Annals of Computer Science and Information Systems
Volume 38

Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering

December 1–2, 2023
MANUU, Hyderabad, India

Pradeep Kumar, Manuel Cardona, Vijender Kumar Solanki,
Tran Duc Tan, Abdul Wahid (eds.)

PTI

Annals of Computer Science and Information Systems, Volume 38

ISSN: 2300-5963

Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering

Pradeep Kumar, Manuel Cardona,
Vijender Kumar Solanki, Tran Duc Tan,
Abdul Wahid (eds.)

PTI
POLSKIE TOWARZYSTWO INFORMATYCZNE
POLISH INFORMATION PROCESSING SOCIETY

Annals of Computer Science and Information Systems, Volume 38

Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering

ISBN 978-83-969601-5-3
ISSN: 2300-5963
DOI: 10.15439/978-83-969601-5-3

Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover photo:
Aleksander Denisiuk,
Elbląg, Poland

Also in this series:

Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**

Volume 36: Position Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-1-5, ISBN USB: 978-83-969601-2-2**

Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8**

Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation **ISBN 978-83-965897-8-1**

Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-965897-6-7, ISBN USB: 978-83-965897-7-4**

Volume 32: Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-4-3, ISBN USB: 978-83-965897-5-0**

Volume 31: Position Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-2-9, ISBN USB: 978-83-965897-3-6**

Volume 30: Proceedings of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-962423-9-6, ISBN USB: 978-83-965897-0-5**

Volume 29: Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics**ISBN WEB: 978-83-962423-7-2, ISBN USB: 978-83-962423-6-5**

Volume 28: Proceedings of the 2021 International Conference on Research in Management & Technovation, **ISBN WEB: 978-83-962423-4-1, ISBN USB: 978-83-962423-5-8**

Volume 27: Proceedings of the Sixth International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-962423-2-7, ISBN USB: 978-83-962423-3-4**

Volume 26: Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-959183-9-1, ISBN USB: 978-83-962423-0-3**

Volume 25: Proceedings of the 16th Conference on Computer Science and Intelligence Systems, **ISBN Web 978-83-959183-6-0, ISBN USB 978-83-959183-7-7, ISBN ART 978-83-959183-8-4**

Dear Reader, we are delighted to share with you a glimpse of the 8th International Conference on Research in Intelligent Computing in Engineering (RICE 2023). RICE 2023 is organized by the Department of Computer Science and Information Technology, School of Technology, Maulana Azad National Urdu University (A Central University), Hyderabad, Telangana, India; Jointly co-organized by Universidad Don Bosco, El Salvador, CA, during December 1st–2nd, 2023.

We are truly thankful to the Polish Information Processing Society (PTI), Poland for approving the proceedings of the 8th International Conference on Research in Intelligent Computing in Engineering (RICE 2023). It is appearing in the Annals of Computer Science and Information Systems series by PTI (ISSN-2300-5963). The series has been submitted to Copernicus, DBLP, Cross Ref, Scholar, BazEkon, Open Access Library, Academic Keys, Journal Click, PBN, and ARIANTE. At this stage, the efforts, whole-hearted support, and suggestions given by Editor-in-Chief Prof. Marcin Paprzycki and Prof. Maria Ganzha are highly applaudable and commendable.

We are pleased to report that various researchers are interested in participating in the 8th edition of RICE 2023. It is a privilege for us to hear five keynote speakers from different countries share their insightful perspectives on conference-related topics. The information is provided below:

- Dr. Le Anh Ngoc, Director of Innovations, Swinburne Vietnam Alliance.
- Dr. A. Govardhan. Senior Professor & Rector, Jawaharlal Nehru Technical University, Hyderabad, India.
- Dr. Tran Duc Tan, Vice Dean, Phenikaa University, Hanoi, Vietnam.
- Dr. Atul Negi, Professor, CSIS, University of Hyderabad, India.
- Dr. Bui Tien Son, Hanoi University of Industry, Hanoi, Vietnam.

Finally, we would like to take this opportunity to express our sincere appreciation to the Advisory Board, Technical Program Committee, Organizing Committee, International Scientific Committee, institutions, industries, and volunteers, who contributed to the success of this conference either directly or indirectly.

Proceeding's Editors – RICE 2023:

Pradeep Kumar, Maulana Azad National Urdu University, Hyderabad, Telangana, India.

Manuel Cardona, Universidad Don Bosco, El Salvador, Central America.

Vijender Kumar Solanki, CMR Institute of Technology, Hyderabad, Telangana, India.

Tran Duc Tan, Phenikaa University, Hanoi, Vietnam.

Abdul Wahid, Maulana Azad National Urdu University, Hyderabad, Telangana, India.

Annals of Computer Science and Information Systems, Volume 38

Eighth International Conference on Research in Intelligent Computing in Engineering

December 1–2, 2023. Hyderabad, India

TABLE OF CONTENTS

The Eighth International Conference on Research in Intelligent Computing in Engineering

rice2023 Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 1–6
DOI: 10.15439/2023R43
ISSN 2300-5963 ACSIS, Vol. 38

Applications of Machine Learning for Diabetes Prediction: A Comprehensive Review

Nayeem Ahmed
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India
naeemahmed0410@gmail.com.

Syed Imtiyaz Hassan
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Gachibowli Hyderabad, India
s.imtiyaz@gmail.com.

Md. Zair Hussain
Polytechnic Hyderabad
Maulana Azad National Urdu University
Gachibowli Hyderabad, India
zair@manuu.edu.in

***Abstract*—The use of machine learning techniques has drawn more attention due to its potential to improve early identification and intervention in diabetes, a critical global health concern. This article offers an extensive overview of the various machine learning algorithms used in diabetes prediction, including ensemble techniques, logistic regression, support vector machines, decision trees, and neural networks. The research closely examines how these algorithms make use of a variety of data sources, including wearable sensor data, electronic health records, clinical data, and genetic information. The report also emphasizes the difficulties that these applications face, including as interpretability, model integration into clinical procedures, and ethical issues. This review elucidates the significant influence of machine learning on diabetes prediction, paving the way for more useful risk assessment, individualized therapies, and improved patient outcomes. It does this by thoroughly examining recent studies and their conclusions.**

***Index Terms*—Machine Learning, Data Mining, Diabetes Prediction, Support Vector Machine, Neural Networks**

I. Introduction

Diabetes has emerged as a global health challenge, with its prevalence reaching alarming levels worldwide. As illustrated in figure.1, the International Diabetes Federation estimates that 463 million individuals worldwide had diabetes in 2019 and that 700 million will have the disease by 2045 [1]. Diabetes causes serious consequences including cardiovascular disease, renal failure, and blindness, placing a significant strain on patients, healthcare systems, and society at large [2].

Early prediction and intervention play a crucial role in managing diabetes effectively and reducing its impact on individuals' health outcomes [3]. Traditional approaches to diabetes prediction have relied on clinical risk scores and biomarkers, but they often lack accuracy and fail to capture the complexity of the disease. A paradigm change in diabetes prediction has been brought about by the development of machine learning (ML) techniques, which have the potential to enhance the precision, effectiveness, and personalized character of predictive models [4].

A substantial corpus of research has been done in the last ten years on using machine learning algorithms to predict diabetes. This research has utilised several ML methods, such as Logistic Regression [5], Decision Trees [6], Support Vector Machines [7], Neural Networks [8], and deep learning models, to create prediction models that can precisely identify people at risk of getting diabetes. By utilizing diverse data sources such as electronic health records (EHRs), genomic data, wearable devices, and behavioral data, ML models can capture the complex interplay of factors contributing to diabetes onset.

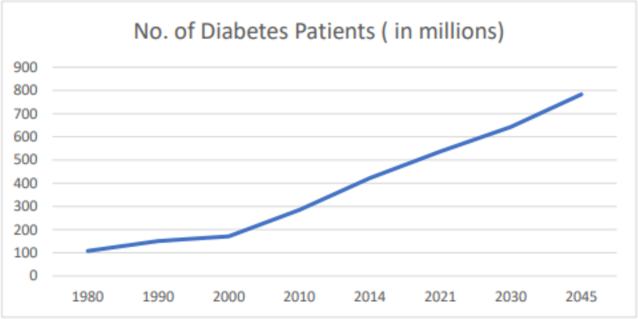


Figure.1: No of Diabetes cases worldwide [1]

The skills of machine learning in predicting diabetes have been demonstrated in a number of significant research studies. As an example, researchers in [4] successfully created a prediction model to evaluate the risk of diabetes by utilizing different classifiers on the PIMA dataset. The effectiveness and accuracy of data mining and machine learning techniques in reducing risk variables were demonstrated in this study. A noteworthy study in [9] combined genomic and clinical data, using deep learning techniques to improve prediction accuracy and reveal the underlying genetic architecture of diabetes.

These key studies and others have paved the way for advancements in diabetes prediction, showcasing the potential of machine learning techniques to revolutionize clinical decision_making and improve patient outcomes. ML models offer the ability to integrate large-scale, heterogeneous data, uncover hidden patterns, and generate accurate risk predictions tailored to individual patients.

Our goal in this large review paper is to provide a thorough summary of the most recent machine learning-based research on diabetes prediction. We will examine the methodology used in these investigations, assess their contributions to the literature, and highlight the most important outcomes and difficulties found. To advance the science of diabetes prediction and direct the creation of more precise and clinically useful ML models, we will synthesize the existing literature to find gaps and opportunities for future study.

The subsequent sections of the paper are structured as follows: Section 2 introduces the machine learning algorithms utilized for diabetes prediction. Section 3 discusses the data sources and features utilized in diabetes prediction. Section 4 focuses on the applications of machine learning in diabetes prediction. The impacts, challenges, and future directions in this field are addressed in Section 5. Finally, Section 6 concludes the paper with a conclusion.

II. Machine Learning Algorithms for Diabetes Prediction

Machine learning algorithms have been extensively applied to diabetes prediction, offering valuable insights and improved accuracy in identifying individuals at risk of developing the disease. Within this section, we will delve into the intricacies of diverse machine learning algorithms, exploring their specific applications in diabetes prediction.

A. *Logistic Regression*

Diabetes can be predicted well using the well-liked method for binary classification problems, logistic regression [5]. This strategy effectively replicates the relationship between independent variables and the probability of a particular outcome. In studies predicting diabetes, researchers have used logistic regression models, frequently integrating clinical and genetic variables.

For instance, researchers [10] used logistic regression models in a study to foretell the onset of diabetes. They gathered a wide range of clinical characteristics, including age, BMI, blood pressure, and genetic markers. They discovered that the addition of genetic markers considerably improved the predictive performance of the model by studying the data from a large cohort of patients. The study demonstrated how logistic regression may use a mix of clinical and genetic data to detect diabetes risk early. This demonstrates the algorithm's potential for early diabetes onset prediction.

B. *Decision Trees and Random Forests*

Decision tree algorithms, such as C4.5 and CART, have been widely used in diabetes prediction due to their interpretability and ability to handle both numerical and categorical data [6]. Decision trees recursively split the data based on features to create a tree-like structure, allowing for easy interpretation of the prediction process. Random forests, an ensemble method based on decision trees, combine multiple decision trees to improve the model's performance and robustness.

Researchers used decision tree algorithms to identify risk variables related with type2 diabetes by examining a large dataset of electronic health records in their work, which was published in [11]. They built decision tree models to identify critical factors linked with diabetes onset by analyzing several clinical variables such as BMI, fasting glucose levels, and family history. Their findings provide light on the fundamental elements that contribute to the development of diabetes.

Random forests have also been utilized to increase the model's accuracy and generalization capability in diabetes prediction. The authors of [12] used a mix of clinical and genetic factors to predict diabetes using random forest models. Their study found that adding both types of data enhanced prediction ability when compared to models that just used clinical variables. The random forest method shows promise in capturing the complicated interactions between multiple risk factors and the onset of diabetes.

C. *Support Vector Machines (SVM)*

Support Vector Machines (SVM) are robust supervised learning algorithms employed for classification tasks, including diabetes prediction. The primary objective of SVM is to determine an optimal hyper plane that effectively separates different classes by maximizing the margin between them [7].

Researchers in [13] employed SVM for diabetes prediction, utilizing a combination of clinical measurements, genetic markers, and lifestyle factors. Their study demonstrated the effectiveness of SVM in accurately classifying individuals at risk of developing diabetes. By integrating diverse data sources, including genetic and lifestyle factors, their SVM model achieved high prediction accuracy.

D. *Neural Networks*

Neural networks, especially deep learning models as referenced in [8], have garnered substantial attention in diabetes prediction owing to their capability to discern intricate patterns from high-dimensional data. Deep learning models consist of multiple layers of interconnected artificial neurons that can extract and process features automatically.

The authors of [14] suggested DiaNet, a revolutionary deep learning network for predicting diabetes from retinal pictures. DiaNet has an outstanding accuracy of more than 84% in detecting important retinal regions and distinguishing the Qatari diabetic cohort from the control group. The study found that retinal images include predictive markers for diabetes and other co morbidities, implying that retinal images could be used in clinical diagnosis in the future.

E. *Ensemble Methods*

Ensemble methods combine multiple weak classifiers to create a strong predictive model. AdaBoost [15] and Gradient Boosting [16] are popular ensemble techniques used in diabetes prediction. These methods iteratively train weak classifiers and assign higher weights to misclassified instances, focusing on the difficult samples.

The researchers introduce eDiaPredict in [17], an ensemble-based system for diabetes prediction that uses a variety of machine learning methods, including Support Vector Machine, Neural Network, Random Forest, XGBoost, and Decision tree. This technique remarkably achieves a 95% accuracy rate when applied to the PIMA Indian diabetes dataset. This emphasizes the value of effective machine learning algorithms in predicting and identifying severe situations in diabetes patients early on.

The items in table 2.1 provide an example of the wide range of machine learning techniques used in diabetes prediction. The ability to correctly identify people who are at risk of getting diabetes has significantly improved thanks to the use of ensemble methods, logistic regression, decision trees, support vector machines, neural networks, and ensem-

Table II. Existing Works On Diabetes Prediction Using Machine Learning Algorithms

Study	Algorithm	Dataset	Features	Accuracy (%)	Key findings
[10]	Logistic Regression	HER data	Clinical Genetic	82.5	Inclusion of genetic makers significantly improves prediction performance.
[11]	Decision Tree	Electronic health records	BMI, Fasting glucose levels, Family history	75.3	Identified key risk factors associated with type 2 diabetes onsets.
[12]	Random Forest	Clinical and genetic data	Clinical Genetic	87.9	Incorporating genetic features enhances prediction accuracy compared to using only clinical data.
[13]	Support Vector Machine	Clinical Measurements, Genetic markers, Lifestyle factors	Clinical Genetic Lifestyle	81.2	SVM accurately classifies individual's diabetes by integrating diverse data sources.
[14]	Neural Networks	Retinal Images	Retinal Images	91.6	Deep learning model using CNN architecture achieves high accuracy in non-invasive diabetes detection.
[17]	Ensemble Methods	Clinical & demographic data	Clinical Demographic	95	Logistic Regression is an efficient algorithm for prediction models with data preprocessing normalization and ensemble techniques enhancing overall performance.
[18]	Logistic Regression	Prima Indians Diabetes Database (PIDD)	Clinical Genetic	95.20	Logistic regression model provides accurate prediction using combined clinical and genetic features.
[19]	Random Forest	Prima Indians Diabetes Database	Clinical Genetic	83.67	When predicting diabetes, RF performed better than the deep learning and SVM techniques.
[20]	Support Vector Machine	Prima Indians Diabetes Database	Clinical Lifestyle	98.7	The proposed architecture using K-means clustering and SVM achieved an accuracy of 98.7% in predicting diabetes patients.
[21]	Artificial Neural Networks	Kurdistan region dataset	Genetic	91	The error rate decreased during training, indicating improved prediction accuracy based on network design.
[22]	Ensemble Methods (AdaBoost)	Diabetes UCI dataset	Clinical	98	Diabetes prediction using the AdaBoost M1 ensemble algorithm has a 98% accuracy rate.

ble approaches. These algorithms, when used in conjunction with sensible feature selection and model optimization, have the potential to enhance clinical decision-making and enable tailored treatments for the management of diabetes.

III. Applications of Machine Learning in Diabetes Prediction

Machine learning has been applied to various specific applications in diabetes prediction, offering valuable insights and potential clinical implications. This section, discusses existing studies that have utilized machine learning for applications such as early detection, risk stratification, and personalized treatment in the context of diabetes prediction.

The ability to quickly intervene and prevent complications depends on the early identification of diabetes. Machine learning algorithms have shown promise in identifying persons at risk of diabetes before clinical symptoms appear [23].

Based on a person's likelihood of getting diabetes or issues associated to diabetes, machine learning models can help group people into various risk categories. This enables tailored treatment regimens and targeted actions [24].

Building models that can analyze patient-specific data and generate specialized predictions for diabetes diagnosis, management, and therapy is a key component of personalized treatment for diabetes prediction using machine learning. Predictive models are created using machine learning algorithms that are trained on a variety of patient variables, including medical history, genetic information, lifestyle factors, and biomarkers [25].

A. Integration of Machine Learning Model into the Diabetes Clinical Workflow

Machine learning models must be smoothly incorporated into the clinical workflow in order to be used for diabetes prediction. As depicted in figure.2, the integration process entails giving careful consideration to data collection, preprocessing, model training, result communication, and decision support. This section focuses on the process for incorporating the clinical workflow for diabetes prediction with the machine learning model.

