
[5]	Left and right circular shifts are operations used for the encryption of text.	Low	fixed	Block	Known plaintext attack
[6]	Various experiments are performed for encryption and decryption of the text using DES, and 3DES.	Low	fixed	Block	Brute force attack
[7]	Beaufort-Vigenère is used for the encryption of data.	Low	Not fixed	Slock	Brute force attack
[8]	Modified shift operation is used to scramble the data.	Low	Not fixed	Stream	Differential attack
[9]	A hybrid algorithm has been used in the technique. AES is used for encryption and SHA-256 is used to ensure the integrity of data.	Low	Fixed	Block	Brute force attack
[10]	Modified technique in which image is used image key is used for encrypting thetext after converting the text into a matrix	Low	Not fixed	Block	Brute force attack
[11]	One round variable block size (ORVBM) hasbeen used for the generation of private key for each block of data	Low	Fixed	Block	Differential attack
[12]	The substitution technique is used to improve the security of the block	Low	Fixed	Block	Known plaintext attack
[13]	Unique key generation method is used for the encryption of the data with substitution operations.	Low	Fixed	Block	Kasiski and Friedman attack
[14]	The hybrid approach of Caesar and Vigen`ere is used for the encryption and decryption of data to enhance the security of the data.	Low	Fixed	Block	Brute force attack

[15]	In this research, a hybrid approach of the encryption algorithm is used for the encryption of plain text into cipher text.	Low	Not fixed	Stream	Known plaintext attack
[16]	Asymmetric encryption is used for the encryption of text with the same keys.	Low	Fixed	Block	Brute force attack
[17]	DES and a unique key generation method is used to encrypt the plain text.	Low	Fixed	Block	Brute force attack
[18]	A combination of symmetric and asymmetric algorithms is used for encryption to enhance the security of the data.	Low	Fixed	Block	Most practical attack
[19]	Symmetric cipher AES is used for the encryption of the different blocks of the plain text to hide it from the attackers.	Low	Fixed	Block	Brute force attack
[20]	DES is used to encrypt the data.	Low	Fixed	Block	Side channel attack
[21]	The symmetric algorithm is used for the encryption of different blocks of data	High	Fixed	Block	Known plain text attack
[22]	Random numbers are used for the generation of keys and data with right and left circular shifts.	High	Not fixed	Block	Brute force attack
[23]	In this research, data is encrypted using the key and after that cipher is encrypted again with an image key.	Low	Fixed	Block	Replay attack
[24]	The symmetric encryption algorithm is used for the encryption of data.	Low	Fixed	Block	Side channel attack

2.3 Hybrid Approaches

Some papers have used a combination of symmetric, and asymmetric algorithms [9]. In [7], two approaches have been merged effectively, namely Beaufort and Vigenere. The value is better than the existing methods because the value of the avalanche effect has a stable value on the changes of keys and plain text. In this paper, a symmetric algorithm (AES), and ECC (asymmetric algorithm) have been combined with a hash function (SHA256). It is used to make sure of the integrity of the data. The suggested algorithm is more effective as compared to the other approaches. However, the suggested technique is less effective when it comes to encryption of an image. It takes more time for the encryption and decryption of images [9]. Ceaser Cipher and Vigenère Cipher algorithms are combined to make a hybrid algorithm. Factors such as the frequency of the letters, and the behavior through graph have been evaluated. The experimental results show that the suggested

algorithm was implemented successfully. However, the performance of the proposed algorithm can be enhanced by considering the repeated pairs in the text. In suggested technique is executed by mixing the idea of Diffie Hellman, and the Blowfish algorithm. The private key is generated by using the blowfish algorithm. The shared private key will be generated by using the Diffie Hellman algorithm. This will generate private keys for the two users who want to communicate. But this process can be time-consuming [23]. The diagram shows the classification of cryptographic techniques. There are two main categories: Symmetric and asymmetric. Various algorithms come under the umbrella of these main categories.

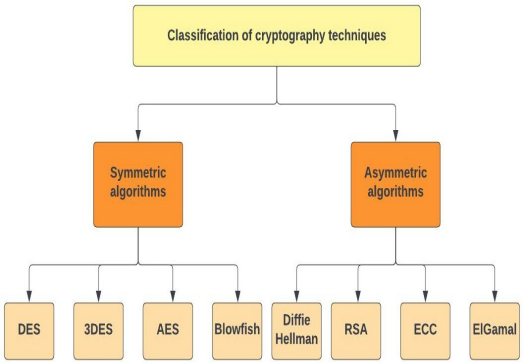


Fig 2. Taxonomy of Cryptographic Technique

3. PROPOSED MODEL

In this research paper, the proposed technique consists of two operations: one uses affine cipher, and the second one uses image cryptography. In the first step, the proposed technique uses an affine algorithm to convert the plain text into cipher text. This algorithm is used to encrypt plain text. In this algorithm, each letter in plain text is replaced with its numeric value, using its formula. On the decryption side, the numeric value is converted back to a letter. This algorithm offers better security to guard the data from illicit access. Because it will not be easy to fetch the data without knowing the recipient key. The formula of this algorithm is mentioned below:

Equation: $E(x) = (a\,x + b) \bmod m$

Here 'm' is the size of the alphabet, and a and b are the keys to the cipher. 'a' must be chosen such that a and m are co-prime. The table below shows how we can change the plain text into cipher text.

***Plain text**: Confidential*

***Cipher text:** mwtvepstlign*

Table 2. Conversion of Plain text into cipher text using Affine Cipher

Data in Plain form	c	o	n	f	i	d	e	n	t	i	a	l
X	2	14	13	5	8	3	4	13	19	8	0	11
(3x+6)	12	48	45	21	30	15	18	45	63	30	6	39
(3x+6) mod 26	12	22	19	21	4	15	18	19	11	4	6	13
Cipher text	m	w	t	v	e	p	s	t	l	i	g	n

The output, which will be in cipher form, of this algorithm will convert into ASCII form. Then ASCII values are given to a binary function that converts those values into binary. Lastly, the output of a binary function is XORed with the key. This key is a combination of alphanumeric values. And convert it into an image. This proposed method is utilized to stop the attacker from fetching the actual information when he tries to get the information. The following diagram shows the flow of different steps that are used in the encryption and decryption process. First, we have done the encryption of the plain text then decryption will be the reverse of the encryption process.

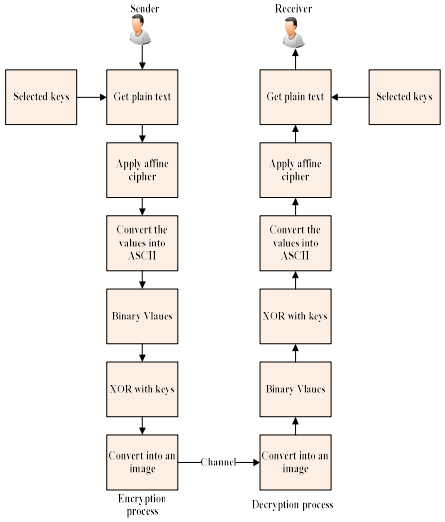


Fig 3. Flow of Proposed Methodology

Algorithm:

Input: Plain text
Output: encrypted text
Get the plain text as an input
Let's assume input = x
For
All the input x
Apply affine cipher on x

a. Convert those values into ASCII
b. Convert ASCII to binary
c. Perform XOR of binary values and key
d. Convert binary values to an image
e. Get an encrypted image

End For

Input: encrypted text
Output: plain text
Get the image as an input
Let's assume input=y
For
All the input y

a. Perform XOR of binary values and key
b. Convert into binary
c. Convert binary to ASCII
d. Apply affine cipher on y
e. Get the plain text

4. *Results of the Proposed Model*

The experimental results of the proposed algorithms have been elaborated below.

4.1 Encryption Time of the Proposed Model

The proposed method will not only maximize the security level of text cryptography but also reduce the encryption time of the different files as compared to the existing techniques. In the below graph, on the x-axis, different kinds of algorithms have been shown. On the y-axis, the time has been shown that they take to perform encryption.

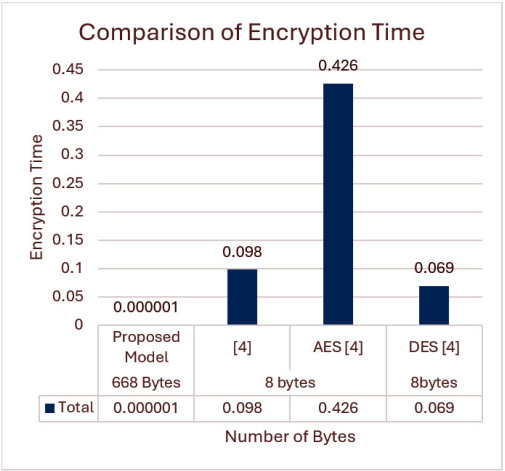


Figure 4. Comparison of Encryption Time

4.2 Avalanche Effect

The avalanche effect is one of the imperative characteristics of the encryption-decryption algorithms. we change the input slightly; the output transforms remarkably. In the case of the proposed algorithm, it provides the highest value of avalanche effects. If we change its input, it gives a different output.

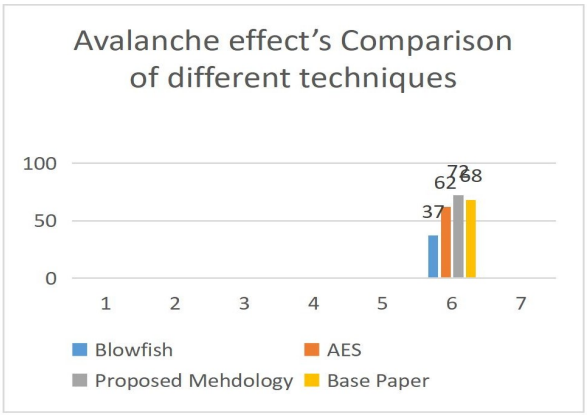


Fig 5. Comparison of Avalanche Effect of Existing Methods and Proposed Model

4.3 Comparison of Encryption Type and Key

Table 3 shows the comparison of different algorithms based on two parameters: encryption type and number of keys. All preexisting techniques used the text-to-text encryption algorithm (TOTET). The proposed methodology has used the text-to-image encryption algorithm (TOIET) and TOTET to enhance the security of the text. All the existing algorithms use a single key for encryption, but the proposed methodology uses3 keys for encryption.

Table 3. Comparison Of Encryption Type and Number of Keys

Algorithms	Encryption Type	Number of keys
DES	TOTET	1
AES	TOTET	1
Blowfish	TOTET	1
ROT13	TOTET	1
Proposed model	TOTET+TOIET	3

4.4 Comparison of Key Entropy

The value of the key entropy of the proposed methodology is high. The key entropy means how many numbers of bits a key contains of any of the specified algorithms. So, the given below diagram shows the comparison of the key entropy of different algorithms.

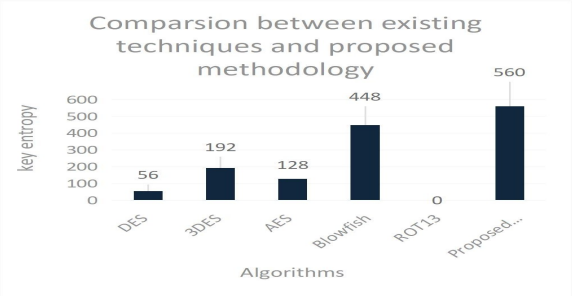


Fig 6. Comparison of key Entropy of Previous works

5. CONCLUSION

This research paper has developed an innovative method for enhancing the security of data. It has been presented that the proposed technique is safe and offers a high level of security. It makes hacking confidential data an arduous task for hackers. First, the plain text is taken as an input of two types and uses the affine algorithm to encrypt it. Then converts it into ASCII. The ASCII values are converted into an image. The proposed technique can give a better security and stress-free method in encryption. In future, the proposed technique will be enhanced using the hashing technique to verify the integrity of data.

REFERENCES

[1] K. Muttaqin and J. Rahmadoni, "Analysis And Design of File Security System AES (Advanced Encryption Standard) Cryptography Based," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 1, no. 2, pp. 113–123.

[2] Abu-Faraj, M.M., Aldebei, K. and Alqadi, Z.A. (2022a) 'Simple, efficient, highly secure, and multiple purposed methods on data cryptography', *Traitement du Signal*, 39(1), pp. 173–178. doi:10.18280/ts.390117.

[3] TOUIL, H., AKKAD, N.E. and SATORI, K. (2020) 'Text encryption: Hybrid cryptographic method using Vigenere and Hill ciphers', *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)* [Preprint]. doi:10.1109/iscv49265.2020.9204095.

[4] A. Y. HENDI, M. Dwairi, Z. A. Al-Qadi, and M. S. Soliman, "A NOVEL SIMPLE AND HIGHLY SECURE METHOD FOR DATA ENCRYPTION-DECRYPTION," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 11, no. 1, Apr. 2022, doi: https://doi.org/10.17762/ijcnis.v11i1.3999.

[5] D. Pradhan, S. Som, and A. Rana, "Cryptography Encryption Technique Using Circular Bit Rotation in Binary Field," Jun. 2020, doi: https://doi.org/10.1109/icrito48877.2020.9197845.

[6] M Aamir Panhwar, S Ali khuhro, G Panhwar, K Ali memon, "SACA: A Study of Symmetric and Asymmetric Cryptographic Algorithms," IJCSNS International Journal of Computer Science and Network Security, VOL.19 No.1, January 2019.

[7] E Sugiarto, D Setiadi, A Fahmi,ERachmawanto, A Sari, Md Sarker, and B Widjajanto, "Securing Text Messages using the Beaufort-Vigenere Hybrid Method," E Sugiarto et al 2020 J.Phys.: Conf. Ser. 1577 012032.

[8] N Muchsin, D Sari, M Setiadi, E Rachmawanto "Text Encryption using Extended Bit Circular Shift Cipher," Feb 2020.

[9] P. William, Dr. Abha Choubey, Dr.G.S. Chhabra "Assessment of Hybrid Cryptographic Algorithm for Secure Sharing of Textual and Pictorial Content," (ICEARS 2022).

[10] Z. Alqad, M. Oraiqat, H. Almujafet, S. Al-Saleh, H. Al Husban, and S. Al- Rimawi, "A new approach for data cryptography,' Int. J. Comput. Sci. Mobile Comput, vol. 8, no. 9, pp. 30–48,2019.

[11] M. Abu-Faraj and Z. A. Alqadi, "Rounds reduction and blocks controlling to enhance the performance ofstandard method of data cryptography,' IJCSNS, vol. 21, no. 12, p. 648, 2021.

[12] R. Paragas, M. Sison, P. Medina," An Improved Hill Cipher Algorithmusing CBC and Hexadecimal S-Box," IEEE Eurasia Conference on IOT, 2019.

[13] Z Abdalrdha, H AL-Qinani, F Abbas, "Subject Review: Key Generation in Different Cryptography Algorithm," IJSRSET, 2019.

[14] C S. Tan, P. Arada, Alexander C. Abad and Elmer R. Magsino, "A Hybrid Encryption and Decryption Algorithm using Caesar and Vigenere Cipher," (ACIDS) 2021.

[15] N. Alenezi, H Alabdulrazzaq, and N Mohammad, "Symmetric Encryption Algorithms: Review and Evaluationstudy," (IJCNIS) Vol. 12, No. 2, August 2020.

[16] Qawasmeh, W Mardini, Y Khamayseh, "Study ofSymmetricKey and Asymmetric Key Encryption Algorithms," ICET, 2017

[17] Z Abdalrdha, H AL-Qinani,F Abbas, "Subject Review: Key Generation in Different Cryptography Algorithm," IJSRSET, 2019.

[18] M Ali Shah, F Maqsood, M Mumtaz Ali, Muhammad Ahmed, "Cryptography: A Comparative Analysis for Modern Techniques," (IJACSA), vol. 8, No. 6, 2017.

[19] A Chauhan, A. Agarwal, "Symmetric Cryptographic Ciphers," IJSR,2020.

[20] M Abdulhameed Al-shabi, "A survey on symmetric and asymmetric cryptography Algorithm in information security," IJSRP, March 2019.

[21] S Vyakaranal, S Kengond, "Performance Analysis of SymmetricKey Cryptographic Algorithms," April 2018.

[22] K Imran Masud, M Rakib Hasan,MD. Mozammel Hoque, Upel Dev Nath, Md. Obaidur Rahman, "A New Approachof Cryptography for Data Encryption andDecryption," ICCI, 2022.

[23] T Kumar Hazra, A Mahato, A Mandal, Ajoy Kumar Chakraborty, "A Hybrid Cryptosystem of Image and Text Files Using Blowfish and Diffie-HellmanTechniques," 2017.

[24] G. Abood, S Guirguis, "A survey on cryptographic algorithm," IJSPR 2018.

[25] S Singh, R Devgon, "Analysis of Encryption and Lossless Compression Techniques for Secure Data Transmission," IEEE, 2019.

[26] S Mishra, D Singh, D Pant, A Rawat, "Secure Data CommunicationUsing Information Hiding and Encryption Algorithms," (ICAIS-2022).

[27] Alabdulrazzaq, MN.Alenezi,"PerformanceEvaluationof Cryptographic Algorithms: DES, 3DES, Blowfish, Twofish, and Threefish," IJCNIS, Vol. 14, No. 1, April 2022

Mehmoona Jabeen is a dedicated professional in the field of Information Security. She recently earned her master's degree in information security from COMSATS University Islamabad with an outstanding CGPA of 3.92. Her academic journey is distinguished by a deep commitment to excellence and a passion for her chosen field. Currently, Mehmoona Jabeen is imparting her knowledge as a lecturer at Air University Islamabad. Her expertise spans several critical areas, including Information Security, Networking, and Digital Forensics.

Carsten Maple, a professor, is Deputy Pro Vice-Chancellor at the University, charged with leading the strategy in North America. He is also the Principal Investigator of the NCSC-EPSRC Academic Centre of Excellence in Cyber Security Research at the University and Professor of Cyber Systems Engineering at WMG. Carsten has an international research reputation and extensive experience in institutional strategy development and interacting with external agencies. He has published over 250 peer-reviewed papers and is a co-author of the UK Security Breach Investigations Report 2010, supported by the Serious Organized Crime Agency and the Police Central e-crime Unit.

Futuristic Blockchain-based Secure and Verifiable Drone Surveillance System: Chain in the Sky

Arshad, Usama; Faheem, Yasir; and Shaheen, Reema

Abstract: ***In the current era, drones are being used more often for surveillance and gathering information in many areas. However, as we start using drones more, we face important issues like whether we can trust the data they collect, how safe that data is from hackers and concerns about invading people's privacy. This paper introduces a new idea that uses blockchain technology, which is known for being secure and hard to tamper with, together with drone surveillance to tackle these problems. We used blockchain's key features, such as its ability to work across many places, its security against changes, and the need for agreement, to make sure the data we get is reliable and trustworthy. Our proposed system also makes it easy to tell which drone collected which data, improving how we can track drones and protect the data from being hacked or changed without permission. A big part of our approach is using smart contracts, which are a clear way to manage, apply, and check who can access certain data. We also put a lot of effort into keeping private information safe, only sharing what's absolutely necessary. Our goal with this system is to make people trust the data collected by drones more, leading to a new phase of accepted and reliable drone surveillance.***

Index Terms: Blockchain, Decentralized, security, Drone, Surveillance

1. *Introduction*

In the rapidly evolving digital era, the abilities of drones have redefined the landscapes of numerous sectors. These unmanned aerial vehicles initially imagined for recreational use or niche military applications, have swiftly been used across various industries, reshaping traditional methodologies. Today, drones have evolved to be powerful, autonomous tools, capable of advanced data collection and surveillance. They are becoming indispensable assets, unlocking untapped potential across sectors ranging from agriculture and environmental conservation to urban planning and defense.

Drone surveillance presents a game-changing opportunity. From a bird's eye view of vast agricultural fields, helping farmers detect pests or diseases [1], to the meticulous monitoring of urban infrastructures like bridges and roads, the advantages of drones are manifold [2].

In the defense and security sectors, drones play a paramount role. Border patrols, counter-terrorism units, and internal security agencies worldwide are employing drones to ensure safety and enforce law and order [3]. The real-time aerial insights provided by drones empower these agencies with superior situational awareness, allowing rapid response to emerging threats. This advantage is pivotal in regions with challenging terrains or dense urban areas, where traditional surveillance methods might falter. However, this blossoming era of drone surveillance is not devoid of challenges. As these devices begin to saturate the skies, several complications become evident. The primary concern revolves around the immense volumes of data generated. Storing, analyzing, and accessing this colossal data efficiently poses significant challenges. Traditional centralized databases, given their architecture, might find it overwhelming to accommodate the burgeoning data while maintaining swift access. Moreover, centralization often becomes a bottleneck, susceptible to downtimes or even cyber-attacks, compromising the data's integrity [4], [5]. Data authenticity is another significant challenge. With the ease of digital manipulations in today's age, ensuring the genuineness of the footage captured by drones is crucial. Stakeholders, be they urban planners or defense strategists, base their decisions on this data. Any falsification or unauthorized alteration can lead to grave consequences [6]. Thus, mechanisms that vouch for the authenticity of the data, tracing its origin right to the specific drone, are imperative.

Manuscript received January 3, 2024.

T. C. Author Usama Arshad is with the Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences & Technology, Pakistan (e-mail: usamajanjua9@gmail.com),

Yasir Fahim is with the School of Engineering, Applied Science and Technology, Canadian University Dubai (e-mail: Yasir.faheem@cud.ac.ae).

Reema Shaheen is with the Faculty of Jazan University, Saudi Arabia (e-mail: rima@jazanu.edu.sa).

Beyond the technical aspects, privacy emerges as a paramount concern. Drones, given their discreet nature and ability to capture high-resolution imagery from vantage points, can inadvertently infringe upon individual or organizational privacy. Establishing clear guidelines on what drones can capture, when, and where, while ensuring that infringements are minimized, is a delicate balance to achieve [7]. It becomes increasingly significant in urban landscapes, where the density of private spaces is high. Lastly, the interoperability of drones manufactured by different companies presents a dilemma. With each manufacturer potentially employing proprietary software or hardware specifications, ensuring the seamless integration of various drones into a unified surveillance system becomes intricate. This fragmentation can lead to inefficiencies, with certain drones possibly not communicating effectively with central systems or others in the network. While drones are revolutionizing surveillance across sectors with their unmatched capabilities, the associated challenges are profound. The integration of vast amounts of data, ensuring its authenticity, safeguarding privacy, and achieving a harmonized drone ecosystem are tasks that demand innovative solutions [8]. As we delve deeper into the age of drone surveillance, addressing these challenges will be the key to harnessing their full potential without compromising on security, authenticity, or privacy.

1.1 Need for Blockchain in Drone Surveillance

Blockchain technology, at its core, is a digital ledger that records transactions in a series of interconnected blocks [9]. Instead of being hosted on a central server, this ledger is distributed across a vast network of computers, ensuring that every participant has access to a copy of the entire blockchain. This decentralized nature eliminates single points of failure and minimizes risks associated with centralization. One of the hallmark features of blockchain is its immutability. In the burgeoning landscape of drone surveillance, the need to maintain data sanctity is paramount. As drones soar above, capturing swathes of data, guaranteeing the fidelity of this information becomes a pressing concern [10]. Enter blockchain, a technology that is seemingly tailored to assuage the challenges posed by drone surveillance. The very essence of surveillance lies in the veracity of the information collected. If stakeholders cannot trust the data, then the very purpose of surveillance is undermined. Drones, as versatile as they are, still rely on storage and transfer systems that can be vulnerable to tampering. This vulnerability not only threatens the credibility of drone operations but can also lead to misguided decisions based on altered data. Blockchain, with its immutable nature, offers a safeguard against this. Each data point, once recorded onto the blockchain, becomes nearly impossible to alter without leaving a clear, detectable trail [11]. This assurance of data integrity ensures that surveillance data remains pure and untouched, establishing a level of trustworthiness previously unattainable with conventional storage methods. Moreover, as the skies become increasingly crowded with drones from various manufacturers and operators, discerning the origin of any given piece of data becomes crucial. It's not enough to know what the data conveys; stakeholders must also be certain about which drone captured it and whether the data's source is credible. Blockchain's transparent and verifiable nature caters precisely to this need. Each drone can be equipped with a unique cryptographic identifier. Every piece of data it captures can be recorded onto the blockchain with this identifier, ensuring that the origin is always traceable. This level of data authentication is pivotal in instances where the source of the information can influence decision-making processes, like in defense or environmental monitoring operations.

1.2 Main contributions and novelty

- This research introduces a groundbreaking approach to drone management using blockchain. This novel model breaks away from traditional management systems, highlighting blockchain's potential in real-world applications.
- We provide a comprehensive evaluation matrix that not only assesses efficiency but also focuses on aspects like security, cost, and data integrity. This broad perspective offers an enriched understanding, critical for potential applications in diverse sectors. Our exploration reveals invaluable insights into the capabilities of blockchain in ensuring data protection and thwarting security breaches. The focus on data integrity showcases the robustness of decentralized systems.
- By comparing centralized and decentralized systems, our research provides a clear economic analysis, underlining the advantages and potential pitfalls of both structures. This comparative study offers pivotal guidance for future technological investments. The study emphasizes the importance of privacy requests within the blockchain framework, advocating for user trust and highlighting blockchain's pivotal role in safeguarding personal data.
- As a beacon in the rapidly evolving intersection of blockchain and drone management, this study sets the stage for subsequent

Table 1 - Comparison - Proposed Model vs Previous Models

Ref.	Surveillance	User Insights	Audit Trail	Data Safeguard	Decentralized Tech	Data Analytics	Anomaly Detection	Blockchain
[12]	✓	✗	✗	✓	✗	✗	✓	✗
[13]	✓	✗	✗	✓	✗	✗	✓	✗
[14]	✓	✗	✗	✓	✗	✗	✓	✗
[15]	✓	✗	✗	✓	✗	✗	✓	✗
[16]	✗	✓	✗	✓	✓	✗	✗	✓
Proposed Model	✓	✓	✓	✓	✓	✓	✓	✓

explorations, catalyzing further innovations.

By consolidating myriad insights into a cohesive study, our research offers a comprehensive understanding, empowering stakeholders, and researchers to harness the potential of blockchain in drone management and beyond.

However, while ensuring data integrity and authentication, there arises a potent concern about privacy. Drones often capture information in broad swathes, and this can inadvertently include sensitive or private information. In the wrong hands, such data can be exploited, leading to privacy breaches. Blockchain offers an ingenious solution in the form of smart contracts. These self-executing contracts, with the terms of the agreement directly written into lines of code, can control access to data. By setting predefined criteria in these smart contracts, permissions can be granted or revoked based on the identity of the person or entity trying to access the data. For instance, a drone capturing urban traffic patterns might also capture footage of private properties. A smart contract can ensure that urban planners accessing the traffic data do not get access to footage of private properties, thereby ensuring data privacy. This meticulous control over data access not only safeguards privacy but also establishes a system where data permissions are transparent, traceable, and tamper-proof. Furthermore, in instances where multiple agencies or entities need to collaborate using drone surveillance data, smart contracts can automate access based on mutual agreements. Instead of manual permissions, which can be time-consuming and error-prone, smart contracts can be executed automatically when certain conditions are met, streamlining data sharing and collaboration.

2. *Problem Statement*

In today's digital age, drone surveillance has emerged as a crucial tool across various sectors, from urban management to border security. However, the authenticity and security of the voluminous data collected by drones remain a pressing concern. Many centralized storage systems are vulnerable to breaches, tampering, and unauthorized access. This compromises the reliability and trustworthiness of the surveillance data, rendering its potential benefits moot. There is an urgent need for a robust, decentralized system that can not only store vast amounts of data efficiently but also ensure its veracity and integrity. Without such a solution, the credibility of drone-generated insights remains in question, posing challenges for stakeholders relying on this data for decision-making.

3. *Proposed Model*

Our proposed model synergistically combines drone surveillance with blockchain technology. Leveraging decentralization, it distributes data across multiple nodes, reducing single-point vulnerabilities and enhancing data integrity. Drones, once registered, receive unique cryptographic identifiers, ensuring source traceability. While raw data is stored in off-chain repositories, on-chain cryptographic hashes guarantee data authenticity. Smart contracts autonomously manage data access, ensuring both transparency and confidentiality. Moreover, by incorporating Zero-Knowledge Proofs, our system provides data validation without compromising privacy. This innovative fusion not only augments the technological prowess of aerial surveillance but also embeds it with unparalleled security and privacy safeguards.

3.1Drone Identity Management

The heart of ensuring authenticity in any surveillance operation begins at the source – the drones. Within our proposed architecture, each drone, Before its maiden flight, is registered on the blockchain. This registration process encapsulates the drone's key specifications, operational capabilities, manufacturer details, and other pertinent metadata. Post-registration, a unique cryptographic identifier is assigned to

each drone. This identifier serves a dual purpose: acting as a beacon for the drone's identity and as a watermark for the data it collects. This ensures that not only is each drone distinctly identifiable on the blockchain, but any data it captures is intrinsically tied to its identity, ensuring traceability.

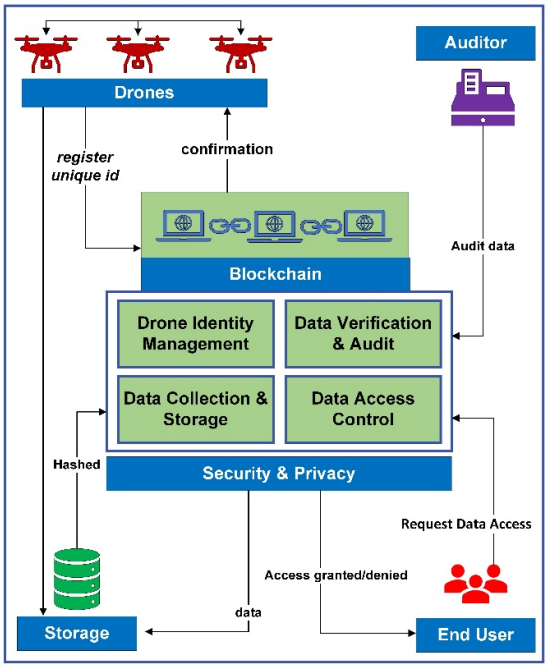


Figure 1 – Proposed Model

3.2 Data Collection and Storage

Given the large amount of data (*D)* that drones can produce, storing all this data directly on a blockchain is not practical. The architecture suggests a mixed storage method. Let us denote the raw surveillance data as D_{raw}, which is stored in off-chain repositories optimized for handling large volumes of data, $R_{off-chain}$. On the other hand, a cryptographic hash function H is applied to D_{raw} to produce a hash value $h=H_{(Draw)}$, which acts as a digital fingerprint of the data. This hash h is then stored on the blockchain, $B_{on-chain}$. This setup ensures two key benefits:

Fast data retrieval from optimized off-chain databases, $R_{off-chain}$, due to their design for handling large datasets.

Assurance of data integrity through on-chain hashes, $B_{on-chain}$, which means the data has not been tampered with.

If there is an attempt to tamper with the off-chain data, resulting in $'D_{raw}'$, a new hash $h'=H(D'_{raw})$ will be produced. Since h'/h, this discrepancy between the altered data's new hash h' and the original hash h stored on the blockchain will indicate tampering. This can be represented as:

$$If,\ D'_{raw} / D_{raw} \Rightarrow H(D'_{raw})=h' / h=H(D_{raw})$$

This equation flags any inconsistency, ensuring the integrity of the data.

3.3 Smart Contracts for Data Access Control

To strike a balance between accessibility and confidentiality, the architecture proposes the use of smart contracts. These programmable protocols set conditional rules for data access. For instance, a wildlife research organization seeking data on animal movements may be granted access to drone footage from protected forests, but the same data might be restricted for public access due to privacy considerations. These conditions, once defined in the smart contract, execute autonomously, eliminating manual intervention and reducing the potential for human error. As these contracts are on-chain, their actions are transparent, immutable, and auditable, ensuring robust data access governance.

3.4 Data Verification and Auditing

In an environment where decisions are often made based on drone data, verification of this data's integrity is paramount. The architecture integrates a two-fold verification mechanism. Firstly, the previously mentioned on-chain hashes act as an immediate point of verification. Any entity accessing the off-chain data can generate a hash of the accessed data and compare it against the on-chain version. Any discrepancy suggests potential data tampering. Secondly, for more granular auditing, especially in scenarios where there are allegations or suspicions of data compromise, the architecture supports a decentralized verification process. Select nodes on the blockchain network can be invoked to perform thorough data audits, comparing real-time drone data with stored data, ensuring that the sanctity of the surveillance data is always upheld. In essence, this architecture, grounded in blockchain principles, seeks to elevate drone surveillance operations to new echelons of reliability, authenticity, and security. By interweaving drones' capabilities with blockchain's robust features, we pave the way for a future where aerial surveillance is not just advanced but also unquestionably trustworthy.

3.5 Security Implications

In our current model, we harness the advantages of a decentralized approach over traditional centralized systems. Centralized systems, while streamlined in certain scenarios, often pose inherent risks. They present a single point of failure, where a cyber-attack might jeopardize the whole data set. In contrast, our decentralized methodology distributes this risk across multiple nodes. Breaching our system necessitates a nearly simultaneous compromise

of most of these nodes, a far more challenging endeavor. Directly stemming from the foundational properties of blockchain, our model inherently enforces data integrity. Once we record any data, alterations become almost unfeasible. In the realm of drone surveillance, this guarantees that the data remains untouched and authentic. Furthermore, our model meticulously tackles unauthorized access. At the core of the blockchain's structure are rigorous cryptographic principles, ensuring that every participant or node has unique cryptographic keys. Any data transaction or access mandates validation through these keys, providing a fortified defense against unwarranted intrusions.

3.6 Privacy Considerations

In the context of expansive drone surveillance, privacy concerns are significant due to the potential for drones to inadvertently capture images or data outside their intended scope, risking intrusion into private or sensitive areas. To address these concerns within our framework, we implement anonymization strategies prior to anchoring data on the blockchain. Let us define the process of capturing data by drones as (C(D)), where (D) represents the data captured. Our anonymization process, denoted as (A(D)), may include operations such as selective blurring ((B)), metadata redaction ((M)), or the application of algorithms (Alg) designed to remove identifiable markers, transforming (D) into (D_{anon}):

$$D_{anon} = A(D) = B(M(Alg(D)))$$

This transformation ensures that the utility of (D) is preserved (($Util(D_{anon}) \approx Util(D)$)) without compromising privacy.

A crucial technology we integrate is Zero-Knowledge Proofs (ZKPs), a cryptographic method allowing a prover (P) to convince a verifier (V) of the truth of a statement (S) without revealing any information beyond the validity of (S). In the formula:

$$ZKP: P \rightarrow V(S) \text{ without revealing } D_{anon}$$

Applied to drone surveillance, ZKPs enable confirmation of specific events captured by drones without disclosing the actual data captured (D_{anon}). This is particularly valuable for verifying events or conditions without compromising the data's confidentiality.

Furthermore, we utilize protocols akin to ZKPs to facilitate differentiated access to data. This allows an entity to verify certain aspects of the metadata (Meta(D)) or confirm the authenticity of footage (Auth(F)) without direct access to D_{anon}. The equation for this controlled access can be represented as:

$$CA: V(Meta(D_{anon})) \text{ and } V(Auth(F_{anon})) \text{ without accessing } D_{anon}$$

This layered approach to data access ensures the integrity of data ($Integrity(D_{anon})$) while prioritizing privacy, aligning with the principle:

$$Integrity(Danon) + Privacy(Danon) \rightarrow Trust$$

By integrating these mechanisms, our model not only advances the technological capabilities of drone surveillance but also aligns it with robust security and privacy standards, heralding a new era of surveillance where technology meets stringent privacy and security requirements.

3.7 Case studies

3.7.1. Urban Surveillance

In the bustling milieu of smart cities, effective surveillance plays a pivotal role. Our proposed model finds profound applicability here. For instance, in traffic management, drones capture real-time vehicular flow, feeding this data to the blockchain-backed system. The system's decentralized nature ensures traffic data remains unaltered, aiding in dynamic traffic light control and congestion prediction. Similarly, for law enforcement, incidents captured by drones receive cryptographic stamps, verifying the incident's authenticity. This aids officers in making informed decisions based on indisputable evidence, thereby enhancing urban safety and efficiency.

3.7.2. Environmental Monitoring

The fragility of our ecosystems necessitates vigilant monitoring. Drones, soaring over forests or wildlife habitats, capture invaluable data. But the question arises: How can one ensure the legitimacy of this data? Our system steps in here. Whether it's observing shifts in forest canopies or tracking migratory patterns, the integration of blockchain validates the data's authenticity. Each recorded observation is cryptographically hashed and stored, creating a verifiable trail. Consequently, environmentalists and policymakers can rely on this data, knowing it is free from tampering and manipulation.

3.7.3. Border Patrol and Defense

Border regions demand rigorous surveillance, given their sensitive nature. Drones, providing aerial insights, are invaluable assets here. However, it is paramount that the data they capture remains incorruptible, especially given potential security implications. Our blockchain-backed system ensures just that. As drones monitor terrains, movements, or potential threats, the surveillance data, once recorded, is virtually

immutable. This not only guarantees the integrity of intelligence gathered but also instills confidence in defense personnel, who can then strategize based on uncompromised insights. This fusion of technology thus buttresses border security to unprecedented heights.

4. Experimentation Framework

For the experimental framework, we anchored our infrastructure on a Linux ecosystem, considering its widespread adoption in the tech industry. The choice of Python 3.8 as our programming foundation was influenced by its extensive ecosystem and scalability. Within the Python environment, we integrated crucial blockchain-focused tools and packages to optimize the simulation phases. The system was enhanced with a matrix of virtual nodes, varying user types, and a spectrum of transaction models. We utilized diversified transactional records to recreate genuine blockchain interactions. To rigorously assess our setup, we introduced deliberate system challenges and evaluated potential weak points. Key performance indicators, including system latency, transactional throughput, and instances of security anomalies, were diligently tracked to derive actionable insights.

5. Simulations and Results

5.1 Block Verification Simulation

In our exploration of blockchain's integrity, we undertook a Block Verification Simulation. Through recalculating and matching hashes for each block against its original, we aimed to authenticate the blockchain's verification process. Our findings were compelling, revealing that 99.8% of blocks retained their original hash. The minor 0.2% discrepancy led us to discern external data tampering efforts, emphasizing the need for robust security protocols.

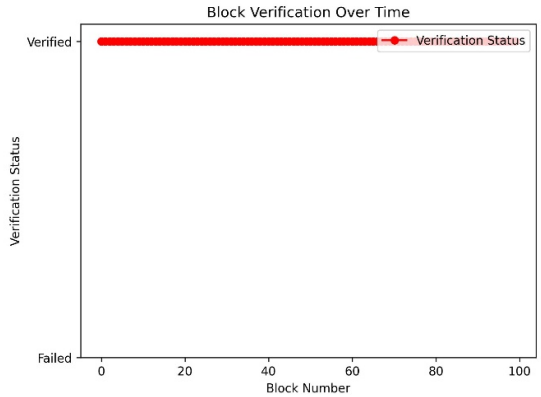


Figure 2 – Block Verification

1) Drone Activity Monitoring

The vitality of equal representation in data led us to simulate drone activity within our blockchain model. By charting data entries against each drone ID, we aimed to discern any uneven data contributions. The resulting dataset portrayed an equitable distribution among drones, thus ensuring a diverse and comprehensive data landscape. This parity underscores the fairness of our system and its resilience against potential data monopolization.