Firstly, patient data is collected, which includes relevant medical records, laboratory results, and patient demographics. This data is then pre-processed to handle missing values, outliers, and standardize the variables. Feature engineering techniques are applied to extract meaningful features from the data, such as glucose levels, body mass index, and medical history. Subsequently, the pre-processed data is utilized

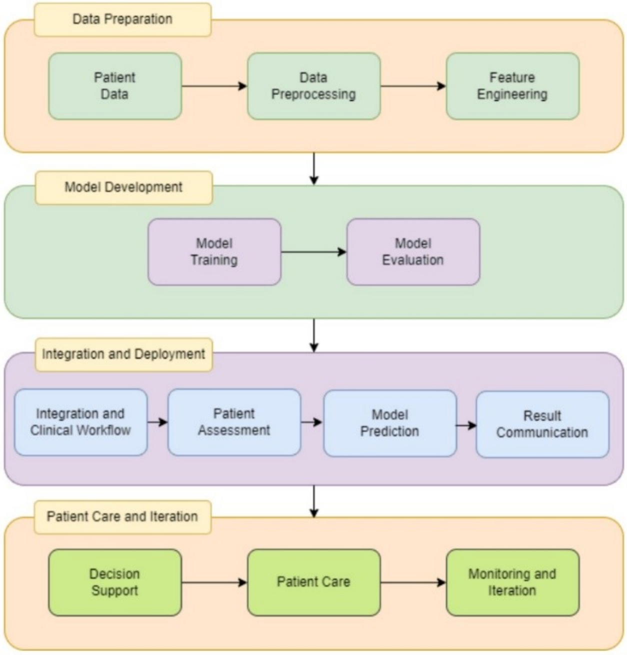


Figure.2: Integration of Machine Learning Model into the Diabetes Clinical Workflow

to train machine learning models. For diabetes prediction, diverse algorithms like support vector machines, logistic regression, and deep learning models can be employed. The trained models are evaluated to assess their performance and determine their accuracy, sensitivity, specificity, and other relevant metrics.

Once the models are validated, they are integrated into the clinical workflow. The models receive patient assessments and generate predictions regarding the likelihood of diabetes. These predictions are communicated to healthcare providers and patients, enabling informed decision-making and personalized care planning.

The model's results serve as decision support for clinicians, aiding in the determination of appropriate treatment plans, lifestyle modifications, or the need for further diagnostic tests. Additionally, the patient's assessment, along with the model's predictions, guides the communication of the results to the patient, fostering shared decision-making and patient engagement. To ensure the continuous monitoring and improvement of the model's performance, regular monitoring and iteration are essential. This includes tracking the model's predictions, evaluating its accuracy over time, and updating the model based on new patient data and feedback.

The integrated clinical workflow for diabetes prediction faces challenges that need to be addressed. Some of the factors that need to be considered include the interpretability and explain ability of machine learning models, the smooth integration of these models into existing clinical workflows, and the ethical aspects related to data privacy and security.

IV. Literature Review

In recent years, machine learning has become a pivotal innovation in medicine, with a promising outlook for the future. This study aims to employ machine learning classifiers to categorize diabetes patients based on their self-reported information and clinical conditions. We provide an overview of research conducted over the past decade to identify shortcomings in existing works related to machine learning classifiers for diabetes treatment strategies.

Sun and Zhang [26] investigated a range of deep learning and classification techniques, such as support vector machines, decision trees, random forests, and artificial neural networks. The authors [27] used a logistic regression-based classification technique to classify diabetes-related data in different research. There were 459 patients in the training dataset and 128 patients in the testing dataset. The logistic regression model is noteworthy for achieving a high 92% classification accuracy. It's important to note, nevertheless, that this model's validation was limited because it wasn't compared to other diabetes prediction models that are currently in use. For training and testing, the dataset was divided in half.

Naive Bayes, decision trees, and SVM learning techniques were used by researchers in their examination of the Pima Indians Diabetes Collection [28]. Notably, when it came to predicting diabetes, the Naive Bayes classifier showed the best accuracy. Using 10 equal pieces of the dataset—nine for training and one for assessment—Sisodia used tenfold cross-validation. Precision, accuracy, recall, and area under the curve were employed in the assessment as conventional evaluation criteria for diabetes prediction.

The authors in [29] evaluated a number of machine learning approaches in their study. In particular, they assessed how well neural networks (NN), random forests, and Naive Bayes performed. The Matthews correlation coefficient was the assessment metric used by the authors to determine how effective these strategies were.

To extract relevant features from the Pima Indians Diabetes Dataset, the authors in [30] used two different feature selection techniques: Principal Component Analysis (PCA) and Linear Discriminant Evaluation. Factor analysis approaches include Principal Component Analysis and Linear Discriminant Analysis. A comparative comparison of attribute selection procedures was also included in the study. The authors used the dataset under examination to test a number of machine learning techniques, such as the AdaBoost, K-Nearest Neighbors (KNN), and Radial Basis Kernel, for the classification job.

A single diagnosis strategy for early-stage diabetes is clearly not very successful, as demonstrated by the Gujral Writing Survey of Diabetes Assumptions findings [31]. Artificial Neural Networks (ANN) incorporate numerous classifiers, such Evolutionary Algorithms, Principal Component.

Analysis, and Support Vector Machines (SVM), to get the best results.

In the study of the Pima Indians Diabetes Dataset, the researchers[32] made noteworthy contributions. The importance of variables including BMI, blood glucose levels, and the number of pregnancies in the dataset was highlighted by their analysis. Using logistic regression and RStudio, they estimated accuracy and obtained an accuracy rate of 75.32%.

In their examination of the Pima Indians Diabetes Dataset, the authors in [33] used a variety of models, such as the multilayer perceptron, Bayes net, Hoeffding tree,and JRip. Both the greedy iterative and the best initial feature selection techniques were used in the research to increase classifier effectiveness. The researchers chose just four characteristics from the complete eight: age, BMI, diabetes pedigree function, and plasma glucose level. With a recall score of 76.2% and an accuracy rate of 75.7%, the Hoeffding tree approach in particular showed remarkable results.

Diabetic complications can be rather serious and the illness spreads quickly. Though trustworthy statistics are hard to come by, early diagnosis lowers risks. With the help of feature selection, hyperparameter optimization, and missing value imputation, the authors in [34] can provide a weighted ensemble of machine learning classifiers (NB, RF, DT, XGB, and LGB) and introduce a SA dataset. The new dataset for reliable diabetes prediction models employing population-level data is beneficial to the ensemble (DT + RF + XGB + LGB) with our preprocessing, as demonstrated by the considerable improvement in prediction (0.735 accuracy and 0.832 AUC).

Support Vector Machine (SVM) and Artificial Neural Network (ANN) models are employed in a fused machine learning (ML) technique presented in another study in [35]. The final diabetes diagnosis made by the fuzzy logic system is based on real-time medical information, and the dataset is split into training and testing data in a 70:30 ratio. With an astounding accuracy of 94.87%, the suggested fused model outperforms earlier approaches.

A low-code Pycaret machine learning approach is used in another study [36] for the categorization, detection, and prediction of diabetes. Gradient boosting emerges as the most accurate when many classifiers are hyper-tuned; it achieves 90% accuracy, outperforming other machine learning classifiers.

V. Discussion and Challenges

A range of machine learning techniques for diabetes prediction are presented in the evaluated literature. The following are significant flaws and difficulties, which include the requirement for larger and more varied datasets, the investigation of deep learning methodologies, and the development of strict model comparison procedures.

- The literature frequently mentions the drawback of diabetes prediction based on a limited set of characteristics. The predictive power of publicly accessible datasets, like the Pima Indians Diabetes Collection, may be limited since they sometimes only include a small number of variables. This emphasizes that in order to improve forecast accuracy, feature engineering or the inclusion of additional data sources are required.
- Some research exclude data that isn't full, which reduces the size of the dataset and could affect how reliable the findings are. Robust handling of missing data is necessary to guarantee the accuracy of forecasts.
- The evaluated research has not made full use of deep learning techniques like recurrent neural networks. More precise and effective diabetes prediction systems may be produced by investigating sophisticated deep learning algorithms.
- Many research restricts the validation and benchmarking of their techniques by failing to compare their suggested models with the current diabetes prediction models. Analytical comparisons may shed light on how well certain methods perform in relation to one another.

VI. Conclusion

The review article has extensively examined the use of machine learning for diabetes prediction, offering a detailed analysis of its applications. The review highlighted various machine learning algorithms, methodologies, and datasets used in previous studies, along with their contributions to the field. The integration of machine learning models into the clinical workflow has shown promising results in improving the prediction of diabetes and its related complications. These models have demonstrated their effectiveness in risk stratification, early detection, and personalized interventions, leading to better patient outcomes and management of the disease.

However, several challenges need to be addressed for the widespread adoption of machine learning-based diabetes prediction models. These include Interpretability and Explainability of the models, seamless integration into clinical workflows, ethical considerations regarding data privacy and informed consent, and ensuring generalizability and external validation across diverse populations. Future directions in this field involve the development of robust and interpretable models, exploring novel data sources and features, assessing long-term outcomes and clinical utility, and promoting personalized risk assessment and continuous monitoring.

References

[1] International Diabetes Federation https://idf.org fetched 20.09.2023

[2] Pitt, Bertram, et al. "Cardiovascular events with finerenone in kidney disease and type 2 diabetes." New England journal of medicine 385.24 (2021): 2252-2263.

[3] Jyotismita Chaki, S. Thillai Ganesh, S.K Cidham, S. Ananda Theertan, Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part B, 2022, Pages 3204-3225, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.06.013.

[4] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord. 2020 Apr 14;19(1):391-403. doi: 10.1007/s40200-020-00520-5.

[5] LaValley, Michael P. "Logistic regression." Circulation 117.18 (2008): 2395-2399

[6] Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." 2011 International conference on innovations in information technology. IEEE, 2011.

[7] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1-3, doi: 10.1109/ICECA49313.2020.9297411.

[8] El_Jerjawi, Nesreen Samer, and Samy S. Abu-Naser. "Diabetes prediction using artificial neural network." (2018).

[9] Aslan MF, Sabanci K. A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data. Diagnostics (Basel). 2023 Feb 20;13(4):796. doi: 10.3390/diagnostics13040796. PMID: 36832284; PMCID: PMC9955314.

[10] Priyanka Rajendra, Shahram Latifi, Prediction of diabetes using logistic regression and ensemble techniques, Computer Methods and Pro-

grams in Biomedicine Update, Volume 1, 2021, 100032, ISSN 2666-9900, https://doi.org/10.1016/j.cmpbup.2021.100032.

[11] W. Chen, S. Chen, H. Zhang and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2017, pp. 386-390, doi: 10.1109/ICSESS.2017.8342938.

[12] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878802.

[13] Patil, Ratna, et al. "A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus." Int. J. Electr. Comput. Eng 12.1 (2022): 524.

[14] M. T. Islam, H. R. H. Al-Absi, E. A. Ruagh and T. Alam, "DiaNet: A Deep Learning Based Architecture to Diagnose Diabetes Using Retinal Images Only," in IEEE Access, vol. 9, pp. 15686-15695, 2021, doi: 10.1109/ACCESS.2021.3052477.

[15] Farajollahi, Boshra, et al. "Diabetes diagnosis using machine learning." Frontiers in Health Informatics 10.1 (2021): 65.

[16] Rufo, Derara Duba, et al. "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)." Diagnostics 11.9 (2021): 1714.

[17] Ashima Singh, Arwinder Dhillon, Neeraj Kumar, M. Shamim Hossain, Ghulam Muhammad, and Manoj Kumar. 2021. EDiaPredict: An Ensemble-based Framework for Diabetes Prediction. ACM Trans. Multimedia Comput. Commun. Appl. 17, 2s, Article 66 (June 2021), 26 pages. https://doi.org/10.1145/3415155

[18] Taha, Altyeb Altaher, and Sharaf Jameel Malebary. "A hybrid meta-classifier of fuzzy clustering and logistic regression for diabetes prediction." Computers, Materials and Continua 71.2 (2022): 6089-105.

[19] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.

[20] Arora, Nitin, et al. "A Novel Architecture for Diabetes Patients' Prediction Using K-Means Clustering and SVM." Mathematical Problems in Engineering 2022 (2022).

[21] Jader, Rasool, and Sadegh Aminifar. "Fast and Accurate Artificial Neural Network Model for Diabetes Recogni-tion." NeuroQuantology 20.10 (2022): 2187-2196.

[22] Yadav, D.C., Pal, S. (2021). An Ensemble Approach for Classification and Prediction of Diabetes Mellitus Disease. In: Mathur, R., Gupta, C. P., Katewa, V., Jat, D. S., Yadav, N. (eds) Emerging Trends in Data Driven Computing and Communications. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore. https://doi.org/10.1007/978-981- 16-3915-9_18

[23] Kopitar, L., Kocbek, P., Cilar, L. et al. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 10, 11981 (2020). https://doi.org/10.1038/s41598-020-68771-z

[24] Tuppad, Ashwini, and Shantala Devi Patil. "Machine learning for diabetes clinical decision support: a review." Advances in Computational Intelligence 2.2 (2022): 22.

[25] Berchialla, Paola, et al. "Prediction of treatment outcome in clinical trials under a personalized medicine perspective." Scientific Reports 12.1 (2022): 4115.

[26] Sun, Y. L. & Zhang, D. L., 2019. Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey. Technical Gazette, Volume 26, p. 872– 880

[27] Qawqzeh, Y. K. et al., 2020. Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling. BioMed Research International, pp. 1-20.

[28] Sisodia, D. & Sisodia, D. S., 2018. Prediction of diabetes using classification algorithms. Procedia Computer Science, Volume 132, p. 1578–1585.

[29] Hussain, A. & Naaz, S., 2021. Prediction of diabetes mellitus: comparative study of various machine learning models. Advances in Intelligent Systems and Computing, Volume 1166, p. 103–115.

[30] Choubey, D. K. et al., 2020. Comparative analysis of classification methods with PCA and LDA for diabetes. Current Diabetes Reviews, 16(8), p. 833–850.

[31] Gujral, S., 2017. Early diabetes detection using machine learning: a review. International Journal for Innovative Research in Science & Technology, 3(10), pp. 45-60.

[32] Tigga, N. P. & Garg, S., 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, Volume 167, p. 706–716.

[33] Mercaldo, F., Nardone, V. & Santone, A., 2017. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science, Volume 112, p. 2519–2528.

[34] Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, Meshref H. Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. *International Journal of Environmental Research and Public Health*. 2022; 19(19):12378. https://doi.org/10.3390/ijerph191912378

[35] U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

[36] Whig, P., Gupta, K., Jiwani, N. et al. A novel method for diabetes classification and prediction with Pycaret. Microsyst Technol 29, 1479–1487 (2023). https://doi.org/10.1007/s00542-023-054732

DOI: 10.15439/2023R63
ISSN 2300-5963 ACSIS, Vol. 38

Classification of Plant Species with Iris Dataset Using ANN, KNN and K-Means Algorithms

M. Hanefi Calp
Department of Management Information Systems
Faculty of Economics & Administrative
Sciences, Ankara Hacı Bayram Veli University
Ankara, Turkey
hanefi.calp@hbv.edu.tr
0000-0001-7991-438X

Vijender Kumar Solanki
CMR Institute of Technology
Hyderabad, TS, India
spesinfo@yahoo.com
0000-0001-5784-1052

***Abstract*—In this study, plant species were classified on the Iris dataset using Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and K-Means algorithms. In this process, models were developed for each method, success rates were obtained, and a model with a minimum error rate was introduced. The dataset of the study was obtained from the Kaggle website. The classification process was applied repeatedly on the iris dataset, and the classification or prediction with the minimum error rate was aimed at the established models. In the study process, first of all, the dataset was obtained, prepared, and visualized. Models were created using the Jupiter Notebook editor via the Anaconda desktop GUI. Then, the models were analyzed and the most successful algorithm was selected. As a result, according to the prediction/classification models, it was seen that the most successful model was obtained with the KNN algorithm, and the most unsuccessful model was obtained with the ANN algorithm.**

***Index Terms*—iris, plant species, ann, knn, k-means, classification.**

I. Introduction

Artificial intelligence is defined as the ability of a computer or computer-assisted machine to perform tasks such as human characteristics, thinking like a human, solving problems, finding solutions, understanding, making sense, generalizing, and learning from past experiences [1].

In general, Artificial Intelligence, inspired by humans and nature and successfully solving real-world problems with mathematical and logical approaches, has increased its preference level as it can be used effectively in many areas. Problems that cannot be solved with traditional methods and techniques or that cannot be achieved at the desired level can be successfully solved by using models created with Artificial Intelligence. At this point, very successful results can be obtained for real-world problems such as prediction, diagnosis, classification, pattern recognition, optimization, and interpretation with AI. Artificial Intelligence is used in many fields such as engineering-based fields, education, economy, military, and health [2-7]. In summary, AI is the general name of technology created with completely artificial tools, which can exhibit human-like behaviors and movements. In other words, AI can fully perform human actions such as feeling, predicting behavior, and making decisions [8].

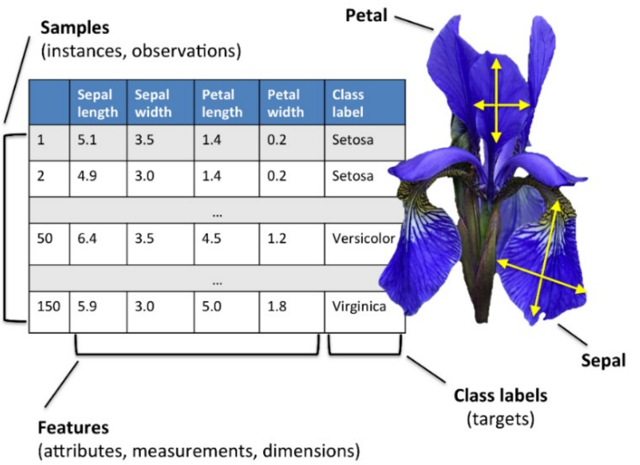
In this context, this study aims to classify plant species on the iris dataset by using ANN, KNN, and K-Means algorithms. In the second part of the study, all the details of the material and method are explained. Then, in the third part, the results and recommendations obtained from the study are given.

II. Material and Method

In the study, ANN, KNN, and K-Means algorithms were used and models were created. Python programming language was used to create the models. Matplotlib and Numpy, Sckit Learn libraries were used for visualization and calculations. All the details of the models are given in this section.

A. Dataset

The dataset used in the creation of the models belongs to “Iris”. The dataset was obtained from the Kaggle website. The Iris dataset contains information on petal length and width, sepal length and width of Iris versicolor, Iris virginica, and Iris setosa plant species. In addition, the dataset includes a total of 150 samples, 50 from each plant species. All values (minimum, maximum, and average values) of the petal and sepal of each species are given in Figure 1.