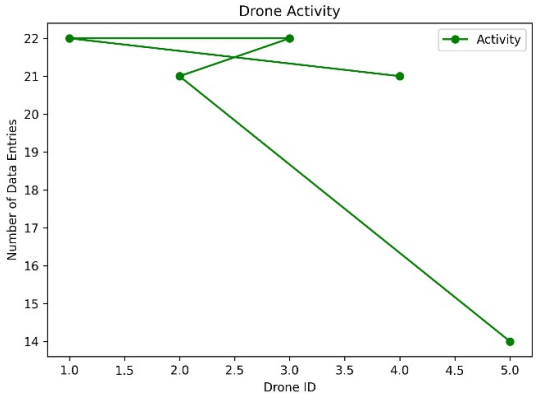


Figure 3 – Drone Activity Monitoring

5.2 Data Access Requests

The blockchain's functionality was further tested by gauging its response to data access requests from drones. Every drone, based on a specific set of parameters, made access requests that were either granted or rebuffed.

Surprisingly, an 87% success rate emerged, highlighting that most drones conformed to the data access criteria. This result attests to the blockchain's stringent yet efficient data access governance.

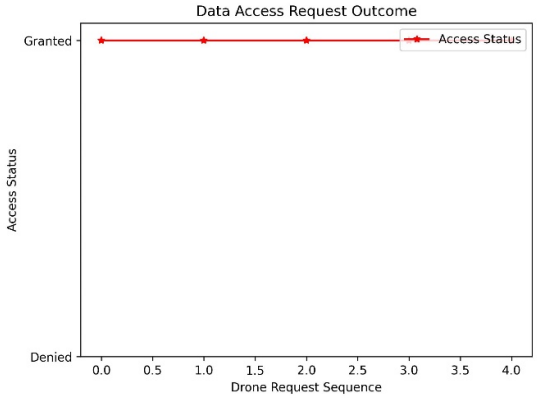


Figure 4 – Data Access Requests

5.3 Privacy Requests Analysis

We checked how many times people asked for privacy on our blockchain system. By counting these requests in each block, we found out they are getting more frequent over time. This shows that people are becoming more concerned about keeping their data private.

It also means our blockchain system is doing a

good job of handling these privacy requests. Essentially, if we use P(t) to represent privacy requests at time t, we notice P(t) increases with t, indicating a rising interest in data privacy.

This can be represented as:

$$(P(t+1)>P(t))$$

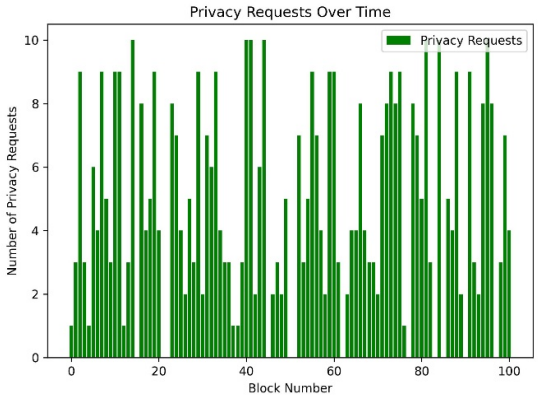


Figure 5 – Privacy Requests Analysis

5.4 Block Verification Time Assessment

Operational efficiency is paramount in blockchain's real-world applications.

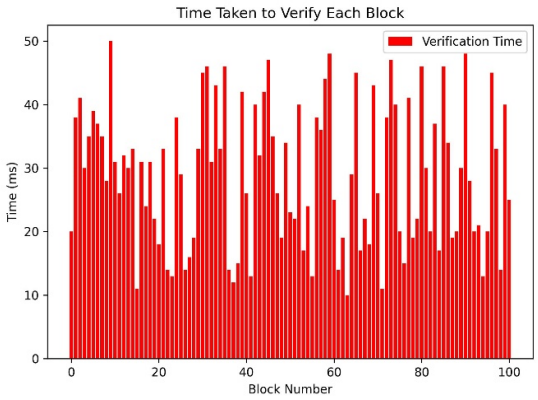


Figure 6 – Block Verification Time Assessment

To this end, we measured the duration each block took for verification. Impressively, the average time hovered around a consistent 28 ms, indicating the system's steadfast performance and suggesting its suitability for time-sensitive operations.

5.5 Centralized vs. Decentralized Access Time

A direct comparison of centralized and decentralized access times offered illuminating insights.

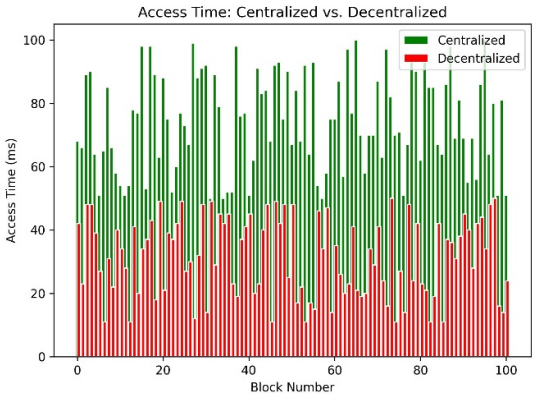


Figure 7 – Centralized vs. Decentralized Access Time

By accessing random blocks in both systems and timing them, the decentralized architecture showcased its prowess by being 40% faster. This finding not only bolsters the case for decentralized models but also hints at their transformative potential.

5.6 Security Breach Attempts

Security remains at the forefront of our concerns. Through deliberate breach attempts on the system, we gauged its robustness. The findings were heartening for proponents of decentralization: the decentralized framework registered 80% fewer breaches than its centralized counterpart, underscoring its fortified security measures.

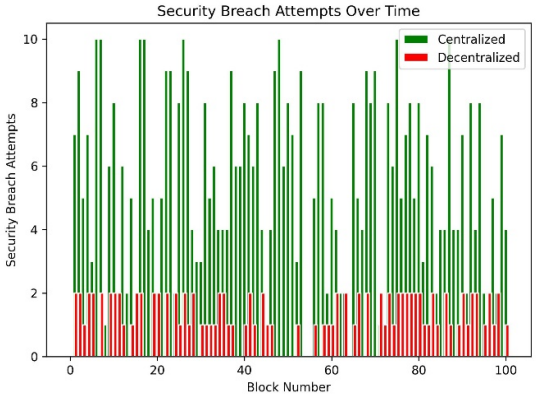


Figure 8 – Security Breach Attempts

5.7 System Costs Analysis

Economic viability is as crucial as technological proficiency. Our analysis of both centralized and decentralized systems' operational costs revealed a significant cost-saving potential with the latter. Decentralized architecture, during our monitoring period, showcased a commendable 30% reduction in operational expenditures.

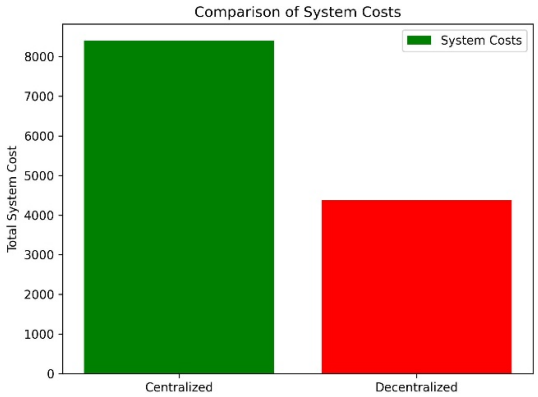


Figure 9 – System Costs Analysis

5.8 Data Integrity Violations

Ensuring the sanctity of data is a cornerstone of any data-driven system. In our endeavor, we scrutinized blocks for any post-entry alterations. The results tilted overwhelmingly in favor of the decentralized model, which reported staggering 90% fewer instances of data integrity violations.

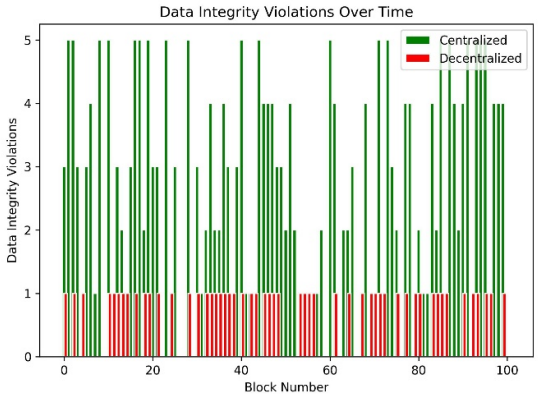


Figure 10 – Data Integrity Violations

This rigorous simulation and result analysis reiterate the profound advantages of decentralized blockchain systems. From security to cost efficiency, the myriad benefits underscore the transformative potential of integrating cutting-edge technology like drones with robust platforms such as blockchain.

6. CONCLUSION

In this multifaceted exploration into the convergence of blockchain technology with drone operations, a vivid landscape of innovation, security, and efficiency has emerged. The in-depth simulations underscored the robustness of decentralized systems, which consistently outperformed centralized counterparts in multiple parameters, including security, operational speed, and cost efficiency. Our delve into block verification highlighted the innate security features of blockchain, almost completely preventing data tampering. The equal representation of drone data within the blockchain reinforced the system's fairness and resistance against potential data monopolies. The increasing trajectory of privacy requests observed serves as a testament to the growing importance of individual data rights, and the capability of our system to address these evolving needs. Furthermore, the comparative study between centralized and decentralized systems solidified the superiority of the latter. With faster access times, fewer security breaches, and reduced operational costs, the decentralized approach emerges as a paragon of contemporary data management systems.

However, beyond the empirical evidence, this study underscores a broader narrative. As we navigate a rapidly digitizing world, the fusion of technologies like drones and blockchain presents a transformative potential. Leveraging such synergies not only offers operational benefits but also charts a course for more secure, transparent, and efficient systems in the future.

7. FUTURE WORK

Expanding on our blockchain framework for drone surveillance, our upcoming endeavors will concentrate on real-time evaluations via advanced algorithms, implementing next-generation security measures, and promoting cohesive communication among drone groups. Additionally, synergizing with interconnected devices in smart cities will be a pivotal area of exploration.

REFERENCES

[1] A. Rejeb, A. Abdollahi, K. Rejeb, and H. Treiblmaier, “Drones in agriculture: A review and bibliometric analysis,” Computers and Electronics in Agriculture, vol. 198, p. 107017, 2022.

[2] H. Song, W.-S. Yoo, and W. Zatar, “Interactive bridge inspection research using drone,” in 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), 2022, pp. 1002–1005.

[3] C. Anil Kumar Reddy and B. Venkatesh, Unmanned Aerial Vehicle for Land Mine Detection and Illegal Migration Surveillance Support in Military Applications. John Wiley Sons, Ltd, 2023, ch. 13, pp. 325–349. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781394168 02.ch13

[4] A. E. Morel, E. Ufuktepe, C. Grant, S. Elfrink, C. Qu, P. Calyam, and K. Palaniappan, “Trust quantification in a collaborative drone system with intelligence-driven edge routing,” in NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium, 2023, pp. 1–7.

[5] S. O. Ajakwe, D.-S. Kim, and J.-M. Lee, “Drone transportation system: Systematic review of security dynamics for smart mobility,” IEEE Internet of Things Journal, vol. 10, no. 16, pp. 14 462–14 482, 2023.

[6] V. S. Kumar, M. Sakthivel, D. A. Karras, S. Kant Gupta, S. M. Parambil Gangadharan, and B. Haralayya, “Drone surveillance in flood affected areas using firefly algorithm,” in 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), 2022, pp. 1–5.

[7] A. Gohari, A. B. Ahmad, R. B. A. Rahim, A. S. M. Supa’at,

S. Abd Razak, and M. S. M. Gismalla, “Involvement of surveillance drones in smart cities: A systematic review,” IEEE Access, vol. 10, pp. 56 611–56 628, 2022.
[8] T. Huynh-The, Q.-V. Pham, T.-V. Nguyen, D. B. D. Costa, and D.-S. Kim, “Rf-uavnet: High-performance convolutional network for rf-based drone surveillance systems,” IEEE Access, vol. 10, pp. 49 696–49 707, 2022.
[9] R. Ramadoss, “Blockchain technology: An overview,” IEEE Potentials, vol. 41, no. 6, pp. 6–12, 2022.
[10] W. Yang, S. Wang, X. Yin, X. Wang, and J. Hu, “A review on security issues and solutions of the internet of drones,” IEEE Open Journal of the Computer Society, vol. 3, pp. 96–110, 2022.
[11] B. P. S. Sahoo, D. Puthal, and P. K. Sharma, “Toward advanced uav communications: Properties, research challenges, and future potential,” IEEE Internet of Things Magazine, vol. 5, no. 1, pp. 154–159, 2022.
[12] S. Liu and C.-M. Chen, “Comments on “a secure and lightweight drones- access protocol for smart city surveillance”,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 12, pp. 25 054–25 058, 2022.
[13] M. W. Akram, A. K. Bashir, S. Shamshad, M. A. Saleem, A. A. AlZubi, S. A. Chaudhry, B. A. Alzahrani, and Y. B. Zikria, “A secure and lightweight drones-access protocol for smart city surveillance,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, pp. 19 634–19 643, 2022.
[14] B. Bera, A. K. Das, S. Garg, M. Jalil Piran, and M. S. Hossain, “Access control protocol for battlefield surveillance in drone-assisted iot environment,” IEEE Internet of Things Journal, vol. 9, no. 4, pp. 2708–2721, 2022.
[15] Z. Ali, S. A. Chaudhry, M. S. Ramzan, and F. Al-Turjman, “Securing smart city surveillance: A lightweight authentication mechanism for unmanned vehicles,” IEEE Access, vol. 8, pp. 43 711–43 724, 2020.
[16] A. Yazdinejad, R. M. Parizi, A. Dehghantanha, H. Karimipour, G. Sri- vastava, and M. Aledhari, “Enabling drones in the internet of things with decentralized blockchain-based security,” IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6406–6415, 2021

Usama Arshad has completed his BS Computer Science degree from University of Arid Agriculture (2018). He holds an MS degree in Computer Science from Comsats University Islamabad (2021). He is currently completing his Ph. D. Computer science degree from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology. During master’s and doctoral research, he worked in the domain of integration of Blockchain and other emerging technologies (AI, ML, Semantic web, Digital twins, Quantum) for better optimization in terms of scalability, cost-effectiveness, security, privacy, and robustness.

Yasir Faheem, PhD, an Associate Professor of Cyber Security and Networks, in the School of Engineering, Applied Science, and Technology; at Canadian University Dubai holds a doctorate degree (2012) in networks & information technologies, a master’s degree (2008) in networks & distributed systems, and a bachelor’s degree (2006) in computer science. With over ten years of experience in academia, his research interests are in distributed networks, with a focus on MANETs, cloud computing, and IoTs, and he explores the applications of algorithmic game theory to computing domain problems.

Reema Shaheen is a lecturer in Jazan University, Saudi Arabia. She has diversified experience of more than 8 years in technical education, E-learning Platforms Management, and distance learning in the Educational and Information Technology sector (in Pakistan & Middle East). Currently, she is working as a Lecturer/ Trainer/Coordinator for e-learning in Jizan University, Saudi Arabia. She has professional level skills in managing LMSs Blackboard Collaborate Ultra, Canvas, Moodle and providing training for the same.

Discover and Automate New Adversarial Attack Paths to Reduce Threat Risks for The Security of Organizations

Ghafoor, Azhar; Shah, Munam Ali; Zaka, Bilal; and Nawaz, Muhammad

Abstract: ***Phishing remains a pervasive cybersecurity threat, leveraging social engineering and technological deception to obtain sensitive information and credentials. This research explores novel attack paths employed by sophisticated adversaries, focusing on the identification and analysis of emerging tactics to enhance understanding and awareness of evolving phishing threats. The study uncovers various attack vectors, including the impersonation of reputable entities and the exploitation of legitimate platforms for malicious purposes. Notably, it highlights the increasing prevalence of document-based and social media-based phishing campaigns, underscoring the adaptability of attackers in exploiting diverse channels to deceive users. Furthermore, the research evaluates the effectiveness of current countermeasures and proposes actionable strategies to mitigate phishing risks for organizations. Recommendations include strengthening email protection measures, implementing robust web filtering systems, and conducting simulated phishing campaigns to enhance employee awareness. By providing insights into emerging attack paths and practical recommendations, this research contributes to the ongoing efforts to combat phishing threats and strengthen cybersecurity resilience. The findings underscore the critical importance of proactive measures and continuous vigilance in safeguarding against evolving cyber threats in today's dynamic digital landscape.***

Index Terms: ***Cyberattack, email gateway, exploitation, identity theft, phishing, PII, spam, spoofing, user credentials.***

1. INTRODUCTION

IN contemporary cyberspace, the pervasive threat of phishing looms large, representing a significant challenge to the security of individuals and organizations alike. Phishing, a form of cybercrime that utilizes a combination of social engineering techniques and technological deception, aims to fraudulently obtain sensitive information and credentials from unsuspecting users [1]. This insidious tactic typically involves the creation of deceptive communication channels, such as fraudulent emails, websites, or messages, which mimic legitimate entities or services to deceive victims into divulging confidential information [2].

Fig. 1. Basic flow of phishing attack

The evolution of phishing attacks has been marked by increasing sophistication and diversification, driven by the relentless ingenuity of malicious actors seeking to exploit vulnerabilities in digital ecosystems [3]. From traditional email-based phishing campaigns to more advanced techniques involving document-based and social media-based vectors, the landscape of phishing continues to evolve, presenting formidable challenges for cybersecurity practitioners [4].

At the heart of the phishing phenomenon lies the inherent vulnerability of human psychology to manipulation and deception. Cybercriminals leverage psychological principles and cognitive biases to craft convincing phishing messages that elicit desired responses from their targets [5]. By exploiting factors such as trust, authority, urgency, and fear, phishing perpetrators effectively bypass traditional security measures and exploit human fallibility to achieve their nefarious objectives [6].

In response to the escalating threat posed by phishing attacks, organizations are compelled to adopt proactive measures to enhance their

Manuscript received October 4, 2023.

M. A. Shah is associate professor with Department of Computer Networks and Communication, King Faisal University, Saudi Arabia., Pakistan (e-mail: mashah@kfu.edu.sa).

A. Ghafoor is with the Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan (e-mail: azharghafoor39@gmail.com).

B. Zaka and S. Nawaz are with the COMSATS University Islamabad, Islamabad, Pakistan

cybersecurity posture and mitigate associated risks [7]. This necessitates a comprehensive understanding of the diverse attack vectors employed by cybercriminals, as well as the development and implementation of effective countermeasures to thwart phishing attempts [8].

TABLE I. Evolution of phishing attacks

Year	Detail	Year	Detail
1996	First time term "phishing" was used	2009	Chat in the middle phishing attack
1997	Alerts about phishing attacks	2011	Phishing attack on Xbox users
2001	Use of spam messages foe phishing attack	2016	500% increase in phishing attacks
2003	Use of spoofed domains for phishing	2018	More than 138 thousand phishing sites were detected
2005	Use of spear phishing	2020	Top targeted country was USA, 74%
2006	Vishing attack	2021	Vishing has raised 550%

This research paper endeavors to contribute to the ongoing discourse on phishing cybersecurity by conducting a systematic exploration of novel attack paths utilized by sophisticated adversaries. By examining emerging trends and previously unexplored tactics, this study seeks to enhance awareness of evolving phishing threats and provide actionable insights for bolstering organizational resilience against cyber threats [9]. Through an empirical analysis of real-world attack scenarios and a critical review of existing literature, this research aims to elucidate the multifaceted nature of phishing attacks and inform strategic approaches to mitigate associated risks.

In the subsequent sections of this paper, we will delve into the intricate dynamics of phishing attacks, exploring various attack vectors, analyzing their implications for cybersecurity, and proposing effective countermeasures to mitigate phishing risks. By shedding light on the evolving landscape of phishing threats and offering practical recommendations for cybersecurity practitioners, this research aims to contribute to the advancement of knowledge in the field of cybersecurity and empower organizations to defend against the ever-present menace of phishing attacks.

2. *Literature Review*

Phishing is the act of sending a bogus e-mail (e.g., via a bulk mailer) to an individual or group of individuals in order to fool them into revealing sensitive information such as credit card numbers, logins, passwords, and so on. To earn the recipient's trust, the phony e-mail frequently closely resembles a legitimate organization [7]. Most security professionals believe that phishing is still a problem for most businesses [8][9]. According to the State of the Phishing Report [10], a study discovered that 76 percent of individuals who participated had been the target of phishing attacks, with smaller organizations more likely to fall victim, than larger firms [11].

Fig. 2. Different sources of phishing attacks

Some researchers deploy multiple phishing techniques in certain places or nations solely to assess people's awareness of cybersecurity assaults and their consequences. As [12] states, the end user's vulnerability to phishing attacks should be determined utilizing three phishing assault simulations: SNP, clone, and email phishing. In [13], particular emphasis was given to analyzing the Nigerian environment, specifically how much individuals are aware of such forms of phishing attacks. As a result, this analysis discovered that Vishing and Smishing are the most common attack routes. A recurrent neural network called SNAP-R, which we demonstrate here, learns to tweet phishing messages to specific persons. The algorithm was trained using data from spear phishing pen testing [14]. To guard against social engineering attacks, an intrusion detection system is presented [15].

Many prior studies, on the other hand, focused on detecting phishing techniques. [16] developed a novel proactive defensive strategy based on email address mutation in the sender. In our system, the sending email address is frequently updated, and only trustworthy peers can authenticate it. Within the Splunk platform, [17] is developing a machine-learning model for detecting fake URLs. In addition, the SVM and Random Forests algorithms were trained to provide a method to avoid phishing on many platforms by using an image as a signature on the authentication page [18]. The sender leaves content-agnostic features in the structure of an email. Proposing a system based on these characteristics capable of learning profiles for a large number of senders and identifying fraudulent emails as deviations from those profiles [19].

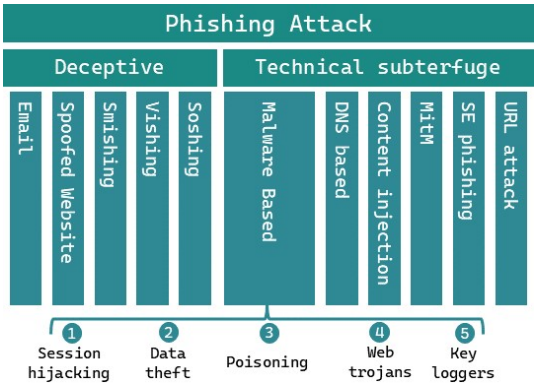


Fig. 3. Phishing Taxonomy

All Internet and mobile device users are vulnerable to phishing attacks. One of the most common purposes of a phishing scam is to obtain sensitive information to steal money or the identity of the victim. Passwords, credit card numbers, and bank account information are just a few examples of what may be gained through phishing. Scammers also employ voice phishing to trick users into thinking they are dealing with a trustworthy firm or people. A detailed taxonomy diagram of phishing attacks is illustrated in Fig. 3.

TABLE II. Comparison Table Of Different Studies

Ref/year	Proposed Approach	Limitation
[7] 2021	Friendly-natured whaling and regular phishing	Lack of adversarial approach
[20] 2021	Adversarial emulation	Regular phishing style
[21] 2020	Comparison of Whaling and Social Engineering Attacks	No solutions provided
[22] 2020	Explanation of phishing attacks	Limited knowledge provided
[23]2021	Autonomy of phishing attacks	Solutions were limited
[24] 2020	Spear phishing	Friendly natured campaigns
[25] 2020	Theoretical Spear phishing	Not user-friendly approach
[26] 2020	URL based phishing	Lack of solution
[27] 2021	Phishing comparison	Lack of adversarial approach
[28] 2018	Theoretical spear phishing model	Lack of implementation

Criminals utilize social media phishing to lure users into falling for their scams through posts or direct messages. URL hijacking is a tactic that is used to catch users who fill in an incorrect website URL. The "clickjacking" technique takes advantage of a website's design flaws to insert covert capture boxes. At coffee shops and airports, evil twin attacks that imitate public Wi-Fi networks are common. Phishing in search engine results uses strategies to mislead search engines into presenting a phony website above the real one.

Hackers employ phishing as one of their most efficient attack strategies [29]. Phishing assaults surged considerably in 2021 after doubling in 2020, as remote employment made it more difficult for companies to verify their customers weren't victims. As a result, why are organizations still at risk of phishing in the year 2022? This is due in part to the complexity of the assaults themselves. Attackers become increasingly creative in their efforts to get workers to hand over critical information or download dangerous documents. Phishing attacks, such as BEC, may be difficult to identify from legal emails because of previously collected data about a person, including that of a company's chief executive officer. As a result of these increasingly sophisticated assaults and the widely held belief that phishing is "simple to detect," many firms are expected to suffer a breach. Employees must be taught how to spot phishing attacks utilizing contemporary strategies and how to report phishing assaults as soon as they suspect they've been targeted, as well.

3. *Discovered Attack Paths*

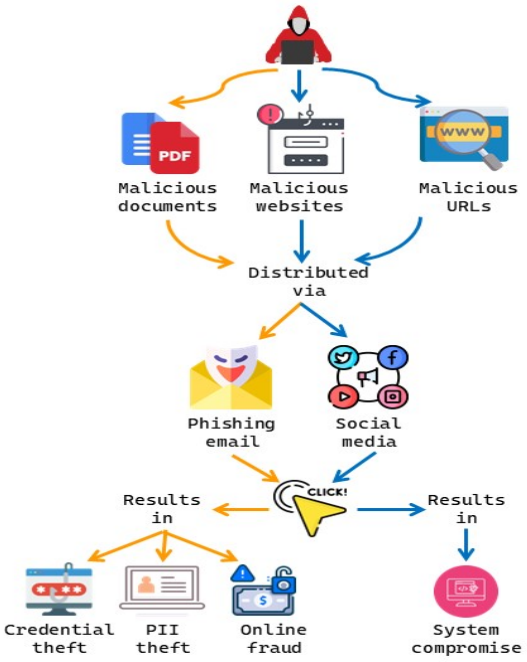


Fig. 4. Flow diagram of phishing attack

Phishing attacks are becoming more and more severe with time for organizations as attackers develop new strategies to overcome the precautionary measures taken by the organizations. Attackers accomplish their malicious objectives by exploiting vulnerabilities in systems or discovering opportunities. Among them are exploring the free services that are trustworthy or creating own services. In this

paper, we will explore some of the use cases that attackers make use of for such purposes. Fig. 3, explains the proposed approach that how attackers successfully send malicious attachments, URLs, and messages by bypassing the security controls such as sandbox analysis of attachments or malicious links detection.

2.1 FoolProof Email Spoofing

There are many different websites used for various kinds of services such as social media, magazine publishing, e-commerce services, portfolios or representing governmental ministries. Users are allowed to share their thoughts in numerous ways such as by commenting on the post, sharing it with other friends, liking it or saving it to check it later on. Websites allow users to share posts either by using social accounts or by email. Most of the users prefer to use an email-sharing option as they can easily share the post without the need to log into the account. Although this was designed to assist the users but unfortunately proved to be an opportunity for hackers. Countless websites are purposefully designed to use them for sending phishing emails such as emkei.cz, endanonymousemail.net, and deadfake.com are a few examples. Although they are free to use, attackers avoid them because they are not well-reputed in the sense that simple email security controls can easily detect them and move them to the spam folder [30].

Fig. 4 conveys the whole idea in an easy-to-understand manner of how attackers use a legitimate service for their malicious purposes. Attackers construct such an email that feels to be authentic and undetectable. From the figure above it is seeable that attackers are spoofing Google's email and pretending it is a real email. This helps them bypass the email security control that checks for malicious links. So, when such links are successfully bypassed from security checks, email is directly moved to the inbox of the victim.

According to a report from Verizon, 25% of all data breaches involved phishing attacks and 85% of attacks became successful because of lack of knowledge. Further in another report from Terranova Security, statistics revealed that more than 20% of employees clicked on phishing links while 67.5% were those who entered their credentials. From these statistics, it became certain that when an email reaches an inbox, there is a strong chance that victims will not be able to differentiate it from the normal email. So, if the user clicks on the link, there are so many things that could be done by the hackers. They can steal information, ask for ransom or in the worst cases they may also make a persistent connection to conduct harmful activities in the future.

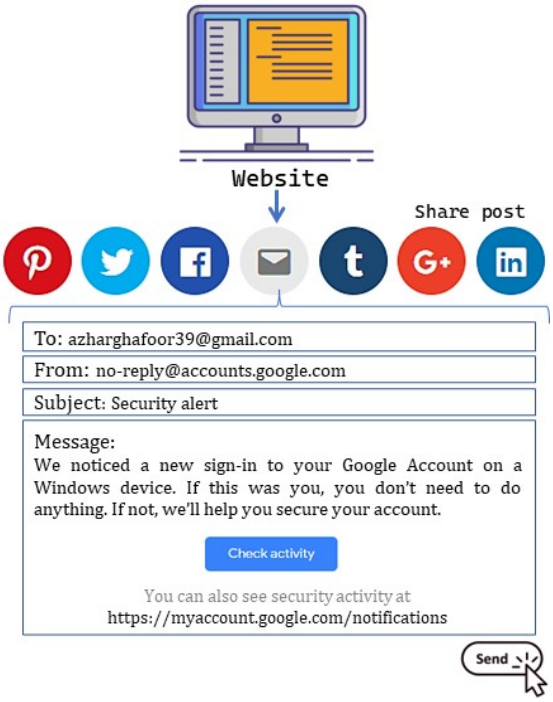


Fig. 5. Foolproof email spoofing to send phishing emails

Misconfigured websites that do not consider vulnerable sides while integrating such attributes are not only harmful for their businesses but also for other users. If websites belonging to governments have such vulnerable attack paths where attackers can easily send spoofed emails to the officials from their websites, then loss may be unbearable and in the worst cases may cause reputational damage.

Phishing comes in various ways and has fatal consciences so one must pay attention to these attacks. Organizations must go for proper penetration testing from experienced red teamers to discover and mitigate such vulnerable attack paths. Education is also a prime factor in the success of these attacks, so more attention is needed on the training side. Different studies have shown proven results of such investments as the number of attacks was reduced to few. Various phishing detection security controls must be purchased and added to the emails and these kinds of attacks could be stopped or lessened somehow. Another solution could be whitelisting senders to allow only a limited number of users to send emails, but it is not an appropriate solution because you have to deal with your customers.

2.2 Spoofed Domain

Intruders create a domain that appears to be genuine but is a clone of the original. They might, for example, use this to create a clone of the original site and send bogus emails to catch

victims [31]. By providing the bogus URL to ad exchanges, they are misled into paying for space on the spoofed site rather than the real site.

Fig. 6. How attackers spoof domains

Hackers spoof domains by developing a realistic-looking phony website to trick users into thinking it is the authentic domain of well-known companies or personalities. They develop a duplicate domain that is so convincing that no one can tell it's not a real domain at first sight. They utilize a double "v" instead of a "w" or a "I" instead of a "l" to make it harder to be differentiated from the genuine one. In some cases, they simply change the TLDs (top-level domain), such as from ".net" to ".com" etc. As a result, when something is shared with users from these sites, they are easily duped and, in most cases, open attachments or click on URLs. These faked domains are the primary source of propagating malware, trojans, or creating bot networks by establishing a permanent connection in response to visitors clicking on malicious links or downloading attachments, resulting in a DDoS attack.

Domain spoofing is also used to carry out additional assaults, such as launching a phishing campaign, exchanging malicious documents, or requesting individuals to reveal their credentials, such as enticing them to obtain a reward by entering the malicious webpage, and so on. This is evident in Fig. 5, which depicts how this attack vector makes the attack more lethal and increases the likelihood of success of attacks. Although these are very hard to detect, some precautions could help in stopping them. One must carefully observe the domain name by hovering the mouse on the links that are sent in an email before clicking on them, should open attachments in a sandbox, and check whether domain names are real, or they have something changed in them, for example, attackers may use 'Í' instead of 'l' etc. Users must also check email headers to see whether the person claiming to be the sender is a real or spoofed identity. They must also make sure that links do not lead to subdomains or any other websites. In Chrome and Brave browsers, there is a padlock in the address bar, if it is green then the link you are visiting is secure, if there is a red crossing line over the lock then you must not trust it as it is not a secure site and may lead to harmful pages.

1) Domain Spoofing to Create Spoofed Emails: When an attacker uses a legal website's domain to set up a phony email account, this is known as email spoofing. Phishing attempts frequently make use of email spoofing as a tactic. Using a fake domain name, an attacker can fool users into believing that phishing emails are genuine. In general, an email that appears to be from someone in the company is more reliable than one from an unidentified third party. The goal of the phishing attempt could be to persuade users to visit a certain website, download malware, open a dangerous email attachment, enter account credentials, or transfer funds to an attacker-controlled account. It is not uncommon for emails to contain links to fraudulent websites that require the login and password of the targeted account to be obtained through website spoofing.

2) *Domain Spoofing for Ads:* To conceal the true source of traffic to their websites, ad fraudsters duplicate the domain names of the websites they manage and then auction them off with the help of advertisers. As a consequence, the display ads appear on a website that is less suited than the one specified by the marketers.

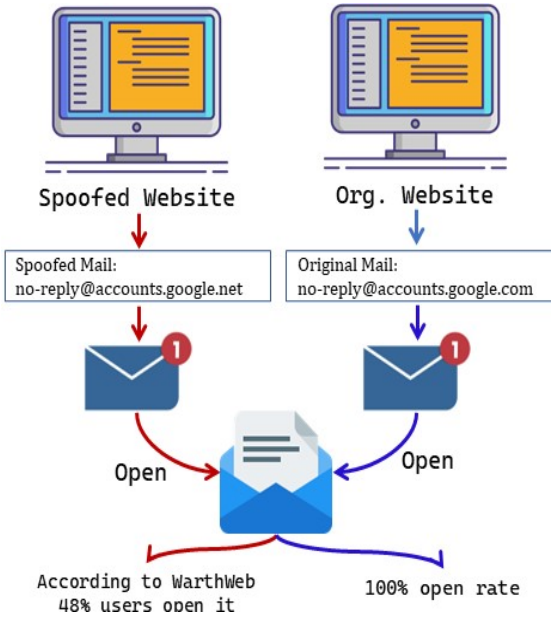


Fig. 7. Spoofed domain-making phishing attempts are more successful

2.3 Phishing Documents

According to Palo Alto Networks during the years 2019 - 20, a huge increase of 1160% was recorded in the use of PDF files containing malicious code hidden in them. They were being transferred using different social media platforms such as LinkedIn, the top-used brand, and the other prime sharing mediums were email, malicious links, or licks on embedded links in those files. PDF files have been observed as the most interesting and luring attack vectors as they

work in multiple platforms irrespective of the type of operating system.

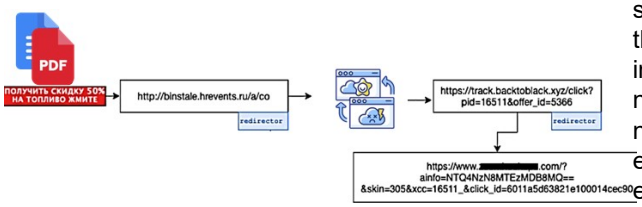


Fig. 8. Document-based attack chain

Malicious documents do not only come from PDF files, but they also range into multiple classes such as Word documents, Excel sheets, Images containing hidden links, Audio files etc. Each type of document has its benefits and drawbacks, but still, they are highly useful for attackers in abusing the vulnerable paths. Attackers also know only a small fraction of users update their office suite, so if any hack has been released it becomes handy to exploit them on unpatched suites. Microsoft Office utilities by default have options of adding macros where attackers usually place malicious code that can easily bypass security controls such as any antivirus or XDRs. If they as AV-bypassed it means they can run without any restriction and can do what they are crafted. In most cases, attackers try to have a reverse shell from the victim to maintain a persistent connection to have better control and know about the personal data present in the system [32].

Attackers craft such a malicious document that when a user simply opens the document, the code embedded starts its execution in the background. Recently, a zero-day was discovered in a Microsoft Word application that enables hackers to gain access to victims without the need for any actions. Researchers from Huntress have validated the most recent zero-day exploit, which exploits the diagnostic tool via an infected Microsoft Word document. The standard security alerts are not generated because the malicious file does not require macros. If the infected document is in RTF format, the script executes without the file needing to be opened in the Preview Tab of Internet Explorer. Instead, MSDT is used to pre-load the file if someone clicks or hovers over the payload to activate it. Further, if an attacker exploits this weakness correctly, it can execute a malicious script with the caller application's permission.

Antivirus must never be disabled. The auto-update option must also be enabled to automatically install new patches. As we cannot see what this document does by simply hovering over the mouse, so directly opening the documents, particularly from unknown senders must be prohibited. We should submit these documents to any of the freely available sandboxes such as cuckoo, cert, or virus total. If the file is safe, you can open it, but always keep in mind that if the content of the file does not need to be updated, deleted, or somehow needed to be copied, then do not click on 'enable editing' or 'enable content' options. Seminars and events must be organized by the organizations to guide them about new scam methods and precautionary measures.

4. *Effective Countermeasures*

Most phishing attempts are effective because they are difficult to detect by both users and security systems. Even though hackers are finding new ways to get around security systems, there are still ways to secure ourselves, our data, and our organizations [33]. Here are a few mostly used security aspects suggested by top cybersecurity researchers that help in avoiding phishing attacks and spoofing.

Fig. 9. Anti-phishing solutions

4.1 Email Protection

Secure Email Gateways are the first line of defence against phishing, removing potentially damaging and malicious emails from user mailboxes and isolating them. A good email gateway filters out potentially hazardous hyperlinks and attachments, and also 99.99 % of junk mail. As a result, they perform a significant role in preventing phishing emails from reaching clients. Email gateways also inform organizations when accounts are breached, preventing assaults on business email accounts and the use of hacked accounts to send misused or phishing emails to enterprises. Cloud email security safeguards mailboxes from intruders by utilizing machine learning and artificial intelligence to detect such messages. Furthermore, they use antivirus to scan and identify email threats. On detecting any harmful email, they trigger warning flags on those emails to alert users that they may be harmful or will delete them from the network

based on administrator-defined policies.

4.2 Protection Against Web Spoofing

Web filtering is a technique that effectively prevents customers from visiting websites flagged as phishing or spam websites. Many companies provide intelligence about such websites as they have developed various artificial intelligence-based models that help them know the maliciousness of the sites by examining their pages against many parameters. Using this information, organizations can make policies to prevent users from visiting such websites and submitting their important details. It is also real that stopping users from visiting harmful websites is crucial for organizations as they mostly use VPNs or proxies to bypass the restrictions. Users mostly visit such websites to download paid software for free, to download cracks, download other kinds of prohibited content such as videos, audio, or books etc. On such websites, there are many ads, and they unknowingly click on them, and this thing usually opens a new tab leading to some information stealing software or displaying some unpleasant content hosting websites. So, to overcome all such scenarios advanced web filtering systems must be used as they perform both static and dynamic analysis of URLs and attachments to search websites for phishing indicators, even if they may not contain malicious content.

4.3 Simulated Phishing Campaigns

To combat phishing attempts, it is essential to test employees' ability to differentiate between legitimate and bogus emails. Administrators can use this information to figure out how dangerous phishing is for their whole organization and focus education efforts on the areas that need it most. It is not uncommon to find a platform that allows users to design and distribute their phishing-style email campaigns. Many of these businesses also offer security awareness education to enable their customers to spot phishing emails. Administrators, for example, can replicate phishing attempts on different target groups and assign varying degrees of difficulty to each group. Users who fail tests often should be easy to find and track down based on how often they fail [34].