Samples
(instances, observations)
Petal
Sepal
Class labels
(targets)
Features
(attributes, measurements, dimensions)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Fig. 1. Dataset [9]

In Table 1, the example 5 rows of input data (input, output) were given. Output variables indicated the plant species. Species were coded 0 (0: Setosa), 1 (1: Versicolor), and 2 (2: Virginica).

TABLE 1. The sample data for input and output

No	Sepal lenght (cm)	Sepal width (cm)	Petal lenght (cm)	Petal width (cm)	Iris Type	
0	5.1	3.5	1.4	0.2	0	2
1	4.9	3.0	1.4	0.2	1	1
2	4.7	3.2	1.3	0.2	2	0
3	4.6	3.1	1.5	0.2	3	1
4	5.0	3.6	1.4	0.2	4	2

B. *ANN*

ANN is a technology that analyzes data by copying the working logic of the human brain and generates new information from this data with different learning algorithms. ANN provides the opportunity to make some decisions or actions using data. Technically, the task of ANN is to produce an output using the dataset entered into the model. In order to achieve this, the network is trained with sample datasets. Then, the network becomes able to make a comment or make a decision, and thus the outputs are determined [10-13]. ANN is widely used in many fields such as prediction, classification, system diagnostics, pattern recognition, robotics, and signal processing [14].

As seen in Figure 2, an artificial neural network was established with plant species, namely Setosa, Versicolor, and Virginia, in the input layer, sepal width, sepal length, petal width, and petal length in the output layer in the iris dataset in which artificial neural networks were applied.

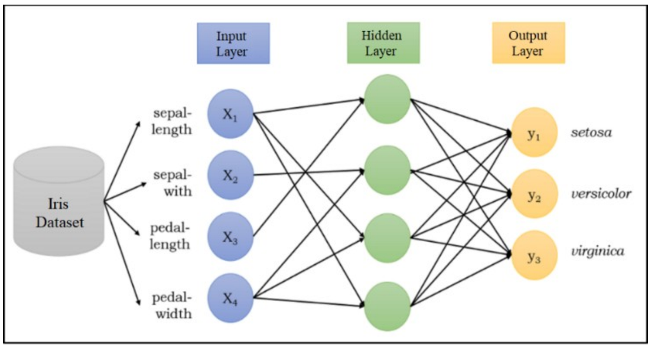


Fig. 2. Classification of the dataset

1) *Data Preprocessing Phase*

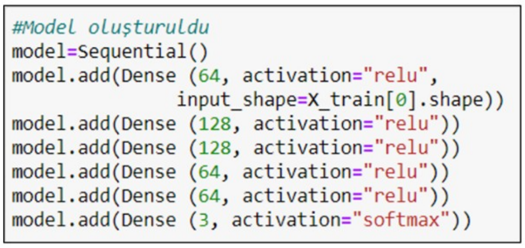
The Data Preprocessing stage is very important before the model setup. At this stage, missing data is corrected and data is normalized. The info() method in the Pandas library is used to find the missing data. In case of missing data, the average value was entered in that row or completely removed. The dataset was divided into two a 10% (15) test and a 90% (135) training set. The "Train Test Split" function is used to split the dataset (input and output values). This function is imported from the "sklearn library". The model was established with the training data, and the model was tested with the test data that the model had never seen.

2) *Normalizing of the Data*

The variance of the data was tested and the data with high variance was normalized. In the output obtained, X_train and Y_test have low variance. Therefore, there is no need for normalization.

3) *Creating of the Model*

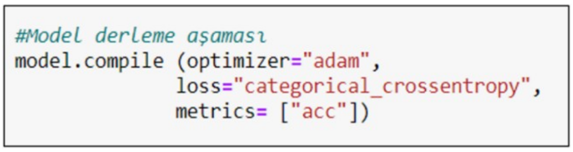
At this stage, layers were added to the model with the model.add() function, and how many neurons were printed into these layers was included in the model.add() function (Figure 3).



```
#Model oluşturuldu
model=Sequential()
model.add(Dense (64, activation="relu",
              input_shape=X_train[0].shape))
model.add(Dense (128, activation="relu"))
model.add(Dense (128, activation="relu"))
model.add(Dense (64, activation="relu"))
model.add(Dense (64, activation="relu"))
model.add(Dense (3, activation="softmax"))
```

Fig. 3. Pseudocode for the model creating

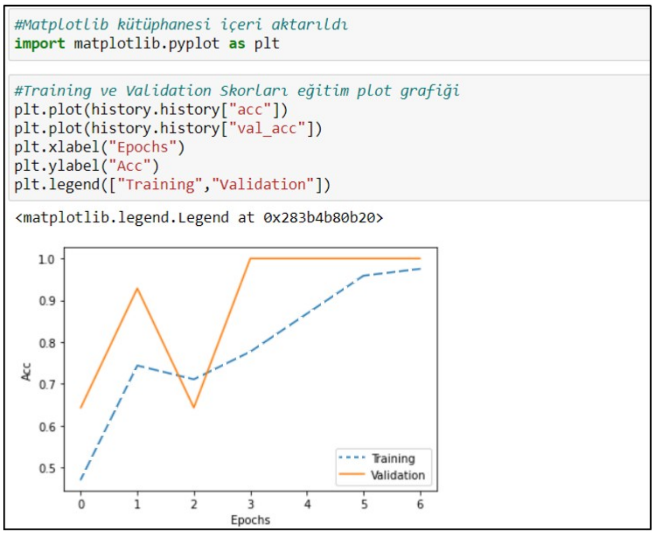
The number of epochs indicates how many times the model will be trained. As the number of epochs increases, the training time of the model also increases. 7 epoch values were entered during the model compilation phase. In the output of the code, acc value: 0.9752, val_acc: 1.000.



```
#Model derleme aşaması
model.compile (optimizer="adam",
              loss="categorical_crossentropy",
              metrics= ["acc"])
```

Fig. 4. Compilation phase of the model

In Figure 5, there is a graph showing the data on the training history. In the graph, the blue lines show the Training scores and the orange lines show the Validation score. With this table, it was concluded that the model learns more with the increase in the epoch value.



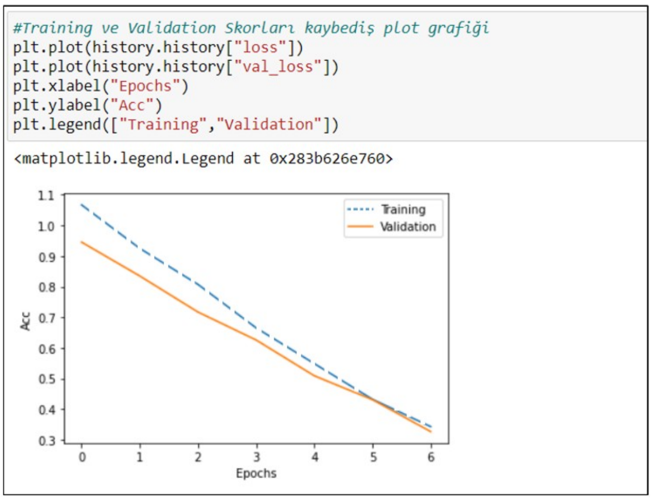
```
#Matplotlib kütüphanesi içeri aktarıldı
import matplotlib.pyplot as plt
```

```
#Training ve Validation Skorları eğitim plot grafiği
plt.plot(history.history["acc"])
plt.plot(history.history["val_acc"])
plt.xlabel("Epochs")
plt.ylabel("Acc")
plt.legend(["Training","Validation"])
```

```
<matplotlib.legend.Legend at 0x283b4b80b20>
```

Fig. 5. Training Graph of the Training and Validation Scores

In Figure 6, the graph of the loss value was drawn. After a certain value, it was seen that both values, namely Training and Validation scores, decreased. In other words, as the Epochs value increased, the model made better predictions and the error decreased.



```
#Training ve Validation Skorları kaybediş plot grafiği
plt.plot(history.history["loss"])
plt.plot(history.history["val_loss"])
plt.xlabel("Epochs")
plt.ylabel("Acc")
plt.legend(["Training","Validation"])
```

```
<matplotlib.legend.Legend at 0x283b626e760>
```

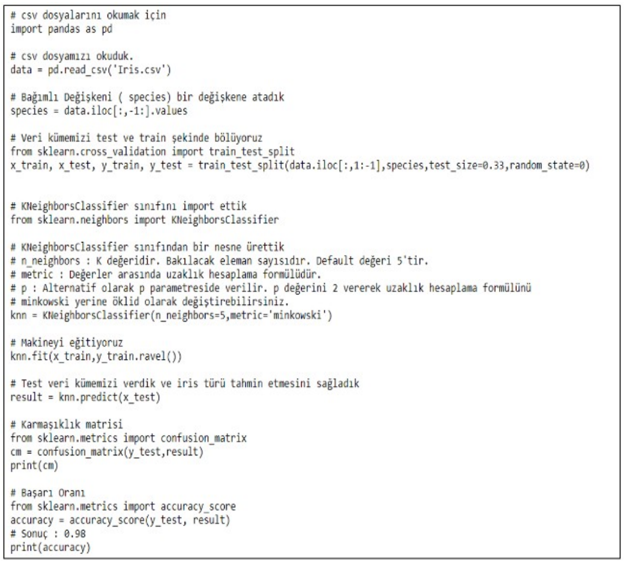
Fig. 6. Loss Graph of the Training and Validation Scores

Accuracy estimation shows the percentage value of how the model will predict test data it has never seen before. As a result, the created ANN model showed 86% success.

C. KNN

KNN is one of the machine learning algorithms that is easy to implement because it produces successful results against noisy training data. KNN is used in classification and regression problems in supervised learning and there is no training phase. KNN is considered the simplest machine learning algorithm. In the KNN algorithm, the K number determined to determine which class the object will be included in shows how many neighbors are closest to it. The distance to these neighbors is calculated with the Euclidean function. The unknown object is included in that class if it is closest to which class in the number K [15-17].

The closeness of the accuracy rate to 1 is extremely important for the accuracy of the dataset. In Figure 7, the codes developed for the creation of the KNN model and their explanations are indicated.



```
# csv dosyalarını okumak için
import pandas as pd

# csv dosyamızı okuduk.
data = pd.read_csv('Iris.csv')

# Bağımlı Değişkeni ( species) bir değişkene atadık
species = data.iloc[:,-1:].values

# Veri kümemizi test ve train şekinde bölüyoruz
from sklearn.cross_validation import train_test_split
x_train, x_test, y_train, y_test = train_test_split(data.iloc[:,1:-1],species,test_size=0.33,random_state=0)


# KNeighborsClassifier sınıfını import ettik
from sklearn.neighbors import KNeighborsClassifier

# KNeighborsClassifier sınıfından bir nesne ürettik
# n_neighbors : K değeridir. Bakılacak eleman sayısıdır. Default değeri 5'tir.
# metric : Değerler arasında uzaklık hesaplama formülüdür.
# p : Alternatif olarak p parametreside verilir. p değerini 2 vererek uzaklık hesaplama formülünü
# minkowski yerine öklid olarak değiştirebilirsiniz.
knn = KNeighborsClassifier(n_neighbors=5,metric='minkowski')

# Makineyi eğitiyoruz
knn.fit(x_train,y_train.ravel())

# Test veri kümemizi verdik ve iris türü tahmin etmesini sağladık
result = knn.predict(x_test)

# Karmaşıklık matrisi
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test,result)
print(cm)

# Başarı Oranı
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, result)
# Sonuç : 0.98
print(accuracy)
```

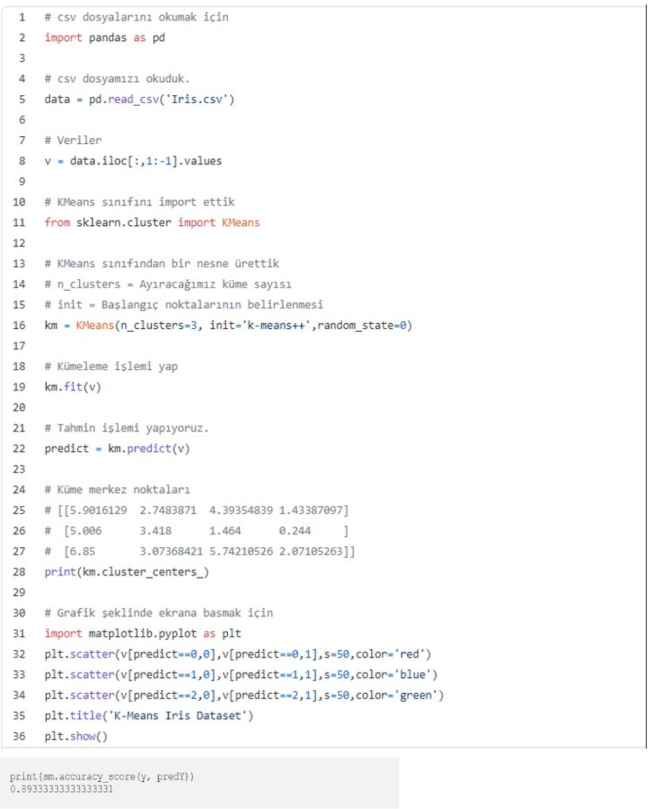
Fig. 7. KNN codes and explanations

At this stage, the Complexity matrix was obtained. According to this result, when the dataset was considered, as a result of the complexity matrix of the application, it was concluded that 48 of the 50 datasets were estimated correctly, while 1 of them was incorrectly estimated. The success rate is 48/50 = 0.96. The Accuracy Score is very close to 1 with 0.96. The closer the obtained value was to 1, the higher the accuracy or performance of the model.

D. K-Means Clustering

K-means clustering is one of the widely used and easy-to-implement clustering algorithms and uses the exploratory data analysis technique [18]. The aim is to ensure that the clusters obtained at the end of the partitioning process have maximum similarities within clusters and minimum similarities between clusters. It can cluster large-scale data quickly and effectively. "K" refers to the fixed number of clusters needed before starting the algorithm. With its iterative partitioner structure, the K-means algorithm reduces the sum of the distances of each data to the cluster it belongs to. The K-means algorithm tries to detect K clusters that will make the square error minimum [19].

The closeness of the accuracy rate to 1 is extremely important for the accuracy of the dataset. In Figure 8, the codes of the algorithm developed for the creation of the K-Means model are given.



```
# csv dosyalarını okumak için
import pandas as pd

# csv dosyamızı okuduk.
data = pd.read_csv('Iris.csv')

# Veriler
v = data.iloc[:,1:-1].values

# KMeans sınıfını import ettik
from sklearn.cluster import KMeans

# KMeans sınıfından bir nesne ürettik
# n_clusters = Ayıracağımız küme sayısı
# init = Başlangıç noktalarının belirlenmesi
km = KMeans(n_clusters=3, init='k-means++',random_state=0)

# Kümeleme işlemi yap
km.fit(v)

# Tahmin işlemi yapıyoruz.
predict = km.predict(v)

# Küme merkez noktaları
# [[5.9016129  2.7483871  4.39354839 1.43387097]
#  [5.006      3.418      1.464      0.244     ]
#  [6.85       3.07368421 5.74210526 2.07105263]]
print(km.cluster_centers_)

# Grafik şeklinde ekrana basmak için
import matplotlib.pyplot as plt
plt.scatter(v[predict==0,0],v[predict==0,1],s=50,color='red')
plt.scatter(v[predict==1,0],v[predict==1,1],s=50,color='blue')
plt.scatter(v[predict==2,0],v[predict==2,1],s=50,color='green')
plt.title('K-Means Iris Dataset')
plt.show()
```

```
print(sm.accuracy_score(y, predY))
0.8933333333333331
```

Fig. 8. Codes of the K-Means Algorithm

In Figure 9, the clustering result of the K-Means algorithm was given. In our model, which was made using K-Means and Hierarchical techniques, the k parameter, that was, the number of clusters was determined as 3. Based on the k parameter, that was, the number of clusters, out of 150 data in the dataset, it was continued as 3 clusters as Setosa, Virginica, and Versicolor. Setosa, Virginica, and Versicolor clusters each contain 33.33% of the Iris dataset. Looking at Figure 9, it was seen that the success rate/accuracy rate of the data in the dataset is 89%. The closer the obtained value was to 1, the higher the performance of the model.

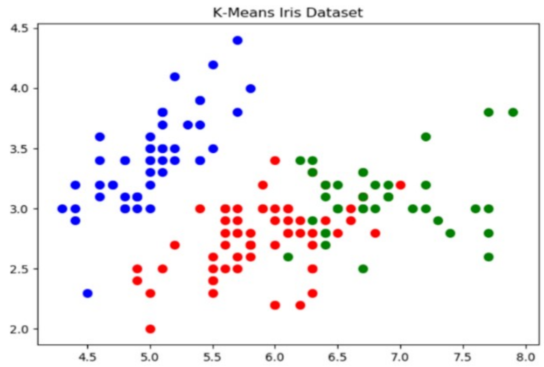


Fig. 9. Clustering result of the K-Means algorithm

In general, when the results obtained from all models were compared, the accuracy rates of 86% in the ANN method, 96% in the KNN method, and 89% in the K-Means method were reached. This situation revealed that the highest performance was obtained with the KNN method.

Finally, different tests were carried out using real data on the proposed models, and the experimental results obtained are given in Table 2.

TABLE 2. The experimental results

No	KNN	Loss (KNN)	K-Means	Loss (K-Means)	ANN	Loss (ANN)
1	97,4	2,6	89,14	10,86	87,4	12,6
2	97,3	2,7	88,25	11,75	85,6	14,4
3	96,25	3,75	89,54	10,46	87,3	12,7
4	95,35	4,65	86,2	13,8	86,5	13,5
5	97,27	2,73	92,5	7,5	90,8	9,2
6	96,14	3,86	90,7	9,3	83,2	16,8
7	96,7	3,3	91,5	8,5	88,4	11,6
8	94,8	5,2	90,8	9,2	87,2	12,8
9	97,6	2,4	89,6	10,4	81,5	18,5
10	97,35	2,65	88,2	11,8	87,6	12,4
Average	96,616	3,384	89,643	10,357	86,55	13,45

When Table 2 is examined, it confirms the success rates obtained from the models. In other words, it can be seen that the most successful results are obtained with the KNN model.

III. Conclusion and Recommendations

There is a significant increase in the amount of data day by day with the development of technology and the internet. This situation has brought about the discovery, analysis, use, and application of the available data in every field. This makes it possible to easily identify some relationships that cannot be detected by classical methods using data, and to make predictions and predictions for the future by performing big data analysis. Big data is used effectively in many current algorithms from artificial intelligence methods to image processing techniques and provides many advantages.

Therefore, in this study, big data from the iris plant was used and the flower type was estimated according to the sepal length, sepal width, petal length, and petal width values of the iris flower using artificial intelligence methods. For this purpose, ANN, KNN (K-Nearest Neighbors Algorithm), and K-Means Clustering methods were used. As a result of the methods used, the lowest success rate was obtained in the ANN method, and the highest success rate was obtained in the KNN method. As a result of the codes written, the method with the highest accuracy rate was KNN. The study is expected to be a sample application for other estimation methods. In future studies, it is planned to make comparisons using different algorithms.

References

[1] V. V., Nabiyev, Yapay Zekâ: İnsan-Bilgisayar Etkileşimi. Baski Yeri: Seçkin Yayıncılık, 2012.

[2] S.J., Russell, P., Norvig, J.F., Canny, J.M. Malik, and D.D., Edwards, “Artificial Intelligence: A Modern Approach”, 2(9): Upper Saddle River: Prentice Hall, 2003.

[3] V.V., Nabiyev, “Yapay Zeka: Problemler-Yöntemler-Algoritmalar”, Seçkin Yayıncılık, 2012.

[4] N., Allahverdi, “Uzman Sistemler: Bir Yapay Zeka Uygulaması”, Atlas Yayın Dağıtım, 2002.

[5] A., Strong, “Applications of artificial intelligence and associated technologies”, Science (ETEBMS-2016), 5(6), 2016.

[6] R., Butuner, I., Cinar, Y. S., Taspinar, R., Kursun, M. H., Calp, & M. Koklu, (2023). Classification of deep image features of lentil varieties with machine learning techniques. European Food Research and Technology, 249(5), 1303-1316.

[7] T., Savaş & S. Savaş, (2021). Tekdüzen kaynak bulucu yoluyla kimlik avı tespiti için makine öğrenmesi algoritmalarının özellik tabanlı performans karşılaştırması. Politeknik Dergisi, 1-1.

[8] S., Ipek, Yapay zekanın toplum üzerindeki etkisi ve yapay zekâ (AI) filmi bağlamında yapay zekaya bakış. Uluslararası Ders Kitapları ve Eğitim Materyalleri Dergisi, 2(2), 203-215, 2019.

[9] S. Raschka: https://sebastianraschka.com/images/blog/2015/principal_component_analysis_files/iris.png, Access Date: 02.02.23.

[10] B. Marr, What is an Artificial Neural Networks?, https://www.bernardmarr.com/default.asp?contentID=2126, Access Date: 10.02.23.

[11] F. Fahrettin, https://medium.com/@fahrettinf/4-1-1-artificial-neural-networks-6257a7a54bb3, Access Date: 21.02.23.

[12] M. F., Keskenler, & E. F., Keskenler, Geçmişten günümüze yapay sinir ağları ve tarihçesi. Takvim-i Vekayi, 5(2), 8-18, 2017.

[13] M. H., Calp, & U., Kose, Estimation of burned areas in forest fires using artificial neural networks. Ingeniería Solidaria, 16(3), 1-22, 2020.

[14] M. Ö., Efe, & O. Kaynak, (2004). Yapay sinir ağları ve uygulamaları. Boğaziçi Üniversitesi.

[15] E. Hatipoğlu, https://medium.com/@ekrem.hatipoglu/machine-learning-classification-k-nn-k-en-yak%C4%B1n-kom%C5%9Fu-part-9-6f18cd6185d, Access Date: 23.02.23.

[16] J., Han, J. Pei & M., Kamber (2011). Data Mining: Concepts and Techniques. Elsevier.

[17] D., Kilinc, E., Borandag, F., Yucalar, V., Tunali, M., Simsek, & A. Ozcift, (2016), Classification of Scientific Articles Using Text Mining with KNN Algorithm and R Language. Marmara Journal of Pure and Applied Sciences, 3, 89-94. https://doi.org/10.7240/mufbed.69674.

[18] https://medium.com/deep-learning-turkiye/k-means-algoritmas%C4%B1-b460620dd02a, Access Date: 15.02.23.

[19] Vikipedi, https://tr.wikipedia.org/wiki/K-means_k%C3%BCmeleme, Access Date: 18.02.23.

DOI: 10.15439/2023R59
ISSN 2300-5963 ACSIS, Vol. 38

An Efficient Ontology Based Drug Prescription Model

Md.Gulzar
Department of Computer Science &
Information Technology
Maulana Azad National Urdu University
Gachibowli Hyderabad, India
gulzar.md572@gmail.com

Muqeem Ahmed
Department of Computer Science &
Information Technology
Maulana Azad National Urdu University
Gachibowli Hyderabad, India
muqeem.ahmed@gmail.com

***Abstract*—Medication is a process of prescribing medicines by knowledgeable physicians. Medicines which are not prescribed when consumed may generate side effects. Some diseases require more than one drug to control the disease. If drugs are not carefully prescribed adverse reactions may happen. Meny people died due to medical errors in prescribing medicines by medical practitioners based on their experience. For avoiding all these adverse effects, we need a recommendation system or a decision support system for efficiently prescribing medicines. Many parameters need to be considered before prescribing the medicines like patient's age, medical history, side effects of drugs, possible allergies, drug-drug interactions, drug-disease interactions and drug-food interactions. Semantic web provides tools and technologies like ontologies to construct recommendation models and can retrieve data using tools like SPARQL and inference mechanisms to infer new knowledge.**

***Index Terms*—drug, medicine, prescription, ontology, interactions.**

I. Introduction

Medicine Recommendation System will give a list of drugs for patients according to their needs or according to diagnosed disease [11]. For the treatment of chronic disease or non-communicable diseases (NCDs) like Diabetes, cardiovascular diseases, and hypertension physicians prescribe more than one medicine and for some diseases like type 2 diabetes doctors may have various options to select drugs to prescript [10]. Sometimes patients can have comorbidities i.e., having more than one disease. Controlling these comorbidities is a challenging task. Medication has to be given frequently like mostly for six months or more. During treatment the patients may asked to undergo some pathological tests. Based on the results drug prescription will be given by considering various parameters. During treatment the prescribed dosage of drugs or may be the drugs can be changed based on pathological results [9]. And one more challenging thing in prescribing medicine is that every year a number of new medicines are released [12]. Hence prescribing medicines is a complex task. For efficient prescription of medicine various information is needed like medical history, drugs currently consuming, Allergies, diagnosis, laboratory results as well as personal information like age, blood group, communication details of patients and care givers [9]. Electronic Health Records (EHR) will contain information of patient. This EHR data can be used for efficient prescription of medicine. And some authors have used ontologies for patient information. Exiting medication recommendation models are MEDBOWLI, MedFinder, GaLenOWL, SemMed and ODDx.

The remaining paper is structured as follows. In section II we discussed semantic web technologies and its standards like ontologies and Knowledge bases. Parameters needed for effective prescription of medicine is discussed in section III. A brief literature survey is given in section IV. Section V contains observation and analysis of studied literature. In section VI we have proposed a knowledge base using various ontologies for effective prescription of medicine. In section VII we have concluded.

II. Semantic Web and Standards

Tim Berners-Lee coined the idea semantic web, which was named as web 3.0. His idea is to build a mechanism for data integration and sharing, where data is scattered on the web. In semantic web the web of documents is replaced web of data [6]. Links are given to data elements on the web page which also called as web of linked data along with semantics. This forms a semantic network which is used to build intelligent applications [1]. Semantic web uses various standards defined in W3C like RDF, RDFS, SPARQL, SWRL, Ontologies, Knowledge bases and Inference engines.

A. Ontology

Semantic web allows data to be integrated from different resources and allows to share. This integration of data from various sources forms an ontology [14]. Ontology consists of features of Artificial intelligence and machine learning [10]. Ontologies are used to describe the knowledge or concepts in a given domain. Ontology describes vocabulary, which is used by researchers to share information in a particular domain. Ontology contains machine interpretable definitions of concepts and relationships among concepts. An ontology consists of classes (concepts), sub classes, instances of classes, properties of a class, and rules on properties. Ontology is created using Web Ontology Language (OWL) which is XML based. The rules are written using RDF syntax using Semantic Web Rule Language (SWRL). The complete domain ontology provides complete information of the domain [3]. Information can be shared and retrieved using SPARQL query language. New information can be inferred using inference engines like JESS, Hermi T reasoner. Ontologies can also be used for various purposes like linking data, data sharing and reuse, decision support using knowledge bases, databases using XML schema and Natural language processing. Many ontologies for medical information are created like Biological Pathways Exchange

(BioPax), SNOMED-CT, GALEN ontology, Gene Ontology and Foundation Model of Anatomy (FMA). These ontologies constructed using standards like BIO Portal, OBO Foundry and ONIONS (Ontological Integration of Naïve Sources).

B. *Knowledge Base*

Ontology along with a set of instances of classes forms a knowledge base. Ontology is a part of knowledge base or knowledge base starts with an ontology. Ontologies and knowledge bases may already exist in electronic form, which can be imported and reused in other applications. Knowledge bases organizes the data as human brain organizes information. Knowledge base allows storage, analysis and reuse of knowledge such that a machine can interpret. Knowledge Base helps search engines and other content retrieval applications to retrieve text and interpret results to advanced queries and helps in decision making by retrieving efficient knowledge. There is an inference layer built on top of knowledge base which is a series of rules and statistical models for interpreting information which helps machines to derive knowledge from Knowledge bases. Knowledge bases can be constructed using graph databases. Graph databases uses W3C standards for describing data and semantics like RDF, SPARQL and SKOS.

III. Parameters to be Considered for Drug Prescription

Medication recommendation system allows physicians to prescribe medicines efficiently by considering various parameters like drug side effects, reactions and risks to patients by considering interaction of drugs with other drugs, diseases and foods. This type of considerations helps the physician to prescribe safe drugs [27].

A. *A. Drug Side Effects*

Drug side effects are known to physicians in advance and patients will be informed about the side effects which they may have to face during treatment. Side effects will resolve as the medication is continued for days, weeks or even months. For example, cetirizine medicine which is used to cure allergies and running nose can have side effects of causing drowsiness [21]. Some drugs can be used due to their side effects. For example, the anoxeric patients who have low body weight due to eating disorders when consumes mirtazapine whose one of the side effects is weight gain can improve their body weight. Generally, mirtazapine is an antidepressant [21]. Medicines are tested for side effects before they are launched into the market.

B. *B. Adverse Drug Reactions (ADR)*

ADR is harmful and unintended reactions which occurs due to medical error, misuse or abuse and usage of unlicensed medicine [19][20]. ADR identifies drugs which should be prevented from future usage or alteration of dosage and withdrawal of the drug. The drugs which cause ADR are generally anti diabetic, antibiotics, anticoagulants and cytotoxins. For example, the fatal ADR of consuming anticoagulant with non-steroidal anti-inflammatory drug (NSAID) may cause hemorrhage [19]. ADRs can be prevented by careful monitoring of previous medication histories to find any such ADRs of prescribed medicines. If it is found these drugs has to be changed to prevent future ADRs. In [7], authors have developed a machine learning and rule-based model for predicting Adverse Drug Reactions (ADR) by considering the drug labels. Authors have used Medical Directory for Regulatory Activities (MedRA) for normalization.

C. *Drug-Drug Interactions (DDI)*

DDI is complex and causes Adverse Drug Events (ADE). In some cases, DDI are intended. Unintended DDIs are very harmful and even can cause deaths [17]. Allergies due to interactions with drug compounds of previously consumed medicines existing in patient’s body with currently consumed medicines can occur. And Even drugs prescribed for controlling any disease can also interact with each other. Drugbank is an online repository which can return drug-drug interactions of at most five drugs [15]. A knowledge base with DDI included ADE can be used as decision support systems for prescribing drugs [17]. For example, the drug Abilify which is also called as aripiprazole which is used to treat the symptoms of psychotic conditions known as schizophrenia. This medicine works by changing the actions of chemical in the brain. When this medicine combinedly consumed with Ativan which is also known as lorazepam which is used to treat anxiety disorders will increase the central nervous system depression. The authors in [16] used data from mayo clinic to find drug-drug interactions for three cardiovascular diseases and verified on Drugbank.

D. *Drug-Disease Interactions*

Most of the diseases require more than one drug to cure it. According to a study people with age above 55 daily takes four medicines. Sometimes the consumed drugs may cause a new disease to born or worsen the condition of existing disease [4]. These are known as Drug Disease Interaction (DDSIs). For example, Donepezil medicine used treat Alzheimer’s disease can cause 3 other diseases and can interact with 9 other drug substances [4]. DDSIs will occur due to negative effects of consuming poorly prescribed drugs. In this case it is necessary to avoid that drug or maximize the dose [18]. These types of drugs are known as contra indications. Managing DDSIs is needed in order to overcome serious harms and to prevent deaths.

E. *Drug-Food Interactions*

The effect of drugs consumed can also depend on foods and drinks. These interactions are known as drug-food interactions [27]. This is the reason why some medicines are asked to take with empty stomach. Some foods and beverages will decrease or delay the effects of medicines. This is the reason of prescribing medicines by limiting the foods and drinks. Drug and food interactions may cause serious side effects. For example, grape juice can interact with all types of drugs [27]. It changes the chemical reactions of drugs. Dairy products and drugs like antibiotics when consumed together can also have interactions. Antibiotics can prevent the absorption of calcium and magnesium of milk products [27]. Dairy products and calcium juices can decrease the absorption of antibiotic drugs like Ciprofloxacin [28]. Consuming Ance medicine with Vitamin A may cause liver failures [1].

IV. Literature Survey

In [1], authors have developed a website to promote safe medical consumption by constructing a medication ontology with 15 categories and 73 sub categories. Authors have not considered drug-drug interaction, reactions and side effects for prescribing medicine. In [2], authors have constructed a rule-based system by considering drug-drug interactions and possible allergies a patient may have due to consumption of the drugs. In [3] authors have developed an Anti-Diabetic Drugs Ontology and Patient Data Ontology for recommending medicines to diabetic patients. Authors have used SWRL for constructing rules and JESS inferring engine for reasoning. In [4], authors have developed a semantic web enabled online system GalenOWL for discovering drug-drug and drug-disease interactions and drug recommendations. The authors have constructed an otology using ICD-10 classification for diseases, and Unique Ingredient Identifier (UNII) to identify active ingredients of the drugs, and Anatomical Therapeutic Chemical Classification (ATC) for classification of drugs. For inferring drug-drug interactions and drug-disease interactions OWLIM reasoner was used. In [5], authors have developed an ontology named as Drug Ontology (DrOn) for providing consistent drug information. Authors have used RxNorm drug terminology constructed by National Library of Medicine (NLM). RxNorm contains several names and relations extracted from knowledge bases. In [6], authors have created an ontology for translational medicine. Translational medicine consists of data from private clinical domain and public pharmaceutical domain. Translational medicines fill the gap between basic research and clinical practice. In [8], authors have developed a recommendation system for anti-diabetic drugs based on the symptoms. Authors have used domain ontologies. They have constructed Anti Diabetic Medicine Ontology to store regulations of drugs from American Association of Clinical Endocrinologists Medical Guidelines for Clinical Practice for the Management of Diabetes Mellitus (AACEMG) and Patient Test Ontology. Authors have created rules of ontology using SWRL from data of AACEMG and used JESS inference engine to retrieve drug information without side effects. In [9], authors have developed an individualized recommendation model for type 2 diabetes. Authors pointed out that one common treatment model is not sufficient to treat any disease because of varying patient profiles, life styles, strengths, weakness, goals, patient's age, diseases duration and affordable cost [10]. Hence individual plan for recommending treatment is needed. Authors have developed two ontologies for patient profiles and anti-diabetic drugs. In [12], authors have constructed ontology-based recommendation model using an Electronic Prescription writer [EPW]. EPW interviews the patient regarding various health care primitives like previous medication, and whether they have mutual negative effects or negative effects with ongoing medications. And EPW prescribe dose of the medicine and may suggest new medicine based on the cost and patient's health condition. In [13], authors have constructed a multi evidence prescription recommendation model using two ontologies namely International Classification of Diseases (ICD) for classifying conditions of different diseases and Anatomical Therapeutic Chemical (ATC) for classifying drug ingredients and functions. Authors have used EHR for patient data. Authors have also used demographic and side effect information.

Following table represents the brief summary of studied literate.

TABLE I. Literature Survey of Medication Recommendation Models

Ref	Ontologies Used	Ontologies Constructed	Technologies Used	Limitations
[1]	Not Used	Medication ontology	OWL, Protégé, SPAQRL	Drug-Disease and Drug-food interactions are not considered.
[2]	Not Used	Rule based system	OWL, SWRL, ODDx inference engine	Not considered side effects, Drug-Disease, Drug-food interactions and adverse drug reactions.
[3]	Not Used	Anti diabetic Drugs Ontology, Patient Data Ontology	SWRL, JESS inference engine, Pellete, Protégé	Does not considered Side effects, Drug-Drug, Drug-Disease, Drug-Food interactions and Adverse Drug Reactions.
[4]	ICD-9, ATC, UNII	GalenOWL	OWLIME-Lite inference engine, SPARQL.	Does not considered drug side effects, and Drug-Food interactions.
[8]	Not Used	Anti Diabetic Medicine Ontology, Patient Test Ontology	SWRL, JESS inference engine, Pellete, Protégé	Does not considered Drug -Drug, Drug-Disease and Drug-Food interactions
[10]	Not Used	Patient Profile Ontology, Anti diabetic drugs Ontology	OWL, SWRL, JESS inference engine	Does not considered Drug -Drug, Drug-Disease, Drug-Food interactions and Adverse Drug Reactions
[12]	Not Used	Not Constructed	OWL	Does not considered Drug-Drug, Drug-Disease, Drug-Food interactions, Adverse Drug Reactions, and side effects.
[13]	ICD-9, ATC, UNII	Not Constructed	Not Used	Does not considered Drug-Disease and Drug-Food interactions.