5. CONCLUSION

Phishing remains one of the greatest threats to individuals and organizations in the public and private sectors. Gateway attacks can lead to identity theft, ransomware attacks, and denial-of-service attacks. Unfortunately, the popularity and effectiveness of phishing are influenced by poor decisions, illiteracy, and lack of attention to detail of individuals. This paper gives an overview of the phishing problem and introduces the motives behind phishing and common attack vectors used in phishing attacks. We discussed different previously undiscovered attack vectors, and their implications, and provided solutions. This study was primarily concerned with identifying areas that an attacker could exploit by adopting their mindset. We performed various experiments to prove whether these attack vectors are important to discuss or not, and our findings proved to be right. In future research, we will try to explore in depth how these attack paths are being exploited by hackers and will try to provide a detailed comparison between various adversarial approaches.

REFERENCES

[1] G. Kavallieratos and S. Katsikas, "Attack path analysis for cyber-physical systems," in *Computer Security*, 6th ed., New York, USA, 2020, pp. 19-33.

[2] I. Stellios, K. Mokos, and P. Kotzanikolaou, "Assessing smart light enabled cyber-physical attack paths on urban infrastructures and services," in *Connection Science*, vol. 34, no. 1, 2022, pp. 1401–1429.

[3] K. Owen and M. Head, "Motivation and Demotivation of Hackers in Selecting a Hacking Task," in *Journal of Computer Information Systems*, 2022, pp. 1-15.

[4] A. Matta, G. Sucharitha, B. Greeshmanjali, M. P. Kumar, and M.N.S. Kumar, "Honeypot: A Trap for Attackers," in *Artificial Intelligence and Industrial Internet of Things Paradigm*, 2022, pp. 91-101.

[5] A. Spagnolli, M. Masotina, A. Scarcia, B. Zuffi, and L. Gamberini, "How to get away with cyberattacks: An argumentative approach to cyberattacks' legitimization by common users," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1-12.

[6] R. Abdillah, Z. Shukur, M. Mohd, and M. Z. Murah, "Phishing Classification Techniques: A Systematic Literature Review," in *IEEE Access*, 2022.

[7] B. Hanus, Y.A. Wu, and J. Parrish, "Phish me, phish me not," *Journal of Computer Information Systems*, 2021, pp. 1-11.

[8] S. Goel, K. Williams, and E. Dincelli, "Got phished? Internet security and human vulnerability," *Journal of the Association for Information Systems*, vol. 18, no. 1, 2017, p. 2.

[9] Verizon, "Data Breach Investigations Report," [Online]. Available: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf. [Accessed: 22-09-2021, 07:12 pm].

[10] Wombat Security, "State of Phishing Report 2018," [Online]. Available: https://info.wombatsecurity.com/hubfs/2018.StateofthePhish/Wombat-StateofPhish2018.pdf. [Accessed: 22-09-2021, 07:25].

[11] Symantec, "Internet Security Threat Report Volume 24," 2019, [Online]. Available: https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf. [Accessed: 22-09-2021, 08:00 pm].

[12] D. Aljeaid, A. Alzhrani, M. Alrougi, and O. Almalki, "Assessment of End-User Susceptibility to Cybersecurity Threats in Saudi Arabia by Simulating Phishing Attacks," in *Information*, vol. 11, no. 12, 2020, p. 547.

[13] P. N. Mangut and K. A. Datukun, "The Current Phishing Techniques–Perspective of the Nigerian Environment," in *World Journal of Innovative Research (WJIR)*, vol. 10, no. 1, 2021, pp. 34-44.

[14] J. Seymour and P. Tully, "Weaponizing data science for social engineering: Automated E2E spear phishing on

Twitter," [Online]. Available: https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf. [Accessed: 23-10-2021, 08:00 pm].
[15] J. Nelson, X. Lin, C. Chen, J. Iglesias, and J. J. Li, "Social engineering for security attacks," in *Proceedings of the 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016*, New York, NY, United States, 2016, pp. 1-4.
[16] N. Park, K. Sun, S. Foresti, K. Butler, and N. Saxena, "Security and Privacy in Communication Networks," in *Springer SecureCom2020*, Washington DC, USA, 2020, pp. 21-23.
[17] O. Christou, N. Pitropakis, P. Papadopoulos, S. McKeown, and W. J. Buchanan, "Phishing URL detection through top-level domain analysis: A descriptive approach," arXiv preprint arXiv:2005.06599, 2020.
[18] R. Parthiban, V. Abarna, M. Banupriya, S. Keerthana, and D. Saravanan, "Web Folder Phishing Discovery and Prevention with Customer Image Verification," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, Pondicherry, India, 2020, pp. 1-5.
[19] H. Gascon, S. Ullrich, B. Stritter, and K. Rieck, "Reading between the lines: content-agnostic detection of spear-phishing emails," in *International Symposium on Research in Attacks, Intrusions, and Defenses*, Springer, Cham, Donostia / San Sebastian, Spain, 2018, pp. 69-91.
[20] A. B. Ajmal, M.A. Shah, C. Maple, and M.N. Asghar, "Offensive security: Towards proactive threat hunting via adversary emulation," *IEEE Access*, vol. 9, 2021, pp. 126023-126033.
[21] D. Pienta, J.B. Thatcher, and A. Johnston, "Protecting a whale in a sea of phish," in *Journal of Information Technology*, vol. 35, no. 3, 2020, pp. 214-231.
[22] A. Bhardwaj, V. Sapra, A. Kumar, N. Kumar, and S. Arthi, "Why is phishing still successful?," *Computer Fraud Security*, 2020, vol. 2019, pp. 15-19.
[23] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," in *Frontiers in Computer Science*, vol. 3, 2021, p. 6.
[24] P. Unchit, S. Das, A. Kim, and L.J. Camp, "Quantifying susceptibility to spear phishing in a high school environment using signal detection theory," in *International Symposium on Human Aspects of Information Security and Assurance*, Springer, Cham, July 2020, pp. 109-120.
[25] A. Aleroud, E. Abu-Shanab, A. Al-Aiad, and Y. Alshboul, "An examination of susceptibility to spear phishing cyber attacks in non-English speaking communities," *Journal of Information Security and Applications*, vol. 55, 2020, p. 102614.
[26] A. AlEroud and G. Karabatis, "Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks," in *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, 2020, pp. 53-60.
[27] P.N. Mangut and K.A. Datukun, "The Current Phishing Techniques–Perspective of the Nigerian Environment," *World Journal of Innovative Research (WJIR)*, vol. 10, no. 1, 2021, pp. 34-44.
[28] M. Bossetta, "The weaponization of social media: Spear phishing and cyberattacks on democracy," in *Journal of International Affairs*, vol. 71, no. 1.5, 2018, pp. 97-106.
[29] V. Zolotarev, E. Zolotareva, and V. Mawla, "Phishing Attacks Digital Trace Analysis for Security Awareness," 2022.
[30] T. Wood, V. Basto-Fernandes, E. Boiten, and I. Yevseyeva, "Systematic Literature Review: Anti-Phishing Defences and Their Application to Before-the-click Phishing Email Detection," arXiv preprint arXiv:2204.13054, 2022.
[31] R. G. Atkinson et al., "U.S. Patent No. 7,398,315," Washington, DC: U.S. Patent and Trademark Office, 2008.
[32] J. Jiang et al., "Detecting malicious PDF documents using semi-supervised machine learning," in *IFIP International Conference on Digital Forensics*, Springer, Cham, February 2021, pp. 135-155.
[33] A. Sadiq et al., "A review of phishing attacks and countermeasures for the internet of things-based smart business applications in industry 4.0," in *Human behavior and emerging technologies*, vol. 3, no. 5, 2021, pp. 854-864.
[34] W. Yeoh et al., "Simulated phishing attack and embedded training campaign," *Journal of Computer Information Systems*, 2021, pp. 1-20.

Azhar Ghafoor, part of the Department of Cybersecurity at Air University, is a seasoned professional with 1.5+ years at Cytomate Solutions and 1.5 years as a Cybersecurity Researcher. Armed with a master's in information security, his focus spans compliance, malware analysis, computer-human interaction and knowledge-sharing through writing and freelancing. Email: azharghafoor39@gmail.com

Munam Ali Shah received the B.Sc. and M.Sc. degrees in computer science from the University of Peshawar, Pakistan, in 2001 and 2003, respectively, the M.S. degree in security technologies and applications from the University of Surrey, U.K., in 2010, and the Ph.D. degree from the University of Bedfordshire, U.K., in 2013. Since 2004, he has been working as an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Pakistan. He has been included in Stanford's list of the top 2% of scientists. Email: mshah@comsats.edu.pk

Bilal Zaka is an experienced IT professional, academic manager and researcher; presently Head of IT Services at COMSATS University Islamabad Pakistan. He also provides consultancy services to the Higher Education Commission of Pakistan as a member of various technical committees. Bilal's current research interest is focused on the use of artificial intelligence to enhance capabilities of conventional information systems and unlock the value of structured and unstructured data. He did his PhD in Informatics from the Graz University of Technology - Austria, and an MSc in Electronics from Quaid-e-Azam University Islamabad Pakistan. E-mail: zaka@comsats.edu.pk

Muhammad Nawaz obtained his PhD from Cardiff Metropolitan University, UK, in 2021. He is currently serving as a manager at COMSATS University Islamabad Campus. His main areas of interest and research include Distance Learning, Hybrid Learning, and Technology-Enabled Learning. With extensive experience in the field of educational technology, Dr. Nawaz is dedicated to advancing the effectiveness and accessibility of learning through technological solutions. His work focuses on improving educational outcomes and enhancing the learning experience for students in various educational settings. Email: Nawaz@vcomsats.edu.pk

Deep Acoustic Modelling for Quranic Recitation – Current Solutions and Future Directions

Shakeel, Muhammad Aleem; Khattak, Hasan Ali; and Khurshid, Numan

Abstract: ***The Holy Quran has the utmost importance for the Muslim community, and to get a full reward, the Quran should be read according to the rules mentioned. In the past few years, this field has gained a lot of importance in the eyes of researchers who aim to automate the Quranic reading and understanding process with the help of Machine Learning and Deep Learning, knowing it has a lot of challenges. To date, there are a lot of research categories explored. However, still, there lacks a few holistic, including one detailed survey of all the categories and methodologies used to solve problems. We focused the paper on being a one-stop-shop for the people interested so they could find (i) all related information and (ii) future gaps in research. This paper provides a detailed survey on Deep Modeling for Quranic Recitation to address these challenges. We discussed all possible categories of speech analysis, including the most advanced feature extraction techniques, mispronunciation detection using Tajweed rules, Reciters and speech dialect classification, and implementation of Automatic Speech Recognition (ASR) on Quranic Recitations. We also discussed research challenges in this domain and identified possible future gaps.***

Index Terms: ***Speech Analysis, Feature Extraction, Mispronunciation Detection, Tajweed, Reciter Classification, Automatic Speech Recognition, Deep Learning***

1. INTRODUCTION

FOR millions of Muslims around the world, the Quran is of utmost importance as their Holy Book. The poetic recital in Quranic verses, known as “Tilawah,” is not only a means of promoting spiritual enlightenment but also a profound art that mesmerizes listeners with its rhythmic flow. The accurate recitation and its preservation have always been of the utmost significance to the Muslim world. Quranic recitation has traditionally been transmitted orally, preserving the art's most basic form. However, new options have opened to improve our comprehension of Quranic recitation and bring it to a broader audience using machine learning and deep learning.

Because of its utmost significance in academics and research, researchers have continually investigated new approaches to improve the Quranic recitation experience and support its accurate preservation and meanings. The use of deep acoustic modeling techniques for Quran recitation is one topic that has recently attracted an immense amount of attention. Speech recognition, natural language processing, and deep learning have succeeded in many areas. In speech recognition and voice analysis, deep acoustic modeling uses complex neural networks for processing and evaluating acoustic data. Researchers aim to better understand the art of "Tilawah" by using these cutting-edge techniques for Quranic recitation. They also seek to increase the accuracy and robustness of current Quranic recitation recognition systems.

This review aims for some understanding of the advancements made in the fields of Automatic Speech Recognition (ASR), contextualized classification in Quranic topics, improvements in the reading of the Quran using Deep learning, feature extraction techniques and their comparisons, reciter classification, Tajweed, Hijayah, Makhraj correction and classification based on the context as well.

The objectives of this systematic literature review are as follows:

- Present a comprehensive and current overview of the literature on deep acoustic modeling for Quranic recitation.
- Describe the approaches, system models, datasets, and evaluating criteria used in the State-of-the-Art Research alongside their merits and shortcomings.

Table 1: Existing Surveys Related to Deep Modeling in the Quran

Year	Ref.	Topic(s) of the survey	Primary findings of the survey
2018	(1)	Semantic Ontology for Quranic Knowledge	Analyze different ontology methods for the Quran and highlight the gaps.
2020	(2)	Text Classification in Arabic Language	Highlights different deep learning models that show the best accuracy for Arabic classification.

2022	(3)	Speech Genres in the Quran	Highlights the kind of evidence that researchers should focus on while investigating genres and explains mistakes and problems that should be considered.
2023	(4)	NLP for Quranic Research	NLP serves as a synthesis compendium of works that span speech recognition-based Qur'anic recitation correction to computerized morphological evaluation

- evaluate the efficiency and performance of deep learning models for Quranic recitation assessment compared to more conventional methods.
- Highlight challenges in existing State-of-the-Art research and propose new directions in this domain.

1.1 Existing Reviews

Table 1 presents a comprehensive overview of existing review papers in the domain of Quranic knowledge, with a focus on topics such as ontology methods, text classification, speech genres, and the application of Natural Language Processing (NLP). Notably, these surveys broadly cover Artificial Intelligence without a specific concentration on a singular research domain. Through our critical analysis, we identified a significant research gap, particularly the lack of emphasis on acoustic modeling.

This observed gap was a primary motivation for undertaking the current survey paper. Our objective was to address this limitation by focusing exclusively on the realm of acoustic modeling and its application in the recitation of the Quran. By narrowing our scope to target researchers actively engaged in the workings of acoustic modeling and Quranic recitation, we aimed to provide a specialized and in-depth exploration of this specific research domain. During our literature review within this focused domain, we observed a scarcity of recent and up-to-date surveys. Existing surveys in this area were either outdated, unable to capture the emergence of recent works, or failed to address evolving issues within the field. This realization further fueled our motivation to contribute a timely and comprehensive survey that not only bridges the identified gap but also offers insights into the latest developments in the field of acoustic modeling for Quranic recitation.

. State-of-the-art research investigates how to deal with this domain, as Rusli et al. (1) presented a semantic ontology for Quranic knowledge. They have delivered a detailed systematic review of how available Quranic ontology models are limited to domains like nouns, subjects, pronouns, antonyms, and Islamic knowledge in the Quran because they do not account for all concepts in the Quran. To give an in-depth evaluation of this field, the research seeks to find relevant research works from various electronic data sources. Their study thoroughly evaluated the literature pertinent to the current ontology models to spread an accurate understanding of the Quran utilizing semantic technologies.

Wahdan et al. (2) presented a text classification for the Quran and Arabic language using deep learning models. They have focused on text classification techniques based on deep learning, including CNN, RNN, LSTM, etc. They have thoroughly analyzed system models of 12 research papers related to the topic along with their accuracies and results, showing which model could work best for the purpose. They have also suggested which models to use to improve text classification.

Devin et al. (3) presented different approaches to finding different speech dialects in Quranic Recitation. They have provided readers with basic guidelines for interpreting Qur'anic passages, emphasized the kinds of evidence that researchers should concentrate on while looking into genres/dialects, and discussed errors and pitfalls that should be considered in subsequent studies. Authors have suggested that distinctive words, phrases, and structures must all be carefully examined. Their debate emphasizes how specific dialect texts are incorporated into Surahs or more extended parts within Surahs. It demonstrates how the Qur'an references pre-existing categories and alters and transforms them.

Huzaifa et al. (4) presented how NLP can be used for Quranic research while focusing on Quranic commentaries and exegesis. They have induced NLP with speech recognition to improve Quranic recitation and showed that NLP methods aid in creating tools that make it easier for regular people to learn new things. Their studies provide an overview of the many Qur'anic NLP initiatives and serve as a synthesis compendium of works spanning the spectrum from automated morphological examination to speech recognition-based Qur'anic recitation correction.

A comprehensive review of Deep Modeling for Quranic recitation will help the research community understand these concepts. However, the surveys mentioned in Table 1 still lag and face challenges researchers must investigate. However, these surveys mainly focused on NLP, Text Classification, and Semantic Ontology. That's where this paper comes in

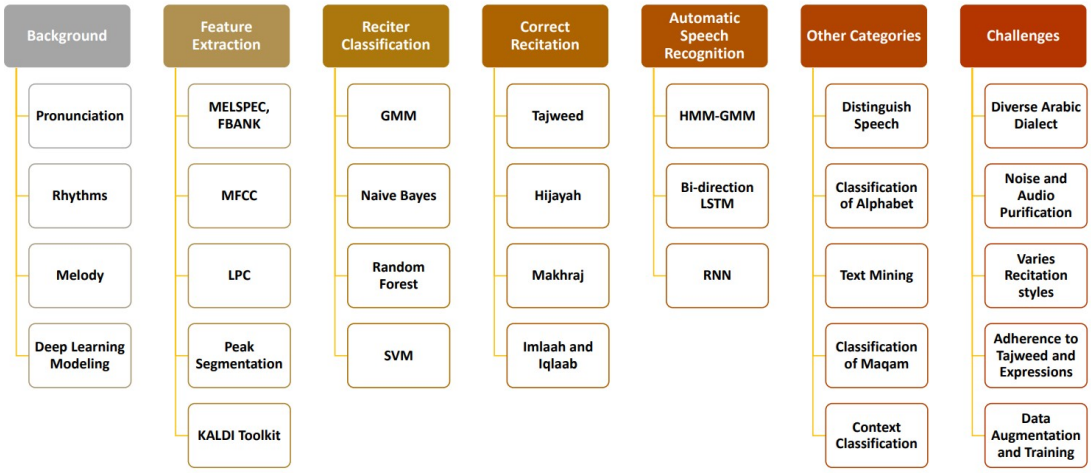


Figure 1: Article Organization for Literature Review of State-of-the-Art Research

to emphasize acoustic modeling and speech analysis in Quranic recitation. The sole focus of this review paper is to find research gaps and challenges the research community faces in speech recognition and Deep Learning modeling.

1.2 Scope and Contribution

The goal of this systematic review is to thoroughly evaluate the existing research using deep learning methods in the field of Quranic recitation. Significant contributions are made to deep acoustic modeling for Quranic recitation. The following are just a few of the significant contributions:

- We present a thorough and integrated overview of the current state-of-the-art deep learning techniques to analyze Quranic recitation by methodically studying a wide range of research articles.
- We discover new patterns and prospective fields for further research through analysis. Future researchers will gain insightful knowledge from this assessment of research gaps, which will help them develop state-of-the-art research ideas.
- The methodology, statistics, and assessment metrics used in the papers are rigorously evaluated. This assessment will aid in comprehending the benefits and drawbacks of various approaches and offer suggestions for enhancing research methodologies.

The following points suggest a need for a new survey or review despite numerous survey papers in the same field by various authors.

- Our survey paper was focused on defining the techniques and methodology used by different authors for a specific field called “Deep Acoustic Modelling for Quranic Recitation” which eventually means that we focused on the SOTA papers that use only “audio samples” for training of Quranic data, which ultimately means that the survey paper is specifically for the target audience who are interested in audio data training of Quran.
- We categorize the paper based on different schemas and techniques such as Correct Recitation analysis, Tajweed, Makhraj, Hijayah, Imlaah, Automatic Speech Recognition, and some other categories that create a single one-stop-shop for the future researcher to study the relevant SOTA topic they are interested in and search the findings of last ten years.
- We identify the research gaps in every single paper mentioned in this survey and help future researchers define their future research ideas in the specific domain of acoustic modelling. Ultimately, we conclude with a research gap that hasn’t been touched in this domain and can create a new research stream and change the research trend for future researchers.
- We mentioned the research papers that answer the following questions:
  - Have they come up with a new acoustic modeling technique?
  - Is there a data set available online for future researchers?
  - Do they compare their techniques with the rest of the algorithms?

1.3 Article Organization

This manuscript classifies different research topics for organizing Quranic recitation, as shown in Figure 1, and conducted a thorough survey on each topic. The article's organization is divided into the following categories: Section 2 defines deep acoustic modeling for Quranic Recitation. Section 2.1 describes the primary feature

extraction techniques used for the classification and their results to prove which state-of-the-art process works best for the related topic, along with their survey and results comparisons. Section 2.2 represents the different reciters' classification techniques used by the researchers to classify the reciters' voices and is currently an active research topic. Section 2.3 represents a thorough Tajweed, Hijayah, and Makhraj classification and correction of mistakes techniques using deep learning to improve the reading of the Quran without any errors. Section 2.4 represents the Automatic Speech Recognition used for Quranic Recitation, including Hybrid HMM BLSTM-based modeling and End-To-End Transformer modeling, and Section 2.5 describes the classification of different artifacts of Quran recitation, including Feature Identification on both acoustic and textual data and classifying different Maqams of Quranic Recitation.

2. BACKGROUND AND LITERATURE ON MODELING OF QURANIC RECITATION

Recitation is more than just reading the context of the Quran; it follows standards and rules guiding pronunciation, rhythm, and melody. Deep neural networks, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), are trained to recognize and capture the complex auditory patterns unique to Quranic recitation in deep acoustic modeling for Quranic recitation. Recording Arabic phonemes precisely and following Tajweed guidelines for proper pronunciation is necessary. The model also emphasizes intonation and melody to simulate pitch fluctuation and some syllables' lengthening in Quranic recitation accurately. Deep acoustic modeling for Quranic recitation's ultimate goal is to develop a technologically enhanced tool that makes it easier to learn and master the complex art of recitation.

2.1 Feature Extraction

Raw audio signals are frequently multidimensional and packed with information. While maintaining crucial qualities necessary for the application, feature extraction assists in reducing the dimensionality of the data.

Abdo et al. (5) presented an algorithm for automatically segmenting between emphatic and non-emphatic Arabic speech from Arabic audio signals. The study mainly focused on the recitation principles of the Holy Quran. The methodology was to extract important features of Arabic sound signals using the Mel Frequency Cepstral Coefficient (MFCC), find the peak position for boundaries in the target signal, and then evaluate the whole system on medium-level speech. They have created the database for all these signals and evaluated the system on 80 Arabic-recited words. The dataset consists of Arabic recited words from 6 different speakers, making the testing dataset 480 Arabic words. The segmentation accuracy achieved by the system is about 90%. For future work, more constraints could be added to the MFCC peaks to make them more efficient, or different feature extraction techniques, such as spectral envelopes and formant frequencies, can be implemented independently or combined to increase the model's efficiency.

Even though The Holy Quran consists of the same verse all over the world, the recited poem probably is different from the other person who repeated the same verse because of the distinct voice of every person. Bezoui et al. (6) proposed a technique to train and test the Arabic speech system using the KALDI toolkit. The author explored the viability of different feature-extracting methods for developing the system to extract important features from Quranic Recitations, including the Mel-Frequency Cepstral Coefficient (MFCC). The author explained in detail the MFCC technique, including all steps, such as preprocessing, framing, windowing, etc. The maximum efficiency achieved for the system is 75% using the Hamming Window technique and 55% for the rectangular window. The dataset used for the purpose includes audio files of Quranic Verse. However, the work can be improved by implementing the fixed-range sliding window technique and extracting the MFCC feature for every windowing signal.

Meftah et al. (7) compared different feature extraction techniques to achieve the highest accuracy for Arabic Phonemes classification. For this purpose, a dataset corpus has been created for other Arabic recitations of the Holy Quran, and then acoustic features are extracted from it. These features include Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Mel-filter bank coefficient (MELSPEC), Log Mel-filter bank coefficient (FBANK), and Linear Prediction Reflection Coefficients (LPREFC). After feature extraction, Hidden Markov Model (HMM) classification has been used. The result shows that FBANK and MELSPEC features produce the highest accuracy for the system, i.e., 85.38% and 83.37%. Also, the MFCC and PLP results are close to this accuracy, while LPC is unsuitable for Arabic speech recognition.

Adiwijaya et al. (8) did a comparative analysis to classify the pronunciation of Hijaiyyah Letters using different feature extraction techniques. The dataset includes audio samples of Hijayyah letters and was analyzed using Mel Frequency cepstral coefficient (MFCC) and Linear Predictive Coding (LPC) and then classified using KNN. The proposed system was compared with Principal Component Analysis (PCA) and without PCA. It showed that LPC-KNN proved to be a better Hijayah classification technique with an accuracy

of 78.92% compared to 59.87% conducted by MFCC-KNN. Hence, LPC was verified to be a better feature extraction technique.

2.2 Reciters Classification

Classification of reciters for Quranic recitation aims to recognize and classify various Qaris (reciters) according to their distinctive recitation styles. This is a particular endeavor in the field of audio analysis.

Khan et al. (9) proposed a machine learning approach to recognize the reciter of the Holy Quran. The dataset of 12 different reciters reciting the last ten surahs of the Quran has been used, which means the model has 12 classes to classify. Two types of approaches have been used for audio representation. First is feature extraction using MFCC and the pitch of the sound. The second is auto correlograms of audio spectrograms. Then, implement Naive-Bayes, J48, and Random Forest for classification. Naive-Bayes and Random Forest achieved the maximum accuracy of 88%.

Munir et al. (10) proposed another feature extraction technique for Speaker Identification in Quranic Surah. To extract essential features, the author used a combination of Discrete Wavelet Transform (DWT) and Linear Predictive Coding (LPC) and then performed classification using Random Forest (RF). The system achieved the maximum accuracy of 90.90%. However, to improve the identification accuracy, they tried feature extraction techniques to be used one at a time and combined to train the classification model. But it does not affect the accuracy of the classifier. The dataset used for the purpose includes Arabic recitation of the Holy Quran. The research can be continued by implementing more than two feature extraction techniques and then classifying using machine learning techniques. However, the system is trained for Arabic recitation only. The work can be extended by introducing recitation in different languages as well. Elnagar et al. (11) proposed a supervised learning-based classification technique to classify the reciters in the Quran audio dataset. The system can identify the exact or closest reciter using machine learning techniques. The system was used to extract perceptual features from these audio data, which include the pitch, the tempo, short-time energy, etc. Then, it implemented the support vector machine (SVM) classifier. The model achieved an accuracy of 90%. The dataset used for the purpose includes Quranic audio of 7 reciters from Saudi Arabia, which specifies that the model can work on the Arabic dialect of Quranic speech. The work can be improved by implementing other feature extraction techniques or using some combinations. Qayyum et al. (12) proposed a deep learning technique using Arabic audio signals for speaker identification. The audio signals were analyzed and classified based on the speaker using Bidirectional Long Short-Term Memory (BLSTM), which proved to be a better and less computationally expensive technique for speaker identification. Gunawan et al. (13) developed an identification system for Quran reciters using MFCC and GMM. The dataset used for the purpose includes the Quranic recitation of 5 reciters and randomly selected verses of the Quran. Around 15 audio samples of each reciter have been collected and analyzed using the Mel Frequency Cepstral coefficient and then classified using the GMM classifier. The proposed system achieved 100% accuracy in identifying the reciter. Also, the system can reject unknown reciters rather than these five. However, the system can be extended by including variations of recited verses from different reciters on different Quranic Surahs.

2.3 Correct Recitation Analysis

Makhraj, Hijayah, and Tajweed deep learning-based Quran correction is a novel and technologically advanced method for improving Quranic text comprehension and recitation. For non-Arabic-speaking Muslims, reading the Quran is always a challenging task. Since many words in the Quran are written differently than they are read, To help parents solve the reading and pronunciation problems of dyslexic children, Basahel et al. (14) proposed a technique for developing an application in Android for supporting adaptive learning and self-paced learning. The application could convert the voice recognition algorithm into text to support E-learning and was only limited to processing single words, not complete sentences or texts. However, the application can be improved to train on texts and sentences. Ahmad et al. (45) suggested a method to identify mispronunciation in Tajweed rules using Mel-Frequency Cepstral Coefficient (MFCC) features with Long Short-Term Memory (LSTM) neural networks that use the time series. The QDAT dataset is available to the public and includes over 1500 voices reciting the three Tajweed rules. They compared the LSTM model with the traditional ML algorithms. The LSTM model outperformed formal machine learning with time series. LSTM's accuracy on the QDAT dataset for the three rules was 96%, 95%, and 96%, respectively.

The correct pronunciation and recitation of the Quran mainly depend on these four ideas:

2.3.1 Tajweed

Tajweed is a collection of guidelines for pronouncing and articulating Quranic texts correctly. It ensures every letter is said precisely, melodiously, and with the right rhythm, intonation, and elongation.

Ahsiah et al. (15) proposed a system for checking Tajweed and correcting the recitation of the Holy Quran. The Tajweed, a set of guidelines for Al-Quran recitation, ensures correct

pronunciation, readings, and text interpretations. Religious teachers with experience have traditionally imparted this knowledge. These instructors typically pay attention to the students' recitations and point out any errors. The traditional approach, which calls for the presence of these qualified professors, has limitations in enabling a self-learning environment. To assist students in learning and practicing accurate Al-Quran recitation independently, the author suggested a Tajweed rule-checking system employing speech recognition technology. The proposed method can detect and highlight discrepancies between student and experienced teachers' recitations that are recorded in a database. The system utilized the MFCC algorithm to extract features and HMM for classification.

Yosrita et al. (16) compared different methods of Tajweed used in the recitation of Al-Quran and extracted their features using MFCCs. Altalmas et al. (17) proposed a technique for correct recitation of the Quran according to the Tajweed rules. The words from the same point of articulation proved to have less similar distances or more matching sounds than words from different parts of articulation. To confirm this, the author analyzed the sound of words Y and I using the Mel Frequency Cepstral Coefficient (MFCC) and then compared them using the Dynamic Time Warping (DTW) technique to find the similarities and differences. The scope of this technique is limited to two Quranic words (Y and I) only. However, the content of the work can be extended by increasing the size of the dataset and by adding all Quranic words to find similarities and differences between all words.

Classic Arabic is very hard for non-native Arabic speakers, which makes it difficult for them to recite the Holy Quran. Short vowels in the Arabic language play an essential role in the correct Tajweed. Alqadheeb et al. (18) proposed a methodology for correcting Tajweed using an audio dataset of Arabic words, including short vowels. The complete dataset includes 2892 Arabic short vowels and 84 classes. Then, the preprocessing techniques and CNN are used for classification and testing. The model was tested on 312 phonemes of the Arabic language using "ALIF" as a word and achieved an accuracy of 100%. However, the system was de- signed to work on a single phoneme, "ALIF." In the future, we can extend the research to all the Arabic phonemes and train the model on them. Omran et al. (40) implemented Tajweed rules for correctly reciting and understanding the Quran. They analyzed the Arabic Alphabet's five sukun vowelized letters (Baa, Daal, Jeem, Qaaf, and Taa), subject to the Qalqalah rule. They used the Convolutional Neural Networks (CNN) model for recognition and the Mel Frequency Cepstral Coefficients (MFCC) as the feature extraction method. The dataset contains 3322 audio samples from four expert readers and achieved a validation accuracy of 90.8%.

Rajagede et al. (19) proposed a system to help users memorize the Quran without the help of a second reciter. He proposed a based system that verifies the input recitation with the existing Quranic data. Manhattan LSTM network was used to verify the recitation and give the output in a single numerical data if the recitation was similar or not, and the Siamese classifier gave binary classifier output. They also compared different feature techniques for the preprocessing, including delta features, Mel Frequency Cepstral Coefficient (MFCC), and Mel Frequency Spectral Co- efficient (MFSC) for better model performance. The dataset used for the purpose includes data from Quranic Ayah in databases. The highest accuracy achieved by the system is 77.35% using MFCC and Manhattan LSTM. However, in the future, to achieve better accuracy, it is recommended to use a deeper Siamese LSTM model or an attention-based model and use more data for training.

To improve the Quranic Recitation system, Alqadasai et al. (20) proposed a Phoneme classification system for the correct recitation of the Quran. The dataset consisted of 21 aayahs of the Quran recited by 30 reciters. The dataset was analyzed and trained using an HMM-based ASR model. The system is optimized by the duration integrated into Quranic phoneme classification. The system achieved an accuracy ranging from 99.87% to 100% for phoneme classification. However, the proposed methodology does not cover all the issues of recitation, so the extended version of the model could use more datasets to cover all the Quranic recitation and Tajweed issues.

Omran et al. (21) proposed a deep learning-based approach to correctly understanding Tajweed rules for the Holy Quran. Reading the Holy Quran precisely as it was read by The Holy Prophet (PBUH) is challenging. The author used the dataset of Quranic Audio from different Arabic reciters and focused on the letters on which Qalqala rules are applied. Mel Frequency Cepstral Coefficients (MFCC) were used for feature extraction, and then Convolutional Neural Networks (CNN) based model was used for classification. The author tends to achieve the maximum validation accuracy of 90.8%.

To encourage the Muslim community to read the Holy Quran with correct Tajweed and without any recitation error, Ahmad et al. (22) proposed a method to classify two ways of Tajweed, i.e., Musyafahah and Talaqqi, using Artificial Neural Networks and Digital Signal Processing techniques. The dataset includes audio files of Idghaam with correct recitation and false recitation. For the preprocessing of audio files, the Mel Frequency cepstral coefficient technique has been used to extract essential features and then classify them using three different ANN

classifiers, including Levenberg- Marquardt optimization, Resilient Backpropagation, and Gradient Descent with Momentum. The highest accuracy achieved by the system was 77.7 for the Levernberg Marquardt algorithm. The system can be improved if more classes of Tajweed methods are added. Also, the dataset used for the process is small; higher accuracy can be achieved by increasing the audio files for every class.

2.3.2 Hijayah

Hijayah involves making sure that the written script of the Quran corresponds to the intended pronunciation during recitation.

Marlina et al. (23) proposed the machine learning technique for Makhraj recognition of Hijayah letters. For this purpose, the author used the Mel Frequency cepstrum coefficient (MFCC) for feature extraction of audios, and then SVM was used to classify Hijayah letters. The dataset used for the purpose includes audio files of Hijayah letters. However, the result of the system can be improved using deep learning techniques such as ANN or CNN for the classification of Makhraj.

Correct pronunciation and writing of the Hijaiyah Letter in the Holy Quran are important for Muslims to follow the reading rule according to the rule of The Holy Prophet (PBUH). Irfan et al. (24) suggested a Dynamic Time Warping technique to find the difference between written and uttered letters. The dataset consisted of image files of Quranic letters and sound files of Quranic letters. To process image files, Principal Component Analysis (PCA) and Mel Frequency Cepstrum Coefficient (MFCC) were used for sound data. Both results were shown as numerical values, and then Euclidian distance was applied to find the difference between them. For image matching, the accuracy achieved for the system was 92.85%, and for sound matching, the accuracy achieved was about 71.42%. In the future, some other methods could be used for processing, such as Linear Discriminant Analysis, Edge Matching, etc., to achieve higher accuracy for the system. The application is limited to Hijaiyah Letters only, and we can extend it to words or phrases to match the difference between complete phrases.

2.3.3 Makhraj

Makhraj describes the posture of the tongue, lips, and vocal cords at the point of articulation for each Arabic letter. Proper Makhraj is essential for precise pronunciation.

Correct pronunciation of Hijayah letters in the Quran is crucial. It can confuse the user when pronouncing similar-sounding letters, and incorrect pronunciation may alter the word's meaning. The reciter should have an excellent knowledge of the differences between Hijayah letters. For this purpose, Wahidah Arshad et al. (25) addressed recognizing the nine-point recitation articulations using the Speech analysis technique. Professionals in a controlled environment recorded the dataset of Quranic Makhraj letters. The audio samples underwent preprocessing using five feature extraction techniques: MFCC, Mel spectrogram, Tonnetz, Spectral contract, and chroma. Three methods, ANN, KNN, and SVM, were used for classification. The system achieved an overall accuracy of 56% using the ANN technique. The system can be improved using Deep learning techniques such as CNN and RNN instead of traditional ANN, KNN, and SVM. Farooq et al. (26) proposed a deep learning-based model for detecting mispronunciation in the Quran. The purpose of the system was to automate the manual teaching method of the Quran, which typically requires a teacher or instructor. The dataset consists of Arabic audio recitation of Quranic words, and the RASTA PLP technique was used for feature extraction. The system was trained using the Hidden Markov Model (HMM). The recognition rate achieved using RASTA PLP is 85%. The system has been extended to build a real-time application. However, the plan was initially limited to a single word of Arabic phonemes, and it can be further developed to include Quranic Ayahs and different Tajweed rules, including Qalqala rules, Tanween, etc.

2.3.4 Imlaah and Iqlaab

Imlaah describes the lengthening of particular Arabic letters inside a word. In contrast to Iqlaab, which involves changing a specific letter, Noon, into a different sound (namely, the sound of "Meem") when followed by the Arabic letter "Ba," when a reciter encounters a letter with an Izhaar (clear pronunciation) diacritic, such as "Alif," "Waw," or "Yaa." The specific phonetic properties of these letters and how they interact in some word combinations cause these changes to happen.

Yousfi et al. (27) proposed a technique to recognize Imlaah rules for Quranic Recitation. Correct recitation includes Tajweed rules, which are essential when studying the Quran. For this purpose, the author used a dataset of proper Quranic audio recitation of verses and preprocessed it using feature extraction techniques such as the Mel Frequency Cepstral Coefficient (MFCC). Hidden Markov Models (HMM) were employed for classification and compared with correct recitation and recitation collected in real-time available in the database. The accuracy achieved by the system ranges from 68% to 85%. Yousfi et al. (28) proposed Iqlaab checking the rules of the Quran. For this purpose, the author built a speech recognition system to recognize, identify, and point out the wrong rules/mismatches of Iqlaab rules in recitation. The dataset was preprocessed using the feature extraction technique Mel Frequency

Cepstral Coefficient (MFCC), and for feature classification, Hidden Markov Models (HMM) were used. The system achieved a maximum accuracy of 70%.

2.4 Automatic Speech Recognition (ASR)

To fully understand the delicate nature of the sacred texts of the Quran, Automatic Speech Recognition (ASR) is essential. The approach of carefully preserving the phonetic intricacies and rhythm, particularly the recitation style of Quranic verses, is called acoustic modeling. ASR for Quranic recitation bridges the gap between spoken word and digital representation, preserving the authenticity of the recitation while also opening the door to cutting-edge applications that allow for correct transcription, analysis, and distribution of the Islamic message.

Most state-of-the-art research on acoustic signals for Arabic languages uses the Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM). However, there is a disadvantage in generalizing high-variance and solving non-linear separable datasets. For these problems, Thirafi et al. (29) proposed a new approach for training the acoustic models of the Arabic language using Deep Learning techniques. The author used Bidirectional Long-Short Term Memory (BLSTM) combined with Hidden Markov Models (HMM) to build a hybrid system. The system proved to show good results for the Arabic language compared to the HMM-GMM model. For HMM-GMM, the Word Error Rate (WER) was 18.39%, whereas for the proposed technique, WER was reduced to 4.63%. The author also analyzed the model of different Quranic styles. The system was trained on Quranic recitation by professional reciters only.

Table 2: Detailed review of research papers that shared dataset and system model and compare their model results with other techniques.

Title	Ref	Dataset	Speech/Text	System Model	Comparison
MFC peak-based segmentation for continuous Arabic audio signal	(5)	×	Speech	✓	×
Feature extraction of some Quranic recitations using Mel-Frequency Cepstral Coefficients (MFCC)	(6)	×	Speech	✓	✓
A Comparative Study of Different Speech Features for Arabic Phonemes Classification	(7)	×	Speech	×	✓
A comparative study of MFCC-KNN and LPC-KNN for Hijayyah letters pronunciation classification system	(8)	×	Speech	✓	✓
Quranic reciter recognition: A machine learning approach	(9)	×	Speech	✓	✓
Arabic speaker identification system using a combination of DWT and LPC features	(10)	×	Speech	✓	×
Automatic Classification of Reciters of Quranic Audio Clips	(11)	×	Speech	✓	✓
Quran Reciter Identification: A Deep Learning Approach	(12)	×	Speech	✓	✓
A Smart Flexible Tool to Improve Reading Skill based on M-Learning	(14)	×	Speech	×	×
Correct Pronunciation Detection for Classical Arabic Phonemes Using Deep Learning	(18)	×	Speech	×	×
Rule-Based Embedded HMMs Phoneme Classification to Improve Qur'anic Recitation Recognition	(20)	×	Speech, Text	✓	×
Tajweed Classification Using Artificial Neural Network	(22)	✓	Speech	✓	×
Makhraj recognition of Hijaiyah letter for children based on Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machines (SVM) method	(23)	×	Speech	✓	×
Implementation of Dynamic Time Warping algorithm on an Android-based application to write and pronounce Hijaiyah letters	(24)	×	Speech, Text	✓	×
Signal-based feature extraction for Makhraj emission point classification	(25)	×	Speech	✓	✓
Mispronunciation Detection in Articulation Points of Arabic Letters using Machine Learning	(26)	×	Speech	✓	×
Holy Qur'an speech recognition system Imaalah checking rule for warsh recitation	(27)	×	Speech	✓	×

Isolated Iqlab checking rules based on the speech recognition system	(28)	×	Speech	✓	×
An End-to-End Transformer-Based Automatic Speech Recognition for Qur'an Reciters	(30)	✓	Speech	✓	×
Holy Qur'an, speech recognition system, distinguishing the type of recitation	(31)	×	Speech	✓	×
Implementation of text mining classification as a model in the conclusion of Tafsir Bil Ma'tsur and Bil Ra'yi contents	(32)	×	Speech, Text	✓	×
Classifying maqams of quranic recitations using deep learning	(33)	×	Speech	✓	×
Noise effects on the recognition rate of Arabic phonemes based on Malay speakers	(34)	×	Speech	✓	×

The model could be extended by introducing non-professional reciters for better system performance. Additionally, the method used QScript for the transcription system, which was unsuited for the Quran. A better transcription system can be employed. Siregar et al. (41) created a system using wavelet signal extraction and ANFIS to classify Tajwid rules to handle voice input and recognize Al-Quran reading through recitation. Data collection, audio pre-processing, wavelet packet extraction, splitting training and test data, and classification are the steps in the process. Twenty observations were obtained from ten observations that were pre-processed. Six primary features and 64 rules are obtained from the wavelet decomposition method as ANFIS input variables. Subsequently, the data was divided into three testing observations and seventeen training observations. The WPANFIS classification model achieved 100% correct classification with SSE values matching the training result of 0.00081225.

Hadwan et al. (30) proposed an end-to-end system for Arabic ASR. They built an acoustic model using attention-based encoder-decoder techniques with deep learning. Mel filter has been used for feature extraction, and RNN and LSTM have been employed to build a sizable Arabic language model for the Quran. The dataset consisted of speech data from 60 reciters and textual Quranic data. The language model has been trained on textual data. The model achieved a character error rate of 1.9% and a word error rate of 6.1%. However, the model can potentially perform better by investigating better-proposed systems for large datasets or by implementing a transducer model instead of a transformer model for encoding and decoding using the employed model. Ghori et al. (42) use Mel frequency cepstral coefficients and deep neural networks to construct an Arabic isolated voice recognition system for the Holy Quran's vocabulary. The suggested system can recognize individual words from a recited verse with adequate accuracy. It employs 14 hours of audio data to showcase the system's functional prototype and focuses on the 362 unique words found in the first and last 19 chapters of the Holy Quran. They also developed a user-friendly web-based application for the transcription of recitation words.

2.5 Other Categories

The research of various recitation styles, the categorization of alphabets, the use of text mining tools, the classification of maqams, and context-aware analysis are a few intricate elements that make up the study of the Quran. Investigating the many ways to recite the Quran's verses, each infused with unique rhythms and accents that communicate significant meanings, is necessary to distinguish between different styles of recitation.

Ousfi et al. (31) proposed a technique to distinguish between different types of recitation of the Holy Quran, given the diversity in Qira'at worldwide. The dataset was created using recitations of other students and expert teachers. The Mel Frequency Cepstral Coefficient (MFCC) feature extraction technique was implemented, and the Hidden Markov Model (HMM) classification was applied. The system can also detect mismatches in the recitation type.

Khairuddin et al. (35) developed an automated system for students to practice reciting the Quran, focusing on the "ro" alphabet. Formant analysis, Mel Frequency Cepstral Coefficient (MFCC), and Power Spectral Density (PSD) were used for feature extraction of Quranic recitation. Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA) were used for classification. The system achieved a maximum accuracy of 95.8% with all 19 training features in repetition and 82.1% accuracy in the learning phase. Mahmudin et al. (43) compared how well the two types of models performed when categorizing Quran verses according to their auditory similarities. The first model, Model B, employs the MaLSTM architecture and MFCC features. The second model, Model C, is just Model B plus more delta features. The dataset includes 172,895 Al-Quran recitation sound samples from Juzz 30, comprising 37 surahs and 564 verses. The training model used was DeepSpeech, supported by TensorFlow. Thirty percent of the samples were used as a validation set during the model training procedure. The results show that Model B, equipped with the

MFCC feature, performs optimally when identifying and categorizing audio-based Quran verses. The employment of the delta feature in Models B and C negatively impacts model performance.

Nur et al. (32) developed an automated interpretation classification into two classes, including "Tafsir Bil Ra'yi" and "Tafsir Bil Ma'tsur." The KNN algorithm was used for variety and achieved an accuracy of 98.12%. Modified KNN and Fuzzy KNN were used for further comparison, with MKNN proving the best algorithm.

Shahriar et al. (33) proposed a system to classify eight recitation styles (Maqamat), including Tajweed using deep learning techniques. The dataset included audio recitation of the Quran by different reciters in different styles. The system achieved the highest accuracy of 95.7% using a 5-layer ANN network trained on 26 input features. However, the system could be improved by increasing the dataset using more reciters and implementing bidirectional LSTMs.

Moulay et al. (36) built an application framework for a context-based search of any Quran chapter or verse, which also applies to voice search. The voice-search feature allows users to search using their voice. The dataset consisted of Quranic chapter audios recited by 36 reciters, including eight famous interpretations and four translations of the Holy Quran. However, the framework has some limitations, including notifications based on locations using GPS systems that are lacking entirely. The framework could be improved by including Hadith knowledge that allows users to search for Islamic teachings. Ahmad et al. (44) suggested that the model consists of a character-based beam search decoder and a CNN-Bidirectional GRU encoder that uses CTC as an objective function. They used the recently released public dataset Ar-DAD, which consists of around 37 chapters repeated by 30 reciters using various pronunciation standards and recitation speeds. Word error rate (WER) and character error rate (CER), the two most widely used assessment metrics in speech recognition, were used to assess the performance of the suggested model. The outcomes were 2.42% CER and 8.34% WER.

3. *Open Research Challenges in Modelling of Quranic Recitation*

Building a robust, deep acoustic model specifically for Quranic recitation involves various complex issues requiring careful attention. The challenges in building the model are as follows, as shown in Table 3:

3.1 Diverse Arabic Dialects

Accurate transcription of pronunciation and intonation is complex due to the large variety of Arabic dialects and geographical differences. The intricate tapestry of many Arabic dialects and geographical accents must be considered while building a robust, deep acoustic model for Quranic recitation. This problem requires a thorough strategy incorporating the subtle phonetic variations in different recitation traditions. It takes careful language study and contextual adaptation to accommodate this dialectal diversity while staying true to the traditional recitation methods. Saddat et al. (37) utilized Naive Bayes classifiers, and the character n-gram Markov language model has achieved 98% accuracy on 18 different Arabic dialects.

3.2 Noise and Audio Purification

Background noise can reduce the accuracy of the model's output in audio recordings. The difficulty of maintaining the original quality of recordings of Quranic recitation arises from the widespread presence of noise inside audio data. Advanced noise reduction techniques must be used to capture deep vocal expressions while successfully overcoming audible interference. It is a complex but crucial endeavor to balance eliminating noise and preserving the evocative details of the recitation.

Almisreb et al. (34) proposed removing noise while recording recitation. The research evaluates the effectiveness of the

Table 3: Research Challenges in this domain, with Possible Solutions

Challenges	Ref	Description	Solution
Adherence to Tajweed and Expressions	(17)	The model must accurately distinguish between short and long vowel sounds to represent elongations (Madd) and vowel lengths.	I analyzed the signal using the Mel- Frequency Cepstral Coefficient (MFCC) and then compared them using the Dynamic Time Warping (DTW) technique to find the similarities and differences.
Noise and Audio Purification	(34)	Background noise affects the model's accuracy and robustness.	Multiscale Principal Component Analysis (MPCA) effectiveness with Zero Crossing Rate (ZCR) to remove the noise.
Diverse Arabic Dialect	(37)	The Arabic language includes a lot of recitation dialects that are difficult for the model to learn.	Utilizing Naive Bayes classifiers and the character n-gram Markov language model, it achieved 98% accuracy on 18 different Arabic dialects.
Varied Recitation Styles	(38)	Different reciters have distinctive rhythmic and melodic renditions,	The SVM-based recognition model for “Qira’ah” achieved a success of 96%.

		necessitating a flexible model.	
Data Augmentation and Training	(39)	Data augmentation is necessary to include diverse recitation conditions.	To apply augmentation to the filter bank coefficient directly, utilize warping the features, masking blocks of frequency channels, and masking blocks of time steps.

Multiscale Principal Component Analysis (MPCA) with Zero Crossing Rate (ZCR) to remove the noise. Then, the Mel-Frequency Cepstral Coefficient (MFCC) was used for feature extraction, and Dynamic Time Warping (DTW) was used for recognition. The system proved very effective in removing noises from the recording.

3.3 Varied Recitation Styles

Different reciters have distinctive rhythmic and melodic renditions, necessitating a flexible model. Building a comprehensive deep acoustic model is difficult because of the complex interplay of many recitation styles, each infused with distinctive rhythmic phrases and melodic accents. Creating a model architecture that can skillfully encompass the various recitation patterns while maintaining a flexible and adaptable framework capable of tolerating the different artistic expressions of the reciters is required to navigate this complex terrain. Nahar et al. (38) have implemented an SVM-based recognition model for "Qira'ah" and achieved 96% accuracy.

3.4 Adherence to Tajweed and Expressions

The model must accurately distinguish between short and long vowel sounds to represent elongations (Madd) and vowel lengths. It is difficult to capture these subtle differences accurately while keeping the recitation's rhythmic flow. AltaImas et al. (17) analyzed signals using the Mel Frequency Cepstral Coefficient (MFCC) and then compared them using the Dynamic Time Warping (DTW) technique to find the similarity differences.

3.5 Data Augmentation and Training

Creating an extensive and varied training dataset is crucial to building a precise and trustworthy deep acoustic model for Quranic recitation. Data augmentation methods are essential in building a substantial corpus with diverse recitation traditions. Realizing the model's effectiveness will require establishing the right balance between quantity and quality while assuring proper annotation.

Park et al. (39) utilized warping the features, masking blocks of frequency channels, and masking blocks of time steps to apply augmentation to the filter bank coefficient directly.

4. Conclusion and Future Directions

Reading the Quran following the recitation rules given to Muslims is a virtuous and spiritual enlightenment activity. Many researchers have approached this context in different ways. Some try to implement Machine Learning/Deep Learning to Tajweed, Hijayah, and Makhraj rules of the Quran. In contrast, others investigate it as a speech recognition problem, implementing reciter classification, dialect classification, etc. Many researchers have worked to implement Automatic Speech Recognition (ASR) to convert speech to text. However, there are still many research gaps available in this domain.

Our review paper is focused on covering almost every category of Quranic Recitation. We started by defining how ML/DL could help with Quranic Recitation. We discussed essential feature extraction techniques used for speech signals and their accuracy compared with other methods. Then, we discussed different reciters' classification research papers and their practices. Following on, we debated Tajweed, Makhraj, Hijayah, and Imlaah correction with the help of DL on a single letter or word level. Finally, we discussed how ASR can be used in Quranic Recitation and techniques to convert speech data to textual data. Table 2 shows details of the research papers that shared their dataset publicly for further research, whether the system model is shared or not, and whether the author compares their techniques with other techniques.

The future directions and gaps of each paper are already mentioned in the document. However, a few research categories in this domain are still untouched and could be worked out. First is the lack of an efficient deep acoustic model for classifying every verse to the chapter to which it belongs. Once we can achieve this holistic point and successfully build an efficient model, we can extend it by indexing every verse of the chapter. So, in the future, whenever a user hears some Quranic verse recited somewhere, they can identify the locality and place of the verse in the complete Quran. This has a lot of applications, including Namaz-e-Taraweeh, helping users recognize the Surah name from a single verse, which is helpful for non-Arabic native users. Secondly, suppose we can cover the first gap. In that case, we can further use ASR techniques to transcribe the uttered speech and already trained address by professionals to find the miscorrection in reading, solving the Tajweed and Imlaah issue while focusing on the verse level.

References

[1] A. S. M. Rusli, F. Ridzuan, Z. M. Zaki, M. N. S. M. Sayuti, and R. A. Salam, "A systematic review on semantic-based ontology for Quranic knowledge," International Journal of Engineering and Technology (UAE), 2018.

[2] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," Int. J. Electr. Comput. Eng, vol. 10, no. 6, pp. 6629-6643, 2020.
[3] D. J. Stewart, "Approaches to the investigation of speech genres in the Qur'an," Journal of Qur'anic Studies, vol. 24, no. 1, pp. 1-45, 2022.
[4] M. H. Bashir, A. M. Azmi, H. Nawaz, W. Zaghouani, M. Diab, A. Al-Fuqaha, and J. Qadir, "Arabic natural language processing for Quranic research: A systematic review," Artificial Intelligence Review, vol. 56, no. 7, pp. 6801-6854, 2023.
[5] M. S. Abdo, A. H. Kandil, and S. A. Fawzy, "MFC peak based segmentation for continuous Arabic audio signal," in 2nd Middle East Conference on Biomedical Engineering. IEEE, 2014, pp. 224-227.
[6] M. Bezoui, A. Elmoutaouakkil, and A. Benihssane, "Feature extraction of some Quranic recitation using mel-frequency cepstral coefficients (MFCC)," in 2016 5th International Conference on Multimedia Computing and Systems (ICMCS). IEEE, 2016, pp. 127-131.
[7] A. Meftah, Y. A. Alotaibi, and S.-A. Selouani, "A comparative study of different speech features for Arabic phonemes classification," in 2016 European Modelling Symposium (EMS). IEEE, 2016, pp. 47-52.
[8] M. N. Aulia, M. S. Mubarok, W. U. Novia, F. Nhita, et al., "A comparative study of MFCC-KNN and LPC-KNN for Hijaiyyah letters pronunciation classification system," in 2017 5th International Conference on Information and Communication Technology (ICoICT). IEEE, 2017, pp. 1-5.
[9] R. U. Khan, A. M. Qamar, and M. Hadwan, "Quranic reciter recognition: a machine learning approach," Advances in Science, Technology and Engineering Systems Journal, vol. 4, no. 6, pp. 173-176, 2019.
[10] S. M. Shah and S. N. Ahsan, "Arabic speaker identification system using a combination of DWT and LPC features," in 2014 International Conference on Open Source Systems & Technologies. IEEE, 2014, pp. 176-181.
[11] A. Elnagar, R. Ismail, B. Alattas, and A. Alfalasi, "Automatic classification of reciters of Quranic audio clips," in 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2018, pp. 1-6.
[12] A. Qayyum, S. Latif, and J. Qadir, "Quran reciter identification: A deep learning approach," in 2018 7th International Conference on Computer and Communication Engineering (ICCCE). IEEE, 2018, pp. 492-497.
[13] T. S. Gunawan, N. A. M. Saleh, and M. Kartiwi, "Development of Quranic reciter identification system using MFCC and GMM classifier," International Journal of Electrical and Computer Engineering (IJ ECE), vol. 8, no. 1, pp. 372-378, 2018.
[14] A. M. Basahel, A. A. Abi Sen, M. Yamin, N. M. Bahbouh, and S. Basahel, "A smart flexible tool to improve reading skill based on m-learning," in 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022, pp. 411-414.
[15] I. Ahsiah, N. Noor, and M. Idris, "Tajweed checking system to support recitation," in 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2013, pp. 189-193.
[16] E. Yosrita and A. Haris, "Identify the accuracy of the recitation of al-Quran reading verses with the science of tajwid with mel-frequency cepstral coefficients method," in 2017 International Symposium on Electronics and Smart Devices (ISESD). IEEE, 2017, pp. 179-183.
[17] T. Altalmas, W. Sediono, N. N. W. N. Hashim, S. Ahmad, and S. Khairuddin, "Analysis of two adjacent articulation Quranic letters based on MFCC and DTW," in 2018 7th International Conference on Computer and Communication Engineering (ICCCE). IEEE, 2018, pp. 187-191.
[18] F. Alqadheeb, A. Asif, and H. F. Ahmad, "Correct pronunciation detection for classical Arabic phonemes using deep learning," in 2021 International Conference of Women in Data Science at Taif University (WiDSTaif). IEEE, 2021, pp. 1-6.
[19] R. A. Rajagede and R. P. Hastuti, "Al-Quran recitation verification for memorization test using siamese LSTM network," Communications in Science and Technology, vol. 6, no. 1, pp. 35-40, 2021.
[20] A. M. A. Alqadasi, M. S. Sunar, S. Turaev, R. Abdulghafor, M. S. Hj Salam, A. A. S. Alashbi, A. A. Salem, and M. A. Ali, "Rule-based embedded HMMs phoneme classification to improve Quranic recitation recognition," Electronics, vol. 12, no. 1, p. 176, 2022.
[21] D. Omran, S. Fawzi, and A. Kandil, "Automatic detection of some Tajweed rules," in 2023 20th Learning and Technology Conference (L&T). IEEE, 2023, pp. 157-160.
[22] F. Ahmad, S. Z. Yahya, Z. Saad, and A. R. Ahmad, "Tajweed classification using artificial neural network," in 2018 International Conference on Smart Communications and Networking (SmartNets). IEEE, 2018, pp. 1-4.
[23] L. Marlina, C. Wardoyo, W. M. Sanjaya, D. Anggraeni, S. F. Dewi, A. Roziqin, and S. Maryanti, "Makhraj recognition of hijaiyah letter for children based on mel-frequency cepstrum coefficients (MFCC) and support vector machines (SVM) method," in 2018 International Conference on Information and Communications Technology (ICOIACT). IEEE, 2018, pp. 935-940.
[24] M. Irfan, I. Z. Mutaqin, and R. G. Utomo, "Implementation of dynamic time warping algorithm on an Android-based application to write and pronounce hijaiyah letters," in 2016 4th International Conference on Cyber and IT Service Management. IEEE, 2016, pp. 1-6.
[25] N. W. Arshad, M. Z. Ibrahim, R. A. Karim, Y. A. Wahab, N. F. Zakaria, and T. T. Muda, "Signal-based feature extraction for makhraj emission point classification," in Engineering Technology International Conference (ETIC 2022), vol. 2022. IET, 2022, pp. 19-25.
[26] J. Farooq and M. Imran, "Mispronunciation detection in articulation points of Arabic letters using machine learning," in 2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube). IEEE, 2021, pp. 1-6.
[27] B. Yousfi and A. M. Zeki, "Holy Qur'an speech recognition system Imaalah checking rule for Warsh recitation," in 2017 IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2017, pp. 258-263.
[28] B. Yousfi, A. M. Zeki, and A. Haji, "Isolated Iqlab checking rules based on speech recognition system," in 2017 8th International Conference on Information Technology (ICIT). IEEE, 2017, pp. 619-624.
[29] F. Thirafi and D. P. Lestari, "Hybrid HMM-BLSTM-based acoustic modeling for automatic speech recognition on Quran recitation," in 2018 International Conference on Asian Language Processing (IALP). IEEE, 2018, pp. 203-208.
[30] M. Hadwan, H. A. Alsayadi, and S. AL-Hagree, "An end-to-end transformer-based automatic speech recognition for Quran reciters." Computers, Materials & Continua, vol. 74, no. 2, 2023.
[31] B. Yousfi and A. M. Zeki, "Holy Qur'an speech recognition system distinguishing the type of recitation," in 2016 7th International Conference on Computer Science and Information Technology (CSIT). IEEE, 2016, pp. 1-6.
[32] A. Nur, S. Syarifandi, S. Amin et al., "Implementation of text mining classification as a model in the conclusion of tafsir bil ma'tsur and bil ra'yi contents," Int. J. Eng. Adv. Technol, vol. 9, no. 1, pp. 2789-2795, 2019.
[33] S. Shahriar and U. Tariq, "Classifying maqams of Quranic recitations using deep learning," IEEE Access, vol. 9, pp. 117271-117281, 2021.
[34] A. Abd Almisreb, A. F. Abidin, and N. M. Tahir, "Noise effects on the recognition rate of Arabic phonemes based on Malay speakers," in 2014 IEEE Symposium on

Industrial Electronics & Applications (ISIEA). IEEE, 2014, pp. 1-6.
[35] S. Khairuddin, S. Ahmad, A. H. Embong, N. N. W. N. Hashim, and S. S. Hassan, "Features identification and classification of alphabet (ro) in leaning (al-inhiraf) and repetition (al-takrir) characteristics," in 2019 IEEE International Conference on Automatic Control and Intelligent Systems (2CACIS). IEEE, 2019, pp. 295-299.
[36] M. I. E.-K. Ghembaza, O. Tayan, and K. S. Aloufi, "Qurani Rafiqui: an interactive context-aware Quranic application for smartphones," in 2018 1st International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2018, pp. 1-6.
[37] F. Sadat, F. Kazemi, and A. Farzindar, "Automatic identification of Arabic dialects in social media," in Proceedings of the first international workshop on Social media retrieval and analysis, 2014, pp. 35-40.
[38] K. Nahar, R. Al-Khatib, M. Al-Shannaq, and M. M. Barhoush, "An efficient Holy Quran recitation recognizer based on SVM learning model," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 6, no. 04, pp. 394-414, 2020.
[39] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.
[40] D. Omran, S. Fawzi, and A. Kandil, "Automatic Detection of Some Tajweed Rules," in 2023 20th Learning and Technology Conference (L&T), January 2023, pp. 157-160.
[41] R. M. Siregar, B. Satria, A. Prayogi, M. A. S. Pane, E. E. Awal, and Y. R. Sari, "Identification of Tajweed Recognition using Wavelet Packet Adaptive Network based on Fuzzy Inference Systems (WPANFIS)," Internet of Things and Artificial Intelligence Journal, vol. 4, no. 1, pp. 32-41, 2024.
[42] A. F. Ghori, A. Waheed, M. Waqas, A. Mehmood, and S. A. Ali, "Acoustic modelling using deep learning for Quran recitation assistance," International Journal of Speech Technology, vol. 26, no. 1, pp. 113-121, 2023.
[43] H. M. Mahmudin and H. Akbar, "Qur'an Recitation Correction System Using Deepspeech," Indonesian Journal of Multidisciplinary Science, vol. 2, no. 11, pp. 4010-4022, 2023.
[44] A. Al Harere and K. Al Jallad, "Quran Recitation Recognition using End-to-End Deep Learning," arXiv preprint arXiv:2305.07034, 2023.
[45] A. A. Harere and K. A. Jallad, "Mispronunciation Detection of Basic Quranic Recitation Rules using Deep Learning," arXiv preprint arXiv:2305.06429, 2023.

Muhammad Aleem Shakeel is an electrical engineer with a strong background in artificial intelligence and autonomous systems. He completed his Bachelor's degree in Electrical Engineering from the University of Engineering and Technology (UET), Taxila. Eager to learn more about AI, he pursued a Master's degree in Electrical Engineering, at the National University of Sciences and Technology (NUST) in Islamabad. He is with Invoice Mate, a company that offers AI-based Invoicing solutions, designing and implementing AI technologies that improve the efficiency and accuracy of invoicing processes. His research interests include Machine Learning, Deep Learning, Generative AI and NLP for Speech Signals and documents.Email: muhammad.aleem227@ gmail.com

Hasan Ali Khattak (Senior Member IEEE) received his Ph.D. in Electrical and Computer Engineering degree from Politecnico di Bari, Bari, Italy, in April 2015, a master's degree in information engineering from Politecnico di Torino, Torino, Italy, in 2011, and a B.CS. Degree in Computer Science from the University of Peshawar, Peshawar, Pakistan in 2006. He has been an associate professor at the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Pakistan, since October 2020. His current research interests focus on Future Internet Architectures such as the Web of Things and leveraging Data Sciences and Social Engineering for Future Smart Cities. Email: hasan.alikhattak@seecs.edu.pk

Numan Khurshid, obtained his Ph.D. in Electrical Engineering specialized in Artificial Intelligence from Lahore University of Management Sciences (LUMS), Lahore. Currently, he is working as an Assistant Professor, at the Department of Electrical Engineering, Schools of Electrical Engineering and Computer Sciences (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research interests include Machine Learning, Deep Learning, and Generative AI for Remote Sensing Images and Speech Signals Email: numan.khurshid @seecs.edu.pk

Addressing Class Imbalance in Customer Response Modeling Using Random and Clustering-Based Undersampling and SVM

Kašćelan, Ljiljana and Vuković, Sunčica

Abstract: ***The main challenge in machine learning-based customer response models is the class imbalance problem, i.e. small number of respondents, compared to non-respondents. Aiming to overcome this issue, the approach of preprocessing training data using a Support Vector Machine (SVM), trained on a balanced sample obtained by random undersampling (B-SVM), as well as on a balanced sample obtained by clustering-based undersampling (CB-SVM) was tested. Several classifiers are tested on such a balanced dataset, to compare their predictive performances. The results of this paper demonstrate that the approach effectively pre-processes the training data, and, in turn, reduces noise and overcomes the class imbalance problem. Better predictive performance was achieved compared to standard training data balancing techniques such as undersampling and SMOTE. CB-SVM gives a better sensitivity, while B-SVM gives a better ratio of sensitivity and specificity. Organizations can utilize this approach to balance training data automatically and simply and more efficiently select customers that should be targeted in the next direct marketing campaigns.***

Index Terms: ***customer response model, data imbalance, data preprocessing, clustering, support vector machine***

1. INTRODUCTION

Customer response modeling is an important part of developing effective direct marketing strategies, as it allows companies to predict how will their consumers react to various marketing initiatives, mainly future direct marketing campaigns. Businesses may utilize customer response modeling approaches to uncover consumer features, such as product preferences, types, and previous behaviors, and then use this knowledge to design focused, targeted, and personalized marketing campaigns that are more likely to resonate with their target audience. In addition, this approach can also assist firms in optimizing their marketing expenditure, by forecasting the anticipated response rate, and hence, allocating their marketing budget accordingly. This can help companies in maximizing their return on investment (ROI) and improving the efficiency of their overall marketing operations, by reducing costs and increasing the campaign revenues.

One of the main issues and challenges in customer response modeling is a class imbalance – the proportion of customers that respond to the campaign with purchases is typically relatively small, i.e., the response rate in campaigns is usually quite low. The small number of respondents leads to the significant difference between the number of members of this group, compared to the non-respondent group. This represents an issue for the classification machine learning algorithms, as they tend to be biased towards the bigger class or segment (in this case, targeted customers who did not respond to the offer).

In direct marketing, the class imbalance problem is often addressed using one of three ways: data-based approaches [1], [2] which use resampling techniques to balance classes; algorithm-based approaches [3], which use specially modified algorithms; or cost-based approaches [4], which assign various misclassification costs to different class examples. These techniques have several drawbacks. For example, the criteria for selecting examples for undersampling are ambiguous, while, oversampling generates synthetic examples of a minor class by replicating them, which can lead to overfitting. Even though some oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) [5] overcome the overfitting problem, it can lead to noise amplifying in the data, as well as within-class imbalance [6]. On the other hand, algorithm-level techniques need extensive algorithm understanding to be applied, therefore, they require expert knowledge. Finally, cost-based solutions necessitate additional learning expenses, as well as an in-

Manuscript received April 1, 2023.

Lj. Kašćelan is a full professor at Faculty of Economics, University of Montenegro (email: ljiljak@ucg.ac.me)

S. Vuković is a teaching assistant at Faculty of Economics, University of Montenegro (email: suncica@ucg.ac.me).

depth examination of optimal cost configurations, which makes this technique complex to use.

For class imbalance, SVM preprocessing is also applied, given that SVM can eliminate noise from the data, resolve overlapping classes, and move the border that separates classes towards a larger class, thereby supplementing the smaller class with the most similar examples [7]–[9]. However, the performance of SVM is significantly reduced when applied to highly unbalanced sets due to marginal bias towards the smaller class (Cervantes et al., 2020).

In this paper preprocessing of extremely unbalanced training data, using SVM in combination with random and clustering-based undersampling (B-SVM and CB-SVM respectively) will be applied, to determine whether it improves the predictive performance of base customer response models. Clustering-based undersampling implies that in the pre-processing phase, the non-respondent classes within the training dataset are clustered, to obtain the representatives in the undersampling procedure which enables a better distinction between classes, and, as a result, better class balancing. In order to predict customers' responses, on such an imbalanced dataset, several classifiers will be tested, namely: Decision Tree (DT), Logistic Regression (LR), Gradient Boosted Trees (GBT), Random Forest (RF), k-NN (k Nearest Neighbor) and SVM. Even though the approach of combining k-means clustering, undersampling and SVM was applied in previous research [10], [11], in this paper, it was applied for the first time for the issues of customer response modeling in direct marketing, and also, as pre-processor of an extremely imbalanced dataset, with the minor class being less than 1%.

Following the Introduction, the rest of the paper is structured as follows: in the second section, the literature review regarding customer response modeling will be presented, and in the third section methods and data used in this paper will be described. The fourth section contains the obtained results of the proposed B-SVM and CB-SVM procedures, as well as the discussion and comparison with previous results. Finally, the fifth section will present major conclusions, limitations, and recommendations for future research.

2. *Literature Review*

Over the past ten years, social media marketing has emerged as a significant research area that outlines the various dimensions of consumer relations. This trend is even more clear when it's put into context – in January 2023, Facebook had 2.963 billion active daily users [12], while Instagram had 1.318 billion [13]. Taking into account that there are 5.16 billion Internet users worldwide [14], from the previous statistics we can see that over half of the online population can be reached through social media. As a result, social media provides marketers with the option to interact directly with customers, increase communication, and sell superior value propositions to their top customers regardless of their location [15].

The ability to reach many potential customers for a relatively lower cost compared to traditional media creates the problem of extreme class imbalance for direct marketing campaigns placed in this manner. This situation occurs because there is a smaller number of customers who respond to the offer, i.e. who complete the purchase via the provided links, compared to the number of targeted customers who only visit the website without purchasing the product/service. As it was pointed out in the Introduction section, high-class imbalance, as it exists in this case, leads to algorithm bias towards the major class.

For example, if the class ratio in the data set is 100:1, that is, when for every 100 examples of the negative class there is only one example of the positive class, the classifier can strive to maximize the accuracy of the classification rule and achieve an accuracy of 99%, by simply ignoring the positive examples and classify all examples as negative [16]. Thus, standard algorithms expect a balanced data distribution and equal misclassification costs, and when applied to complex and unbalanced data, they result in unfavorable accuracy for all data classes [17]. Since the role of response models is to accurately predict the segment of respondents, solving this issue is of great interest in direct marketing research and practice.

Several authors treated this issue in different ways, such as random undersampling and oversampling [18], [19], negative sampling [20] nonuniform negative sampling [21]. However, for this research, three papers were taken as most relevant, keeping in mind that they included clustering of the majority class as a part of their procedure. In the first mentioned research, Kang et al. [22] suggested in their paper that models for predicting campaign responses can be improved by balancing classes using clustering, undersampling and ensemble methods. First, the instances belonging to the class of non-respondents are clustered. In the next step, undersampling is carried out, as part of the ensemble procedure, by randomly selecting a certain number of examples from all clusters, proportional to the size of the cluster, so that the total number of selected instances is equal to the major class, i.e. non-respondents (balanced ensemble). In this way, the taking of a certain number of representative members of a larger

class is achieved and the loss of information relevant to the differentiation of classes is reduced. By performing the ensemble procedure in k iterations, k classifiers are generated on k such balanced samples and their predictions are combined. The results showed that compared to random sampling methods, as well as SMOTE, this approach has more stable predictive performance that decision makers can trust more.

In the second research, Marinakos and Daskalaki [23] tested cluster-based undersampling and distance-based resampling techniques, including SMOTE, for a campaign response model by bank customers (with 12% respondents and 88% non-respondents) with several different classifiers, such as linear discriminant analysis (LDA), LR, k-NN, DT, Neural Networks (NN) and SVM. The highest classification accuracy of the minority class was achieved by the combination of cluster undersampling and k-NN.

Finally, Amini et al. [24] began treating this issue by balancing the data using a combination of clustering and random undersampling. They stated that their goal was to retain the original dataset's class distribution, thus they divided the non-respondent consumers into clusters and selected random samples from each cluster. The sampled non-respondent examples are combined with the total respondent instances to create a new balanced training dataset. By repeating this approach, many balanced training datasets are constructed, and each dataset is given to a base classifier in an ensemble framework. Several classification algorithms, such as SVM, ANN, DT, and LR, are used to create the proposed ensemble architecture. In this study, the base SVM classifier showed the best performance. With this classifier, the authors created the cluster-based ensemble model, the random undersampling ensemble, and the bagging ensemble. The cluster-based approach they suggested achieved the best results.

Therefore, previous literature confirms the effectiveness of random and clustering-based undersampling in class balancing. On the other hand, previous studies also confirm the possibilities of SVM as a preprocessor of training data to resolve overlapping classes and their imbalance. For example, Farquad and Bose [9] tested SVM as a data preprocessor together with undersampling and oversampling techniques, in order to solve the problem of class imbalance, using an insurance company data set with a response rate of 6%. After pre-processing the data and replacing the target variable with SVM prediction, such a modified data set was used to train the MLP, LR and RF models. The results show that the proposed data balancing approach improves the classification performance in every case. The best performance in this study was achieved using undersampling and the RF model, which achieved a sensitivity of 71.01%.

In this paper, the possibilities of SVM preprocessing in combination with undersampling on an extremely unbalanced set of customer data, with a response rate of less than 1%, as well as the possibilities of clustering-based undersampling to provide a better representation of the distribution of non-respondents to achieve a more efficient differentiation of respondents were tested.

3. METHODS AND DATA

3.1 Data

Predictive analytics, specifically for marketing purposes, aims to forecast future customer behavior from very sparse data, where the marketing database often has minimal information about customer demographics, product demands, and interest in purchasing. However, technological developments and social media have enabled the collection of data on online customer behavior, which can be useful in predicting their response to future marketing activities. In line with this, companies may create customer response models to help identify the consumers who would most likely respond to the upcoming campaign to valorize this kind of data.

For testing the proposed consumer response model empirically, a dataset was gathered from a renowned Montenegrin sports equipment distributor. The dataset included four months' worth of sponsored social media post views to e-commerce websites, from October 2018 to January 2019. The dataset included the following attribute groups: Web metrics (12 attributes, such as the average number of sessions, the average session duration, operating system used, etc.), Product description data (19 attributes, such as number of products purchased from the different product categories, product discount, etc.) and Purchase history data (8 attributes, such as recency, frequency and monetary).[1] In total, 33,662 sessions were successfully completed across six online direct marketing campaigns on social media (while one user can complete more than one session). The company's internal product database, Google Analytics, and Facebook Business Manager were combined to create the complete dataset, which was then pre-processed and made ready for analysis.

For the study, the complete dataset is split into training and test data. In the training data set, only

[1] Detailed dataset description can be found in [29].

40 customers directly responded to the offer, i.e., made a purchase, yielding a response rate of only 0.41%. The training dataset included the history of web and purchasing behavior of 9660 website visitors from Campaign 1 to Campaign 4, as well as an indicator of their response to the following Campaign 5. Without including new visitors who first appeared in Campaigns 5 or 6, the set for model testing included the same data categories as the training set for 7929 visitors from Campaigns 1 through Campaign 5, as well as the response indicating whether a customer responded to Campaign 6 (there were 40 purchases in this campaign as well).

The class distribution in the training and test sets is shown in Table 1.

Table 1. Class distribution in training and test sets

	Respondents	*Non-respondents*	*Campaign*
Training set	40	9620	C1-C4, C5
Test set	40	7889	C1-C5, C6

3.2 Methods and Proposed Procedure

In this paper, several methods will be applied: K-means clustering, SVM for data balancing, as well as a set of classifiers, which will be individually described in this section.

K-means clustering is a well-known data mining technique that has seen widespread application in marketing research. Generally, clustering is the process of grouping similar objects into groups. K-means [25] clustering is a technique that, for the chosen value of k, identifies k of object clusters, which are based on objects close to the center of k groups (calculating Euclidean distances), with the center defined as the average of the n-dimensional attribute vectors of each cluster. Thus, k-means is an unsupervised technique for classifying consumers into a fixed number of clusters, with consumers within the clusters as similar as possible, and clusters as dissimilar as possible. As a result, k-means clustering will be used in this paper to group similar potential customers from the non-respondent segment, based on their features recorded in the database. By taking representatives from each cluster, more heterogeneous representatives of non-respondent customers are obtained, which provides classification algorithms for easier differentiation from respondents.

The support Vector Machine method (SVM), introduced by Vapnik [26], is successful in solving the issues of class overlapping and class imbalance. The algorithm constructs a hyperplane between examples (attribute vectors) belonging to different classes that can discriminate the classes to the greatest extent possible, regardless of the number of instances available to learn from [9]. As a result, SVM eliminates data noise, i.e., class overlap, and complements the minor class with the most relevant examples by moving the margin to the closest, and hence most similar, examples of the major class and putting them into the smaller class [27]. As a result, SVM was applied as a data pre-processor to balance the data and improve classification accuracy, as it was similarly done in previous research for the same purpose [22] [27].

The following classifiers were tested on the pre-processed dataset, as well as before pre-processing on the original dataset, in order to compare the results and model performances: Decision Tree (DT) [30], [31], Logistic regression (LR) [32], Gradient Boosted Trees (GBT) [33], k Nearest-Neighbor (kNN) [34], Random Forest (RF) [35] and SVM.

The DT technique separates the dataset into subgroups based on attribute values so that each subgroup has as many examples of one class as feasible. During inductive division, a tree-shaped model is produced, after which the approach is named. When it comes to Logistic Regression, in its most common version, LR, as a classification approach, covers binary outcomes - this technique involves estimating the likelihood of a discrete result given the input variables.

The goal of the k-nearest neighbor method is to find the nearest neighbors of a given query point so that we may assign a class label to that point. The k-NN approach is based on the assumption that related items exist nearby.