V. Observation and analysis

From the above table we can analyze that authors have used SWRL for constructing rules, SPARQL queries for retrieving information and for inferencing new knowledge various inference engines like JESS, ODDx and OWLIME-Lite are used. From the 8 articles only four authors of [1],[8],[10] and [13] considered side effects. Three authors [4],[8] and [13] have considered adverse drug reactions, authors of [2], [4] and [13] have considered Drug-Drug interactions, only one author in [4] have considered Drug-Disease interactions and none of the author has considered Drug-Food interactions as shown in below Fig. 1.

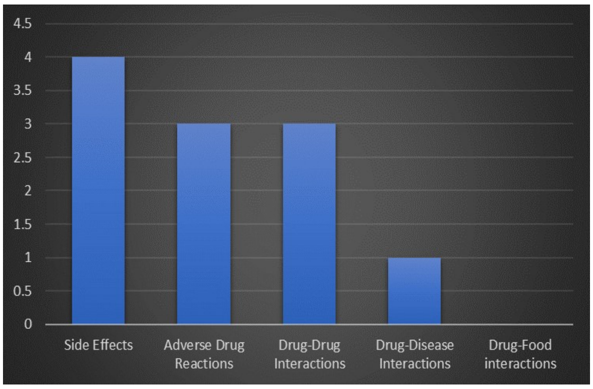


Fig. 1. Analysis of parameters considered

VI. Proposed Drug Prescription Model

Drugs can be efficiently prescribed by considering side effects, reactions, interactions of drug-Disease with drugs, disease and foods. For doing this we proposed to construct a knowledge base by considering various ontologies like Food Interactions with Drugs Evidence Ontology (FIDEO)[22], The Drug-Drug Interactions Ontology(DINTO)[23], Pharmacovigilance Ontology(PVONTO)[24], and other online materials like drugs.com for identifying side effects[25] and interaction checker at drugbank.com[26]. We can use inference engine like JESS to infer the information of drugs from constructed knowledge base. This mechanism is shown in following Fig. 2.

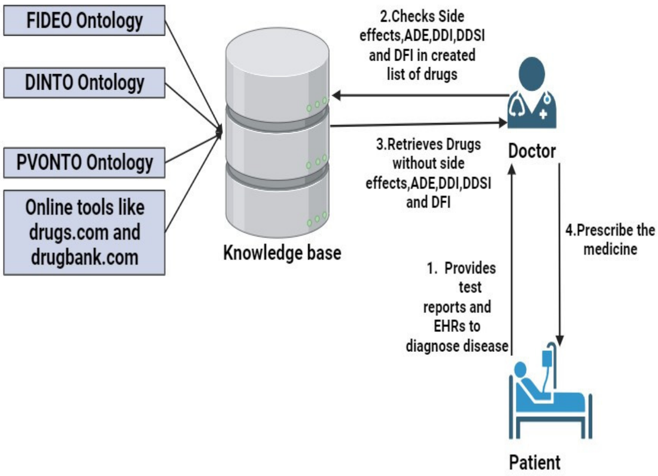

FIDEO Ontology
DINTO Ontology
PVONTO Ontology
Online tools like drugs.com and drugbank.com
Knowledge base
2.Checks Side effects,ADE,DDI,DDSI and DFI in created list of drugs
3.Retrieves Drugs without side effects,ADE,DDI,DDSI and DFI
Doctor
4.Prescribe the medicine
1. Provides test reports and EHRs to diagnose disease
Patient

Fig. 2. Construction of knowledge base

The Proposed drug prescription model tries to prescribe drugs without interactions. The model works as follows.

- The physician diagnoses the disease and prepares a tentative list of drugs to prescribe.
- The Physician infers drug side effects information and removes those drugs from tentative list.
- The physician the infers adverse drug reaction (ADR)information of available drugs. If ADR is present those drugs are replaced or removed from tentative list
- Further the physician infers drug interactions like Drug-Drug, Drug-Disease and Drug-Food interactions one after another and removes or changes such drugs if any.
- Finally, physician gets a list of drugs which are safe to prescribe.

This working method is shown in following Fig. 3.

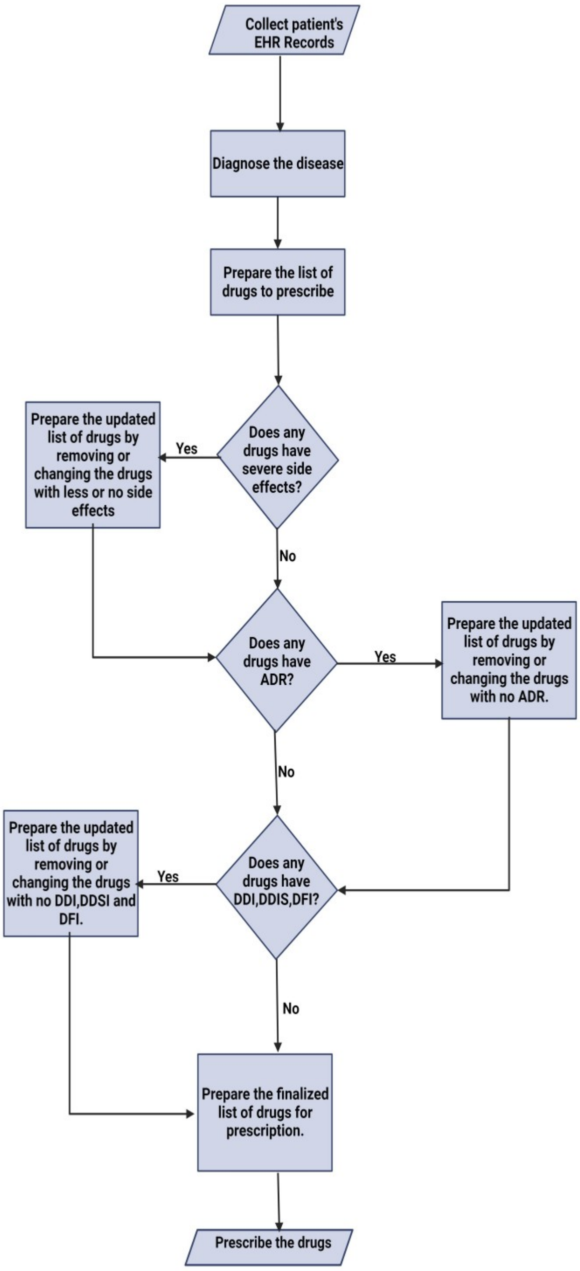

Collect patient's EHR Records
Diagnose the disease
Prepare the list of drugs to prescribe
Does any drugs have severe side effects?
Yes
Prepare the updated list of drugs by removing or changing the drugs with less or no side effects
No
Does any drugs have ADR?
Yes
Prepare the updated list of drugs by removing or changing the drugs with no ADR.
No
Does any drugs have DDI,DDIS,DFI?
Yes
Prepare the updated list of drugs by removing or changing the drugs with no DDI,DDSI and DFI.
No
Prepare the finalized list of drugs for prescription.
Prescribe the drugs

Fig. 3. Working flow of Proposed Drug Prescription model.

VII. Conclusion

In this article we have studied works of various authors in the field of drug prescription using semantic web technologies. We have identified very important parameters like Drug side effects, Adverse Drug Reactions, Drug-Drug,

Drug-Disease, Drug-Food interactions. The consideration of these parameters in prescription of drugs will help the physicians to effectively prescribe correct drugs. We also proposed a recommendation model for prescribing drugs by constructing one knowledge base by using various ontologies used in medical domain and information available on the websites. In future we try to implement proposed model by using various semantic web technologies.

References

[1] Yoosooka, Burasakorn, and Suvil Chomchaiya. "Medication recommender system." in 2015 International Conference on Science and Technology (TICST), pp. 313-317. IEEE, 2015.

[2] Rodríguez, Alejandro, Enrique Jiménez, Jesús Fernández, Martin Eccius, Juan Miguel Gómez, Giner Alor-Hernandez, Rubén Posada-Gomez, and Carlos Laufer. "Semmed: Applying semantic web to medical recommendation systems." in First International Conference on Intensive Applications and Services, pp. 47-52. IEEE, 2009.

[3] Chen, Rung-Ching, Cho-Tsan Bau, and Yun-Hou Huang. "Development of anti-diabetic drugs ontology for guideline-based clinical drugs recommend system using OWL and SWRL." in International Conference on Fuzzy Systems, pp. 1-6. IEEE, 2010.

[4] Doulaverakis, Charalampos, George Nikolaidis, Athanasios Kleontas, and Ioannis Kompatsiaris. "GalenOWL: Ontology-based drug recommendations discovery." in Journal of biomedical semantics 3, no. 1, pp 1-9.2012

[5] Hanna, Josh, Eric Joseph, Mathias Brochhausen, and William R. Hogan. "Building a drug ontology based on RxNorm and other sources." in Journal of biomedical semantics 4,pp 1-9.2013

[6] Machado, Catia M., Dietrich Rebholz-Schuhmann, Ana T. Freitas, and Francisco M. Couto. "The semantic web in translational medicine: current applications and future directions." Briefings in bioinformatics 16, no. 1,pp.89-103,2015.

[7] Hur, Junguk, Cui Tao, and Yongqun He. "A 2018 workshop: vaccine and drug ontology studies (VDOS 2018)." BMC bioinformatics 20, no. 21, pp. 1-5,2019.

[8] Chen, Rung-Ching, Yun-Hou Huang, Cho-Tsan Bau, and Shyi-Ming Chen. "A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection." Expert Systems with Applications 39, no. 4, pp. 3995-4006,2012

[9] Sae-Ang, Apichat, Sawrawit Chairat, Natchada Tansuebchueasai, Orapan Fumaneeshoat, Thammasin Ingviya, and Sitthichok Chaichulee. "Drug Recommendation from Diagnosis Codes: Classification vs. Collaborative Filtering Approaches." International Journal of Environmental Research and Public Health 20, no. 1,2022.

[10] Mahmoud, Nesma, and Heba Elbeh. "IRS-T2D: Individualize recommendation system for type2 diabetes medication based on ontology and SWRL." In Proceedings of the 10th International Conference on Informatics and Systems, pp. 203-209. 2016.

[11] Venkat, T., N. Rao, A. Unnisa, and K. Sreni. "Medicine recommendation system based on patient reviews." International journal of Scientific & Technology research 9, no. 2, pp. 3308-3312,2020.

[12] Puustjärvi, Juha, and Leena Puustjärvi. "Improving the Quality of Medication by Semantic Web Technologies." *New Developments in Artificial Intelligence and the Semantic Web*,2016.

[13] Yao, Zijun, Bin Liu, Fei Wang, Daby Sow, and Ying Li. "Ontology-aware Prescription Recommendation in Treatment Pathways Using Multi-evidence Healthcare Data." ACM Transactions on Information Systems 41, no. 4 . pp. 1-29,2023.

[14] Shobowale, K. O. "Ontology in Medicine as a Database Management System." Ontology-Based Information Retrieval for Healthcare Systems, pp. 69-90,2020

[15] Drug interaction checker https://go.drugbank.com/drug-interaction-checker

[16] Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. "Mining drug-drug interaction patterns from linked data: A case study for warfarin, clopidogrel, and simvastatin." In 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 23-30. IEEE, 2013.

[17] Noor, A., Assiri, A., Ayvaz, S., Clark, C., & Dumontier, M. "Drug-drug interaction discovery and demystification using Semantic Web technologies" in Journal of the American Medical Informatics Association, 24(3), pp. 556-564,2017.

[18] Diesveld, Maaike ME, Suzanne de Klerk, Pieter Cornu, Dorothea Strobach, Katja Taxis, and Sander D. Borgsteede. "Management of drug-disease interactions: a best practice from the Netherlands." International journal of clinical pharmacy 43, pp. 1437-1450,2021.

[19] Coleman, Jamie J., and Sarah K. Pontefract. "Adverse drug reactions." Clinical Medicine 16, no. 5,pp. 481,2015.

[20] Edwards, I. Ralph, and Jeffrey K. Aronson. "Adverse drug reactions: definitions, diagnosis, and management." The lancet 356, no. 9237 pp. 1255-1259,2000.

[21] Cetirizine Drug Side effects, https://www.drugs.com/sfx/cetirizine-side-effects.html

[22] Bordea, Georgeta, Jean Nikiema, Romain Griffier, Thierry Hamon, and Fleur Mougin. "FIDEO: food interactions with drugs evidence ontology." In 11th International Conference on Biomedical Ontologies. 2020.

[23] Herrero-Zazo, Maria, Isabel Segura-Bedmar, Janna Hastings, and Paloma Martinez. "DINTO: using OWL ontologies and SWRL rules to infer drug–drug interactions and their mechanisms." Journal of chemical information and modeling 55, pp. 1698-1707,2015.

[24] Pharmacovigilance Ontology (PVONTO), https://bioportal.bioontology.org/ontologies/PVONTO

[25] Drug side effects checker, https://www.drugs.com/sfx/

[26] Drug interaction checker, https://go.drugbank.com/drug-interaction-checker#results

[27] Bushra, Rabia, Nousheen Aslam, and Arshad Yar Khan. "Food-drug interactions." Oman medical journal 26, no. 2 ,pp. 77,2011

[28] Food interaction checker,https://go.drugbank.com/food-interaction-checker#results

DOI: 10.15439/2023R25
ISSN 2300-5963 ACSIS, Vol. 38

Machine Learning Algorithms for Retinal Image Analysis and Glaucoma Detection

Syed Akhter Hussain
Computer Science and Engineering Department
Hi-Tech Institute of Technology
Aurangabad Maharashtra India
akhter.it@gmail.com

Pratap Mohanrao Mohite
Computer Science and Engineering Department
Hi-Tech Institute of Technology
Aurangabad Maharashtra India
Pratapmohite1989@gmail.com

Sandip Eknathrao Ingle
Computer Science and Engineering Department
Hi-Tech Institute of Technology
Aurangabad Maharashtra India
seingle@gmail.com

Mohammad Waseem Ahmed Siddiqui
Computer Science and Engineering Department
Hi-Tech Institute of Technology
Aurangabad Maharashtra India
wassid09@gmail.com

Mohammed Zeeshan Raziuddin
Computer Science and Engineering Department
CSMSS College of Engineering
Aurangabad Maharashtra India
zeeeshan.shaikh@gmail.com

***Abstract*—In recent years machine learning technology are widely used in modern biomedical imaging systems to recognise and classify a wide range of human disorders.The development of a retinal image analysis system requires precise segmentation.in this work we are using fully connected conditional random filed model to overcome energy minimization problem as compare to potts model which is limited for elongated retinal structure as it takes pairwise potential which in turn low priority for vessel segmentaion .in this work parameters learned automatically by structured output support vector machine and gives structured predictions we use publically available data sets DRIVE,STARE,HRF and CHASEDB1 to train our system, after segmentation we are classifify them with the help of suport vector machine and K-nearest neghbour machine learning algorithms to get accurate results.we compare and validate our result with respect to sensitivity,specifity,precision,time complexity and F1 score performance metrics.**

***Index Terms*—Retinal image, SVM, K-NN, segmentation, FC-CRF.**

I. Introduction

Glaucoma is leading eye disease in today's world that leads to vision lost glaucoma occurs due to increase in intra ocular pressure of human eye which damages optic nerve head. It manages visual information to the brain. When intraocular pressure rises due to hypertension or a malfunction of the eye drainage system, aqueous humor flows between the cornea and lens, and vitreous humor is present in a rare part of the eye ball. Aqueous humor nourishes and removes the wastage and responsible to maintain intraocular pressure, vitreous humor holds the eyeball and maintains its shape and size. If pressure continues it damages optic nerve head, glaucoma will cause everlasting vision lost. Clinical screening had certain flaws, and according to the WHO, 285 million individuals were visually endangered, with 39 million being blind and 246 million having poor vision [1]. In India, there are approximately 11.2 million people suffering from glaucoma and other eye diseases. 6.48 million People are expected to have primary open angle glaucoma, and 2.54 million people have primary angle-closure glaucoma. Around 65% of people with low vision and 82% of people who are older than 50 years are blind [2]. 26 million people in Latin America had limited eyesight, and 3.2 million were blind [3].Similar types of data are seen throughout Europe [4]. Researchers estimated in their study [5, 6] that the rest of the globe suffers from glaucoma, diabetic retinopathy, and age-related macular degeneration. Fundus photo graphs are a type of medical imaging technique that is used to identify eye diseases. They are non-invasive and simple to execute since they are computer-aided [7].Manual analysis is a time-consuming method since ophthalmologists must perform and verify multiple fundus photographs with varying degrees of parameters, and diagnosis may change from one expert to the next based on their expertise and skills [8]. The objective of retinal image analysis is to generate qualitative data for clinical evaluation. Ophthalmologists require both qualitative and quantitative retinal vascular impressions.

II. Litrature Review

Over the last decade, there has been a lot of focus on the topic of automated segmentation of retinal blood vessels [9].all previous work on vascular segmentation is based on supervised and unsupervised approaches. Unsupervised approaches are dominated by matched filtering, vessel tracking, morphological changes, and model-based algorithms. A 2-D linear structuring element is used to generate a Gaussian intensity profile of the retinal blood vessels. Employing Gaussians and their derivatives for vessel enhancement [10]. The structuring element is rotated 8-12 times to suit the vessels in various configurations in order to extract the border of the vessel.. Because a halting condition is examined for each end pixel, this approach has a significant temporal complexity. Another vessel tracking approach is [11]. Gabor filters are used for detecting and extracting blood vessels. Because of the insertion of a significant number of erroneous edges, this approach suffers from over detection of blood vessel pixels [12].A morphology-based technique for center-line identification combines morphological modifications with curvature information and matched-filtering. Because of the center-line detection of the vessel and the subsequent vessel filling operation, this approach has a high temporal complexity and is vulnerable to false edges caused by bright area edges such as optic discs and exudates describes perceptual transformation algorithms for segmenting veins in retinal pictures with bright and red lesions [13]. Active contour models are used in another model-based vascular segmentation method suggested in [14], although it again has computational complexity issues. Additionally, neighborhood analysis and gradient-based data are used by multiscale vessel segmentation algorithms presented in [15-16] to identify the vessel pixels. All of these unsupervised techniques are either computationally demanding or sensitive to abnormalities in the retina.Pixels are divided into vessel and nonvessel groups using the supervised vessel segmentation methods. Authors in [17] presented, a 31-feature set collected using the derivatives of Gaussians for k-nearest neighbor (*k*-NN) classifier. The method described in [18] was enhanced by the use of ridge-based vessel detection. Here, the picture is divided by naming the nearest ridge member for each pixel. A 27 feature set is then calculated for each pixel and utilised by a *k*-NN classifier. The vastness of the feature sets slows down both of these procedures. Additionally, these techniques depend on training data and are vulnerable to spurious edges. Another approach described in [19] makes use of a Gabor-wavelet-extracted six-feature set and a Gaussian mixture model (GMM) classifier. This approach also depends on training data, and it takes hours to train GMM models using a mixture of 20 Gaussians. Line operators and a support vector machine (SVM) classifier with a three-feature set per pixel are used in the technique described in [20]. Due to the SVM classifiers, this approach is computationally demanding and particularly sensitive to the training data. In several applications, conditional random fields (CRFs) are widely utilised for picture segmentation [21,22,23]. To our knowledge, they have never been used to segment blood vessels in fundus images.. This is probably because the elongated structures that make up a vascular segmentation are given a low prior by the common pairwise potentials, like those in a Potts model. Due to this feature, we developed a unique blood vessel segmentation approach based on completely linked CRFs [24].

III. Method

Early glaucoma detection and classification will enable patients to receive appropriate care and assistance from their eye surgeons, thereby improving their standard of living.

A. Vascular Segmentation by CRF

Conditional Random Fields approach is a statistical modelling technique in which pixel mapping is done in graph form. In the CRFs model, each pixel is considered as a node and is connected to other nodes that form the edge according to connectivity rules [25, 26, 27], so the segmentation task in this technique is posed as an energy minimization problem. Local neighborhood based CRFs vary from Fully Connected CRFs in that earlier, 4 pixel neighborhood connectivity is followed by each node [28], but in the latter, every node is understood to be related to every other pixel in the fundus picture.

Here y = {yi} a labeled pixels of the image I in label space L = {−1, 1}, where 1 is considered as retinal blood vessels and -1 extra class. Characterization of conditional random field (I, y) is done by Gibbs distribution:

$$p(Y|I)=\frac{1}{Z(I)}\exp\left(-\sum_{c\in C_G}\varphi_C(Y_C|I)\right) \qquad (1)$$

Z(I) is the normalization constant, G is the image graph associated with I, and CG is the set of cliques., with potential ϕc [23]. Gibbs energy function can be derived from following Equation:

$$E(Y|I)=\sum_{C\in C_G}\varphi_{C(Y_c|I)} \qquad (2)$$

Energy minimization is performed for labeling that is the maximum a posteriori simply it is called MAP:

$$y^*=arg\min_{Y\in L}E(Y|I) \qquad (3)$$

Binary segmentation of the vasculature derives from the minimization of E(Y|I). To denote ϕc(Y$_c$|I), we use $\psi_c(y_c)$. Additionally, unary and pairwise potentials energy decompositions are regarded as higher order potentials [29]. Total energy is obtained by adding the unary potential and pairwise potential.

$$E(Y)=\sum_i\psi_u(y_i,x_i)+\sum_{(i,j)\epsilon C_G}\psi_p(y_i,y_j,f_i,f_j) \qquad (4)$$

k$^{(m)}$ is defined as rigid function based on arbitrary feature, linear combination of weight is defined as f $^{(m)}$, wp $^{(m)}$ and μ(y$_i$, y$_j$) defined the label compatible function. f$^{(m)}$ is traced similarity in between connected pixels determine by Gaussian kernels. It is obtained by connectivity rule applied on pixels neighbors by using conditional random filed formulation. Compatible function is defined by μ, Parameters w$_u$, wp$^{(m)}$ employed to manage the unary features' weight as well as pairwise kernels with respect to energy function in addition, and learning of bias is defined by w$_β$.

Furthermore Gridiron diagrams above define LNB-CRFs. Accordingly, each pixel is considered to be connected to its four related neighbors through an edge in this technique. According to paired potentials provided as an mth pairwise feature, the function is derived as follows:

$$K^{(m)}\left(f_i^{(m)},f_j^{(m)}\right)=\frac{\left|f_i^{(m)}-f_j^{(m)}\right|}{2^{\theta}_{(m)}} \qquad (5)$$

where (m) is a bandwidth that controls the weight of pixel feature differences. By using the mincut/max-flow strategy, the grid approach's energy consumption is minimised.

Finally the FC-CRF model is represented in graphical form, where each pixel in the image is connected to other pixels this is the highest ordered potential it is also used to trace long-range interactions between image pixels. It is advantageous in the segmentation process because it helps to improve accuracy, but it has limitations with inference. Recently authors in [30] presented a competent inference approach by taking pairwise potential and mean field approximation of CRF, which helped to perform accurate segmentations in a matter of seconds. The FC-CRF's pairwise kernels are as follows:

$$k^{(m)}\left(f_i^{(m)},f_j^{(m)}\right)=\exp\left(\frac{-\left|p_i-p_j\right|^2}{2\theta_p^2}-\frac{\left|f_i^{(m)}-f_j^{(m)}\right|^2}{2\theta_m^2}\right) \quad (6)$$

In this equation pi and pj are defined as co-ordinate vectors of pixels respectively. Kernel widths control the degree weight defined as θp and θ(m). For instance, when increase in lengthy interactions observed by θp and vice versa for local neighborhoods. In the same way, when θ(m) increases tolerance is higher with respect to mth feature and vice versa for lowering the tolerance successively.

B. *Learning of CRFs by SOSVM*

In order to learn w = (w_u, w_β, w_p), where w_u,w_β and w_p are the weights of unary potential over bais term and pairwise kernel the approach is not suitable for high dimensional features. To overcome this we are employing SVM, a supervised learning approach that enforces the 1-slack formulation in terms of margin rescaling, as recommended in [31]. S = { (s $^{(1),}$ y$^{(1)}$) , ..., (s $^{(n)}$, y$^{(n)}$) } is the training set. Here, y$^{(i)}$ is ith image, x$^{(i)}$ is unary potential feature set, f$^{(i)}$ is pairwise potential feature set, s$^{(i)}$ is training set containing both unary and pairwise features. Hence to get weights W we have to use following formulation:

$$\min_{w,\xi\geq 0}\frac{1}{2}\|w\|^2+C\xi \quad (7)$$

Put through

$$\forall\left(y^{-(1)},y^{-(n)}\right):\textstyle\sum_{i=1}^{n}\langle w,\psi\left(s^{(i)},y^{(i)}\right)-\psi\left(s^{(i)},y^{-(i)}\right)\rangle\geq \sum_{i=1}^{n}\Delta\left(y^{(i)},y^{-(i)}-\xi\right) \quad (8)$$

C is defined as regularization constant; ξ is a slack variable with respect to constraints y$^{-(i)}$, for labeled y feature map is defined as ϕ(s, y); and to measures loss of function we defined Δ(y, y¯), Δ is nothing but Hamming loss formulate as follows:

$$\Delta(y,y^-)=\textstyle\sum_i[y_i\neq y_i^-] \quad (9)$$

Contained in the brackets is Prediction of labeling with respect to segmentations gold standard, feature map is formulated as follows:

$$\varphi(s,y)=\left(\textstyle\sum_k\varphi u(x_k,y_k)\,,\sum_k y_\beta(\beta,y_k),\sum_k\sum_{j<k}\varphi_p\left(y_k,y_j,f_k,f_j\right)\right) \quad (10)$$

Φu unary feature map sum, ϕβ bias feature map, ϕp pairwise feature map.

Binary vector ϕy(y$_i$) ϵ {0,1}$^{|L|}$ as:

$$\varphi_y\left(y_i\right)=\begin{cases}(1,0)\,if\ y_i=-1\\(0,1)\,if\ y_i=1\end{cases} \quad (11)$$

Feature maps obtained individually defined as follows:

$$\varphi_u\left(x_k,y_k\right)=x_k\otimes\varphi_y\left(y_k\right) \quad (12)$$

$$\varphi_\beta\left(\beta,y_i\right)=\beta\varphi_y\left(y_i\right) \quad (13)$$

$$\forall m:\left[\varphi_p\left(y_k,y_j,f_k,f_j\right)\right]_m=\mu\left(y_i,y_j\right)k^{(m)}\left(f_i^{(m)},f_j^{(m)}\right) \quad (14)$$

⊗ Is kronecker product, hence using cutting plane technique is applied on equation (7) as author presented this efficient technique [30-31].

C. *Features for Segmentation and Classification*

We employed a different strategy for feature extraction for blood vessel segmentation in retinal vessels [32]. Demonstrated multiscale line detectors with regard to 2D Gabor wavelets applied to unary potentials, as in [33] vessels are amplified in fundus pictures for paired potentials as indicated in [34]. Similarly, authors in [35] discuss many features. Authors in [36,37] provide a systematic explanation of the feature extraction procedure, which is based on a grey level scale. Furthermore, authors recommended range of view for selected characteristics to minimise erroneous detection. As a result, the FOV mask is in charge of removing erroneous detection

D. *Scaling Retinal Images to Various Resolution*

We adjust the weights of unary features as well as pairwise kernels for effective characterization of retinal vasculature, as we know that retinal structure has low dense potentials, so obtained features are more sensitive during calibration with respect to retinal image pixel. The 2D Gabor wavelet is used for scaling. Similarly, Authors in [32] present a Line detection algorithm, and authors in [34] presents enhancing strategy is proportional to the linear structured component because poor resolution these parameters are set for DRIVE dataset [18]. Because the method is not proportionately scaled, performance will suffer if these settings are used to high quality Images The benefit of these characteristics is the shift in orientation caused by Change in angle has no effect on the pixel resolution of the retinal fundus picture the same performance may be predicted for preprocessing of feature parameters such as the measurement of the median filter for background estimation, or range of the opening pretend by boundary development. For this parameterθ_p, which is used in pairwise potential connections with qualifying distance of each image pixel, FC-CRF is impacted by pixel image resolution considered with pairwise potential.

E. *Machine Learning Algorithms for Classification*

In this presented work we used SVM and K-NN classification algorithms for glaucoma detection.

1) SVM

In our work, we can define support vector machines as support vector classifiers. Kernels (similarity quantifiers) are used to broaden the feature space in this case. The classification and regression analysis are done using the supervised learning approach and the analysed data. In our scenario, non-linear classification is required to deal with large dimensions. We tweak the following settings for this goal.

- Kernel parameter:-determines whether the separation is linear or non-linear.
- Regularization parameter:- This parameter is in responsible for SVM optimisation during the training phase, calculating the number of misclassifying avoided spots.

- Gamma parameter:-defines the influence in the training phase that is low (far) or high (near).
- Line separation for high class points using the margin parameter.

2) k-NN

It is a supervised classification strategy that trains the closest feature space and separates datasets into two sets. Each row has k closest training sets of pixels resolution, and categorization is done using a majority of votes. *K*-NN works by calculating distances between training and testing data vectors using the Euclidian distance formulation the number K denotes related identified neighbors. When k=1, we name it the nearest neighbor algorithm since we acquire the closest training samples.

IV. Materials and Assesment

Evolution and validation of our work is explained in this section.

A. Datasets

We have used different types of data set for training purpose mentioned following.

TABLE I. Datasets

Dataset	Capturing Angle	Resolution
DRIVE[18]	45◦ FOV	565 × 584 with 8 bits per color channel
CHASEDB1 [38]	35◦ FOV	700 × 605 pixels with 8 bits per color channel
STARE [10]	30◦ FOV	1280 × 960 pixels with 8 bits per color channel
HRF [39, 40]	60◦ FOV	resolution 3304 × 2336 pixels

B. Gold Standard Metric for Evolution

We compared our segmentation results to the gold standard labeling available for datasets. Quality of findings based on seven specific dimensions in terms of true positive, true negative, false positive, and false negative, which are described as TP,TN,FP, and FN concurrently and taking into account pixels existing inside field of view.

$$S_e = \frac{TP}{TP+FN} \qquad (15)$$

$$S_p = \frac{TN}{TN+FP} \qquad (16)$$

$$P_r = \frac{TP}{TP+FP} \qquad (17)$$

$$Tx = (x\text{-}1) \qquad (18)$$

$$F_1 = \frac{2.P_r.R_e}{P_r+R_e} \qquad (19)$$

V. Results

In this section, we present a detailed study of our automated segmentation system for the detection of various eye-related diseases, as shown in Figure 1. We perform segmentation on the DRIVE dataset, and moving forward, we apply it to the CHASEDB1 dataset, as depicted in Figure 2.The segmentation results for the STARE dataset are presented in Figure 3, and for the HRF dataset, segmentation results are shown in Figure 4.In the initial stage, we conducted a comprehensive study of different eye diseases, understanding their occurrence and their impact on human eyes. Afterward, we calibrated them using image processing techniques, collecting all the necessary parameters for our work, which are summarized in TABLE II and TABLE III here we observe that K-NN Classification algorithm works more effectively then SVM Classification algorithm as shown in Figure 5 and Figure 6 respectively .For retinal vessel segmentation, we employed a fully connected conditional random field model, specifically chosen due to the dense and elongated structure of retinal vessels, where unary and pairwise potentials are crucial. We then classified the segmented vessels using supervised learning techniques, employing both Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN) classification algorithms.

A. Segmentation results on DRIVE data set.

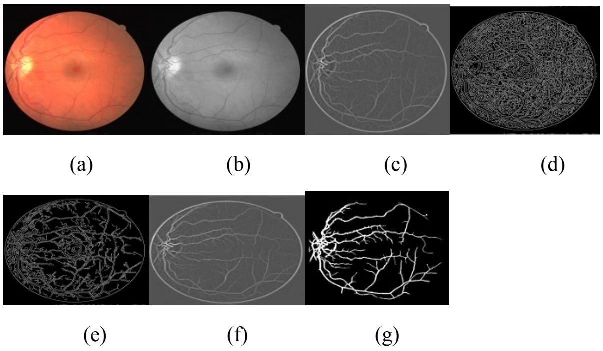


Figure 1. Segmentation Results on DRIVE (a) Input image (b) gray scale image (c) Eigen enhanced image (d) Wavelet enhanced image output (e) local enhanced image output (f) Background normalization (g) Vessel segmentation result.

B. Segmentation results on CHASEDB1 dataset

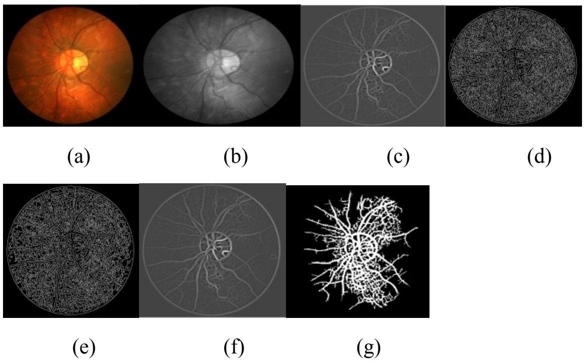


Figure 2. Ssegementation Results on CHASEDB1(a) Input image (b) gray scale image (c) Eigen enhanced image (d) Wavelet enhanced image output (e) local enhanced image output (f) Background normalization (g) Vessel segmentation result.

C. Segmentation results on STARE dataset

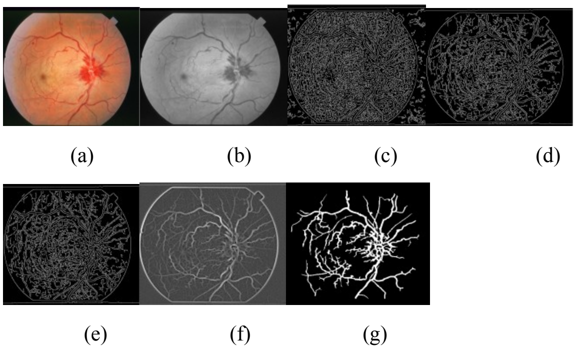


Figure 3. Segementation Results on STARE (a) Input image (b) gray scale image (c) Eigen enhanced image (d) Wavelet enhanced image output (e) local enhanced image output (f) Background normalization (g) Vessel segmentation result.

D. Segmentation results on HRF dataset

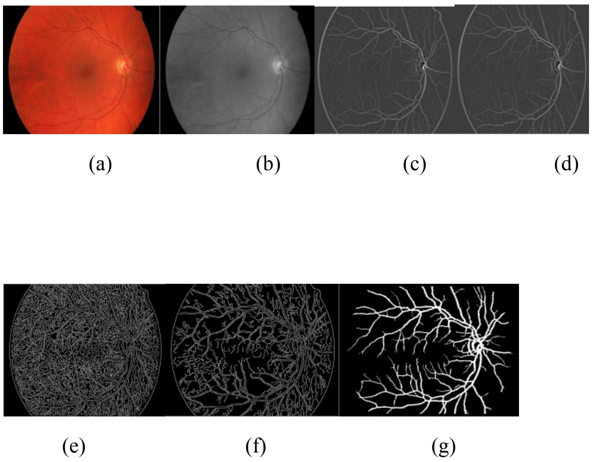


Figure 4. Segementation Results on HRF (a) Input image (b) gray scale image (c) Eigen enhanced image (d) Wavelet enhanced image output (e) local enhanced image output (f) Background normalization (g) Vessel segmentation result.