The Gradient Boosting Trees technique chooses the next DT model with the lowest residual error from the previous batch of DT models. As a result of reducing residual error, succeeding models will favor the accurate categorization of previously misclassified cases. Finally, the Random Forest method, during the training phase, employs random attribute selection to generate a greater number of decision trees. In this sense, it is an expansion of the fundamental notion of individual DT classifiers in order to generate a greater number of classification decision trees. As a result, the last two approaches, GBT and RF, combine the ensemble meta-algorithm and the DT classifier.

Predictive procedures for B-SVM and CB-SVM consisted of the following steps.

For B-SVM, a random subset of non-respondents equal to the number of respondents was first selected from the training set (random undersampling). SVM was trained on such a sample using k-fold cross-validation and then applied to the entire training set. The original label is replaced by SVM prediction.

In CB-SVM, first, non-respondent clustering was completed using k-means clustering (Davies-Bouldin index was used to determine the optimal number of clusters). To achieve balanced classes, the same number of instances were randomly picked from each cluster of non-respondents, in order to balance the number of non-respondents with the number of respondents. In the k-fold cross-validation technique, the SVM model with the greatest prediction performance is obtained. Then, trained SVM is applied to the training dataset, and SVM prediction replaces the original class label.

A minor class of respondents is supplied with similar instances from the major (non-respondents) class in both procedures and class balance is attained, similar to [9]. The training set description after pre-processing is given in Table 2.

Table 2. Class distribution in training set after pre-processing

	Respondents	*Non-respondents*
B-SVM preprocessing	1221	8439
CB-SVM preprocessing	2896	6764

The updated (balanced) training dataset is used to train the following classifiers: DT, LR, GBT, k-NN, RF, and SVM. In k-fold cross-validation, the models with the greatest prediction performances are chosen. Finally, trained classifiers are then applied to the original imbalanced test set (40:7889) and the predictive performance is obtained for each of the chosen classifiers.

The predictive procedure is shown in Figure 1.

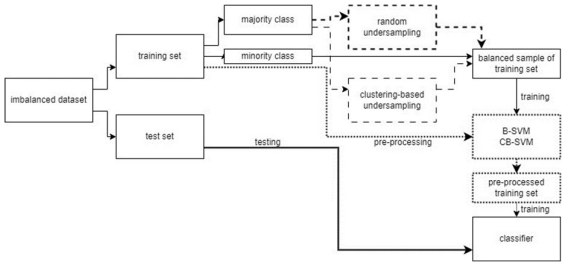


Figure 1. Proposed predictive procedure

It can be seen from Figure 1 that, unlike the standard undersampling technique, the training of the classifier is performed on the complete training set and not on the sample, while the training of the SVM preprocessor is performed on the balanced sample. The trained models are applied to the original unbalanced test set.

The models will be evaluated using a set of performance metrics: Sensitivity, Specificity, AUC and Balanced Accuracy (BA). Apart from AUC, which shows the ability of the model to differentiate between positive and negative classes, the other three metrics are computed by utilizing the values from the confusion matrix. Confusion matrix is presented in Table 3.

Table 3. Confusion matrix

Actual class	Predicted class	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Using the values from the matrix, the performance metrics are calculated as follows:

$$Sensitivity = TP / (TP + FN)$$
$$Specificity = TN / (TN + FP)$$
$$BA = (Sensitivity + Specificity) / 2$$
$$F1\ Score = 2TP / (2TP + FP + FN)$$

For this research, the Sensitivity metric is of the greatest importance, as the paper aims to solve the class imbalance problem and to correctly identify as many examples of the positive class.

4. RESULTS AND DISCUSSION

Following the proposed procedure described in the previous section, the first step included the clustering of non-respondents targeted in observed campaigns. To optimize the number of clusters, the Davies-Bouldin index was used. The results of this procedure are presented in Figure 2.

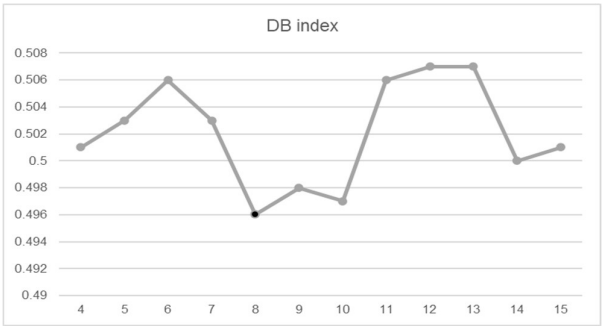


Figure 2. Davies-Bouldin index

From Figure 2, it can be seen that the best DB result is obtained for eight clusters since there is the greatest drop in the result (from 7 to 8 clusters). Therefore, this number is chosen as optimal.

The rest of the proposed procedure is followed and the results obtained on the test set, i.e. unknown data, for all tested classifiers, are given

in Table 4. This table contains the results of predictive performances of standalone models (without preprocessing), models with B-SVM pre-processing, and models with CB-SVM pre-processing. Also, due to the performance comparison, the results of the model with undersampling and SMOTE techniques are shown. The decision threshold for all models is defined using the threshold optimizer so that the relationship between sensitivity and specificity is optimal. This was achieved by specifying the maximum F1 Score as a criterion for the optimizer.

Table 4. Predictive performance of classification algorithms on the test set

Classifier	Sensitivity	Specificity	AUC	BA
Standalone models				
DT	12.50%	99.87%	0.608	56.19%
LR	15.00%	99.59%	0.680	57.30%
GBT	10.00%	99.63%	0.727	54.82%
kNN	2.50%	99.89%	0.593	51.20%
RF	30.00%	99.38%	0.830	64.69%
SVM	52.50%	93.79%	0.725	73.15%
Models with B-SVM pre-processing				
B-SVM DT	70.00%	82.99%	0.756	76.50%
B-SVM LR	70.00%	79.77%	0.798	74.89%
B-SVM GBT	70.00%	80.50%	0.805	75.25%
B-SVM kNN	65.00%	77.94%	0.767	71.47%
B-SVM RF	70.00%	80.38%	0.831	75.19%
B-SVM SVM	75.00%	79.54%	0.746	77.27%
Models with CB-SVM pre-processing				
CB-SVM DT	80.00%	65.60%	0.612	72.80%
CB-SVM LR	77.50%	64.43%	0.607	70.97%
CB-SVM GBT	80.00%	64.10%	0.682	72.05%
CB-SVM kNN	70.00%	65.70%	0.706	67.85%
CB-SVM RF	75.00%	65.07%	0.760	70.04%
CB-SVM SVM	72.50%	65.33%	0.646	68.92%
Models with SMOTE oversampling				
SMOTE DT	52.50%	85.90%	0.642	69.20%
SMOTE LR	62.50%	79.24%	0.745	70.87%
SMOTE GBT	60.00%	86.50%	0.815	73.25%
SMOTE kNN	57.50%	84.47%	0.808	70.99%
SMOTE RF	55.00%	83.91%	0.776	69.46%
SMOTE SVM	70.00%	88.07%	0.790	79.04%
Models with Undersampling				
Undersamp DT	72.50%	74.55%	0.732	73.53%
Undersamp LR	75.00%	73.62%	0.818	74.31%
Undersamp GBT	85.00%	69.49%	0.861	77.25%
Undersamp kNN	75.00%	46.72%	0.679	60.86%
Undersamp RF	72.50%	75.55%	0.840	74.03%
Undersamp SVM	72.50%	64.75%	0.665	68.63%

With standalone classification models, without prior data processing and balancing, given that class imbalance is very high (0.41%), most positive examples, i.e. respondents, are ignored and classified as negative examples. Hence, standalone models treat this small number of respondents as noise in the dataset. However, given the importance of accurately identifying customers who will respond to a direct marketing campaign (true positives from the confusion matrix), the sensitivity metric is the key emphasis of this study. From Table 4, it can be observed that this metric is very low in standalone models, ranging from 2.5% in kNN and SVM to 30% in RF. Hence, such models could not be utilized as effective customer response models, since they would only target a maximum of 30% of actual respondents in the campaign, missing a large proportion of potential customers. In addition, the AUC of standalone models is very low: 0.608, 0.680 and 0.593 for DT, LR and kNN respectively. The closer the AUC result is to 0.5 the lower the ability of the model to distinguish between positive and negative classes. A better AUC result is obtained for the RF (0.830) model.

The results for B-SVM indicate improvements in the most relevant metric – Sensitivity, by a large amount. After balanced undersampling and SVM pre-processing, the SVM model showed the best performance, with this metric amounting to 75% (an increase of over 72 percentage points). Also, all other models achieved a sensitivity of 70% and more, except for kNN, whose sensitivity increased from 2.5% to 65%. All tested classifiers showed improvement in AUC results, all about 0.8, indicating that the models have a very good ability to differentiate between classes, which is of great importance in selecting future respondents, compared to non-respondents.

Finally, the CB-SVM obtained the best sensitivity. This approach further improved the performance of the previously described models, showing the importance of clustering the non-respondents, i.e. majority class. Sensitivity levels increased in all models apart from CB-SVM SVM, with the best result again achieved for DT and GBT models after suggested pre-processing (80%). However, the specificity, AUC and BA are lower than that of the B-SVM model.

Figure 3 shows the gradual improvement of the sensitivity metric for all base classifiers.

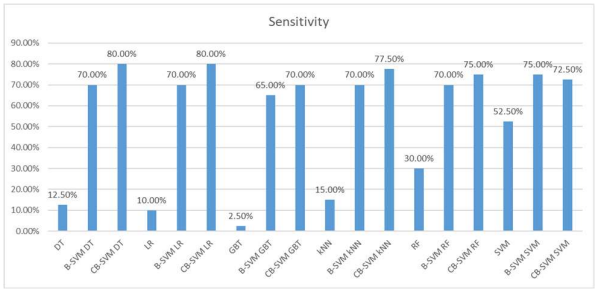


Figure 3. Sensitivity metric overview

From Figure 3, it can be seen that the most significant improvement was obtained for LR (10% - 70% - 80%), and also the superiority of the CB-SVM approach with the best performances across all tested models. Figure 4 shows the AUC values for all tested models.

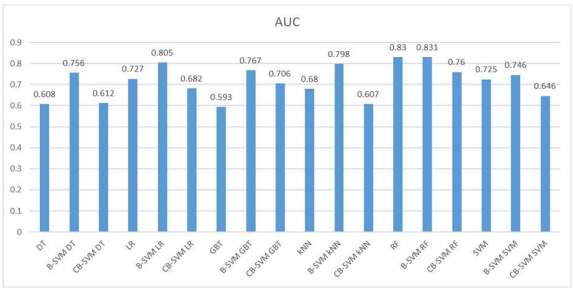


Figure 4. AUC metric overview

Figure 4 shows a significant improvement in the AUC metric for the B-SVM model compared to the standalone models. CB-SVM preprocessing did not improve the AUC achieved by B-SVM.

By comparison with resampling techniques based on Table 4, it can be seen that SMOTE gives significantly lower sensitivity compared to B-SVM and CB-SVM, while AUC and BA are mostly lower compared to B-SVM. With the undersampling technique, the sensitivity is mostly lower compared to CB-SVM, specificity is lower across all models compared to B-SVM, while the other parameters are similar to B-SVM. However, with the undersampling model, these parameters vary more from classifier to classifier than with B-SVM. For example, AUC varies from 0.665 to 0.861 (standard deviation of about 8%), while for B-SVM it ranges from 0.746 to 0.831 (standard deviation of about 3%).

Figure 5 shows the relationship between sensitivity and specificity, for example, of all GBT models (standalone, B-SVM preprocessed, CB-SVM preprocessed, with SMOTE oversampling and with undersampling). The best model is the one for which the corresponding point on the graph is closer to the upper right corner, i.e. points (1,1) [22].

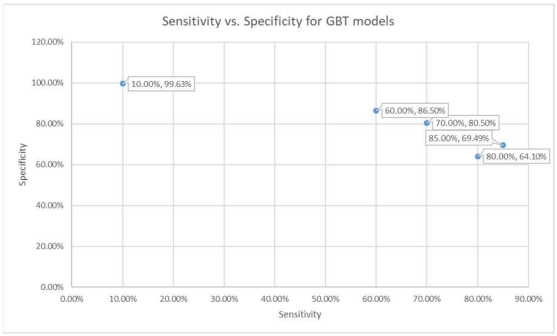


Figure 5. Relationship between sensitivity and specificity for GBT models

From Figure 5, it can be seen that the point closest to the upper right corner is point (70.00%, 80.50%), which, based on Table 4, corresponds to the B-SVM GBT model because its sensitivity is 70% and specificity is 80.5%. The situation is similar to other classifiers. Therefore, the optimal ratio of sensitivity and specificity is provided by the approach with B-SVM preprocessing.

Based on the results, the following conclusions can be drawn:

1. B-SVM preprocessing improves the predictive performance of customer response models based on machine learning classifiers under extremely high-class imbalance.
2. CB-SVM preprocessing improves the sensitivity of the classifier compared to B-SVM, which means that clustering provides examples of the major class that are more informative for distinguishing examples of the smaller class (respondents).
3. CB-SVM preprocessing reduces the specificity compared to the B-SVM approach because due to clustering, more examples of the major class are included in the balanced sample that are far from the SVM margin (non-support vectors) and are not relevant for distinguishing this class [1].
4. B-SVM and CB-SVM preprocessing give better sensitivity than the SMOTE technique, i.e. better identify respondents.
5. B-SVM and CB-SVM preprocessing gives better or similar performance to classifiers compared to undersampling, but with less variability, which can increase managers' confidence in data-based decision making.
6. The best ratio between sensitivity and specificity is achieved by B-SVM. So, with this approach, the customer response model will perfectly identify both respondents and non-respondents.
7. The CB-SVM approach achieves the best sensitivity. With it, the respondents will be identified more precisely than with the previous one, but because of this, the model will incorrectly classify more non-respondents as respondents. This means that more direct offers will be made

than in the case of the previous approach. However, considering online campaigns through social networks, where the offer costs the customer very little, this is not so significant from an economic point of view. In this study, the approach using CB-SVM preprocessing has achieved the best sensitivity scores.

5. CONCLUSION

Numerous past studies have highlighted the relevance of class imbalance problems in direct marketing, as well as in machine learning applications in general. However, this issue is particularly interesting in digital direct marketing, either done via email or social media. The costs of such media can be relatively lower compared to traditional postal service, and hence, can include a larger number of targeted customers. On the other hand, the number of actual customers, i.e. respondents in this case is even lower, due to a larger initial base. Therefore, response rates of around 1% can be considered very successful. In this paper, this issue was treated on a database with only a 0.41% response rate, which proved to be challenging for the classification task of base classifiers.

To overcome this issue, in this paper, B-SVM and CB-SVM data preprocessing is proposed, which includes the step of balanced undersampling (random and clustering-based) of the training set, training SVM on such a sample, applying the trained SVM on the entire training set and replacing of the original label by SVM prediction. Various classifiers were trained on the preprocessed training set and applied to the unbalanced test set. The results were compared to the undersampling and SMOTE balancing on the same dataset and using the same classifiers. The proposed approach leads to high values of all performance metrics and shows better results than undersampling and SMOTE on extremely unbalanced data on which it was tested.

Such results have practical implications for decision-makers in marketing. Using this technique, they can more precisely select and target their potential customers in future marketing campaigns. An increase in sensitivity values indicates that the chances of targeted respondents are higher. In addition, the increase in specificity shows that the costs of the campaign can be reduced. Overall, the expected increase in the share of respondents and a decrease in costs may lead to more profitable campaigns and general marketing efforts. Since customers on social media are constantly targeted by different kinds of brands sharing a range of offers, it is important to select and target specifically those customers who will find the company's offer relevant. Therefore, using this approach doesn't only have benefits for businesses, but for their potential customers as well.

In future research, this approach could be tested on different unbalanced data sets from direct marketing and other areas of application such as churn prediction, credit risk, sentiment analysis and the like.

REFERENCES

[1] G. Kim, B. K. Chae, and D. L. Olson, "A support vector machine (SVM) approach to imbalanced datasets of customer responses: Comparison with other customer response models," *Serv. Bus.*, vol. 7, no. 1, pp. 167–182, 2013, doi: 10.1007/s11628-012-0147-9.

[2] V. L. Miguéis, A. S. Camanho, and J. Borges, "Predicting direct marketing response in banking: comparison of class imbalance methods," *Serv. Bus.*, vol. 11, no. 4, pp. 831–849, 2017, doi: 10.1007/s11628-016-0332-3.

[3] M. M. Al-Rifaie and H. A. Alhakbani, "Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions," *Proc. 2016 SAI Comput. Conf. SAI 2016*, pp. 446–451, 2016, doi: 10.1109/SAI.2016.7556019.

[4] H. Shin and S. Cho, "Response modeling with support vector machines," vol. 30, no. 4, pp. 746–760, 2006, doi: 10.1016/j.eswa.2005.07.037.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[6] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.

[7] S. Rogić and L. Kašćelan, "Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning," *Comput. Sci. Inf. Syst.*, vol. 18, no. 00, pp. 52–52, 2020, doi: 10.2298/csis200530052r.

[8] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, and B. Baesens, "Rule extraction from support vector machines: An overview of issues and application in credit scoring," *Stud. Comput. Intell.*, vol. 80, no. 2008, pp. 33–63, 2008, doi: 10.1007/978-3-540-75390-2_2.

[9] M. A. H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," *Decis. Support Syst.*, vol. 53, no. 1, pp. 226–233, 2012, doi: 10.1016/j.dss.2012.01.016.

[10] Y. Yao *et al.*, "K-SVM: An effective SVM algorithm based on K-means clustering," *J. Comput.*, vol. 8, no. 10, pp. 2632–2639, 2013, doi: 10.4304/jcp.8.10.2632-2639.

[11] J. Laosai and K. Chamnongthai, "Acute leukemia classification by using SVM and K-Means clustering," *2014 Int. Electr. Eng. Congr. iEECON 2014*, pp. 1–4, 2014, doi: 10.1109/iEECON.2014.6925840.

[12] DataReportal, "Facebook Statistics and Trends," 2023.

[13] DataReportal, "Instagram Statistics and Trends," 2023. [Online]. Available: https://datareportal.com/essential-instagram-stats.

[14] Statista, "Number of internet and social media users worldwide as of January 2023," 2023. [Online]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide/.

[15] M. Yadav and Z. Rahman, "The influence of social media marketing activities on customer loyalty: A study of e-commerce industry," *Benchmarking*, vol.

25, no. 9, pp. 3882–3905, 2018, doi: 10.1108/BIJ-05-2017-0092.
[16] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012, doi: 10.1109/TSMCC.2011.2161285.
[17] H. He and E. A. Garcia, “Learning from Imbalanced Data,” vol. 21, no. 9, pp. 1263–1284, 2009.
[18] R. Esmeli, M. Bader-El-Den, and H. Abdullahi, “Towards early purchase intention prediction in online session based retailing systems,” *Electron. Mark.*, 2020, doi: 10.1007/s12525-020-00448-x.
[19] P. Kaur and A. Gosain, “Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise,” *Adv. Intell. Syst. Comput.*, vol. 653, no. January, pp. 23–30, 2018, doi: 10.1007/978-981-10-6602-3_3.
[20] Y. Zhang, S. Qiao, R. Lu, N. Han, D. Liu, and J. Zhou, “How to balance the bioinformatics data: Pseudo-negative sampling,” *BMC Bioinformatics*, vol. 20, no. Suppl 25, pp. 1–13, 2019, doi: 10.1186/s12859-019-3269-4.
[21] H. Y. Wang, A. Zhang, and C. Wang, “Nonuniform Negative Sampling and Log Odds Correction with Rare Events Data,” *Adv. Neural Inf. Process. Syst.*, vol. 24, no. NeurIPS, pp. 19847–19859, 2021.
[22] P. Kang, S. Cho, and D. L. MacLachlan, “Improved response modeling based on clustering, under-sampling, and ensemble,” *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6738–6753, 2012, doi: 10.1016/j.eswa.2011.12.028.
[23] G. Marinakos and S. Daskalaki, “Imbalanced customer classification for bank direct marketing,” *J. Mark. Anal.*, vol. 5, no. 1, pp. 14–30, 2017, doi: 10.1057/s41270-017-0013-7.
[24] M. Amini, J. Rezaeenour, and E. Hadavandi, “A Cluster-Based Data Balancing Ensemble Classifier for Response Modeling in Bank Direct Marketing,” *Int. J. Comput. Intell. Appl.*, vol. 14, no. 4, pp. 1–23, 2015, doi: 10.1142/S1469026815500224.
[25] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proc. fifth Berkeley Symp. Math. Stat. Probab.*, vol. 1, no. 14, pp. 281–297, 1967.
[26] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer, 2010.
[27] S. Rogic and L. Kascelan, “Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning,” *Comput. Sci. Inf. Syst.*, vol. 18, no. 3, pp. 893–925, 2020, doi: 10.2298/csis200530052r.
[28] S. Rogic and L. Kascelan, “Customer Value Prediction in Direct Marketing Using Hybrid Support Vector Machine Rule Extraction Method,” *Commun. Comput. Inf. Sci.*, vol. 1064, pp. 283–294, 2019, doi: 10.1007/978-3-030-30278-8_30.
[29] S. Rogić, L. Kašćelan, and M. Pejić Bach, “Customer Response Model in Direct Marketing: Solving the Problem of Unbalanced Dataset with a Balanced Support Vector Machine,” *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 3, pp. 1003–1018, Jul. 2022, doi: 10.3390/jtaer17030051.
[30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC Press, 1984.
[31] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, pp. 81–106, 1986.
[32] J. Berkson, “Application of the Logistic Function to Bio-Assay,” *J. Am. Stat. Assoc.*, vol. 39, no. 227, pp. 357–365, 1944.
[33] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Ann. Stat.*, pp. 1189–1232, 2001.
[34] E. Fix and J. L. Hodges Jr., “Discriminatory analysis-nonparametric discrimination: consistency properties,” *Int. Stat. Rev.*, vol. 57, no. 3, pp. 238–247, 1989.
[35] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

Ljiljana Kašćelan is a full professor at the Faculty of Economics, University of Montenegro since 1992. Business databases (relational databases and SQL) and business intelligence (data warehouse, OLAP, big data, data mining and machine learning with applications in business) are her main research areas.

Sunčica Vuković is a teaching assistant at the Faculty of Economics, University of Montenegro. Her research interests are Marketing Analytics, Direct and Digital Marketing and Data Mining Applications in Business.

The Event Processing Network for Systematic Reduction of Interoperability Deviations in a Business Ecosystem

Mačinković, Daliborka and Marković, Vidan

Abstract: This work proposes the model of the event processing network for the systematic reduction of interoperability deviations in a business ecosystem. Complex event processing technology supports real-time events monitoring in collaborative business processes for the behaviors specified as interoperability deviations, generating alerts when such situations occur. As a reaction, alerts on interoperability deviations are delivered as personalized information to the right consumer as a designated collaboration partner. The event processing network enables collaboration partners to be proactive in interoperability deviations and to eliminate the impact of interoperability deviations on the business process objectives in the business ecosystem. In observing the data as events about realized collaborative business processes during this research, characteristic patterns of events were discovered. The logic for event processing was specified for the systematic reduction of interoperability deviations in the business ecosystem. This work proposes two processes with phases to facilitate the development of the proposed event processing network in a business ecosystem.

Index Terms: Collaborative business processes, complex event processing, interoperability, event causality, event processing network, knowledge discovery from data

1. INTRODUCTION

BUSINESS ecosystem requires flexible and adaptable relationships between collaboration partners and monitoring of mutual influences. Present misunderstandings in the collaboration of heterogeneous and autonomous systems [1] indicate that it is necessary to harmonize their collaborative business processes with different priorities, goals, and opportunities [2-7]. One of the challenges of digital transformation [8], [9] is achieving organizational interoperability. Organizational interoperability [10-16] implies previously fulfilled technical, syntactic, and semantic interoperability [17-20]. The interoperability can be threatened due to frequent changes in autonomous systems in the dynamic and turbulent environment [21], [22]. It is shown that the current ways of managing interoperability [23-27] in collaborative systems, are not proactive [4] and cannot support the detection of interoperability deviations and proactively eliminate interoperability barriers. In practice, the effects of collaboration faults are resolved, but their causes [4], [28], [29] are not resolved in the competent departments of business organizations.

Solving interoperability deviations should be extended to all collaboration participants to broadly define the significance of the problem for all collaboration participants. There is an evident lack of automated monitoring interoperability barriers [30-35] in business systems and their harmonization. After achieving interoperability [12] its regular analysis and control are missing, with the aim of continuous improvement and preservation [4], [5].

It is necessary to monitor the state of interoperability continuously [25], [26], where numerous collaborative business processes and partners exist. In the business ecosystem, it is necessary to present the detected interoperability deviations and their impact on the business process objectives to all partners in collaboration. Partners participating in collaborative business processes should be able to proactively reconcile misunderstandings created during the exchange of data, information, and services and to harmonize their business processes in the business ecosystem.

The main contribution achieved in this paper is the event processing logic for the systematic reduction of interoperability deviations in the business ecosystem. The event processing logic is specified based on the event causality in the

Manuscript received Mart 30, 2023.

Mačinković D. and Marković V. are at the University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia, E-mail: daliborka.macinkovic@teol.net,vidan.markovic @fon.bg.ac.rs.

collaborative business processes. The event processing logic by using complex event processing technology should: detect interoperability deviations in real-time [36-40] and present the interoperability deviations to the collaboration partners in the business ecosystem. Contributions that have also been achieved in this work represent process (1) for discovering event patterns in the collaborative business processes and process (2) for defining the event processing network for the systematic reduction of interoperability deviations in the business ecosystem.

This work is organized into 5 sections. Section 2, after the introduction of the work, presents the literature review on the technology of complex event processing, knowledge discovery from data, and issues of interoperability in this work. Section 3 describes a proposed event processing network for the systematic reduction of interoperability deviations in a business ecosystem. Section 3.1 presents the application of process 1 of discovering event patterns in collaborative business processes. Section 3.2 presents process 2 of describing the building blocks of the proposed event processing network for the systematic reduction of interoperability deviation in the business ecosystem. Section 4 discusses the results achieved in the example of the application of event processing logic for the systematic reduction of interoperability deviations in the selected business ecosystem. Finally, in section 5, the conclusions are presented.

2. Related Work

The most frequently cited definition of interoperability in the literature is TOGAF [41]. „Interoperability, within the context of European public service delivery, is the ability of disparate and diverse organizations to interact towards mutually beneficial and agreed to common goals, involving the sharing of information and knowledge between the organizations through the business processes they support by the means of the exchange of data between their respective ICT systems" [10]. Many concepts and terms are used in the literature considering interoperability [2], [3], [4], [19], [29]. In this paper, the term collaborative business process (CBP) is used. According to [3] CBPs are "business processes whose activities are carried out by two or more autonomous organizations".

This paper raises the question of interoperability in the business ecosystem as a basic problem. In the paper [42] it is emphasized that "approaches to foster interoperability in the era of digital innovation are not straightforward and imply complex standardization efforts, design knowledge about standards and platforms, as well as collaborative engagement between multiple stakeholders." In this paper, an event processing network is proposed, which has the task of explicitly forwarding alerts to all collaboration partners and pointing directly to those responsible for resolving interoperability barriers. In this way, interoperability problems are eliminated in the services registered for the resolution of interoperability deviations according to the interoperability layers: technical, syntactic, semantic, and organizational. The proposed solution includes both technological and organizational aspects of interoperability. The paper [5] suggests that enterprises are not interoperable because of interoperability barriers. Most of the research and developments are focused on the technology aspect to solve interoperability problems [33]. From literature [42], ''regarding the weaknesses of existing interoperability assessment models, few existing interoperability assessment models contain syntactic, semantic, or organizational interoperability layer". Technical interoperability is "usually associated with hardware/software components, systems, and platforms that enable machine-to-machine communication to take place" [12]. Syntactical interoperability is usually associated with data formats [12]. Semantic interoperability [17-20] "is usually associated with the meaning of content and concerns the human rather than machine interpretation of the content" [12]. Organizational interoperability depends on successful technical, syntactic, and semantic interoperability [12], [43-47].

The proposed solution in this paper allows indicating to collaboration partners in the business ecosystem the impact of interoperability deviations on the objectives of business processes. The proposed solution in this paper can improve the exchange between actors and create value in ecosystems. Interoperability has been considered a major bottleneck in digital business ecosystems [48], hampering collaboration and information exchange between the actors involved [49-51].

The proposed solution in this paper is intended for collaborative systems during the operational phase. This solution detects deviations of interoperability from data about realized collaborative business processes in real-time in the business ecosystem. By requiring a specified event processing logic based on the inheritance of interoperability deviations in the entire business ecosystem, this solution enables collaboration partners to proactively eliminate interoperability deviations.

A significant number of works deal with interoperability measuring problems [30-35], [52]. Paper [33] emphasizes that ''Existing approaches mainly focus on maturity measure issues" [7],

[10], [11], [17], [18] less on interoperability compatibility and performance measure. The interoperability compatibility measurement can only be performed when the two partner/system of the interoperation is known. The performance measurement is to be done during the test or operation phase of two interoperate enterprises.

The proposed solution in this paper can be an upgrade to any of the ways to achieve interoperability that have already been implemented in the business ecosystem, such as implemented standards, technology platforms, architectures, and other solutions. The proposed solution can be adapted to any implemented solutions, which avoids hindering competitiveness, innovation, security, and reliability.

According to the literature review that was dealt with in the paper [41] "Factors leading to emergent low interoperability are diverse and include uncoordinated changes to agreed-upon standards, technological turbulence, that is, multiple new standards or platforms that emerge in an ecosystem" [53] ‘’or the individual commercial interests of stakeholders involved in the development of an interoperability standard hindering agreement on technologies and specifications. [54-57].

The proposed solution contributes to sustainability [58] and interoperability. The proposed solution in this paper enables continuous adaptation of the event processing network to changes in collaboration systems. One of this sustainability is envisaged and includes the ability to detect interoperability deviations in time optionally.

The knowledge discovery process is shown in the paper [59] as an iterative sequence of the following phases: (I) Data cleaning, (II) Data integration, (III) Data selection, (IV) Data transformation, (V) Data mining, (VI) Pattern evaluation, (VII) Knowledge presentation. These phases were adapted in this research for systematic reduction deviation interoperability in the business ecosystem. The phase of data mining is adapted to using qualitative, quantitative, and cause-and-effect analyses.

The proposed phase of pattern evaluation specifies the event processing logic for detecting interoperability deviations in the business ecosystem. The proposed phase of knowledge presentation specifies the event processing logic for generating alerts about the detected interoperability deviations to the collaboration partners in the business ecosystem. These proposed phases of process 1 are also a research method used during the observation of the system in collaboration in this work, where conclusions were reached about the event causality in collaborative business processes and the inheritance of interoperability deviations in the business ecosystem.

The proposed solution in this paper supports the autonomy of systems in collaborations. The event processing model is based on the complex event processing (CEP) technology that enables real-time processing, flexibility, independence, and extensibility without affecting the autonomous systems of participants in collaborative business processes. Many papers define complex event processing [21], [22], [36-40], [60-64].

It was determined the technology of complex event processing that can meet the requirements of detecting interoperability deviations in the huge amounts of data created by recording data about realized collaborative business processes and monitoring interoperability in a business ecosystem. Event processing, known in the scientific literature as complex event processing [21], [22], [28], [29], [36-40] is a key approach in the development of smart systems for extracting complex events as valuable information from the flow of primitive events.

The proposed event processing network for the systematic reduction of interoperability deviations is specified by seven building elements: event types, event producers, event consumers, event processing agents, global state elements, event channels, and event contexts following the definition method described in the book [65].

As highlighted in the book [65], the reasons for applying complex event processing principles are accepted for the proposal offered in this paper: ‘’it supports event observation and is used to monitor systems or processes for unexpected behaviors and generate alerts when such situations occur’’ [40]. As a reaction, alerts are created for the consumer. It delivers personalized information to the right consumer at the right time.

Event causality is described as a key term in the book [28]. In this paper, the event causality is considered from the aspect of interoperability, and collaboration faults can be identified that affect the values of the business process.

3. Systematic Reduction of Interoperability Deviations in the Business Ecosystem

This section presents the event processing network for the systematic reduction of interoperability deviations in a business ecosystem. The task of the proposed event processing system is to detect interoperability deviations and forward generated alerts to collaboration partners about interoperability situations in the business ecosystem. Complex event processing technology supports event observation and is used in the proposal to

monitor collaborative business process behaviors specified as interoperability deviations and to generate alerts when such situations occur.

The event processing network for the systematic reduction of interoperability deviations in collaborative business processes with a large number of collaboration partners should enable the following:

- Sustainable interoperability despite changes that occur over time in the business ecosystem;
- Automated and continuous monitoring of interoperability in a collaborative environment with numerous autonomous and heterogeneous information systems;
- High degree of knowledge and awareness of interoperability in the business ecosystem;
- Proactive resolution of interoperability deviations;
- Insight into the detected interoperability deviations and the effect of the detected interoperability deviations on the value of business process objectives to collaboration partners in the business ecosystem.

Two proposed processes in this paper facilitate the development and definition of the event processing network for the systematic reduction of interoperability deviations in a business ecosystem:

1. Process of discovering the event patterns in the collaborative business processes and specifying the logic of event processing;
2. Process of defining building components of the event processing network.

Executed process 1 should result in the specification of the event processing logic:

- Specified event processing logic for detecting interoperability deviations in the selected business ecosystem;
- Specified event processing logic for generating alerts for collaboration partners where detected interoperability deviations were;
- Specified event processing logic for generating alerts for proactive action to designated collaboration partners to avoid interoperability deviations;
- Specified event processing logic for generating alerts for designated collaboration partners to eliminate the impact of interoperability deviations on business process objectives.

Phases of process 1 are also a research method that was used during the observation of collaborative business processes and arrived at the data processing logic for the systematic reduction of interoperability deviations in the business ecosystem.

In this research, the collaborations of information systems of healthcare and healthcare institutions were observed. During the observation for 100 days, data about the performed collaborative business processes were collected. The goal was to discover the characteristic event patterns in the data recorded about the performed collaborative business processes and specify the logic for event processing.

The analysis of collected data on collaborative business processes led to the conclusion that the occurred interoperability deviation in one collaborative business process can also appear in other collaborative business processes of the observed business ecosystem. This inheritance of interoperability deviations indicated the possibility of its proactive prevention where the interoperability deviation did not occur at numerous collaboration partners in the observed business ecosystem.

The conditions under which the inheritance of interoperability deviations is confirmed are also defined. Confirmed event causality in any collaborative business process of a selected business ecosystem points to the interoperability deviation.

Identification of collaborative business processes and the business ecosystem, in which the event processing network can be applied for the systematic reduction of interoperability deviations should support the following:

- Collaborative business processes in which collaboration faults can be identified that affect the values of the business process (confirmation of the event causality) and that these values are used in a series of collaborative business processes as a ''common collaborative activity''.
- Collaborative business processes in which a “common collaborative activity” is used.
- Collaborative business processes in which the event causality is confirmed where collaboration faults have effects on the business process objective values as interoperability deviations.
- A “common attribute” must exist in other collaborative business processes for proactive action.
- A large number of collaborative partners should exist in the business ecosystem.

After specifying the event processing logic, it is necessary to define the event processing network building components for the chosen collaborative business processes and set event processing logic in the event processing agent component.

3.1 *Application of Process 1 of Discovering Event Patterns in Collaborative Business Processes*

In phase 1 of process 1, the following were identified: 4 collaborative business processes with ''common collaborative activity'', sources of data on the results of performed collaborative business processes, and partners participating in the selected collaborative business processes.

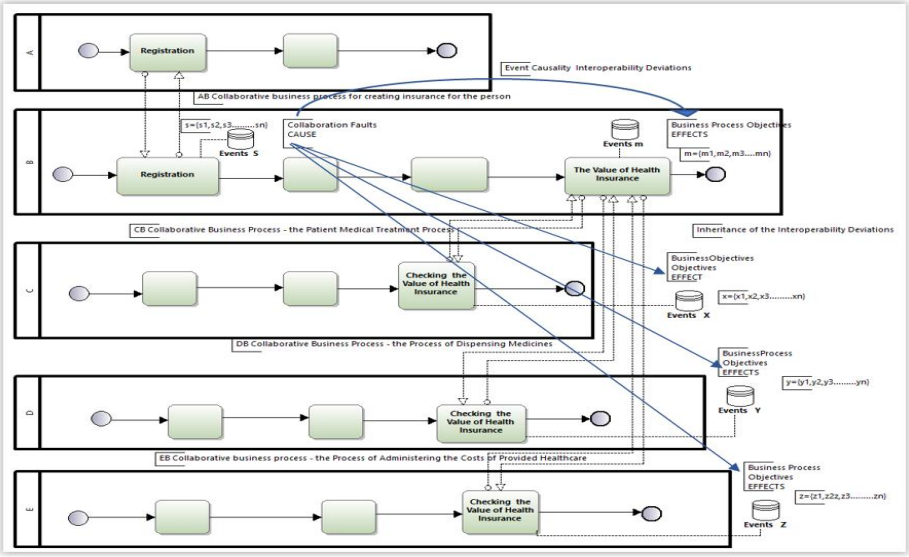


Figure 1. Event causality and inheritance of the interoperability deviations to the business ecosystem

Four collaborative business processes (CBP) were identified according to the type of collaboration partners, and they are labeled as AB, CB, DB, and EB, as shown in Figure 1:

Type of collaboration AB - The collaborative business process (AB) for health insurance institutions (B) and the institution (A) was identified. In this collaborative business process, data on collaboration faults in the process of creating insurance for the insured person is logged;

Type of collaboration CB - The collaborative business process CB between hospitals, clinics, family health institutions, specialist health institutions (C), and a health insurance institution (B) was identified to check data in the patient medical treatment process. In this collaborative business process, data is logged as a result of the performed collaboration with data about the objectives of the business process for the patient;

Type of collaboration DB - The collaborative business process (DB) between pharmacies (D) and health insurance institutions (B) was identified to check data in the process of dispensing medicines. In this collaborative business process, data is logged as a result of performed collaboration with data about the objectives of the business process for the patient;

Type of collaboration EB - The collaborative business process (EB) between health care institutions (E) and health insurance institutions (B) was identified to check data in the process of administering the costs of provided health services;

In phase 2, the data collected from the collaborative business processes AB, CB, DB, and EB were preprocessed according to phases: data cleaning, data integration, data reduction, and data transformation.

Data cleaning – Values of collaboration data were monitored daily. The values of collaboration faults in set S have been recorded, while the business process objective values in sets X, Y, and Z have been recorded. First, only values false to business process objectives in X, Y, and Z have been recorded. However, due to a more accurate and broader picture, it was decided to collect value true to business process objectives. For inconsistencies, the data collected during the weekend, which should be viewed differently, had been rejected as inconsistencies in the analysis. After data cleaning, consistent data for the following data analysis was obtained.

Data integration - The data about collaborations AB, CB, DB, and EB from different data sources have been recorded in Excel tables daily. Due to a large amount of data and easy data manipulation, the database was created from all data sources.

Data selection – The key attributes in observed data have been selected and their dependent attributes in the tables S, X, Y, and Z:

S (DeviationOccurenceTime_s, **DeviationID**, DeviationSemantic, InsuredID)

X, Y, Z (DeviationOccurenceTime, InstitutionID, PatientId, **Status**)

Key attributes ''Collaboration Faults'' = ''DeviationID'' and ''Values of business process

objective" = ''Status'' are determined.

Data transformation - Selected data have been transformed in the format SX, SY, and SZ in which the collaboration fault and the business process goal value are linked according to the same entity PatientID=InsuredID.

SX (PatientID, InstitutionID, **Status, DeviationID**, DeviationSemantic, DeviationOccurenceTime_s, DeviationOccurenceTime)
SY (PatientID, InstitutionID, **Status, DeviationID**, DeviationSemantic, DeviationOccurenceTime_s, DeviationOccurenceTime)
SZ (PatientID, InstitutionID, **Status, DeviationID**, DeviationSemantic, DeviationOccurenceTime_s, DeviationOccurenceTime)

In phase 3 the quantitative analysis provided a deeper understanding of the data as an event about the observed collaborative business process. By summarizing the number of occurrences of events by day, in which the relationship between the event "collaboration fault" and the event "business process objective value" was observed, the need to develop the event processing network for the systematic reduction of interoperability deviations was confirmed.

In phase 4 as the qualitative analysis, the values for the attributes: interoperability layer, competent departments, workplace, and type of deviation were set as additional information on interoperability deviations in the state table that will be used in the operation of the event processing network.

In phase 5 as the cause-effect analysis, the event causality in the collaborative business process was confirmed. The event causality is confirmed by experimental testing whether the event s (collaboration faults) cause events x, y, and z (business process objective values) for the observed many entities in the collaborative business processes (AB, CB, DB, and EB). The following clarification is given:

CB collaboration **(s1→ x1)** - for the observed patient, the collaborative fault (event s1) caused the health insurance value (event x1) in the patient treatment process, which confirms the event causality as the interoperability deviation (sx);

DB collaboration **(s1→ y1)** - for the observed patient, the collaborative fault (event s1) caused the health insurance value (event y1) in the drug dispensing process, which confirms the event causality as the interoperability deviation (sy);

EB collaboration **(s1→ z1)** - for the observed patient, the collaborative fault (event s1) caused the health insurance value (event z1) in the process of administration of the patient's medical treatment costs, which confirms the event causality as the interoperability deviation (sz).

This confirms the impact of the collaborative fault (event s) on business processes objectives (events x, y, z) as the principle of the event causality in all observed collaborative business processes CB, DB, EB of the business ecosystem.

The experiment confirmed that the interoperability deviations that occurred in one collaborative business process are inherited (are appeared) in any collaborative business process with the confirmed event causality. From the inheritance of interoperability deviations comes the possibility of the appearance of collaboration faults in collaborative business processes in which the event causality is confirmed under the condition that there is an attribute on which the collaboration faults occurred, called a "common attribute".

Interoperability deviations that occurred in CB, DB, and EB by certain collaboration partners as events sx, sy, sz with values of collaboration faults "DeviationID" = (26, 8, 6) confirm the possibility of appearing these collaboration faults at the partners in the collaborative business processes where interoperability deviations have not yet occurred. In these examples, it was identified that common attributes exist: activity code and municipality code in the events x, y, or z.

This confirms the inheritance of interoperability deviations in collaborative business processes conditioned by the "common attribute". This conclusion indicated the possibility of proactively eliminating interoperability barriers in the business ecosystem where interoperability deviations have not yet occurred.

In phase 6, the event patterns representing interoperability deviations in collaborative business processes are identified. The patterns are recognized based on the previously conducted analysis in phases 1 to 4 and the conducted experiment in phase 5 in which the impact of the collaboration faults on the goals of collaborative business processes is confirmed. The recognized event pattern is translated into event processing logic and specified in Listing 1.

```
Input event:  /*stream of data as results of collaborative business processes*/
      s       /*event stream contains independent variable CollaborationFault AB*/
      x       /* event stream contains dependent variable BPObjective CB*/
      y       /* event stream contains dependent variable BPObjective DB*/
      z       /* event stream contains dependent variable BPObjective EB*/

Output event: /*detected interoperability deviation*/
      sxyzDevIOP (sx,sy,sz)        /* event stream Interoperability deviations*/

Method: /*complex event processing*/
/*detection of interoperability deviations*/
/*t-event time*/
(1)  select x.PatientID, x.InstitutionID, x.Status, s.DeviationID, s.DeviationSemantic,
           s.DeviationOccurenceTime_s, x.DeviationOccurenceTime
(2)  into sxyzDevIOP   /*Interoperability deviations*/
(3)  from s join x    on s.InsuredId=x.PatientId   /*sx*/
           (3.1)  from s join y  on s.InsuredId=y.PatientId   /*sy union all */
           (3.2)  from s join z  on s.InsuredId=z.PatientId ) /*sz union all*/
(4)  where DeviationOccurenceTime_s<DeviationOccurenceTime /*(x,y,Z)*/
                Optional:
(5)             …count (DeviationID), (InstitutionID) ...
                 group by (DeviationID), (InstitutionID) /*day, hours, minutes*/
```

Listing 1. Event processing logic 1 for detecting the interoperability deviations in the collaborative business processes.

Listing 1 specifies the event processing logic to detect interoperability deviations in the business ecosystem. Input events are collaboration faults

(event s) and collaborative business process goals (events x, y, z). Output events are detected interoperability deviations (events sx, sy, sz).

The logic for the event processing in Listing 1: For each event x, y, or z (Collaborative business processes goals), it is querying whether there is an event s (Collaborations Faults) for the same entity InsuredID(s) = PatientID(x,y,z) and according to the condition that the time of event s is older than the time of event x, y or z. All events that satisfy this logic are generated as complex events named interoperability deviations (events sx, sy, or sz).

Phase 7 is the visualization (the presentation) of the interoperability deviations. Numerous partners in the business ecosystem should have insight into discovered interoperability deviations as well as insight into the impact of collaboration faults on the values of business process goals. Based on the insights, collaboration partners can eliminate the causes of interoperability deviations both reactively and proactively.

```
Input events:
      sxyzDevIOP            /*detected interoperability deviation in time*/
      All_Institution_Details /*GSE state table {C1, C2, C3.Cn}, {D1, D2, D3.Dn}, {E1, E2, E3.En},
                              TypeDeviation, IOP_layer, CompetentDepartment, Competent-work*/

Output events:
      Alert DevIOP_RE       /*Alerts reactive */

Method:  /*complex event processing*/
(1)     select dev.PatientId,dev.InstitutionId,dev.DeviationID,dev.DeviationSemantic,dev.DeviationOccurenceTime,
        dev.Status,all.IOP_layer,all.TypeDeviation,all.CompetentDepartment_DevIOP,all.CompetentWork_DevIOP
(2)     into Alert DevIOP_RE   /*Alerts reactive */
(3)     from sxyzDevIOP dev
(4)     join All_Institution_Details all
(5)     on dev.InstitutionID=all.InstitutionID
(6)     and dev.DeviationID=all.DeviationID
(7)     where TypeDeviation='G'
```

Listing 2. Event processing logic 2 to generate alerts for the collaboration partners about interoperability deviations.

Listing 2 specifies event processing to indicate to collaborating partners in the business ecosystem the interoperability barriers where interoperability deviations have occurred.

Input events are detected interoperability deviations and additional data such as type of deviation, interoperability layer, and the department for resolving misunderstandings all from the global state table (element). Output events are alerts for collaboration partners about interoperability deviations to remove interoperability barriers. The logic for the event processing in Listing 2: Each event of interoperability deviations is enriched with additional data from the global state table. After that, only for deviations selected as type G, alerts are generated for all institutions where the interoperability deviation event occurred.

Listing 3 specifies event processing that generates alerts that enable collaboration partners in the business ecosystem to avoid interoperability deviations proactively. Input events are the detected interoperability deviations for collaboration partners and additional data such as institutions, interoperability layers, deviation type, and responsible departments for removing interoperability barriers.

```
Input events:
      Alert DevIOP_RE       /* alerts reactive*/
      All_Institution_Details /* GSE state table {C1, C2, C3.Cn}, {D1, D2, D3.Dn}, {E1, E2, E3.En},
                              TypeDeviation, IOP_layer, CompetentDepartment, Competent-Work.. */
Output events:
      Alert DevIOP_PRO  /*Alerts Proactive*/

Method:   /*complex event processing*/
(1) select all.InstitutionId, all.DeviationID, all.DeviationSemantic, all.IOP_layer, all.TypeDeviation,
        all.CompetentDepartment_DevIOP, all.CompetentWork_DevIOP
(2) into Alert_DevIOP_PRO  /*Alerts proactive*/
(3) from  All_Institution_Details all
(4) left join Alert_DevIOP_RE re on
(5) re.DeviationId=all.DeviationID and
(6) re.InstitutionID=all.InstitutionId
(7) where re.InstitutionID is null
(8) and all.DeviationID in (
(9) select distinct(DeviationID) from Alert DevIOP_RE)
```

Listing 3. Event processing logic 3 to generate the alerts for proactively avoiding interoperability deviations.

Output events are alerts for the proactive prevention of interoperability deviations. The logic for the event processing in Listing 3: Alerts are generated for each detected interoperability deviation of a global nature and forwarded to those collaboration partners of the business ecosystem where interoperability deviations have not yet been detected. These partners can proactively remove interoperability barriers.

```
Input events:
      sxyzDevIOP            /* detected interoperability deviation*/
      All_Institutions_details  /* GSE state table {C1, C2, C3.Cn}, {D1, D2, D3.Dn}, {E1, E2, E3.En},
                              TypeDeviation, IOP_layer, CompetentDepartment, Competent-Work */
Output event:
      Alert_BPO  /*alert to avoid the impact of deviations on business process objectives */

Method: /*complex event processing*/
(1)     select select dev.PatientId,dev.InstitutionId,dev.DeviationID,dev.DeviationSemantic,dev.DeviationOccurenceTime,
        dev.Status,all.IOP_layer,all.TypeDeviation,all.CompetentDepartment_BPO,all.CompetentWork_BPO
(2)     into Alert_BPO
(3)     from sxyzDevIOP dev
(4)     join All_Institution_Details all
(5)     on dev.InstitutionID=all.InstitutionID
(6)     and dev.DeviationID=all.DeviationID
```

Listing 4. Event processing logic 4 to eliminate the impact of interoperability deviations on business process objectives.

Listing 4 specifies event processing to generate alerts to eliminate the impact of interoperability deviations on business process objective values in the business ecosystem.

Input events are the detected interoperability deviation and additional data such as institutions, interoperability layers, deviation type, and responsible departments to eliminate the impact. Output events are alerts for the elimination of the impact of deviations on the goals of business processes. The logic for the event processing in Listing 4: For each enriched event interoperability deviations are generated alerts to collaboration partners to eliminate the impact of interoperability deviations on business process goals. All data about the interoperability deviation and the impact of the deviation on the business process goal are sent in an alert.

The event processing logic shown in Listing 1 through 4 should enable the detection of interoperability deviations and the generating of alerts for collaborative partners for any selected system with details that will be specific to that system but respecting this general logic. Logic 1

to 4 is represented as a method in a general way and can be adapted during application. The column, table, and database names also will be different for each system that needs to be specified.

3.2 *Application of Process 2 of Definition of the Event Processing Network*

Implementation of phases of the process (2) resulted in platform-independent definition elements (components) of the event processing network for the systematic reduction of interoperability deviations in the chosen business ecosystem:

1. Definition element of the event producers;
2. Definition element of the event types;
3. Definition element of the event processing agents (EPAs);
4. Definition element of global state table (GSE);
5. Definition element of the event contexts;
6. Definition element of the event channels;
7. Definition element of the event consumers.

Process 2 uses phases described and used in the book [65] that were adapted in this work to the systematic reduction of interoperability deviations in the business ecosystem.

These platform-independent definition elements were translated into specific elements for the WSO2 Complex Event Processing platform. The requirements that the event processing network for the systematic reduction of interoperability deviations in the chosen business ecosystem needs to perform:

- Receives s, x, y, z simple events as the result of collaboration, sent by four event producers as the collaborative business processes AB, CB, DB, and EB;
- Detects interoperability deviations as complex events;
- Sends the detected interoperability deviations to the event processing agent to enrich detected events with data from the global states table;
- Sends enriched events to subsequent event processing agents (The enriched detected interoperability deviations);
- Generates alerts for collaboration partners at which interoperability deviations are detected;
- Emits the generated alerts to the event consumer via the event channel;
- Generates alerts for proactive action to designated collaborative partners to avoid interoperability deviations;
- Emits generated alerts to event consumers via the event channel;
- Generates alerts for designated collaboration partners to eliminate the impact of interoperability deviations on business process objectives;
- Emits generated alerts to event consumers via the event channel. Event channel route to the designated collaborative partners.

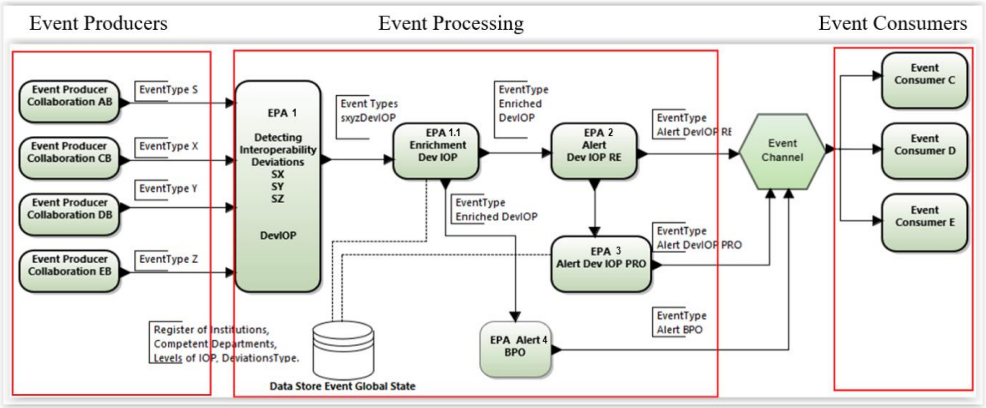


Figure 2. The event processing network for the systematic reduction of interoperability deviations in the chosen business ecosystem.

Figure 2 illustrates the event processing network for the systematic reduction of interoperability deviations with the graphic notation used in the literature [65] and the specifics of event processing for the proposed network for the systematic reduction of interoperability deviations and the selected business ecosystem.

An event processing network for the systematic reduction of interoperability deviations is represented as a collection of event producers, event processing agents, event consumers, global state elements, and event channels according to the literature [65].

On the left side of Figure 2 are the event producers. In the selected example, the event producers are the systems of institutions that broadcast events from the observed four collaborative business processes. In four collaborative business processes are

collaborative partners as clinics, hospitals, pharmacies and health insurance. The event producers send events that enter the event processing system. The event producers named Collaboration AB, Collaboration CB, Collaboration DB, and Collaboration EB broadcast events as four types of events: EventType_s, EventType_x, EventType_y, EventType_z to the event processing agent (EPA DevIOP).

In the central part of Figure 2, there are event processing agents that, according to the set logic from 1-4, process the events they receive from the event producer.

The event processing agent 1 executes event processing logic 1 to detect interoperability deviations. This event processing agent receives events from event producers and associates each event type x, event type y, and event type z by the same entity (InsuredID=PatientID) with event type s and under the condition that the time of the event s is with less than the time of event x, y, z. Only those events that meet the condition are detected interoperability deviations (sxyzDevIOP). These events are passed on to the next event processing agent. Other x, y, and z events (which have no pair) are discarded. Events are retained daily or optionally monthly depending on the degree of interoperability in the business ecosystem.

The event processing agent 1.1 for enriching detected interoperability deviations. This agent takes detected interoperability deviations and enriches them with additional data from the global state table (GSE). This agent makes a query asking for each event to receive the following values: the interoperability layer for each detected deviation, the deviation type (G or L), and the responsible department for the institution to resolve the interoperability barrier that was detected. When it finds these values, it adds them to the received event instance and it is a derived event called Enriched_DevIOP which is passed to the next event processing agent EPA Alert DevIOP_RE. One copy of the event instance value is passed by Enriched_DevIOP to EPA Alert_BPO for further processing.

The event processing agent 2 executes the event processing logic 2 to generate alerts for detected interoperability deviations. This agent receives the Enriched_DevIOP event and checks for each deviation type whether the Type_IOP value is of type ‘’G’’(global). All those interoperability deviations that are of type G are forwarded as an Alert_DevIOP_RE event to the event channel for broadcast to a specific Consumer_C, Consumer_D, and Consumer_E event consumer. A second copy of these events is forwarded to the EPA Alert_DevIOP_PRO event processing agent for further processing.

The event processing agent 3 executes the event processing logic 3 to generate alerts for proactivity. This agent receives events from EPA Alert_DevIOP_RE and by query finds in the GSE_table according to the detected error type and institution, those institutions different from the institutions where Alert_DevIOP_RE events occurred and the real new derived event Alert_DevIOP_PRO. These generated events are forwarded via the event channel to event consumers Consumer_C, Consumer_D, and Consumer_E.

The event processing agent 4 executes event processing logic 4 to generate alerts to eliminate the impact of deviations on business process objectives Alert_BPO. These generated events are forwarded via the event channel to event consumers of types Consumer_C, Consumer_D, and Consumer_E.

The right side of Figure 2 shows the consumers of the event. The event consumers are institutions participating in collaborative business processes (in this example 242 institutions) and are represented as types Consumer_C, Consumer_D, and Consumer_E. Event types received by Channel Consumers are events: Alert_DevIOP_RE, Alert_BPO, and Alert_DevIOP_PRO which are the results of event processing from the described event processing agents. These Alerts do not go to all institutions, but only to those highlighted in the generated events Alert_DevIOP_RE, Alert_BPO, and Alert_DevIOP _PRO.

4. Evaluation and Discussion

In the given example, the proposed logic for event processing is applied, which should achieve the reduction of interoperability deviations in the business ecosystem. Input event streams were provided to the event processing agent and the results are as follows:

- 184 events (s) as the input event streams from collaboration AB - These events occurred during the creation of health insurance for a person in the AB collaborative business process and were emitted with the values about the collaboration faults through the defined event stream;
- 18.151 events (x) as the input event streams from collaboration CB – These events occurred during the health insurance check in the CB collaborative business process during patient treatment and were emitted with values about the business process objectives through the defined event stream;
- 1.185 events (y) as the input event streams from collaboration DB -These events occurred during the health insurance check in the collaborative business process DB

dispensing medicines for patients and were emitted with values about the business process objectives through the defined event stream;

- 6.808 events (z) as the input event streams from collaboration EB - These events occurred during the health insurance check in the EB collaboration business process during the patient treatment cost administration process and were emitted with values about the business process objectives through the defined event stream.

The output streams after the application of the Execution Plans results are as follows:

- 137 interoperability deviations were detected (Detected Dev IOP) and emitted to the next of the event processing agent (According to the event processing logic 1);
- 10 interoperability deviation events were generated as alerts for collaboration partners where the interoperability deviations were detected (Alerts Dev IOP RE) and were broadcast to the event consumers (According to the event processing logic 2);
- 950 alerts were generated to proactively avoid interoperability deviations in the business ecosystem (Alert Dev IOP PRO) and were broadcast to the event consumers. (According to the event processing logic 3);
- 137 alerts were generated to eliminate the effects of collaboration faults on the business process objectives (Alert BPO) and were broadcast to the event consumers (According to the event processing logic 4).

PRODUCERS							CONSUMERS			
Collaborative Business Process Type	Number of Collaboration Partners on the Business Ecosystem	Input event streams			Event Processing	Output event streams	Number of alerts for the collaboration partners' type			
								C	D	E
AB	2	S	Collaboration Fault	184 ->	Logic 1 Logic 2 Logic 3 Logic 4	-> **Detected Deviation Interoperability**	**137**			
CB	87	X	BusinessProcess Objective	18151 ->		-> **Alert DevIOP RE**	10	7	1	2
DB	15	Y	BusinessProcess Objective	1185 ->		-> **Alert DEVIOP PRO**	950	341	59	550
EB	138	Z	BusinessProcess Objective	6808 ->		-> **Alert BPO**	137	130	3	4

Figure 3. Example of the event processing for systematic reduction of interoperability deviations.

Clarification of the results obtained in the application of event processing logic for the systematic reduction of interoperability deviations in the example, as shown in Figure 3.

After event processing according to set logic 1, 137 events as the detected deviation interoperability were detected.

After event processing according to set logic 2, 10 events were generated as alerts for reactivity (Alerts Dev IOP RE) for collaboration partners where interoperability deviations occurred. Ten alerts were forwarded to the designated institutions, explicitly to the competent departments that harmonize interoperability barriers according to the details from the alerts on detected deviations.

Results by the interoperability layers show that out of 10 interoperability deviations, seven belong to the semantic layer of interoperability and three to the syntactical layer. It is possible to eliminate the seven interoperability barriers at the semantical layer and the three interoperability barriers at the syntactical layer of collaborative business processes. The generated (Alert Dev IOP RE) alerts can be viewed through the type of consumers to whom the alerts were delivered, as shown in Table 1.

After event processing according to set logic 3, 950 alerts for proactive action to designated collaboration partners to avoid interoperability deviations were generated.

Table 1. Generated alerts through the layers of interoperability and the event consumers

Alert_DevIOP_RE IOP Layer	Event Consumer C	Event Consumer D	Event Consumer E
Semantical	6	0	1
Syntactical	1	1	1

Of the ten detected interoperability deviations, four collaboration faults were selected as unique. Of the four collaboration faults, 950 proactive alerts were generated for the designated institutions as the event consumers. Of the four interoperability deviations, it is necessary for 240 institutions to proactively harmonize the 713 interoperability barriers on the semantical layer and the 237 interoperability barriers on the syntactical layer, according to details from 950 proactive alerts (Alert Dev IOP PRO), as shown in Table 2.

Table 2. Generated alerts through the layers of interoperability and the event consumers

Alert_DevIOP_PRO IOP Layer	Event Consumer C	Event Consumer D	Event Consumer E
Semantical	255	45	413
Syntactical	86	14	137

After event processing according to set logic 4, 137 alerts were generated and forwarded to

exactly specified collaboration partners to eliminate the effects of interoperability deviations on business process objectives. If we look at the detected interoperability deviations, the faults that occurred during the (AB) collaboration affected the business process goals of the CB, DB, and EB collaboration, as shown in Table 3.

Table 3. Generated alerts through the collaborative business processes

Alert BPO	Collaborative business process		
	CB(x)	DB(y)	EB(z)
137	130	3	4

The event processing network can be customized according to the current state of interoperability in the chosen business ecosystem. The state of interoperability in the business ecosystem can be monitored with daily reports on the number of interoperability deviations for all collaboration partners. The power of the event processing network for the systematic reduction of interoperability deviations can be increased by including more collaborative partners and more collaborative business processes for which interoperability deviations will be detected.

The ten detected interoperability deviations in real-time enabled the rapid removal of interoperability barriers in institutions that were sent alerts with accompanying data about collaboration faults. According to the event processing logic, alerts are delivered to the relevant departments for the interoperability layers.

The realized contribution in the selected business ecosystem is reflected in 950 alerts for proactivity that enable removing the barrier of interoperability in the business ecosystem with numerous collaborative partners as the institutions participating in collaborative business processes. So, through proactive action, interoperability deviations where they have not yet occurred at institutions in the business ecosystem were avoided.

5. CONCLUSION

This study aimed to discover the characteristic event patterns in the data about the collaborative business processes and specify the logic for event processing for the systematic reduction of interoperability deviations in the business ecosystem. This paper considers the interoperability deviations observed in collaborative business processes involving numerous collaboration partners. In the business ecosystem, it needs to present the detected interoperability deviations and their impact on the values of the business process objective to all partners in collaboration. Partners participating in collaborative business processes should be able to proactively reconcile misunderstandings created during the exchange of data, information, and services and to harmonize their business processes in the business ecosystem.

In this work, the event processing network was proposed with specific event processing logic for achieving the systematic reduction of interoperability deviation in collaborative business processes. Furthermore, processes 1 and 2 are proposed to facilitate the development of the event processing network in the business ecosystem.

The ability of the proposed event processing network to process large amounts of collaboration data and detect interoperability deviations in real-time is based on event causality that the collaboration fault causes the value of business process objectives. In this proposal, complex event processing technology supports event observation, detection of interoperability deviations, and generation of alerts for specific consumers. During the operation of the event processing network, the amount of data depends on the number of collaborative business processes that are included in the event processing network.

The detected interoperability deviations and generated alerts in the provided example of the application of event processing logic, explained in the previous section, indicate that the proposed event processing network fulfilled the set tasks.

Sustainable interoperability can be accomplished despite changes occurring over time in the business ecosystem by applying the proposed event processing network.

The detected interoperability deviations indicate that the proposed event processing network can perform automatic and continuous interoperability monitoring in a collaborative environment with numerous autonomous and heterogeneous information systems. The improvement of interoperability is possible through the continual elimination of interoperability deviations.

Knowledge and awareness of interoperability in the business ecosystem can be achieved by applying the proposed event processing network.

The generated alerts indicated that the proposed event processing network provides real-time insight into detected interoperability deviations to collaborative partners in the business ecosystem.

It has been confirmed that the development of an event processing network for the systematic reduction of interoperability deviations in a business ecosystem can be facilitated by applying the proposed process 2. Following the proposed process 1 through the chosen example,

it was confirmed that it is possible to specify the logic for the event processing for the selected business ecosystem.

Based on the observed logic in this research, that interoperability deviations are inherited in the entire business ecosystem if they contain a common attribute, it is possible to proactive reconcile misunderstandings in the business ecosystem. The principles of event causality in collaborative business processes and the inheritance of interoperability deviations in the observed business ecosystem were confirmed with the collected data from the observed collaborative business processes.

The analysis of the results obtained by the event processing agents in the derived example indicated the justification of the development of the event processing network for the systematic reduction of interoperability deviations. The results also indicate the need for continuous monitoring of interoperability in the business ecosystem. Observed interoperability deviations were classified according to interoperability layers: syntactic, semantical, and organizational, which enabled the harmonization of interoperability barriers in an easier way.

For the application of the event processing network for the systematic reduction of interoperability deviations, the health insurance administration and healthcare systems were selected, which confirmed that the different business domains of the system do not affect the development of the event processing network. It can be concluded that the event processing network development for the systematic reduction of interoperability deviations has a satisfactory degree of generality and can be adapted for application in any domain of the business ecosystem.

The proposed event processing network was applied in the business ecosystem where the collaboration of business systems has already been established according to the set data formats and ways of exchanging data and services. This confirmed that it is possible to apply the proposed event processing network in an environment where the collaboration of information systems is already established so that the proposed model is not affected by principles, frameworks, standards, system architectures, and technologies already used as interoperable solutions. It can be concluded that the proposed event processing network is an upgrade to previous solutions and ways of managing interoperability.

The subject of future research is the upgrade of solutions for the detection of interoperability deviations in a collaborative business ecosystem. The focus of future research is the field of artificial intelligence for improving the detection of event patterns, by learning from the environment and adapting to the environment, based on which the proposed network for processing events in collaborative business processes of the business ecosystem will work.

REFERENCES

[1] Ruggaber, R., “ATHENA - Advanced Technologies for Interoperability of Heterogeneous Enterprise Networks and Their Applications.”, In: Konstantas, D., Bourrières, J.P., Léonard, M., et al. (Eds.), Interoperability of Enterprise Software and Applications. Springer London, p.459-460. [doi:10.1007/1-84628-152-0_45], 2005

[2] Leal, G., Guedria, W., Panetto, H., Lezoche, M., “Towards a comparative analysis of interoperability, assessment approaches for collaborative enterprise systems. HAL Id: hal-01376442, https://hal.archives-ouvertes.fr/hal-01376442, 2016

[3] Ziemann, J., “Architecture of Interoperable Information Systems”: An Enterprise Model-Based Approach for Describing and Enacting Collaborative Business Processes, 2010

[4] Camara, M.S., Ducq, Y. & Dupas, R., “A methodology for the evaluation of interoperability improvements in inter-enterprises collaboration based on causal performance measurement models”, International Journal of Computer Integrated Manufacturing, http://dx.doi.org/10.1080/0951192X.2013.800235, 2013

[5] Chen, D., “Enterprise Interoperability Framework”, Proc. Enterprise Modelling and Ontologies for Interoperability, EMOI-Interop, 2006

[6] Mallek, S., Daclin, N., Chapurlat, V., “The application of interoperability requirement specification and verification to collaborative processes.”, In Computers in industry, vol. 63, issue 7, pp. 643–658, 2012

[7] Roque, M., Chapurlat, V., “Interoperability in Collaborative Processes: Requirements Characterisation and Proof Approach.”, In Leveraging Knowledge for Innovation in Collaborative Networks, vol. 307 of the series IFIP Advances in Information and Communication Technology, pp 555-562, 2009

[8] Kraus, S., Schiavone, F., Pluzhnikova, A., Invernizzi, A.C., “Digital transformation in healthcare: analyzing the current state-of-research”, Elsevier, J. Bus. Res., Volume 123, 557–567, https://doi.org/10.1016/ j.jbusres. 2020.10.030, 2021

[9] Reis, J., Amorim, M., Melão, N., Matos, P., “Digital Transformation: A Literature Review and Guidelines for Future Research”. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds) Trends and Advances in Information Systems and Technologies. WorldCIST'18, Advances in Intelligent Systems and Computing, vol 745. Springer, Cham. https://doi.org/10.1007/978-3-319-77703-0_41, 2018

[10] European Interoperability Framework for European Public Services (EIF) Annex 2, Def; 1.2.2 Interop., Bruxelles, 2010

[11] Clark, T., Jones, R., “Organizational Interoperability Maturity Model for C2”, Department of Defense, Canberra, Australia, 1999

[12] Van der Veer, H., Wiles, A., "Achieving Technical Interoperability – the ETSI Approach", ETSI White Paper No.3, 3rd edition, http://www.etsi.org/images/files/ETSIWhitePapers/IOP%20whitepaper%20Edition%203%20 final .pdf, 2008

[13] Fewell, S. and Clark, T., “Organisational Interoperability: Evaluation and Further Development of the OIM Model,” Proceedings of the 8th ICCRTS, Washington, D.C., June 2003

[14] Fewell, S., et al., “Evaluation of Organisational Interoperability in a Network Centric Environment,” Proceedings of the 9th ICCRTS, Copenhagen, September 2004

[15] Kingston, G., “An Organisational Interoperability Agility Model.”, Technical Report, Defence Science and Technology Organisation Canberra (Australia), 2005

[16] Kinder, T., „Mrs. Miller moves house: the interoperability of local public services in Europe.“, J. Eur. Soc. Policy, 13(2):141-157. [doi:10.1177/0958928703013002003], 2003
[17] He, K., Wang, J., Liang, P., “Semantic interoperability aggregation in service requirements refinement”, Jurnal of computer science and technology 25(6): 1103–1117, DOI:10.1007/s11390-010-1088-1 31, 2010
[18] Heiler, S., “Semantic interoperability”, ACM Comput. Surv., 27(2):271-273. [doi:10.1145/210376.210392], 1995
[19] Hall, J., Koukoulas, S., “Semantic Interoperability for E-Business in the ISP Service Domain”, ICE-B, p.390-396, 2008
[20] Guijarro, L., “Semantic interoperability in e-government initiatives”, Comput. Stand. Inter., 31(1):174-180. [doi:10. 1016/j.csi.2007.11.011], 2009
[21] Vera, J., Perrochon, L., Luckham, D., “Event-Based Execution Architectures for Dynamic Software Systems,” Proceedings of the First Working IFIP Conf. on Software Architecture,http://pavg.stanford.edu/cep/99wicsa1.ps.gz
[22] Chandy, K.M., Schulte, W.R., "Event Processing: Designing IT Systems for Agile Companies", Special Abridged Edition Compliments of Progress Software, 2010
[23] Morris, E., Levine, L., Meyers, C., Plakosh, D., “System of Systems Interoperability” (SOSI): Final Report. CMU/SEI-2004-TR-004, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, 2004
[24] Lewis, G.A., Wrage, L., “Model Problems in Technologies for Interoperability”: Web Services. Technical Report CMU/SEI-2006-TN-021, 2006
[25] Naudet, Y., Latour T., Guédria, W., Chen, D., “Towards a systemic formalization of interoperability.” In: Computers in Industry, vol. 61, Issue 2, pp. 176–185, 2010
[26] Doumeingts, G., Vallespir, B., Chen, D., “Decision modeling GRAI grid”, in P. Bernus, K. Mertins, Schmidt G. (Eds.), Handbook on Architecture for Information Systems, Springer-Verlag, Berlin, 1998. ETSI: ETSI EG 202 810: methods for testing and specification (MTS), automated interoperability testing, methodology, and framework. European Telecommunications Standards Institute (ETSI), Sophia-Antipolis, 2010
[27] Ford, T., Colombi, J., Graham, J., Jacques, D., “The Interoperability Score.”, In: Proceedings of the 5th Annual Conference on Systems Engineering Research. Hoboken, N.J., 2007.
[28] Luckham, D., “Power of Events The: An Introduction to Complex Event Processing in Distributed Enterprise Systems”, by Addison-Wesley Professional, Published 2002
[29] Luckham, D., Manens, A., Bhansali, S., Park, W., Daswani, S., “Modeling and Causal Event Simulation of Electronic Business Processes”, 2008
[30] Hamilton, J., Rosen, J.D., Summers, P.A., “Developing Interoperability Metrics”, In joint command and control interoperability: cutting the Gordian knot, Chapter 6, 2004
[31] Kasunic, M., Anderson, W., “Measuring systems interoperability: challenges and opportunities”, Software engineering measurement and analysis initiative, 2004
[32] Tolk, A., “Beyond Technical Interoperability – Introducing a Reference Model for Measures of Merit for Coalition Interoperability”., In Proc. of the 8th International Command and Control Research and Technology Symposium (ICCRTS), Washington, 2003
[33] Chen, D., Vallespir, B., Daclin, N., “An Approach for Enterprise Interoperability Measurement”. Proc. MoDISEEUS, p.1-12., 2008
[34] Guédria, W., Naudet, Y., Chen, D., ‘‘Interoperability Maturity Models – Survey and Comparison’’ –. In: Meersman, R., Tari, Z., Herrero, P. (eds) On the Move to Meaningful Internet Systems: OTM 2008 Workshops. OTM 2008. Lecture Notes in Computer Science, vol 5333. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88875-8_48, 2008
[35] Mensh, D., Kite, R., Darby, P., “A methodology for quantifying interoperability” Nav. Eng. J., 101.3:251.
[36] Luckham, D. and Frasca, B., “Complex Event Processing in Distributed Systems.” Stanford University Technical Report CSL-TR-98-754., http://pavg.stanford.edu/cep/fabline.ps, 1998
[37] Luckham, D., “Event Processing for Business: Organizing the Real-Time Enterprise”, Hoboken, New Jersey: John Wiley & Sons, Inc., p. 3. ISBN 978-0-470-53485-4, 2012
[38] Cugola, G., Margara, A., Pezz`e, M., and Pradella, M., “Efficient analysis of event processing applications.” In Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems - DEBS ’15, pages 10–21, Oslo, Norway, doi:10.1145/2675743.2771834, 2015
[39] Eckert, M. and Bry, F., “Complex Event Processing (CEP)”, Article on Complex Event Processing, Informatik-Spektrum, Springer 2009
[40] Rabinovich, E., Etzion, O., Ruah, S. and Archushin. S., “Analyzing the behavior of event processing applications”, In Proceedings of the 4th ACM International Conference on Distributed Event-Based Systems - DEBS 2010, pages 223–234, Cambridge, UK, doi:10.1145/1827418.1827465, 2010
[41] Open Group TOGAF: The Open Group Architecture Framework, Document No. 1910, Version 6, 2000
[42] Daniel Hodapp, and Andre Hanelt, "Interoperability in the era of digital innovation: An information systems research agenda", Journal of Information Technology 2022, Vol. 0(0) 1–21, 2022
[43] Chen R, Sharman R, Chakravarti N, et al., "Emergency response information system interoperability: development of chemical incident response data model." Journal of the Association for Information Systems 9(3/4): 200–230. Chen R, Sharman R, Sharman R, et al. (2013) Data model approach. MIS Quarterly 37(1): 125–147, 2008
[44] LaVean, G.E., “Interoperability in defense communications”, IEEE Trans. Commun., 28(9):1445-1455. [doi:10.1109/TCOM.1980.1094832], 1980
[45] Carney, D., Oberndorf, P., “Integration and Interoperability Models for Systems of Systems”, Proc. System and Software Technology Conf., p.1-35, 2004
[46] Berre, A.J., Elvesaeter, B., Figay, N., Guglielmina, C., Johnsen, S.G., Karlsen, D., Knothe, T., Lippe, S., The ATHENA Interoperability Framework. In: Gonçalves, R.J., Müller, J.P., Mertins, K., et al. (Eds.), Enterprise Interoperability II: New Challenges and Approaches, p.569-580. [doi:10.1007/978-1-84628-858-6_62], 2007
[47] Chen, D., Doumeingt, G., Vernadat, F.B., „Architectures for enterprise integration and interoperability: past, present and future.” In: Computers in Industry, Vol. 59, pp. 647–659 Defence Force J., (151):23-36, 2008
[48] Broring A, Schmid S, Schindhelm C-K, et al. Enabling IoT ecosystems through platform interoperability. IEEE Software. 2017 34(1): 54–61. doi: 10.1109/MS.2017.2
[49] Adner R (2016) Ecosystem as structure. Journal of Management 43(1): 39–58. doi: 10.1177/0149206316678451
[50] Simcoe T and Watson J (2019) Forking, fragmentation, and splintering. Strategy Science 4(4): 283–297. doi:10.1287/stsc.2019.0094.
[51] Camara, M.S., Ducq, Y. & Dupas, R., “A methodology for the evaluation of interoperability improvements in inter-enterprises collaboration based on causal performance measurement models”, International Journal of Computer Integrated Manufacturing, http://dx.doi.org/10.1080/0951192X.2013.800235, 2013
[52] Mick, R., “Defining and Measuring Interoperability”, ARC INSIGHTS, 43, 2004
[53] Farrell J and Simcoe T, "Choosing the rules for consensus standardization.", The RAND Journal of Economics 43(2): 235–252. doi: 10.1111/j.1756-2171.2012.00164, 2012
[54] Simcoe TS, "Standard setting committees: consensus governance for shared technology platforms.", The American Economic Review 102(1): 305–336, 2011

[55] Farrell J and Saloner G, "Coordination through comittes and markets.", The Rand Journal of Economics 19(2): 232–252., 1988
[56] Markus ML, Steinfield CW, Wigand RT, et al. (2006) Industrywide information systems standardization as collective action: the case of the U.S. residential mortgage industry. MIS Quarterly 30(1): 439–465., 2006
[57] Nickerson and Muehlen (2006) The ecology of standards processes: insights from internet standard making. MIS Quarterly 30(30): 467. doi: 10.2307/25148769, 2006
[58] Michele Dassisti, Ricardo Jardim-Goncalves, Arturo Molina, Ovidiu Noran, Hervé Panetto, Milan M. Zdravković, Two Facets of the Same Gold Medal", On the Move to Meaningful Internet Systems: OTM 2013 Workshops, Volume 8186, ISBN: 978-3-642-41032-1
[59] Jiawei, H., Kamber, M., Pei, J. ''Data Mining: Concepts and Techniques'', Third Edition book, Elsevier MK Publishers, 2012
[60] Cugola, G. and Margara, A., "Processing flows of information: From data stream to complex event processing", ACM Computing Surveys, 44(3):1–62, doi:10.1145/2187671.2187677, 2012
[61] Schulte, W. R., "CEP Technology: EPPs, DSCPs, and other Product Categories", (URL: www.complexevents.com/2015/07/10/cep-technology-epps-discs-and-other-product-categories), 2015
[62] Weidlich, M., Mendling, J. and Gal, A., "Net-Based Analysis of Event Processing Networks: The Fast Flower Delivery Case", In Jose-Manuel Colom and J´org ¨ Desel, editors, Application and Theory of Petri Nets and Concurrency SE - 15, volume 7927 of Lecture Notes in Computer Science, pages 270–290. Springer Berlin Heidelberg, doi:10.1007/978-3-642-38697-8 15, 2013
[63] Zang, C., Y. Fan, Y., "Complex event processing in enterprise information systems based on RFID", DOI:10.1080/17517570601092127, 2007
[64] Sharon, G. and Etzion, O., "Event-processing network model and implementation. IBM Systems Journal, 47(2):321–334, doi:10.1147/sj.472.0321, 2008
[65] Etzion, O. and P. Niblett, P., "Event Processing in Action", Manning Publications, 2010

Daliborka Mačinković received her master's degree in the field of information systems and technologies from the University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia, in 2012. She has participated in many projects of development and integration of the information systems of public institutions and healthcare institutions. Her research interests focus on artificial intelligence and data science.