TABLE II. Average Performance Analysis of *K*-NN Algoritham

Data Set	S_e	S_p	P_r	F1	T_x (sec)
DRIVE	.9334	.9049	.8512	.8903	55
CHASEDB1	.9334	.9063	.8518	.8907	52
STARE	.9334	.9063	.8560	.8929	54
HRF	.9334	.9045	.8544	.8921	52

TABLE III. Average Performance Analysis of SVM Algoritham

Data Set	S_e	S_p	P_r	F1	T_x (sec)
DRIVE	.4600	.7818	.8000	.5841	203
CHASEDB1	.4534	.9063	.7809	.5737	194
STARE	.4667	.7818	.8000	.5894	75
HRF	.4800	.7818	.8000	.6000	75

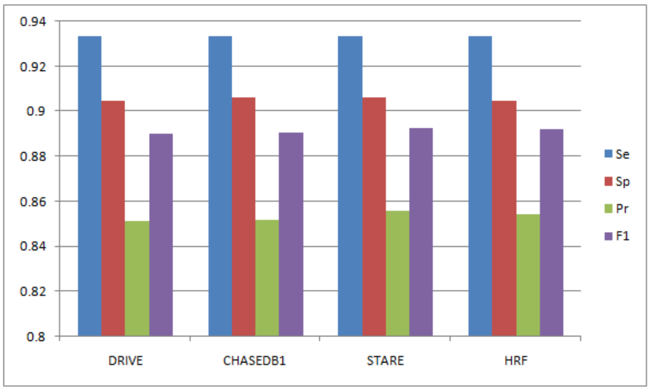


Figure 5. Performance graph of *K*-NN Algoritham

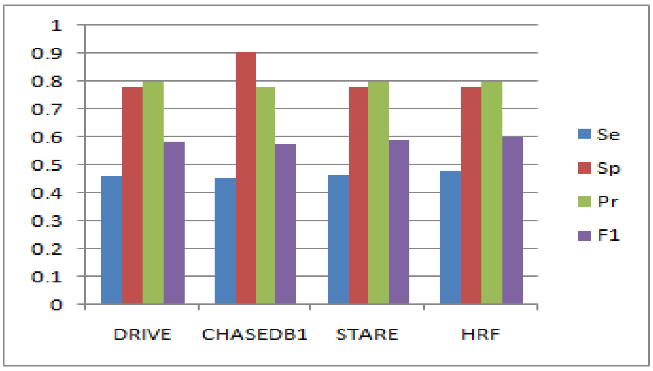


Figure 6. Performance graph of SVM Algoritham

VI. Conclusion

This paper presents the retinal Image analysis and comprehensive machine learning algorithm for segmentation and detection of glaucoma using fully connected random filed model, feature extraction and retinal vasculature reconstruction is more effectively obtained than using unary potential or a local neighborhood based conditional random filed. The efficiency is assessed in terms of sensitivity, specificity, and precision. In this presented work *K*-NN algorithm is worked superior then SVM algorithm. Our retinal vessel segmentation results on DRIVE, STARE, CHASEDB1 and HRF shows expressive performance on dense potentials. Further they can be used for numerous biomedical and biological applications for identification and detection of various problems.

Acknowledgment

I would like to express my heartfelt appreciation to my research supervisor, Dr. Ratnadeep R Deshmukh, Professor, Department of Computer Science and IT. Dr. Babasaheb Ambedkar of Marathwada University, Aurangabad, for his ongoing support for research.

References

[1] Sixty-Sixth World Health Assembly "Towards universal eye health: a global action plan 2014–2019" of WHO https://www.emro.who.int/.

[2] R.George,Ramesh S.ve,Lingam Vijaya "Glaucoma in India:estimetaed burden of disease" in Journal of Glaucoma august 2010 volume 19 issue 6 p 391-397.

[3] C.F.Etienne "Reducing avoidable blindness and visual impairment in the region of the America"in Pan American Journal of Public Health,37(1):1-3,2015.

[4] E.Prokofyeva and Eberhartz zerenner "Epidemology of major eye diseases leading to blindness in Europe: A literature review" in Opthalmic Research,47(4):171-188,2012.

[5] Jannet Leasher,Rupert Bourne,Seth R flaxman and Jos B Jonas"Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 1990-2010" in Diabetes care 39(9):1643-1649.

[6] Y.C.Than et al. "Global prevalence of glaucoma and projection of glaucoma burden through 2040: a systematic review and meta-analysis" in Ophthalmology, 121 (11):2081-2090, 2014.

[7] M.D.Abramoff,Mona K Garvin and Milan Sonka"Retinal imaging and image analysis" in IEEE Reviews Biomedical Engineering, 3:169-208, 2010.

[8] M.D.Abramoff and Nemijer"Mass screening of Diabetic retinopathy using automated methods" in Springer pp 41-50 September, 2015.

[9] M. Fraz et al."Blood vessel segmentation methodologies in retinal images a survey," in Computer. Methods and Programs Biomedicine., vol. 108, no. 1, pp. 407–433, 2012.

[10] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Trans. Med. Imag., vol. 19, no. 3, pp. 203–210, Mar. 2000.

[11] R. Rangayyan, F. Oloumi, F. Oloumi, P. Eshghzadeh-Zanjani, and F. Ayres, "Detection of blood vessels in the retina using gabor filters," in Proc. Canadian Conf. Electr. Comput. Eng., 2007, pp. 717–720.

[12] A. Mendonca, and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," in IEEE Trans. Med. Imag., vol. 25, no. 9, pp. 1200–1213, Sep. 2006.

[13] Benson S.Y. Lam, Yongasheng Gao and Alan Wee-Chung Liew "General retinal vessel segmentation using regularization-based multiconcavity modeling," IEEE Trans. Med. Imag., vol. 29, no. 7, pp. 1369–1381, Jul. 2010.

[14] B. Al-Diri, A. Hunter, and D. Steel, "An active contour model for segmenting and measuring retinal vessels," in IEEE Trans. Med. Imag., vol. 28, no. 9, pp. 1488–1497, Sep. 2009.

[15] A. Budai, G. Michelson, and J. Hornegger, "Multiscale blood vessel segmentation in retinal fundus images," in Proc. Bildverarbeitung fr die Med., pp. 261–265, Mar. 2010

[16] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," in Int. J. Biomed. Imag., article 154860, 2013.

[17] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in Proc. SPIE,vol. 5370, pp. 648–656, 2004.

[18] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," in IEEE Trans. Med. Imag., vol. 23, no. 4, pp. 501–509, Apr. 2004.

[19] J. Soares, J. Leandro, R. Cesar, H. Jelinek, and M. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification" in IEEE Trans. Med. Imag., vol. 25, no. 9, pp. 1214–1222, 2006.

[20] E. Ricci, and R. Perfetti, "Retinal blood vessel segmentation using line operators and support vector classification," in IEEE Trans. Med. Image., vol. 26, no. 10, pp. 1357–1365, Oct. 2007.

[21] Xuming He, Richard Zemel and M.A Carriera perpinan "Multiscale conditional random fields for image labeling," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2. IEEE, 2004, pp. II–695.

[22] S. Kumar and M. Hebert, "Discriminative random fields," in Int. J. Comput. Vision, vol. 68, no. 2, pp. 179–201, Jun. 2006.

[23] S. Z. Li, "Markov Random Field Modeling in Image Analysis" in 3rd ed. Springer, 2009.

[24] J. I. Orlando and M. Blaschko, "Learning fully-connected CRFs for blood vessel segmentation in retinal images," in MICCAI 2014.

[25] P. Kr¨ahenb¨uhl and V. Koltun. "Efficient inference in fully connected CRFs with Gaussian edge potentials" in Advances in Neural Information Processing Systems, pp. 109–117. 2012.

[26] S. Kumar and M. Hebert. "Discriminative random fields" in International Journal of Computer Vision, 68(2):179–201, 2006. ISSN 0920-5691.

[27] J.Lafferty,A Macculam and F.Preira"Conditional random fields: Probabilistic models for segmenting and labelling sequence data"in Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann Publishers Inc., 2001.

[28] Y. Boykov and V. Kolmogorov "An experimental comparison of min-cut/maxflow algorithms for energy minimization in vision"in IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9):1124–1137, 2004.

[29] N. Komodakis,Nikos Paragois and Gergios Tizirtas"MRF energy minimization and beyond via dual decomposition" in IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(3):531–552, 2011.

[30] Philiip Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected ¨ CRFs with Gaussian edge potentials," in Advances in Neural Information Processing Systems, 2012, pp. 109–117.

[31] Jose Ignacio Orlando*, Elena Prokofyeva, and Matthew B. Blaschko" A Discriminatively Trained Fully Connected Conditional Random Field Model for Blood Vessel Segmentation in Fundus Images"in IEEE Transactions On Biomedical Engineering, Vol. X, No. X, Month 2015.

[32] U. T. Nguyen,Alauddin B.,Laurence A.F Park and Kotagiri R"An effective retinal blood vessel segmentation method using multi-scale line detection" in Pattern Recognition, 46(3):703–715, 2013.

[33] J. V. Soares,J.J.G Leandro,R.M.Cesar,H.F.Jelinek and M.J Cree "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification" in IEEE Transactions on Medical Imaging, 25(9), 2006.

[34] F. Zana and J.-C. Klein."Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation" in IEEE Transactions on Image Processing, 10(7):1010–1019, 2001.

[35] J. I. Orlando and M. del Fresno "Reviewing preprocessing and feature extraction techniques for retinal blood vessel segmentation in fundus images" in Mec´anica Computacional, XXXIII (42):2729–2743, 2014.

[36] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov"Trainable COSFIRE filters for vessel delineation with application to retinal images" in Medical Image Analysis, 19(1):46–57, 2015.

[37] D. Mar´ın, A. Aquino, M. E. Geg´undez-Arias, and J. M. Bravo"A new supervised method for blood vessel segmentation in retinal images by using graylevel and moment invariants-based features"in IEEE Transactions on Medical Imaging, 30(1):146–158, 2011.

[38] M. M. Fraz et al., "An ensemble classification-based approach applied to retinal blood vessel segmentation," Biomedical Engineering, IEEE Transactions on, vol. 59, no. 9, pp. 2538–2548, 2012

[39] J. Odstrcilik et al., "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database" in IET Image Processing, vol. 7, no. 4, pp. 373–383, 2013.

[40] J. Odstrcilik,J.Jan,J.Gazárek and R.Kolář "Improvement of vessel segmentation by matched filtering in colour retinal images," in World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany. Springer, 2009, pp. 327–330.

DOI: 10.15439/2023R51
ISSN 2300-5963 ACSIS, Vol. 38

A Review on Software Engineering: Perspective of Emerging Technologies & Challenges

Kulashekar Inkollu
University College of Engineering and Technology, Krishna University
Rudravaram, Machilipatnam, India
kulashekarinkollu@gmail.com

Sai Kiran Gorle
University College of Engineering and Technology, Krishna University
Rudravaram, Machilipatnam, India
saikirangorle909@gmail.com

Sai Ram Kondabattula
University College of Engineering and Technology, Krishna University
Rudravaram, Machilipatnam, India
kondabattulasairam@gmail.com

Pagalla Bhavani Shankar
University College of Engineering and Technology, Krishna University
Rudravaram, Machilipatnam, India
ORCID - 0000-0001-5935-7758

M. Babu Reddy
University College of Arts & Sciences
Krishna University
Rudravaram, Machilipatnam, India
m_babureddy@yahoo.com

***Abstract*—Software Engineering is constantly evolving to meet the demands of emerging technologies. In this paper, we explore the challenges and perspectives of software engineering in the context of emerging technologies like blockchain, cloud computing, deep learning, game development. This paper discusses these challenges and provides insights into the software engineering practices that can adapt to meet the demands of these rapidly evolving fields. Each of these domains presents unique challenges and opportunities for software developers, necessitating adaptive approaches to software engineering. This abstract offers a comprehensive overview of the key challenges inherent to each technology and explores the evolving perspectives, methodologies, and best practices essential to tackle them effectively. Emphasizing the interplay between these technologies and the demand for cross-disciplinary collaboration, this paper serves as a valuable resource for altogether varieties of participants, offering enhanced insight into the challenges encountered by software engineers in the realm of emerging technologies.**

***Index Terms*—Software Engineering, blockchain, cloud computing, deep learning, game development.**

I. Introduction

The rapid advancement of technology has led to the emergence of several disruptive and transformative domains, each posing distinct challenges to software engineering. By comprehending these challenges, developers can better navigate the complexities of these domains and deliver successful software solutions.

Deep Learning (DL), a subset of artificial intelligence, has witnessed remarkable progress in areas like CV (Computer Vision) and NLP(Natural Language Processing). Yet, the development and deployment of deep learning models are fraught with challenges related to data quality, model complexity, and ethical considerations. Prearrange of the current developments in ML, we are also keen-sighted the trades starting to increasingly take benefit of the cited practices, expressly in the large technology firms such as Google, Apple, and Facebook. Google had spread over DL techniques to the enormous volumes of data collected in services such as the Google Translator, Android's voice recognition & depiction, Google's Street View, and their Search service [10]. Apple's virtual personal assistant Siri offers a variety of services such as weather reports, update of sports news, and generic question-answering by utilizing techniques such as DL [11].

Blockchain technology has disrupted traditional paradigms of trust and decentralized systems. However, building secure, scalable, and interoperable blockchain applications remains a formidable challenge for software engineers. In the ancient years, a lot of kindness has been paid to the incipient concepts of blockchain and smart contract. Some spectators are even talking of the dawn of a new era [5] and about the likely of reforming the contemporary financial services, technical infrastructure [6,7]. Ever in the meanwhile digital currencies started to represent a real monetary value, also hacks and attacks started. The key was the MtGox attack and another amazing exploit was that sustained by the DAO organization in June 2016. Concerning software development, the scenario is that of a sort of competition first-come-first-serve (FCFS) which does not pledge neither software quality, nor that all fundamental perceptions of software engineering are taken into the predefined account and liable to the justification.

Cloud computing, on the other hand, has revolutionized the way businesses deploy and manage IT resources. But with its widespread adoption comes the complex task of ensuring data security, efficient resource management, and seamless scalability in the cloud environment. For eras, when officialdoms needed to upsurge their computer systems' data and the capacity of computation, the organization faced a choice between purchasing additional hardware or improving the efficiency of their IT operations. Cloud computing offers a distinct alternative by providing resources to organizations without the need for them to worry about maintaining computing resources [8]. The field of cloud computing engineering disciplines that pertain to cloud computing . within this domain, a systematic approach is adopted to tackle standardization, commercialization, and governance concerns [9].

Game development, a thriving industry in the digital era, pushes the boundaries of software engineering with demands for realism, performance optimization, and multiplayer networking. These challenges require innovative solutions to craft engaging and immersive gaming experiences. The gaming industry requires effective engineering practices that can accommodate its diverse characteristics, such as multimedia asset management and captivating gameplay. The video game industry is faced with the challenge of adapting its software engineering methods to keep up with the escalating complexity of games and the heightened expectations of consumers. There are many ways for game developers to improve their processes. Despite the inherent difficulty and flexibility of established software engineering processes, their application to video game development holds promise for improved project management and risk mitigation. This study examines current game development practices to identify specific challenges and the corresponding SE principles that can help developers overcome them.

In this exploration of software engineering challenges, we delve into the intricacies of each domain, dissecting the hurdles that software engineers encounter as they strive to create robust and innovative solutions. By understanding these challenges and developing strategies to address them, software engineers can navigate the complexities of these cutting-edge fields and pave the way for the next generation of technology solutions.

II. Literature Review

In the realm of software engineering, various challenges and evolving paradigms have emerged, necessitating a comprehensive understanding of the field's dynamics. This literature review synthesizes key insights from a selection of relevant papers to shed light on the software engineering challenges associated with emerging technologies and paradigms.

Deep Learning and Software Engineering (Arpteg, et al., 2018 [1]): In recent years, deep learning, a branch of machine learning, has seen a remarkable increase in popularity. Arpteg et al. (2018) highlight the unique challenges posed by integrating deep learning into software engineering [14,15]. Their work underscores the importance of adaptability and specialized knowledge to effectively incorporate deep learning techniques into software development processes.

Blockchain-Oriented Software Engineering (Porru, et al., 2017 [2]): Blockchain technology [16] has garnered considerable attention, especially in the context of software engineering. Porru et al. (2017) investigate the challenges and new directions in blockchain-oriented software engineering. They emphasize the necessity for innovative development methodologies and tools tailored to blockchain applications, addressing issues of security, scalability, and consensus mechanisms.

Cloud Environment Challenges (Kashfi, 2017 [3]): Kashfi (2017) delves into the software engineering challenges within the cloud environment. From a software development lifecycle view, the paper identifies complexities associated with cloud adoption, such as managing scalability, data privacy, and service orchestration. Understanding these challenges is vital for efficient cloud-based application development.

Game Development Challenges (Kanode & Haddad, 2009 [4]): Kanode and Haddad (2009) explore software engineering challenges in the context of game development. This niche domain presents unique challenges, including real-time rendering, physics simulations, and content creation. The paper underscores the necessity for specialized SE(Software Engineering) practices to ensure the successful development of complex games.

Blockchain in Finance (Swan, 2015 [1]; Unicredit, 2016 [6]; Aymerich et al., 2009 [7]): Blockchain technology's impact on the financial sector has been studied extensively. Swan (2015), Unicredit (2016), and Aymerich et al. (2009) offer insights into blockchain's financial applications [16]. They emphasize the need for security, scalability, and regulatory compliance in blockchain-based financial systems.

Cloud Computing in Software Engineering (Grundy et al., 2012 [8]; Shan, 2011 [9]): Grundy et al. (2012) discuss the implications of cloud computing on software engineering, highlighting the importance of adapting software engineering practices for cloud environments. Shan (2011) presents the concept of "Smart Cloud Engineering" and its significance in achieving optimal cloud-based solutions.

Deep Learning and its Practical Application (Jones, 2014 [10]; Efrati, 2013 [7]): Deep learning's practical application, as discussed by Jones (2014) and Efrati (2013), demonstrates the real-world relevance of deep learning techniques. Apple's use of deep learning showcases its potential for enhancing software applications, thereby contributing to the broader field of software engineering.

Software Development Life Cycle Models (Bhuvaneswari & Prabaharan, 2013 [13]): Bhuvaneswari and Prabaharan (2013) provide a comprehensive survey of software development life cycle models. Understanding various SDLC models is crucial for software engineers to select and adapt methodologies that suit specific project requirements.