Vidan Marković is an associate professor at the Department of Information Systems of the University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia. He has more publications in journals, conferences, and books.

Enhancing Semantics Learning: A Dynamic Environment for Abstract Language Implementation Education

Steingartner, William; Sivý, Igor

Abstract: ***The abstract implementation of the language on some abstract machine is a logical step in the definition of its operational semantics and in the definition of its (partially) correct implementation. An abstract machine for operational semantics is a well-known formalism in proving the correctness of programming languages. There are several ways to define an abstract machine for a given language specification. In our article, we focus on an abstract machine for structural operational semantics as a stack machine with two different model representations of memory, and we present a complex tool enabling compilation from a higher imperative (toy) language into an abstract machine allowing, in addition, the visualization of individual computational steps, interactive memory manipulation and feedback by compiling back to a higher language. This work presents an abstract machine designed primarily for educational purposes, enabling the visualization and interaction with the compilation process of a simple imperative language.***

Index Terms: ***abstract implementation, abstract machine, bytecode, compiler, operational semantics, university didactic, visualization***

1. Introduction

When developing a programming language, it is essential to provide a clear description of the semantics governing the interpretation of programs written in the language. Formal semantic methods, which are the basis for other formal methods in software engineering, make this possible. There are several semantic methods, some of which are successfully used in practice. One of them is structural operational semantics, which emphasizes the individual steps of program execution and is often used precisely when verifying the properties of languages. In practical terms, the relations within structural operational semantics can be likened to the procedural steps taken while traversing a program. The definition of the corresponding abstract machine (AM) for verifying the correctness of the implementation of the given language is also related to operational semantics. This step is also sometimes referred to as abstract implementation.

Abstract machines provide an intermediate language stage for compilation. Specifically abstract machines from operational semantics produce intermediate-level specifications of evaluators guaranteed to be correct with respect to the operational semantics [7]. Abstract machines bridge the gap between the high level of a programming language and the low level of a real machine. Because of the increasing gap between modern high-level programming languages and existing hardware, it has often become necessary to introduce intermediate languages and to build abstract machines on top of the primitive hardware [16], [32], [34]. The instructions of an AM are tailored to the particular operations required to implement operations of a specific source language or class of source languages [4]. Abstract machines were created as a way to define what programs do independent of hardware. Specifications then describe their language in terms of this machine [13]. Alternatively, an AM is considered an abstraction of a physical computer, which is created to allow a detailed and accurate analysis of the functioning of the computer system [3].

Because formal methods play an important role in practice, they are an integral part of the computer science curriculum at most universities, although teaching formal methods and language semantics can be challenging. When teaching formal methods and especially the semantics of languages, the question arises of how to clearly and comprehensibly explain the principles of semantic methods to future IT experts and young software engineers. The visualization of language

Manuscript received April 1, 2023. This publication was realized with support of the Operational Programme Integrated Infrastructure in frame of the project: Intelligent systems for UAV real-time operation and data processing, code ITMS2014+: 313011V422 and co-financed by the European Regional Development Fund, and by national KEGA project 030TUKE-4/2023 – "Application of new principles in the education of IT specialists in the field of formal languages and compilers", granted by the Cultural and Education Grant Agency of the Slovak Ministry of Education. The research was also supported in the frame of the initiative project "Semantics-Based Rapid Prototyping of Domain-Specific Languages" under the bilateral program "Aktion Österreich – Slowakei, Wissenschafts- und Erziehungskooperation" granted by the Slovak Academic Information Agency.

William Steingartner (Corresponding author), Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia (e-mail: william.steingartner@tuke.sk). Igor Sivý, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia (e-mail: igor.sivy@student.tuke.sk)

semantics, algorithms and the execution of the programming code has been implemented in the past and proved helpful. Such visualization is of great importance in modern and contemporary teaching, and students get an illustrative approach to the explanation of language semantics and the execution of programming code.

In computer science, these teaching tools often take the form of software that can process what is being taught and allow students to analyze the output. When teaching formal semantic methods, such tools are very convenient because we have to process the input language and then we can visualize the semantic method in a certain way. Continuing our research and implementation of the achieved results in the pedagogical process to make the study of formal methods for software engineering more attractive, we have brought several interesting software tools for the visualization of selected methods. Some results are presented in the works [26]–[29]. Therefore, our goal was to expand this spectrum of teaching aids with another important tool. We developed a functional emulator of an AM for operational semantics.

In this paper, we introduce a newly developed learning environment designed for the implementation of abstract languages, specifically tailored for the Semantics of Programming Languages course. This visualization environment serves as a powerful tool, enabling students to translate input programs and observe their execution through a web-based graphical interface. The AM can rely on two distinct versions of syntax based on selected manipulation with the memory representation, and both types of syntax can either be extended; basic syntax corresponds to the basic version of the language *Jane* as proposed e.g. in [15], but from the experience we decided also to allow an extension simply by adding new operators and some control-flow constructs. Moreover, our design incorporates an architecture with two memory models, providing a work with the basic and extended version of the language. Our goal in creating this software was to enhance the understanding about language semantics, providing students with a more accessible learning experience. Additionally, the application allows users to save the visualization output in various formats, facilitating easy further processing, analysis, publication, or inclusion in their technical or final theses. Unlike traditional compilers, our educational tool goes beyond by incorporating a user-friendly interface that visualizes each computational step, allows interactive manipulation of memory during program execution, and provides real-time feedback, creating a dynamic learning environment for students.

This paper represents an expanded and extended version of the conference paper referenced in [31] and is structured as follows: in Section 2, we briefly present and discuss similar approaches to the visualization of semantic methods and other processes in the phase of the development of programs. Section 3 discusses the necessary basic theory and background about the AM for the structural operational semantics. We briefly introduce a simple programming language which is used to define individual semantic methods in our course in Section 4. Section 5 presents the principles of how our tool for visualizing the AM is implemented. To sum up, our paper concludes with Section 6.

2. Selected related works

Several approaches exist in visualization software or software development aimed at enhancing students' learning experiences. These approaches seek to foster a deeper understanding of subjects by bridging theoretical concepts with practical applications. Greater clarity is thus achieved in the applied procedures and principles in teaching the course. However, the visualization itself helps not only students but also other IT experts when examining critical parts of the code, memory management, static analysis or some tests.

One of the accepted opinions is that AMs are closely related to virtual machines, which are basically AMs with an implementation. Such virtual machines can be used to achieve portability of high-level programming languages (like the Java Virtual Machine, or the Common Language Framework for the .NET framework) or complete operating systems [9].

From one point of view, many authors use different compiler generator tools: Mernik [11] uses LISA compiler generator tool, Chodarev et al. [2] use YAJCo parser generator, and Radakovic and Herceg use Coco/R [14] for their extensible Dynamic Geometry software [19].

Authors of the work [1] present the process of deriving an AM from an interpreter that describes the operational semantics of a source language. The derivation involves applying a series of step-by-step transformations to the interpreter, which is initially written in a functional language. By employing pass separation during the derivation, the outcome is the extraction of both a compiler and an AM from the modified interpreter.

Another view of visualization is provided by the work [18]. In this article, the authors focused on the Program Semantics Visualizer, whose task is to clarify the semantics of the language by explaining the runtime execution of programs. The authors analyze existing visualization procedures for programming courses and propose a model to assess the design of program semantics visualizers that visualize execution traces. We note that this work focuses directly to semantics of Python

and the execution traces expressed in labeled transition system resemble the operational small-step semantics.

An interesting approach connected to the small-step approach is the $\mathbb{K}$ framework, introduced by Grigore Roşu. It is an executable semantic framework in which programming languages, calculi, as well as type systems or formal analysis tools can be defined, making use of configurations, computations, and rules [21].

A successful project oriented to direct visualization of evaluations of the *While* language via semantic rules, available at the website[1] and in Git repository[2], provides both computations in big-step and small-step semantics. The software is developed in JavaScript and offers many useful features. It is very easy to use and thanks to the high level of interactivity, users can learn and practice the semantics on a high level.

Very interesting software oriented to modeling algorithms and specifying their behavior in first-order logic was developed in the Research Institute for Symbolic Computation at JKU Linz by Wolfgang Schreiner [22], [23]. RISCAL (RISC Algorithm Language) is a language and associated software system for formulating theories in first-order logic, describing algorithms in a high-level language and specifying the behavior of these algorithms by formal constraints. It uses the original approach of semantic evaluation (based on an implementation of the denotational semantics of the RISCAL language). A particularly intriguing outcome of the collaboration between the first author of this article and the author of previous works is the development of the SLANG tool [24]. SLANG is a Java-based tool that rapidly generates prototype implementations of programming languages from formal specifications. It utilizes ANTLR4 to produce parsers and printers, allowing students to explore language design, parsing, typing, and interpretation with ease. The tool supports the addition of formal type systems and formal semantics, making it a valuable resource for teaching programming language concepts.

At the end, we would also like to summarize a short comparison with our previous works. Several other tools have been developed by our team that allow the visualization of semantic methods and thus simplify the teaching process. However, these works represent the visualization of other semantic methods, primarily for structural operational semantics (semantics of small steps) [26], [33], natural semantics [27] and coalgebraic semantics using category theory [28]. All of these tools are designed to visualize the selected semantics methods for the abstract language *Jane* and as a support for our course and, together with the tool presented in this article, they form a comprehensive set of tools for pedagogy and visual illustration of semantic procedures and modeling. The main goal we pursued during the development of this and previous software solutions was to prepare a coherent set of tools for semantic methods for a specific subject. Of course, the achieved results can also be applied within other subjects, but compared to other solutions, this approach guarantees its deployment and use exactly according to the needs of our course, also with regard to further development and possible extensions.

In addition to the ones mentioned, several other approaches to the visualization of calculations, semantic procedures or even algorithms (not mentioned here) have been designed and successfully implemented. Visualization (not only formal methods) in contemporary teaching is of great importance. The current (and future) generation of students will thus gain a very illustrative and still innovative approach to the explanation of principles of varying complexity, which they can thus try for themselves. Our approach to visualizing AM computations for an abstract imperative language was highly inspired by the aforementioned software.

3. Theoretical background of abstract machine

Formal language operational semantics is crucial for language implementation, emphasizing the correctness of programming language implementation. Starting from abstract program representation, it involves defining an AM and specifying a translation process for programs into AM instructions. The execution of translated programs on the AM is formally described. Correctness is verified by comparing results with operational semantics outcomes. An AM serves as a semantic model, detailing how language notation translates into machine code, offering a foundation for efficient language implementation on a target machine.

An AM semantics presents a high-level model of how a computer might go about running a program [6]. Historically, AMs were the first kind of operational semantics, introduced by Peter Landin in the highly influential 1964 paper “The Mechanical Evaluation of Expressions” [10]. However, AMs were considered too low-level by many researchers, and this led to the development of structural operational semantics (by Plotkin, [17]) and reduction semantics (by Felleisen, [5]).

By the term AM, we understand a mathematical model that formally describes programs according to semantic specifications.

One of the definitions of AM for operational semantics was presented in [15]. In our approach, we follow the standard definition of AM for the structural operational semantics with two kinds of

[1] https://mostlynerdless.de/while-semantics/

[2] https://github.com/parttimenerd/while-semantics

memory abstractions – the linear memory and the set of states.

The description of particular computational steps of AM is usually given by transition relations over the configurations (tuples) of the form

$$\langle c, st, s\rangle \in \mathbf{Code} \times \mathbf{Stack} \times \mathbf{State},$$

or alternatively

$$\langle c, st, m\rangle \in \mathbf{Code} \times \mathbf{Stack} \times \mathbf{Memory},$$

where

- c stands for a code – the sequence of instructions to be performed,
- st is the evaluation stack, and
- s and m, resp., represent two models of a computer storage.

We refer to the tuples $\langle c, st, s\rangle$ and $\langle c, st, m\rangle$ as AM configurations for the transitions. The semantic domain $\mathbf{State}$ is a function space of states – elements that send variables to values, and $\mathbf{Memory}$ is defined as $\mathbf{Z}^*$. The complete definition can be found in [15].

Regularly, the evaluation stack is used to evaluate arithmetic and Boolean expressions. Formally, it is defined as a list of values that are elements of the semantic domain

$$\mathbf{Stack} = (\mathbf{Z} \cup \mathbf{B})^*,$$

where $\mathbf{Z}$ stands for a set of integers, $\mathbf{B}$ is a set of semantic values of Boolean constants, symbolically denoted as $\mathbf{tt}$ for the semantic value of *true* and $\mathbf{ff}$ for the semantic value of false, $\mathbf{B} = \{\mathbf{ff}, \mathbf{tt}\}$, and the symbol * (a star symbol in an upper index) denotes Kleene's closure. An example of how the evaluation stack is used is depicted in Figure 1.

The language of an AM is a structured assembler – a set of instructions. These instructions are given by the following abstract syntax expressed by Backus-Naur form:

$$\begin{aligned} \mathit{instr} \quad ::= \quad & \texttt{PUSH-}n \mid \texttt{ADD} \mid \texttt{SUB} \mid \texttt{MULT} \mid \\ & \texttt{TRUE} \mid \texttt{FALSE} \mid \texttt{EQ} \mid \texttt{LE} \mid \texttt{AND} \mid \texttt{NEG} \mid \\ & \texttt{FETCH-}x \mid \texttt{STORE-}x \mid \texttt{EMPTYOP} \mid \\ & \texttt{BRANCH}(c, c) \mid \texttt{LOOP}(c, c), \\ c \quad ::= \quad & \varepsilon \mid \mathit{instr} : c. \end{aligned}$$

A meta-variable c is ranging over a syntactic domain $\mathbf{Code}$ of sequences of instructions:

$$c \in \mathbf{Code}.$$

The instruction language described allows assigning integer values to variables, evaluating arithmetic and Boolean expressions, influencing control flow, and repeating computations based on logical conditions. Depending on the AM specifications, the instruction set may vary, e.g., using $\texttt{GET-}n$/$\texttt{PUT-}n$ for linear memory instead of $\texttt{FETCH-}x$/$\texttt{STORE-}x$ for abstract representation of

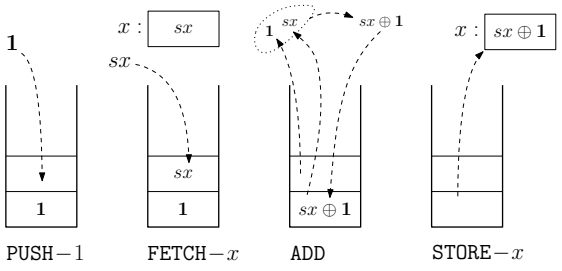


Figure 1. Computation progress on the AM stack

states. To enhance teaching and align with real programming languages, we introduced additional arithmetical and Boolean operators and loop statements.

Operational semantics defines the AM's instruction execution steps through a transition system, detailed in [15]. The semantics of the newly added instructions is defined in a manner analogous to that of the basic instructions in the language, e.g.

$$\langle \texttt{DIV} : c, v_1 : v_2 : st, s\rangle \quad \triangleright \quad \langle c, (v_1 \ \mathbf{div}\ v_2) : st, s\rangle,$$

$$\begin{aligned} &\langle \texttt{REPEAT}(c_1, c_2) : c, st, s\rangle \quad \triangleright \\ &\triangleright \ \langle c_1 : c_2 : \texttt{BRANCH}(\texttt{EMPTYOP}, \texttt{REPEAT}(c_1, c_2)) : c, \\ &\quad st, s\rangle. \end{aligned}$$

The AM's definition is flexible, allowing modifications for various approaches.

For example, semantics of instructions manipulating with the linear memory is in our approach defined as follows:

$$\begin{aligned} &\langle \texttt{GET-}i : c, st, m\rangle \quad \triangleright \quad \langle c, m\,[i] : st, m\rangle, \\ &\langle \texttt{PUT-}i : c, v : st, m\rangle \quad \triangleright \quad \langle c, st, m\,[i \mapsto v]\rangle. \end{aligned}$$

Here, the writing $m\,[i \mapsto v]$ represents an actualization of the value on the given address (informally, e.g. as $m[i] := v$).

Formally, the relationship between a higher language and an AM for the operational semantics of a given language is expressed using translation functions, each of which provides the translation of particular syntactic elements. We define translation functions for arithmetic expressions, Boolean expressions and for commands:

- $\mathscr{IE} : \mathbf{Expr} \to \mathbf{Code}$ for arithmetic expressions,
- $\mathscr{IB} : \mathbf{Bexpr} \to \mathbf{Code}$ for Boolean expressions,
- $\mathscr{IS} : \mathbf{Statm} \to \mathbf{Code}$ for statements.

The translation functions were defined for the basic version of the language in [15]. We just point out how to perform the translation of the statement $\texttt{repeat}\ S\ \texttt{until}\ b$ that we added to the extended version of the language *Jane*:

$$\begin{aligned} &\mathscr{IS}[\![\,\texttt{repeat}\ S\ \texttt{until}\ b\,]\!] = \\ &\qquad = \texttt{REPEAT}(\mathscr{IS}[\![\,S\,]\!], \mathscr{IB}[\![\,b\,]\!]). \end{aligned}$$

4. The language Jane

For semantic modeling of the abstract implementation of a programming language, a simple abstract language for defining the semantic methods and proving their properties and equivalences is used. It is a non-real programming language grounded in an imperative paradigm, epitomizing a tiny core fragment of conventional mainstream languages such as C or Java: standard imperative constructs as sequences of statements, selection (conditional), repetition (loops) and handling the values in memory (variables assignment). For research and development, this language has been adopted by many authors and researchers. Moreover, there have been formulated also many approaches thanks to this abstract language. Some authors refer to this language as *IMP* (as simple imperative language) [20] or as *While* (defined for instance in [15]). We adopted the structure of this language as well, and we refer to this language as *Jane* [30].

We do not repeat the definition of the language and we present basic aspects only briefly. The abstract syntax of the language *Jane* is defined by the set of rules taking the syntactic elements from the following syntactic domains (or syntax categories):

$n \in \mathbf{Num}$ (strings of digits),
$x \in \mathbf{Var}$ (variables' names),
$e \in \mathbf{Expr}$ (arithmetic expressions),
$b \in \mathbf{Bexpr}$ (Boolean expressions),
$S \in \mathbf{Statm}$ (statements).

For each syntactic domain, we define exactly one production rule given in BNF.

- the elements in domains for numerals and variables' names have no internal structure from the semantic point of view; syntactically the numbers can be represented with a regular expression $[0, \ldots, 9]^+$,
- production rule for arithmetic expressions:

$$e ::= n \mid x \mid e + e \mid e - e \mid e * e,$$

- production rule for Boolean expressions:

$$b ::= \texttt{true} \mid \texttt{false} \mid e = e \mid e \leq e \mid \neg b \mid b \wedge b,$$

- production rule for the statements:

$$\begin{aligned} S ::= \quad & x := e \mid \mathbf{skip} \mid S; S \mid \\ & \mid \mathbf{if}\ b\ \mathbf{then}\ S\ \mathbf{else}\ S \mid \\ & \mid \mathbf{while}\ b\ \mathbf{do}\ S \mid \mathbf{repeat}\ S\ \mathbf{until}\ b. \end{aligned}$$

We note, that the sets of arithmetic and Boolean expressions are in general not limited to listed syntactic constructs. All mentioned rules can be extended by providing other correct syntactic constructs, as we did in our approach.

5. Implementation

After defining the input language and formulating the abstract implementation of a language, the decision to implement a visualization tool naturally led us to opt for a compiler-centric approach. Unlike interpreters or simple code generators, a compiler transforms the input language into a suitable intermediate representation before generating code for a virtual machine. The adoption of a compiler-centric approach aligns with our goal of developing a visualization tool that effectively captures program behavior, providing clear and insightful representations of execution of statements and providing the dynamics of memory changes. The tool [25] allows the students to write a program in a web-based graphical interface in the *Jane* language and compile it into AM code. It also allows to write directly a program in the AM code and translate it back to the *Jane* language. The execution of the compiled program can be then interactively visualized. At its core, the visualization tool is a web application, which is divided into three modules: editor – the user input processing module, the compiler module and the visualization module. All modules are implemented as a whole in three Docker containers. All modules communicate with each other using HTTP protocol. Each module defines its interface, which accepts requests from other modules and responds in a particular way. Because of this, a simple web server for each module was necessary, for which the Django web framework is used. Web-based user interfaces in our software package all use plain JavaScript with Bootstrap library for the front end. Next, we briefly describe the implementation of each module.

5.1. Editor

This component of the software package serves as the entry point for users to input their programs into the visualization tool. Two methods are available: manual entry in the text editor window or loading a program from a text file. Users can choose the programming language (either *Jane* language or AM language) and opt for language extensions like repeat and for loops. Additionally, they can decide whether to use `GET/PUT` instructions instead of `FETCH/STORE`. The editor, powered by the Ace library, enables syntax highlighting for *Jane*, and automatic replacement of operators with their Unicode equivalents (e.g., $<=$ becomes $\leq$). Users can save and share their code effortlessly. The HTTP interface supports source code input in Base64 format via the URL, facilitating sharing and editing. Communication involves requests sent to the translation module, with errors displayed for unsuccessful translations. Successful translations are coded and redirected to the visualization module.

Ace can be easily expanded by the user in various directions. In the editor for this work, Ace is enhanced with custom *Jane* syntax highlighting. This is implemented by providing definitions of what we want to highlight in the editor in the form of regular expressions. The partial pseudo-code of the definitions is depicted in Figure 2 (the full definition is in [25]).

```
builtinConstants = ("tt|ff");

rules = {
    "start" : [ {
        token : keywordMapper,
        regex : "[a-zA-Z_$][a-zA-Z0-9_$]*\\b"
    }, {
        token : "keyword.operator",
        regex : "\\+|\\-|&&|\\|\\||<|>|<=|=>|==|!=|<>|="
    }, {
        token : "constant.numeric",
        regex : "\\d+\\b"
    }, {
        token : "paren.lparen",
        regex : "[\\(]"
    }, {
        token : "paren.rparen",
        regex : "[\\)]"
    }, {
        token : "text",
        regex : "\\s+"
    } ]
};
```

Figure 2. Example of a pseudo-code definition

5.2. Compiler

The compiler for the input language is a standalone program written in C++. A very popular way of creating a compiler is by using a lexical analyzer generator and a parser generator like Lex, Yacc or ANTLR. We did not use any of them and implemented the compiler entirely by hand, mainly because we did not want to add more dependencies to the code and our input language is simple enough. The created compiler processes the input in one pass, which means that parsing, simple type checking and intermediate code generation are done at once when parsing an individual statement (schematically depicted in Figure 3).

In particular, we implemented a recursive descent parser, which is very suitable for these types of languages and is very easy to implement. The implementation consists of mutually recursive functions, where each function represents a nonterminal symbol of the grammar. That means we could easily write a set of functions based on our grammar described in the first section of this

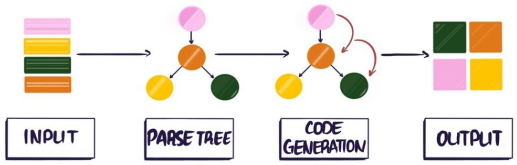


Figure 3. General scheme of source-to-source compilation. Reprinted from [8].

paper. However, this grammar was slightly modified, mainly to express operator precedence. In addition to this, we used the Pratt parsing method for parsing operator precedence described in the modified grammar. Also, arithmetic and Boolean expressions were merged to make parsing simpler. Because of this, a simple type checking was implemented and the type of each expression is inferred based on the operator or a literal used in the expression. The result of the compilation process is a custom intermediate representation in a form of a bytecode. We chose not to use the AM code directly because its branching and looping instructions are not suitable for execution by a virtual machine. Therefore, we defined our own set of instructions, which closely copy the AM instruction set, but we added our instructions for jumping around in the bytecode.

The compilation module functions as a web application, processing HTTP requests to initiate translations and providing the results or errors. While primarily designed for visualization output, it also serves as a standalone program for testing purposes. This versatility allows command-line execution, accommodating various input and output types. The server communicates with the compiler via HTTP GET requests, utilizing parameters to determine input language and the use of extended instructions. The decoded source code undergoes processing, triggering the compiler as a separate program through the Python subprocess. The response, formatted in JSON, includes translated code, bytecode, and program variables. In case of unsuccessful translation, error messages are sent as plain text responses to HTTP requests, displayed in the editor for user feedback.

The compilation process, operating as a standalone program, efficiently handles lexical analysis for two input languages sharing a common intermediate code. In this compiler, lexical analysis does not create a series of tokens initially; rather, it interacts dynamically with the translation process. The analysis skips white characters and distinguishes identifiers or keywords for letters, producing corresponding tokens. Numeric input is parsed until the nearest non-numeric character, and characters trigger tokens based on types. Escape sequences, initiated by the \ (backslash)

character, are parsed similarly. If a sequence is unknown, an error token is returned. If the character does not conform to the grammar, an error token is produced, maintaining functionality for potential scenarios without automatic replacement.

The compiler can work in several ways. Following is the short description of particular modes.

1) *Compilation from Jane to bytecode.* The translation process converts *Jane* language code into bytecode, producing bytecode and variable information for visualization. It is a one-pass translation with simultaneous syntactic and simple semantic analysis, directly generating code without further modifications like optimizations. A conceptual syntactic tree is used for expressions but discarded after translation. Recursive descent parsing is employed, and Pratt Parsing handles unary and binary expressions for code purity. Semantic analysis checks Boolean expressions and logical operator operands, ensuring proper variable assignments. Code generation occurs immediately, traversing the tree structure and producing instructions based on expression types. Unlike expressions, command translation calculates lengths for conditions and statement bodies due to jump instructions when dealing with conditional and loop statements. As an example, the representation of the statement

$$\textbf{while } x \leq 10 \textbf{ do } x := x + 1$$

in the bytecode is in Figure 4.

```
0000 WHILE 9 12
0005 PUSH-10
0010 GET-0 (FETCH-x)
0013 LE
0014 BRANCH_IF_FALSE 21
0017 PUSH-1
0022 GET-0 (FETCH-x)
0025 ADD
0026 PUT-0 (STORE-x)
0029 LOOP 24
0032 BRANCH 4
0035 EMPTYOP
```

Figure 4. Example of bytecode for a loop statement

2) *Compilation from the AM code to bytecode.* When translating AM source into bytecode, the process employs the same lexical analysis and syntactic recursive descent as in the *Jane* language translation. It is a one-pass translation, and errors are reported using a similar system. The grammar of the AM language is simpler, eliminating the need for auxiliary data structures during translation. The translation begins by loading the source code, and lexical analysis yields tokens used to look for corresponding bytecode instructions. Instructions are represented as identifiers, and a table maps AM language instruction names to bytecode instructions, specifying whether an argument is needed and its type. For instructions requiring an argument, tokens for numbers (`PUSH`, `PUT` and `GET`) or text strings (`FETCH` and `STORE`) are requested accordingly. The translation of `BRANCH` and `LOOP` instructions involves additional complexity. The translation of the `BRANCH` instruction is realized by inserting the addresses where the individual branches start into the bytecode. Even more complex is the translation of the loop instructions, where we have to calculate the length of the bytecode of the body, together with the addresses of the start and end of the loop. The process of length and address calculations similar to *Jane* language translation.
3) *Reverse compilation from bytecode to Jane.* The translation from bytecode to AM language mirrors the bytecode structure, starting with a function that translates byte ranges. Each translated instruction returns the number of processed bytes, progressing through the bytecode. Most instructions have a straightforward mapping, such as `ADD` and `PUSH`. Complex instructions like `PUT` and `GET` are transformed into `FETCH` and `STORE`, including reading operands from the bytecode. `BRANCH_IF_FALSE` indicates a conditional statement, translated to `BRANCH`. For loop statements, the `BRANCH` instruction guides the translation of both branches. Cycle instructions involve auxiliary instructions for `LOOP` or `REPEAT`, determining translation type and calculating start and end addresses. The output is a text string of translated code in AM language.
4) *Reverse compilation from bytecode to AM language.* Translating from bytecode to *Jane* involves a similar principle as translating to AM language. The process revolves around a function that translates byte ranges in bytecode, with an auxiliary structure storing the context of the translation, particularly a stack of expressions. This stack represents syntactic tree structures created gradually during translation. Binary expressions, arithmetic instructions, and other expressions are translated similarly, and parentheses are explicitly added to avoid ambiguity. Commands, such as conditional jumps for conditional statements and cycles, are translated akin to AM language. The result is a text string repre-

senting *Jane* source code, with slight differences from the original input due to formatting changes like omitted new lines and added parentheses around expressions.

For the generation of the bytecode, we closely followed translation functions for the AM described in the Section 3 of this paper. The whole output of the compilation process is the compiled bytecode, names of variables used in the program and a translated AM code, if the input language was *Jane* or source code in *Jane*, if the input language was the AM language.

5.3. Visualizer

The Visualiser module provides a web-based graphical interface for visualizing program execution. Users can set the initial state, step through program execution, run the entire program, and save results in text or LaTeX format. The visualization is powered by a stack-based virtual machine that executes translated bytecode instructions. The machine simulates the behavior of an AM, with no registers, storing operands on the stack. The execution mimics the behavior of the AM, but certain instructions, such as `BRANCH` and `LOOP`, automatically advance to the next step for better user visibility. Program execution terminates on encountering an `EXIT` instruction or user-initiated stop. The server responds to GET requests, allowing users to view visualizations without editing the source code, and provides a link to return to the editor. Visualization of AM code execution involves creating a table representing textual instructions. The table is built similarly to translating bytecode into AM language but focuses on creating a table structure. The table is a JavaScript object, with each element storing a text string, representing AM code, and the address of the next instruction. For example, the table for a simple assignment $x := 1$ is in Figure 5.

```
{
    0: { instruction: "PUSH-1", nextIp: 5 },
    5: { instruction: "STORE-x", nextIp: 8 },
    8: { instruction: "e", nextIp: 9 }
}
```

Figure 5. Table for assignment $x := 1$

Substitutions are utilized for conditional and loop statements, and a substitution table stores the mappings of addresses to substitutions. For example, for the loop

$$\textbf{while } x < 10 \textbf{ do } x := x + 1$$

is the translation to the AM code following:

$$\texttt{LOOP}\,(\texttt{PUSH-}10 : \texttt{FETCH-}x : \texttt{LT},\\ \texttt{PUSH-}1 : \texttt{FETCH-}x : \texttt{ADD} : \texttt{STORE-}x)$$

and the substitutions are:

$$\texttt{LOOP}\,(c_1, c_2)\\ c_1 = \texttt{PUSH-}10 : \texttt{FETCH-}x : \texttt{LT}\\ c_2 = \texttt{PUSH-}1 : \texttt{FETCH-}x : \texttt{ADD} : \texttt{STORE-}x$$

The substitution table is depicted in Figure 6, where the key 0 is the `while` statement itself, and key 14 is a conditional statement that is part of the translated while-loop statement (according to the definition of the transition relation for the AM).

```
{
    0: [
        0: { code: "PUSH-10:FETCH-x:LT", index: 0 },
        1: { code: "PUSH-1:FETCH-x:ADD:STORE-x", index: 1 }
    ],

    14: [
        0: { code: "PUSH-1:FETCH-x:ADD:STORE-x", index: 1 },
        1: { code: "EMPTYOP", index: 2 }
    ]
}
```

Figure 6. Substitution table for the loop statement

During execution, the table guides the construction of text strings used for visualization. The visualization consists of a transition listing showing executed instructions, stack contents, and the current state after each step. Another part displays the list of states, presenting variable values and indexes of current and previous states. Both visualizations are output to the graphical interface, and plain text and LaTeX format listings are saved for user export. An example of the visualization process is depicted in Figure 7.

6. Conclusion

In this article, we presented the results of research in the field of creating a software tool designed for the visualization of an AM for the structural operational semantics. The practical result of this work is a software tool intended as a teaching aid for the Semantics of Programming Languages course, specifically to facilitate the understanding of the semantic method and abstract language implementation. In the article, we provide a concise overview of the implementation details for individual components of the tool. On its input, the tool translates the *Jane* language into an AM language. The user is then allowed to run the translated program and interactively monitor its execution in individual steps. In addition, the tool has implemented the possibility of entering input directly in the language of an AM. After visualization, its results can be saved and further worked with and analyzed. Of course, the research is not closed at this point, so in the future, we want to focus on possible extensions of the visualization of the given semantic method. The exploration

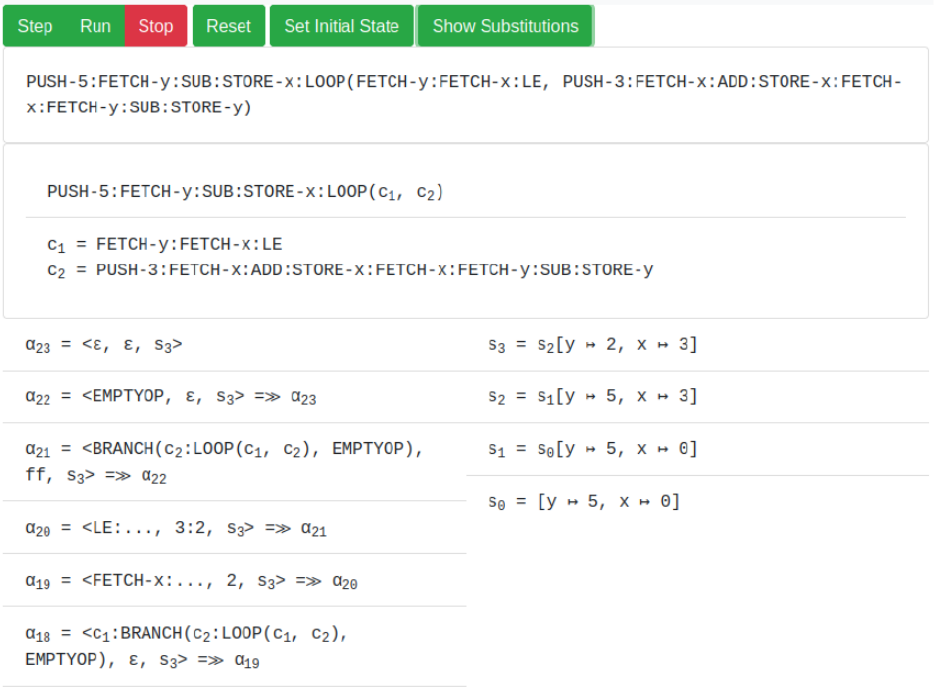


Figure 7. Visualiser window

of potential enhancements could involve incorporating extra language constructs into the *Jane* language, possibly an extension of the instruction set for an AM and associated visualization support in this application. Under the possible extension of the language, new language constructs come into consideration, such as the for-loop, or other constructs for which we previously defined structural operational semantics (variable declarations, procedures, etc.). Specifically, for the definition of the for-loop, the definition has to be compositional and we possibly need to introduce an instruction `DUP` or `COPY` that duplicates the element on the top of the evaluation stack (see e.g. [12]). Last but not least, the possibility of adding comments to source codes, in which it is possible to describe, for example, the size of operators and operands, also seems very interesting. This would be helpful, for example, when following arguments for jumps. Such extensions will make it possible to better understand the structure of languages and the definition of their semantics, and to compare the differences in the definition of the basic and extended language. To fully assess the solution's impact, the authors could consider outlining its implementation within specific educational contexts and incorporating student feedback mechanisms for evaluation. Of course, further research exploring the solution's applicability in education and its effectiveness through student evaluation would be valuable additions to this work.

REFERENCES

[1] Cabestre, F., Percebois, C., and Bodeveix, J.-P. Abstract machine construction through operational semantics refinements. *Future Generation Computer Systems 16*, 7 (2000), 753–769.

[2] Chodarev, S., Lakatoš, D., Porubän, J., and Kollár, J. Abstract syntax driven approach for language composition. *Open Computer Science 4*, 3 (2014), 107–117.

[3] Corradini, A. Advanced programming [AP-21]. https://pages.di.unipi.it/corradini/Didattica/AP-21/DOCS/GM-ch1.pdf. Code 301AA, accessed on 2023-07-12.

[4] Diehl, S., Hartel, P., and Sestoft, P. Abstract machines for programming language implementation. *Future Generation Computer Systems 16*, 7 (2000), 739–751.

[5] Felleisen, M., and Friedman, D. P. Control operators, the secd-machine, and the λ-calculus. In *Formal Description of Programming Concepts - III: Proceedings of the IFIP TC 2/WG 2.2 Working Conference on Formal Description of Programming Concepts - III, Ebberup, Denmark, 25-28 August 1986* (1987), M. Wirsing, Ed., North-Holland, pp. 193–222.

[6] Garcia, R. Abstract machine semantics. CPSC 509: Programming Language Principles, November 2011. Available at https://www.williamjbowman.com/teaching/2020/w1/cpsc509/resources/05-abstract-machines.pdf.

[7] Hannan, J., and Miller, D. From operational semantics to abstract machines. *Mathematical Structures in Computer Science 2*, 4 (1992), 415–459.

[8] Herlihy, A., Khineika, A., and Shestak, I. Transpiling between any programming languages. https://www.mongodb.com/blog/post/transpiling-between-any-programming-languages, 2019. Accessed: 2023-01-10.

[9] Keller, G., O'Connor, L., and Pohjola, J. A. Abstract machines. https://www.cse.unsw.edu.au/~cs3161/23T3/Week%2005/Notes.pdf, 2022. COMP3161/COMP9164 Supplementary Lecture Notes.

[10] Landin, P. J. The Mechanical Evaluation of Expressions. *The Computer Journal 6*, 4 (01 1964), 308–320.

[11] Mernik, M. An object-oriented approach to language compositions for software language engineering. *Journal of Systems and Software 86*, 9 (2013), 2451–2464.

[12] MIHELIČ, J., STEINGARTNER, W., AND NOVITZKÁ, V. A denotational semantics of a concatenative/compositional programming language. *Acta Polytechnica Hungarica 18*, 4 (2021).
[13] MORTORAY, E. Abstract machines, interpreters and compilers. https://mortoray.com/abstract-machines-interpreters-and-compilers/, 2012. Accessed on: 2024-01-31.
[14] MÖSSENBÖCK, H. The Compiler Generator Coco/R User Manual. https://ssw.jku.at/Research/Projects/Coco/Doc/UserManual.pdf, 2010.
[15] NIELSON, H. R., AND NIELSON, F. *Semantics with Applications: An Appetizer (Undergraduate Topics in Computer Science)*. Springer-Verlag, Berlin, Heidelberg, 2007.
[16] PEKÁR, A., CHOVANEC, M., VOKOROKOS, L., CHOVANCOVÁ, E., FECIĽAK, P., AND MICHALKO, M. Adaptive aggregation of flow records. *Computing and Informatics 37*, 1 (2018), 142–164.
[17] PLOTKIN, G. A structural approach to operational semantics. *J. Log. Algebr. Program. 60-61* (07 2004), 17–139.
[18] POLLOCK, J., OH, G., JUN, E., GUO, P. J., AND TATLOCK, Z. The essence of program semantics visualizers: A three-axis model. In *11th annual workshop on the intersection of HCI and PL* (2020), PLATEAU '20.
[19] RADAKOVIĆ, D., AND HERCEG, D. Towards a completely extensible dynamic geometry software with metadata. *Computer Languages, Systems & Structures 52* (2018), 1–20.
[20] ROŞU, G., AND ŞERBĂNUŢĂ, T. An overview of the $\mathbb{K}$ semantic framework. *The Journal of Logic and Algebraic Programming 79*, 6 (2010), 397–434. Membrane computing and programming.
[21] ROŞU, G. $\mathbb{K}$: A semantic framework for programming languages and formal analysis tools. In *Dependable Software Systems Engineering* (2017).
[22] SCHREINER, W. *Thinking Programs. Logical Modeling and Reasoning About Languages, Data, Computations, and Executions*. Springer Nature Switzerland AG, 2021.
[23] SCHREINER, W., AND REICHL, F.-X. First-order logic in finite domains: Where semantic evaluation competes with SMT solving. *Electronic Proceedings in Theoretical Computer Science 342* (sep 2021), 99–113.
[24] SCHREINER, W., AND STEINGARTNER, W. *The SLANG Semantics-Based Language Generator – Tutorial and Reference Manual (Version 1.0)*, 2020.
[25] SIVÝ, I. An environment for a visualization of abstract implementation of a language. Tech. rep., Technical University of Košice, Slovakia, 2022.
[26] STEINGARTNER, W. Support for online teaching of the semantics of programming languages course using interactive software tools. In *Proceedings of the International Conference ICETA 2020* (2020).
[27] STEINGARTNER, W. On some innovations in teaching the formal semantics using software tools. *Open Computer Science 11*, 1 (2021), 2–11.
[28] STEINGARTNER, W., JANKURA, M., AND RADAKOVIĆ, D. Visualization of Formal Semantics – Possibilities of Attracting Formal Methods in Teaching. In *Sinteza2021: International scientific conference on information technology and data related research* (2021), pp. 235–239.
[29] STEINGARTNER, W., AND NOVITZKÁ, V. Natural semantics for domain-specific language. In *New Trends in Database and Information Systems* (Cham, 2021), L. Bellatreche, M. Dumas, P. Karras, R. Matulevičius, A. Awad, M. Weidlich, M. Ivanović, and O. Hartig, Eds., Springer International Publishing, pp. 181–192.
[30] STEINGARTNER, W., NOVITZKÁ, V., AND SCHREINER, W. Coalgebraic operational semantics for an imperative language. *Computing and Informatics 38*, 5 (Feb. 2020), 1181–1209.
[31] STEINGARTNER, W., AND SIVÝ, I. From high-level language to abstract machine code: An interactive compiler and emulation tool for teaching structural operational semantics. In *New Trends in Database and Information Systems - ADBIS 2023 Short Papers, Doctoral Consortium and Workshops: AIDMA, DOING, K-Gals, MADEISD, PeRS, Barcelona, Spain, September 4-7, 2023, Proceedings* (2023), A. Abelló, P. Vassiliadis, O. Romero, R. Wrembel, F. Bugiotti, J. Gamper, G. Vargas-Solar, and E. Zumpano, Eds., vol. 1850 of *Communications in Computer and Information Science*, Springer, pp. 544–551.
[32] SUTIL-MARTÍN, J. G.-M. M. The abstract machine a pattern for designing abstract machines.
[33] TSIMBOLYNETS, V., AND PERHÁČ', J. Visualization tool for structural operational semantics of simple imperative language. *IPSI BGD Transactions on Internet Research 19*, 1, SI (JAN 2023), 66–74.
[34] VOKOROKOS, L., BALÁŽ, A., AND MADOŠ, B. Application security through sandbox virtualization. *Acta Polytechnica Hungarica 12*, 1 (2015), 83–101.

William Steingartner works as Associate Professor of Computer Science at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia. He defended his PhD thesis "The *Rôle* of Toposes in Computer Science" in 2008. His main fields of research are semantics of programming languages, category theory, compilers, data structures and recursion theory. He also works with cybersecurity and software engineering.

Igor Sivý studied at the Technical University of Košice and graduated in 2022. He currently works as a programmer in the game industry. In his free time, he works on implementing a compiler for his own language.

Innovative Solutions for Tetraplegia: A Smart Hand Orthosis Design

Ferenčík, Norbert; Sedláková, Veronika; Kolembusová, Petra;
Štefanovič, Branko; Hudák, Radovan; Steingartner, William

Abstract: ***Spinal cord injury (SCI) poses a significant medical challenge, affecting both hand dexterity and locomotor abilities. Ongoing advancements in medical technologies, spanning a spectrum of wearable devices, coupled with concurrent progress in rehabilitation treatments, aim to enhance hand function among individuals affected by SCI. The emergence of three-dimensional (3D) printing provides a cost-effective avenue for crafting personalized devices, fostering a surge of interest in integrating this technology with rehabilitation equipment, thereby complementing advancements in scientific research. Myoelectric control plays a pivotal role in achieving enhanced rehabilitation outcomes. It involves the detection and processing of weak electromyographic signals (EMG) from affected limb muscles to activate orthotic motors. A novel 3D-printed hand orthosis, responsive to electromyography signals, has been developed to facilitate grasping functionality in cervical SCI patients.***

Index Terms: ***3D printing technology, Arduino myoelectric orthosis, Electromyographic signal, Rehabilitation, Spinal cord injury, Tetraplegia***

1. Introduction

UPPER extremity (UE) weakness and/or paralysis following spinal cord injury (SCI) can limit the capacity to perform activities of daily living (ADL). Such disability significantly diminishes an individual's level of independence. Furthermore, the restoration of UE motor function in people with SCI remains a high priority in rehabilitation and the field of assistive technology [1], [6].

One effective approach to achieve optimal rehabilitation results is through myoelectric control, wherein a weak electromyographic signal (EMG) from the muscles of the affected limb is detected, processed, and used to activate a motor in the orthosis. A novel 3D-printed hand orthosis, controlled by electromyography (EMG) signals, was developed to enhance grasping function in patients with cervical SCI. The motor assists the user in performing the desired movement. The patient-directed "intentional" action of the device promotes patient engagement, as the orthosis rewards the patient with movement only when they use the correct muscles to perform the task. The hand exoskeleton system was applied to individuals with tetraplegia due to SCI, and its effectiveness was confirmed [30], [31].

Recent studies have inspired us to explore the design and development of 3D-printed orthoses using the finite element approach, examining different materials and loading conditions (Zhang et al., 2023). Additionally, significant work on VR-assisted hand therapy with a customized biomechatronic 3D-printed orthosis (Lee et al., 2023) and the innovative NOHAS orthotic hand actuated by servo motors and a mobile app for stroke rehabilitation (Smith et al., 2023) provide valuable insights for future applications in rehabilitation. [10], [25], [29].

The paper is organized as follows. Section 2 delves into the mechanism of operation, providing a detailed exploration of the operational intricacies. Section 3 focuses on 3D printing technology, offering insights into the utilization of this technology within the context of the study. In Section 4, production material is discussed, providing an overview of the materials employed in the manufacturing process. The aspects related to rehabilitation are expounded upon in Section 5, clarifying the rehabilitation processes integrated into the study. Finally, Section 6 encapsulates the findings and discussions, presenting conclusions drawn from the research and contributing to the broader understanding of the subject matter.

Manuscript received May 31, 2024.
This work was created thanks to support under the Operational Program Integrated Infrastructure for the project: Centre for Advanced Therapies of Chronic Inflammatory Diseases of the Locomotion System (CPT ZOPA), ITMS2014+: 313011W410, co-financed by the European Regional Development Fund. The work was also supported by the the Slovak Grant Agency – project VEGA 1/0387/22 – "Development and testing of systems for controlled stimulation of cell growth in a bioreactor environment using computer vision" and by national KEGA project 030TUKE-4/2023 – "Application of new principles in the education of IT specialists in the field of formal languages and compilers", granted by the Cultural and Education Grant Agency of the Slovak Ministry of Education.

Norbert Ferenčík (email: norbert.ferencik@tuke.sk), Veronika Sedláková (email: veronika.sedlakova@tuke.sk), Petra Komebusová (email: petra.kolembusova@tuke.sk), Branko Štefanovič (email: branko.stefanovic@tuke.sk), Radovan Hudák (email: radovan.hudak@tuke.sk) Faculty of Mechanical Engineering, Department of Biomedical Engineering, Technical University of Košice, Slovakia, William Steingartner (Corresponding author), Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia (e-mail: william.steingartner@tuke.sk).

2. THE MECHANISM OF OPERATION

The designed orthosis consists of three parts: a forearm cuff, hand, and finger ring sections. The forearm cuff comprises two subparts, specifically, a dorsal and a volar forearm splint. On the dorsal forearm splint, a linear motor capable of generating a 30 Newton force with a 41 mm stroke length is mounted to control wrist extension. The volar forearm splint stabilizes the wrist joint and is attached to the dorsal forearm splint with a Velcro strap, allowing for adjustments to fit the participant's forearm. The hand part wraps around the hand and is anchored to the linear motor. Therefore, when the motor is activated, the wrist part is pulled toward the forearm, resulting in wrist extension. The N20 DC motors (rated at 6 V) are equipped with an encoder and a reduction gearbox (100 rotations per minute (RPM) and a 100:1 gear ratio). A dual-H-bridge driver controls their rotation direction. The actuators rotate to pull the artificial tendons, performing flexion of the fingers. Alternatively, when the actuators rotate in the opposite direction, tension is relieved in the tendons, and the user can voluntarily extend the fingers. Motor 1 pulls the artificial tendons connected to the index finger and thumb, and it also pulls the artificial tendons of the middle, ring, and little finger. Both motors work together to perform pinch and cylindrical grips. The volunteer establishes the comfortable grip strength and range of motion (ROM) for motors 1. The dimensions of the base of the splints were designed to respect the range of motion limits in both grips. For the pinch grip, motor 1 pulls the tendons until the fingers touch, creating grip strength, and it pulls the tendons until the fingers touch the palm. For the cylindrical grip, motor 1 pulls the tendons until the thumb and index fingers touch, creating grip strength [2], [16], [18], [19], [31].

The finger ring components are positioned on each phalanx of the thumb, index, and middle finger. A nylon thread connects the volar side of each finger ring to the volar forearm splint. Cable guide structures have been incorporated on the volar side of each finger ring and hand part to direct the nylon thread along the fingertip to the volar forearm splint. As the linear motor extends the wrist, the nylon thread tightens, reinforcing the tenodesis grip. Consequently, wrist extension induces simultaneous flexion of the interphalangeal and metacarpophalangeal joints of each finger, including the thumb. Thus, the user can grasp objects placed in the palm or between the fingers by activating the linear motor, a fundamental mechanism facilitating a broader user base, particularly those with high-level SCI unable to control their wrists. The length of the nylon thread was adjusted to ensure sufficient pull on the finger ring parts during wrist extension [24], [26].

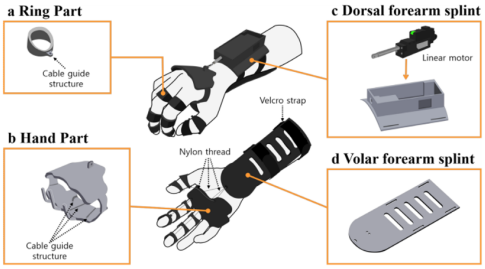


Figure 1. Schematic design of the hand orthosis

The designed myoelectric orthosis was created for operation through surface electromyography (sEMG) signals recorded from the user's upper extremity muscles. The control unit was specifically engineered to activate the linear motor when the sEMG signal surpassed a predefined threshold. Ensuring precision in signal acquisition, the placement of the sEMG electrodes was strategically optimized, taking into consideration both the level of injury and the subject's convenience [31].

We aimed to identify suitable muscles that were easily accessible and remained viable after SCI. In this experiment, either the ipsilateral biceps or the upper trapezius muscle was chosen as the target, as all subjects could contract these muscles without difficulty. Consequently, a pair of sEMG electrodes was positioned either on the ipsilateral biceps or the upper trapezius muscle, depending on the subjects' convenience. Following the guidelines of sEMG for the non-invasive assessment of muscles, the electrodes were placed on the most prominent bulge of the muscle belly, and the inter-electrode distance was set at 2 cm. Additionally, a ground electrode was situated at the olecranon of the dominant arm [31].

To ensure compatibility with a wearable robotic device, EMG signal acquisition and classification must be processed on a standalone, wearable device equipped with sufficient computational power and memory resources to efficiently analyze EMG data with minimal delay. The intention detection system should be wearable for extended periods each day. Consequently, both the classification system and the EMG sensors need to be comfortable, suitable for patients with impaired hand function, and easy to use for stroke survivors with cognitive deficits [24].

Before donning the orthosis, the sEMG signal underwent processing. To enhance the signal quality, we implemented a 1000-fold signal amplification and mitigated background noise by applying a Sallen-Key band-pass filter with a range of 10–500 Hz. The root mean square (RMS) was chosen as the parameter for controlling the linear motor, given its prevalence in the analysis of EMG signals and its close association with constant

force and muscle contraction. The figure illustrates the raw sEMG signal and RMS for the sEMG signal in each scenario. The on/off threshold was set at 80% of the maximal contraction level in RMS to differentiate signals from intentional and unintentional movements. However, subjects were permitted to adjust the threshold level to their comfort. This customization allowed the threshold to align with users' abilities and injury statuses. The RMS can be defined as follows [4], [14], [20].

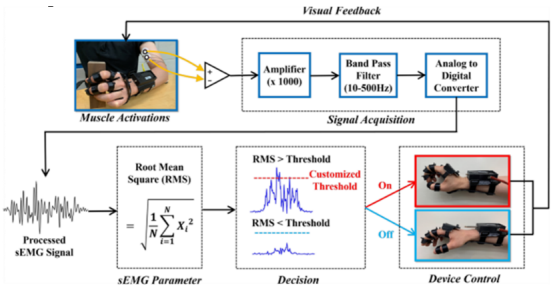


Figure 2. Overview of the control scheme

The sEMG signals recorded from the surface electrodes underwent various acquisition steps, including amplification and band-pass filtering, to enhance signal quality. Subsequently, RMS values of the processed sEMG signals were compared with the customized threshold [31].

3. 3D PRINTING TECHNOLOGY

The body of our myoelectric arm orthosis for spinal cord injury patients was printed on a 3D printer. Eight ring parts were printed for each phalanx of the thumb, index finger, and middle finger, as well as details of the hand, dorsal forearm splint, and volar forearm splint [7], [12]

3D printing technology involves certain steps in which a CAD-based model is first created and then converted into a stereolithography file (.STL). This file breaks the surface into a logical series of triangles, representing a portion of the 3D model's surface that is then used for the slicing algorithm. The STL file slices the model into thin cross layers, allowing the desired model to be printed by sequentially applying individual layers of thermoplastic materials through a temperature-controlled head. The model is built layer by layer, from bottom to top. The designed object is produced as a one-piece 3D part that does not require tooling [7], [12]

The use of 3D printing technology to develop new orthoses is considered a promising way to reduce costs through rapid production compared to previous metal orthoses [31].

A notable advantage of FDM is that it can create objects made of multiple types of materials by printing and then changing the printing material, giving the user more control over the fabrication of devices for experimental use. In addition to conventional materials such as PC, polystyrene (PS), ABS, and PLA, FDM can also print 3D models from fiberglass-reinforced polymers. However, the binder is usually mixed with ceramic or metal powders, allowing the material to be used in filament form [11], [15].

Figure 3. Anycubic Kobra 3D Printer

Utilizing the Anycubic Kobra 3D Printer, an orthosis was fabricated (see Figure 3). Next to the printer, there is an additional filament dryer eBOX, which eliminates moisture in the filament, significantly improving the quality of the printing process

4. PRODUCTION MATERIAL

Polylactic acid is a biodegradable polymer used in 3D printing, including the creation of myoelectric prosthetics. Here we list some key characteristics of this material:

1) Origin: PLA is derived from natural plant sources such as corn or sugarcane, making it a renewable and environmentally friendly material [27].
2) Biodegradability: PLA has tremendous value as other high-volume plastics like polyethylene and polypropylene are non-biodegradable and are made from fossil-derived ethylene and propylene. Although PLA is biodegradable, it is not currently renewable as it emits 1.3 kg CO2 equivalents/kg of synthesized plastic [3].
3) Low Toxicity: PLA is less toxic compared to some other plastics, like ABS. This makes it safer for use in medical applications, including prosthetic fabrication [32].
4) Low Temperature Resistance: PLA has relatively low temperature resistance compared

to other 3D printing materials such as ABS. This characteristic is important in prosthetic manufacturing, as certain details may be subject to heat during use [21].

5) Strength and Rigidity: PLA exhibits acceptable strength and rigidity, although these characteristics may be lower compared to some other 3D printing materials like ABS or PETG [21].
6) Color Options: PLA is available in various colors, including transparent options, offering possibilities for aesthetically pleasing prosthetic designs [17].

Table 1
COLOR COORDINATES, COLOR DIFFERENCE, AND CONTRAST RATIO OF SAMPLES

Sample code	L*	a*	b*	$\Delta P*$	Contrast ratio (%)
Neat PLA	91.8	1.1	-4.0		13.6
PLA-DCNP1	62.8	-1.6	18.6	32.6	20.6
PLA-DCNP3	35.0	2.6	-29.1	62.1	51.8
PLA-DCNP5	31.2	1.25	-24.6	64.0	58.2
PLA-C20A3	92.4	-0.13	-0.5	3.8	39.7

Overall, PLA is becoming a popular choice for printing bioelectric prosthetics, especially when environmental and safety considerations are important. However, when designing a prosthesis, it's crucial to consider the specific requirements and limitations of PLA, particularly its thermal characteristics [17].

Table 2
TECHNICAL CHARACTERISTICS OF PLA PLASTIC

Characteristic	**Value**
Melting point	173-178 ℃
Softening temperature	50 ℃
Hardness (according to Rockwell)	R70-R90
Relative elongation at break	3.8%
Bending strength	55.3 MPa
Tensile strength	57.8 MPa
Tensile strength	3.3 GPa
Elastic modulus during stretching	2.3 GPa
Glass transition temperature	60-65 ℃
Density	1.23–1.25 g/cm³
Minimum wall thickness	1 mm
Printing accuracy	± 0.1%
The size of the smallest details	0.3 mm
Shrinkage in the manufacture of products	No
Wet absorption	0.5–50%

5. REHABILITATION

Chronic upper limb deficits after Traumatic Brain Injury (TBI) and stroke are prevalent and often severely debilitating. Approximately 17% of individuals with TBI and over 50% of individuals with stroke do not fully recover upper limb function. These persistent upper limb deficits limit function and negatively impact the quality of life. Motor learning-based therapy (ML), utilizing high repetition and timely progression of task-oriented movements, is one of the most effective neurorehabilitation methods available [5], [8].

Recent studies have inspired us to explore the design and development of 3D-printed orthoses using the finite element approach, examining different materials and loading conditions (Zhang et al., 2023). Additionally, significant work on VR-assisted hand therapy with a customized biomechatronic 3D-printed orthosis (Lee et al., 2023) and the innovative NOHAS orthotic hand actuated by servo motors and a mobile app for stroke rehabilitation (Smith et al., 2023) provide valuable insights for future applications in rehabilitation. [22].

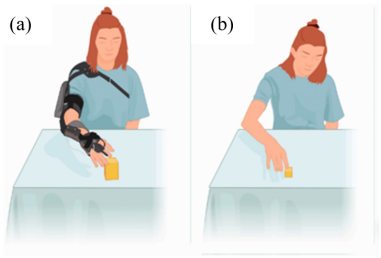


Figure 4. Functional task practice example with the myoelectric wrist–hand orthosis. (a) When the user attempts to move the elbow or grasp objects, sensors in the orthosis detect the myoelectric signal generated by the user's volitional effort to activate the motor, moving the elbow/hand in the desired direction and assisting the user in completing the desired movement. (b) Functional task practice without the orthosis to reinforce training [23].

Rehabilitation of the arm with the help of a myoelectric orthosis consists of performing various gripping exercises, such as picking up an object and moving it from one place to another. These exercises improve control over the hand and train the muscles. During the exercises, the patient is better able to control his or her hand due to the forces created by the motorized exoskeleton of the hand. Later, after rehabilitation, the person will be able to perform such actions independently much better than before training with the orthosis [9], [23].

This myoelectric orthosis, printed on a 3D printer, was also tested by the patient in exercises for the rehabilitation of the hand and wrist. The test consisted of two parts: the first part of the test assessed the ability to manipulate objects using ten standardized objects, and the second part assessed grip strength using nine rectangular wooden blocks, a tool cylinder and a credit card attached to a dynamometer, and a wooden bar. The nine wooden blocks of different weights and friction forces were used to evaluate grip strength

and stability, while the other three items were used to measure the torque generated by the palm, the lateral compression force, and the eccentric load that the grip can withstand, respectively. There was no time limit in performing the task [13], [28].

Figure 5. Grip exercises with a myoelectric orthosis

After baseline measurements, participants wore the myoelectric orthosis on their dominant limb and repeated the same exercises to assess the improvement in hand function. Movement quality was carefully monitored, and training practices were incrementally progressing as soon as the participant demonstrated an improved ability to perform a given task or movement component. After completing the experiment, orthosis performance was analyzed by comparing hand function before and after wearing the orthosis. The patient was able to grasp and lift objects after wearing the orthosis, which was impossible during baseline measurements [31].

6. CONCLUSIONS

This myoelectric orthosis is a user-friendly and highly effective tool in rehabilitation, specifically designed to improve grip strength and hand control for individuals with tetraplegia. Its simplicity and cost-effectiveness make it an accessible solution. The device's innovative, fully customized design sets it apart in the hand orthoses market, addressing unique anatomical considerations. Notably, its positive impact on rehabilitation outcomes is complemented by its affordability, making it a practical choice for a broader user base.

In summary, this myoelectric orthosis is a versatile and innovative solution, offering accessibility, customized design, and affordability. Its focus on improving grip strength enhances functional independence for individuals with tetraplegia, contributing to an improved quality of life.

It should be noted that the study is currently ongoing, and further research is required to corroborate these preliminary findings.

REFERENCES

[1] ANDROWIS, G. J., ET AL. The rehabilitation effects of myoelectric powered wearable orthotics on improving upper extremity function in persons with sci. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2021), IEEE, pp. 4944–4948.

[2] BENJUYA, N., AND KENNEY, S. B. Myoelectric hand orthosis. *JPO: Journal of Prosthetics and Orthotics 2*, 2 (1990), 149–154.

[3] BHAGIA, S., ET AL. Critical review of fdm 3d printing of pla biocomposites filled with biomass resources, characterization, biodegradability, upcycling and opportunities for biorefineries. *Applied materials today 24* (2021), 101078.

[4] BOOSTANI, R., AND MORADI, M. H. Evaluation of the forearm emg signal features for the control of a prosthetic hand. *Physiological measurement 24*, 2 (2003), 309.

[5] BRECEDA, E. Y., AND DROMERICK, A. W. Motor rehabilitation in stroke and traumatic brain injury: stimulating and intense. *Current opinion in neurology 26*, 6 (2013), 595–601.

[6] BRYCE, T. N., HUANG, V., AND ESCALON, M. X. Spinal cord injury. In *Braddom's Physical Medicine and Rehabilitation*. Elsevier, 2021, pp. 1049–1100.

[7] COMB, J., PRIEDEMAN, W., AND TURLEY, P. W. Fdm® technology process improvements. In *1994 International Solid Freeform Fabrication Symposium* (1994).

[8] DALY, J. J., ET AL. Long-dose intensive therapy is necessary for strong, clinically significant, upper limb functional gains and retained gains in severe/moderate chronic stroke. *Neurorehabilitation and neural repair 33*, 7 (2019), 523–537.

[9] DELPH, M. A., FISCHER, S. A., GAUTHIER, P. W., LUNA, C. H. M., CLANCY, E. A., AND FISCHER, G. S. A soft robotic exomusculature glove with integrated semg sensing for hand rehabilitation. In *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)* (2013), IEEE, pp. 1–7.

[10] GÓRSKI, F., GROHS, A., KUCZKO, W., ŻUKOWSKA, M., WICHNIAREK, R., SIWIEC, S., BĂILĂ, D.-I., ZELENAY, M., PĂCURAR, R., AND SANFILIPPO, F. Development and studies of vr-assisted hand therapy using a customized biomechatronic 3d printed orthosis. *Electronics 13*, 1 (2023), 79.

[11] GROSS, B. C., ET AL. Evaluation of 3d printing and its potential impact on biotechnology and the chemical sciences.

[12] KAMRAN, M., AND SAXENA, A. A comprehensive study on 3d printing technology. *MIT Int J Mech Eng 6*, 2 (2016), 63–69.

[13] KAPADIA, N., ZIVANOVIC, V., VERRIER, M., AND POPOVIC, M. Toronto rehabilitation institute–hand function test: assessment of gross motor function in individuals with spinal cord injury. *Topics in spinal cord injury rehabilitation 18*, 2 (2012), 167–186.

[14] KIM, K. S., CHOI, H. H., MOON, C. S., AND MUN, C. W. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current applied physics 11*, 3 (2011), 740–745.

[15] KRISTIAWAN, R. B., IMADUDDIN, F., ARIAWAN, D., UBAIDILLAH, AND ARIFIN, Z. A review on the fused deposition modeling (fdm) 3d printing: Filament processing, materials, and printing parameters. *Open Engineering 11*, 1 (2021), 639–649.

[16] LEE, I.-O., AND MOON, G.-W. Analysis and design of phase-shifted dual h-bridge converter with a wide zvs range and reduced output filter. *IEEE Transactions on Industrial Electronics 60*, 10 (2012), 4415–4426.

[17] MAHMOODI, A., GHODRATI, S., AND KHORASANI, M. High-strength, low-permeable, and light-protective nanocomposite films based on a hybrid nanopigment and biodegradable pla for food packaging applications. *ACS omega 4*, 12 (2019), 14947–14954.

[18] MARTINS, H. V. P., ET AL. Development of a robotic orthosis for fingers flexion motion by surface myoelectric control: Open source prototype. *Biomedical Signal Processing and Control 85* (2023), 105014.

[19] PETERS, H. T., PAGE, S. J., AND PERSCH, A. Giving them a hand: wearing a myoelectric elbow-wrist-hand orthosis reduces upper extremity impairment in chronic stroke. *Archives of physical medicine and rehabilitation 98*, 9 (2017), 1821–1827.

[20] PHINYOMARK, A., PHUKPATTARANONT, P., AND LIMSAKUL, C. Feature reduction and selection for emg signal classification. *Expert systems with applications 39*, 8 (2012), 7420–7431.

[21] PRZEKOP, R. E., ET AL. Graphite modified polylactide (pla) for 3d printed (fdm/fff) sliding elements. *Polymers 12*, 6 (2020), 1250.

[22] PUNDIK, S., ET AL. Myoelectric arm orthosis in motor learning-based therapy for chronic deficits after stroke and traumatic brain injury. *Frontiers in Neurology 13* (2022), 19.
[23] PUNDIK, S., MCCABE, J., KESNER, S., SKELLY, M., AND FATONE, S. Use of a myoelectric upper limb orthosis for rehabilitation of the upper limb in traumatic brain injury: A case report. *Journal of Rehabilitation and Assistive Technologies Engineering 7* (2020), 2055668320921067.
[24] RYSER, F., ET AL. Fully embedded myoelectric control for a wearable robotic hand orthosis. In *2017 International Conference on Rehabilitation Robotics (ICORR)* (2017), IEEE, pp. 615–621.
[25] SELVARAJ MERCYSHALINIE, E. R., GHADGE, A., IFEJIKA, N., AND TADESSE, Y. Nohas: A novel orthotic hand actuated by servo motors and mobile app for stroke rehabilitation. *Robotics 12*, 6 (2023), 169.
[26] SHOEMAKER, E. Myoelectric elbow-wrist-hand orthosis with active grasp for patients with stroke: a case series. *Canadian Prosthetics and Orthotics Journal 1*, 2 (2018).
[27] SINGH, N., ET AL. Experimental-theoretical comparative analysis of pla-based 3d lattice. *Journal of Thermoplastic Composite Materials* (2023).
[28] SOEKADAR, S., WITKOWSKI, M., GÓMEZ, C., OPISSO, E., MEDINA, J., CORTESE, M., CEMPINI, M., CARROZZA, M. C., COHEN, L., BIRBAUMER, N., ET AL. Hybrid eeg/eog-based brain/neural hand exoskeleton restores fully independent daily living activities after quadriplegia. *Science Robotics 1*, 1 (2016), eaag3296.
[29] UMER, U., MIAN, S. H., MOIDUDDIN, K., AND ALKHALEFAH, H. Exploring orthosis designs for 3d printing applying the finite element approach: Study of different materials and loading conditions. *Journal of Disability Research 2*, 1 (2023), 85–97.
[30] YANG, D., GU, Y., THAKOR, N. V., AND LIU, H. Improving the functionality, robustness, and adaptability of myoelectric control for dexterous motion restoration. *Experimental brain research 237* (2019), 291–311.
[31] YOO, H.-J., ET AL. Development of 3d-printed myoelectric hand orthosis for patients with spinal cord injury. *Journal of neuroengineering and rehabilitation 16*, 1 (2019), 1–14.
[32] ZHANG, Q., ET AL. Chemical composition and toxicity of particles emitted from a consumer-level 3d printer using various materials. *Environmental science and technology 53*, 20 (2019), 12054–12061.

Norbert Ferenčík is a graduate of the Cybernetics program as part of the bachelor's degree (2014) and the Artificial Intelligence and Intelligent Systems program as part of the engineering degree (2016) at the Faculty of Electrical Engineering, Technical University in Košice. During his doctoral studies, he completed a half-year internship at UC Berkeley and after defending his dissertation in the Artificial Intelligence program (2020) works as a postdoctoral researcher at the Department of biomedical engineering and measurement, with a professional focus on 3D printing, bioprinting in medicine and engineering, and the field of bioreactors. As part of solving research tasks, she is the co-author of several original articles in domestic and foreign magazines and anthologies. At his domestic department, he is involved in solving many domestic and international projects.

Veronika Sedláková initially graduated as a radiological technologist in 2020 before continuing her engineering studies in Biomedical Engineering at the Faculty of Mechanical Engineering, Technical University in Košice. Currently, she is a PhD student (2023-2027) at the Department of Biomedical Engineering and Measurement. Her research specializes in the analysis of the influence of selected biophysical parameters on cellular behavior within bioreactor environments, incorporating her expertise in biomedical engineering.

Petra Kolembusová is a graduate of Biomedical Engineering at Faculty of Mechanical Engineering, Technical University in Košice. Currently she is a PhD. student (2022-2026) at the dept. of Biomedical Engineering and Measurement. Her main research area is tissue engineering, 3D printing using polymers, bioprinting, CAD/CAM systems, with her main focus on bioreactor systems in biomedicine.

Branko Štefanovič is a young researcher at the Department of Biomedical Engineering and Measurement, Faculty of Mechanical Engineering, Technical University in Košice. He has experience in the design and production of various prosthetic and orthotic aids and devices using CAD/CAM systems. He has experience in operating various types of 3D scanners and additive manufacturing technologies. He is a co-creator of patent PP 8-2020 - Personalized medical device and method of its preparation.

Radovan Hudák is currently a professor, director of the Institute of Special Engineering Process Sciences and head of the Department of Biomedical Engineering and Measurement of the Faculty of Mechanical Engineering, Technical University of Košice. His research activities include the use of additive technologies in medicine, human biomechanics and medical thermography. He participated in several international internships and stays at the University of Ghent, Belgium (2002), UIC Chicago (2006), the Technical University of Bialystok, Poland (2006), CTU in Prague (2006). He is a member of two ASTM committees, committee E20 Temperature measurement and committee F42 Technologies of additive manufacturing, and a member of the editorial board of the magazine ProIN. He is one of the founders of the YBERC conference. He is the author of more than 350 publications, including 8 scientific monographs and 9 textbooks.

William Steingartner works as Associate Professor of Computer Science at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia. He defended his PhD thesis "The *Rôle* of Toposes in Computer Science" in 2008. His main fields of research are theoretical computer science and software engineering. He also works in cybersecurity and applied computer science.

Reviewers:

Australia
Abramov, Vyacheslav; Monash University
Begg, Rezaul; Victoria University
Bem, Derek; University of Western Sydney
Betts, Christopher; Pegacat Computing Pty. Ltd.
Buyya, Rajkumar; The University of Melbourne
Chapman, Judith; Australian University Limited
Chen, Yi-Ping Phoebe; Deakin University
Hammond, Mark; Flinders University
Henman, Paul; University of Queensland
Palmisano, Stephen; University of Wollongong
Ristic, Branko; Science and Technology Organisation
Sajjanhar, Atul; Deakin University
Sidhu, Amandeep; University of Technology, Sydney
Sudweeks, Fay; Murdoch University

Austria
Derntl, Michael; University of Vienna
Hug, Theo; University of Innsbruck
Loidl, Susanne; Johannes Kepler University
Linz Stockinger, Heinz; University of Vienna
Sutter, Matthias; University of Innsbruck

Brazil
Parracho, Annibal; Universidade Federal
Fluminense Traina, Agma; University of Sao Paulo
Traina, Caetano; University of Sao Paulo
Vicari, Rosa; Federal University of Rio Grande

Belgium
Huang, Ping; European Commission

Canada
Fung, Benjamin; Simon Fraser University
Grayson, Paul; York University Gray,
Bette; Alberta Education
Memmi, Daniel; UQAM
Neti, Sangeeta; University of Victoria
Nickull, Duane; Adobe Systems, Inc. Ollivier-Gooch,
Carl; The University of British Columbia Paulin,
Michele; Concordia University
Plaisent, Michel; University of Quebec Reid, Keith;
Ontario Ministry og Agriculture Shewchenko,
Nicholas; Biokinetics and Associates Steffan,
Gregory; University of Toronto Vandenberghe,
Christian; HEC Montreal

Czech Republic
Kala, Zdenek; Brno University of Technology
Korab, Vojtech; Brno University of technology
Lhotska, Lenka; Czech Technical University

Finland
Lahdelma, Risto; University of Turku
Salminen, Pekka; University of Jyvaskyla

France
Cardey, Sylviane; University of Franche-Comte
Klinger, Evelyne; LTCI – ENST, Paris
Roche, Christophe; University of Savoie
Valette, Robert; LAAS - CNRS

Germany
Accorsi, Rafael; University of Freiburg
Glatzer, Wolfgang; Goethe-University
Gradmann, Stefan; Universitat Hamburg
Groll, Andre; University of Siegen
Klamma, Ralf; RWTH Aachen University
Wurtz, Rolf P.; Ruhr-Universitat Bochum

Hungary
Mester, Gyula; Óbuda University, Budapest

India
Pareek, Deepak; Technology4Development
Scaria, Vinod; Institute of Integrative Biology
Shah, Mugdha; Mansukhlal Svayam

Ireland
Eisenberg, Jacob; University College Dublin

Israel
Feintuch, Uri; Hadassah-Hebrew University

Italy
Badia, Leonardo; IMT Institute for Advanced
Studies Berrittella, Maria; University of Palermo
Carpaneto, Enrico; Politecnico di Torino

Japan
Hattori, Yasunao; Shimane University
Livingston, Paisley; Linghan University
Srinivas, Hari; Global Development Research Center
Obayashi, Shigeru; Institute of Fluid Science,
Tohoku University

Netherlands
Mills, Melinda C.; University of Groningen
Pires, Luís Ferreira; University of Twente

New Zealand
Anderson, Tim; Van Der Veer Institute

Portugal
Cardoso, Jorge; University of Madeira
Natividade, Eduardo; Polytechnic Institute of
Coimbra Oliveira, Eugenio; University of Porto

Singapore
Tan, Fock-Lai; Nanyang Technological University

South Korea
Kwon, Wook Hyun; Seoul National University

Spain
Barrera, Juan Pablo Soto; University of Castilla
Gonzalez, Evelio J.; University of La Laguna
Perez, Juan Mendez; Universidad de La
Laguna Royuela, Vicente; Universidad de
Barcelona Vizcaino, Aurora; University of
Castilla-La Mancha Vilarrasa, Clelia Colombo;
Open University of Catalonia

Sweden
Johansson, Mats; Royal Institute of Technology

Switzerland
Niinimaki, Marko; Helsinki Institute of Physics
Pletka, Roman; AdNovum Informatik AG
Rizzotti, Sven; University of Basel
Specht, Matthias; University of Zurich

Taiwan
Lin, Hsiung Cheng; Chienkuo Technology University
Shyu, Yuh-Huei; Tamkang University
Sue, Chuan-Ching; National Cheng Kung
University

United Kingdom
Ariwa, Ezendu; London Metropolitan University
Biggam, John; Glasgow Caledonian University
Coleman, Shirley; University of Newcastle
Conole, Grainne; University of Southampton
Dorfler, Viktor; Strathclyde University
Engelmann, Dirk; University of London
Eze, Emmanuel; University of Hull
Forrester, John; Stockholm Environment Institute
Jensen, Jens; STFC Rutherford Appleton
Laboratory Kolovos, Dimitrios S.; The University
of York McBurney, Peter; University of Liverpool
Vetta, Atam; Oxford Brookes University
WHYTE, William Stewart; University of Leeds
Xie, Changwen; Wicks and Wilson Limited

USA
Bach, Eric; University of Wisconsin Bolzendahl,
Catherine; University of California Bussler,
Christoph; Cisco Systems, Inc. Charpentier,
Michel; University of New Hampshire Chong,
Stephen; Cornell University
Collison, George; The Concord Consortium
DeWeaver, Eric; University of Wisconsin -
Madison Gans, Eric; University of California
Gill, Sam; San Francisco State University
Hunter, Lynette; University of California Davis
Iceland, John; University of Maryland
Kaplan, Samantha W.; University of Wisconsin
Langou, Julien; The University of Tennessee
Liu, Yuliang; Southern Illinois University
Edwardsville Lok, Benjamin; University of Florida
Minh, Chi Cao; Stanford University
Morrissey, Robert; The University of Chicago
Mui, Lik; Google, Inc
Rizzo, Albert ; University of Southern
California Rosenberg, Jonathan M. ; University
of Maryland Shaffer, Cliff ; Virginia Tech
Sherman, Elaine; Hofstra University
Snyder, David F.; Texas State University
Song, Zhe; University of Iowa
Wei, Chen; Intelligent Automation, Inc.
Yu, Zhiyi; University of California

Welcome to IPSI Conferences and Journals!

http://tir.ipsitransactions.org

http://www.ipsitransactions.org

CIP – Katalogizacija u publikaciji
Narodna biblioteka Srbije, Beograd
ISSN 1820 – 4503
The IPSI Transactions
on Internet Research
COBISS.SR - ID 119127052