Software Engineering Body of Knowledge (Swebok) (Abran et al., 2004 [8]): Swebok, as presented by Abran et al. (2004), serves as a guide to the software engineering body of knowledge. It offers a structured framework to understand the core principles and concepts underpinning software engineering.

In conclusion, the reviewed literature demonstrates the evolving landscape of software engineering [14], shaped by emerging technologies like deep learning, blockchain [16], cloud computing, and specific application domains such as finance and gaming. These insights will inform the development of comprehensive and adaptive software engineering practices in the face of evolving challenges and opportunities.

III. The Software Development Life Cycle

The primary aim of software engineering [14] is to establish models and processes that enable the production of the software with comprehensive documentation and effortless maintainability. A software life cycle is a series of identifiable stages that a software product undergoes during its development. Within the realm of software engineering, there are multiple software development lifecycle models[13]. Various levels of the lifecycle are:

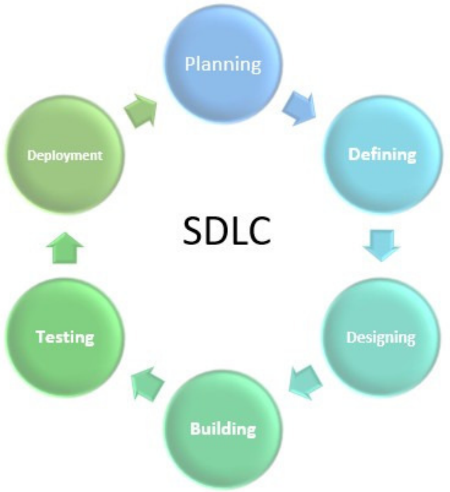


Fig. 1. The Software Development Life Cycle (SDLC)

A. *Planning*

This is a primary level plays a vital role in the life cycle. Planningleads to determine the requirements gathering of a predefined or defined project. It undergone by the experts of the fellow members of the project. It gives the well-defined pre planned guidelines to the fellow members in a specified manner.

B. *Defining*

Once the planning and requirement analysis have been completed, the subsequent phase involves precisely outlining and documenting the product requirements, seeking approval from either the customer or the market analysts. Throughout the project lifecycle, a comprehensive software requirement specification(SRS) document is utilized to ouline and define all the product requirements that will be designed and developed.

C. *Designing phase*

In this phase the design of the software is created. Based on the requirements specified in SRS document the team will develop the design for the software

D. *Building*

Also known as implementation in this phase the design is implemented in code.It is essential that developers abide by the coding standards and guidelines outlined by their organization.

E. *Testing*

This phase includes the testing of the software thoroughly for errors and bugs and to ensure that it meets the requirements and functions correctly.

F. *Deployment*

After successful testing, the software is released into the market and made available to end-users.post-launch maintenance is essential for products once they hit the market.

IV. Emerging Domains

Software Engineering is a well bound creator in the all fields of the engineering aspects. It vary in the different domains as per the expertise based on the challenges and based on the consumption of the engineering process. Deep Learning, Blockchain, Cloud Computing and Game development are the emerging technologies were successfully adopted the software engineering techniques in as per the needs of domain knowledge.

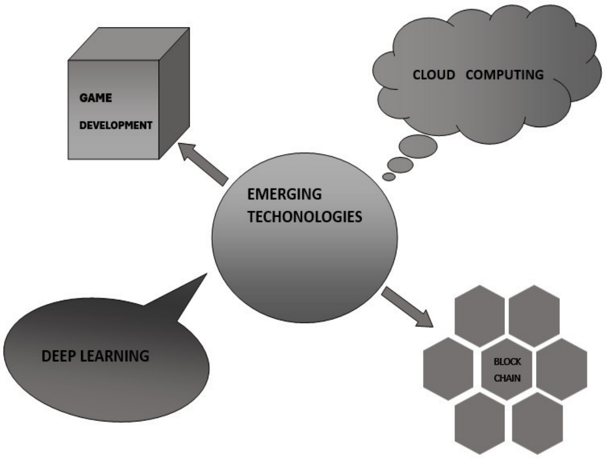


Fig. 2. Emerging Domains In the field of Software Engineering

V. Deep Learning

With its capability to handle complex tasks like image recognition, natural language processing, and decision making, deep learning the branch of machine learning, has gained extensive recognition. Many fields and research domains have embraced the extensive use of deep learning technology [17]. Nevertheless, the development of deep learning models is far from straight forward. Challenges in this domain are categorized into three types [1].

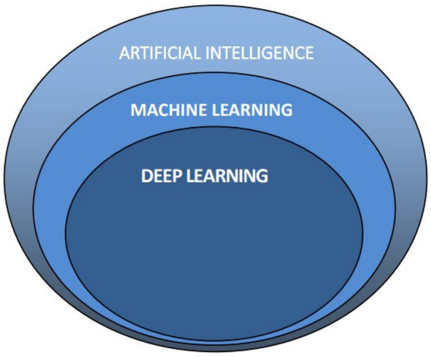


Fig. 3.Intranet Classification of Deep Learning

A. *Deep Learning Challenges*

The authors [1] conducted the experiment on seven real world Machine Learning (ML) projects, the successful execution of these ML projects in conjunction with companies of various sizes and types has facilitated valuable learning experiences and various challenges. These challenges are categorized into three types: development, production, and organizational challenges. The most commonly occurring challenges are stated below with the help of a pie chart. The given below pie chart shows the scope of each challenge that can occur in the project.

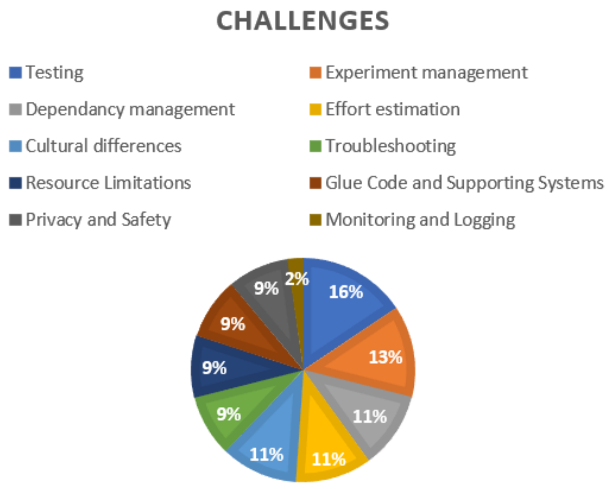


Fig. 4. Pie Chart of Various Challenges in Machine Learning projects

VI. Blockchain

Blockchain technology[16] has gained immense popularity for its transformative potential in areas like finance, supply chain management, and healthcare, blockchain technology has seen a skyrocketing rise in popularity.. In simple words it is like a mathematical structure that stores data or digital transactions, the utilization of blockchain involves an unalterable and distributed digital ledger, comprising interconnected blocks that are safeguarded by virtually unhackable cryptographic signatures, thereby minimizing the risk of tampering or disruption solutions. However the implementation of blockchain solutions comes with a unique set of challenges:

A. Blockchain Challenges

The key elements define a blockchain as a data structure [2]. The authors [2] identify the most relevant Blockchain-Oriented Software Engineering (BOSE) and the consequent issues that arise. To effectively address these challenges, they refer to relevant excerpts from the SWEBOK [8]. to provide a comprehensive understanding of the related problems. The challenges in the blockchain-oriented Software Engineering are listed in the tabular form given below.

Table I. List Of Blockchain Technology Challenges In The Realm Of Software Engineerings

CHALLENGE	EXPLANATION
New Professional Roles	The increasing importance of blockchain has led to the emergence of new professional roles [2], such as intermediaries bridging the gap between business-focused individuals and IT experts, requiring expertise in finance, law, and technology.
Security And Reliability	Blockchain-based systems (BOS) must prioritize security and reliability throughout the software development lifecycle, with a focus on testing suites for smart contracts (SCT) and blockchain transactions (BTT) to ensure trustworthiness and integrity.
Architecture Of The Software	In BOS development, software architects should define selection criteria for blockchain implementations and consider advanced data representations like Object Graphs to improve operational efficiency.
Modelling Languages	In BOS development, specialized graphical modelling languages and adaptations of existing models, such as UML diagrams, are often needed to accurately represent the unique characteristics of the BOS environment, as traditional diagrams may fall short.
Metrics	For BOSE Systems, specialized metrics are needed, and the Goal/Question/Metric (GQM) method can be adapted to measure complexity, communication, resource consumption (e.g., gas in Ethereum), and overall performance in the distributed blockchain environment.

VII. Cloud Computing

Cloud computing has become a fundamental component of modern software engineering. It offers agility, scalability, and cost-efficiency, enabling software engineers to focus on building innovative applications while relying on cloud providers for infrastructure management and support.

However, engineers must also address security, resource management, and vendor-related considerations [3] when adopting cloud solutions.

Customers can avail the services offered by the cloud computing models in three ways:

- Software as a service (SaaS): This model delivers on solicit claims over the internet (network).
- Platform as a service (PaaS): It supplies a framework.
- Infrastructure as a service (IaaS): This archetypal offers solicit infrastructure possessions, often in the form of virtual machines.

Depending on the service models offered, software development encounters may be associated with various roles. On Par with challenges cloud computing also has its own security threats and risks [19].

A. Cloud Computing Challenges

Cloud computing has revolutionized the way software is developed and deployed. However, it introduces its own set of challenges:

Table Ii. List Of Challenges And Considerations In The Cloud Computing Environment

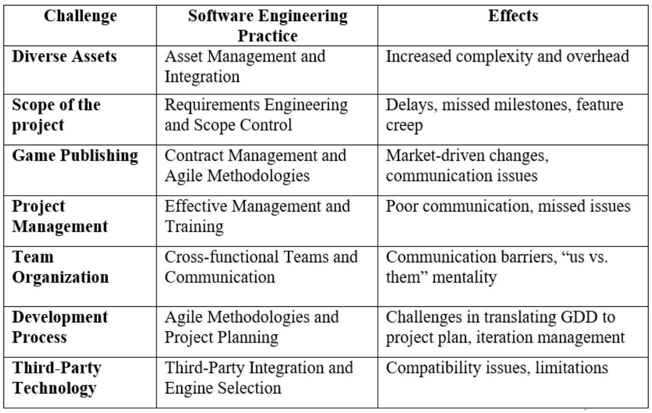
Challenge	Considerations
Software Requirements	Functional Requirements: Prioritizing specific requirements
	Non-Functional Requirements: Security and Privacy, Reliability, Delay, Scalability, Availability
	Other Requirements: SLA, Vendor Lock-in, Lack of Standards for development, Cloud Evaluation, Consumption patterns
Design	Choosing an appropriate design pattern, Platform problems, Parallel design, Design for Errors
Implementation	The cost of data transmission to the cloud, Topological dependencies problems, Implementation risks, Virtual machine's communications, Billing strategies
Testing	Security test, Expandability and performance test, Integrity related test, Innovation in testing, Testing tools
Maintenance and Support	Development support, Service Level Agreement, Resource and cost optimization

VIII. Game Development

Game development is a multidisciplinary field that requires collaboration among artists, designers, programmers, and testers. It combines technical expertise with creativity and innovation to create interactive experiences that captivate players.

The Video game industry, with its unique characteristics like managing multimedia assets and creating engaging gameplay experiences, requires tailored engineering practices [4]. As games become more intricate and players expectations rise, game developers must adapt by enhancing their software engineering methods. This involves implementing proven software engineering processes and practices that are both rigorous and adaptable to effectively manage projects and minimize risks in game development. This research explores the specific challenges in game development and how sound software engineering practices can assist developers in addressing these challenges effectively. The main challenge that all the game developers faces on the testing and the testing should be automated, as it helps the game developers[18], including the all types of stakeholders.

Table III. The Challenges and Effects in the Domain of Game Development



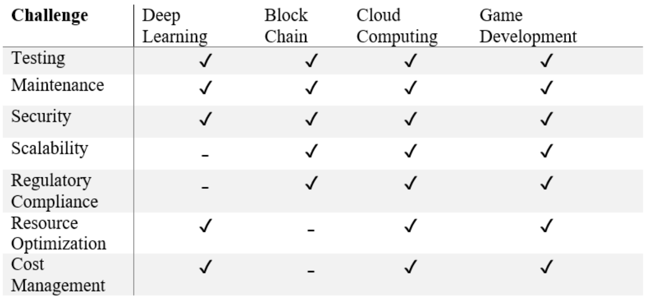
Challenge	Software Engineering Practice	Effects
Diverse Assets	Asset Management and Integration	Increased complexity and overhead
Scope of the project	Requirements Engineering and Scope Control	Delays, missed milestones, feature creep
Game Publishing	Contract Management and Agile Methodologies	Market-driven changes, communication issues
Project Management	Effective Management and Training	Poor communication, missed issues
Team Organization	Cross-functional Teams and Communication	Communication barriers, "us vs. them" mentality
Development Process	Agile Methodologies and Project Planning	Challenges in translating GDD to project plan, iteration management
Third-Party Technology	Third-Party Integration and Engine Selection	Compatibility issues, limitations

IX. Conclusion

The rise of cutting-edge technologies like blockchain, cloud computing, deep learning, and game development has brought exciting opportunities, but also significant challenges, to enterprise software development. This paper delves into these challenges throughout the software development lifecycle (SDLC), examining each stage individually

Given below table provides the different types of challenges that all the domains will face in the software development process. This includes various challenges some of the challenges are commonly faced by all the domains of software development, these are mainly testing and maintenance challenges.

Table IV. The Different Types of Challenges That All the Domains Will Face in the Software Development



Challenge	Deep Learning	Block Chain	Cloud Computing	Game Development
Testing	✓	✓	✓	✓
Maintenance	✓	✓	✓	✓
Security	✓	✓	✓	✓
Scalability	-	✓	✓	✓
Regulatory Compliance	-	✓	✓	✓
Resource Optimization	✓	-	✓	✓
Cost Management	✓	-	✓	✓

The above Table IV shows the comparison between the different challenges that occur in the emerging technologies in the domain of software engineering.

X. Future Work

The objective of examining challenges from this particular standpoint is as follows:

- Categorizing the encounters to offer an optimal solution based on the advance phases of future work.
- Advising a new-fangled tactic for software developers to face the tests.

Clear benefits can be observed in utilizing emerging technologies for software development, despite the presence of challenges. Hence, forthcoming work will aim to propose an appropriate approach to effectively address these existing challenges.

References

[1] Arpteg, Anders, et al. "Software engineering challenges of deep learning." 2018 44th euromicro conference on software engineering and advanced applications (SEAA). IEEE, 2018.

[2] Porru, Simone, et al. "Blockchain-oriented software engineering: challenges and new directions."2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C). IEEE, 2017.

[3] Kashfi, Hanieh. "Software engineering challenges in cloud environment: Software development lifecycle perspective."International Journal of Scientific Research in Computer Science, Engineering and Information Technology 2.3 (2017): 251-256.

[4] Kanode, Christopher M., and Hisham M. Haddad. "Software engineering challenges in game development."2009 Sixth International Conference on Information Technology: New Generations. IEEE, 2009.

[5] M. Swan, Blockchain: Blueprint for a new economy. O'Reilly Media, Inc., 2015.

[6] Unicredit, "Blockchain technology and applications from a financial perspective,"2016.

[7] F. Aymerich, G. Fenu, and S. Surcis, "A real time financial system based on grid and cloud computing," in Proceedings of the ACM Symposiumon Applied Computing, 2009, pp. 1219-1220.

[8] Grundy, J., Kaefer, G., Keong, J., and Liu, A., "Guest Editors' Introduction: Software Engineering for the Cloud", IEEE Software, vol. 29, pp. 26-29, 2012.

[9] Shan, T., "Smart Cloud Engineering, Nomenclature, and Enablement", In Proceedings of the 1st International Conference on Cloud Computing and Services Science, Noordwijkerhout, Netherlands, 7-9 May, 2011.

[10] N. Jones, "The learning machines," Nature, vol. 505, no. 7482, p. 146, 2014.

[11] A. Efrati, "How deep learning works at apple,"2013.

[12] A. Abran, J. W. Moore, P. Bourque, R. Dupuis, and L. Tripp, "Swebok: Guide to the software engineering body of knowledge 2004 version," IEEE Computer Society, Los Alamitos, California, 2004.

[13] Bhuvaneswari, T., Prabaharan, S., "A Survey on Software Development Life Cycle Models", International Journal of Computer Science and Mobile Computing, Vol. 2, Issue.5, pp. 262 – 267, May 2013.

[14] Pagalla Bhavani Shankar, et al. "Effort Estimation of Software (EEOS) Using Machine Learning Approached Algorithms", International Journal for Innovative Engineering and Management Research, Vol.12, Issue.02, pp.162-165.

[15] Pagalla Bhavani Shankar, M. Babu Reddy, "Significant Strategies to Assess Software Effort Estimation: A View of Functional Point", I-manager's Journal on Software Engineering, Vol. 17, No. 3, pp. 33-37.

[16] Pagalla Bhavani Shankar, "Blockchain: The Essential Future of Modern Internet", International Journal for Modern Trends in Science and Technology, 6(10): 60-64, 2020.

[17] Dong, Shi, Ping Wang, and Khushnood Abbas. "A survey on deep learning and its applications." *Computer Science Review* 40 (2021): 100379.

[18] Politowski, Cristiano, Yann-Gaël Guéhéneuc, and Fabio Petrillo. "Towards automated video game testing: still a long way to go." *Proceedings of the 6th International ICSE Workshop on Games and Software Engineering: Engineering Fun, Inspiration, and Motivation*. 2022.

[19] Alouffi, Bader, et al. "A systematic literature review on cloud computing security: threats and mitigation strategies." *IEEE Access* 9 (2021): 57792-57807.

DOI: 10.15439/2023R29
ISSN 2300-5963 ACSIS, Vol. 38

Detection of Copy-Move Image Forgery Using Local Binary Pattern from Detailed Wavelet Coefficient

Daljeet Kaur
Department of Computer Science & Engg
Gyan Ganga College of Technology
Jabalpur, India
daljeetkaur@ggct.co.in

Ajay Lala
Department of Computer Science & Engg
Gyan Ganga College of Technology
Jabalpur, India
ajaylala@ggits.org

Kamaljeet Singh Kalsi
Department of Computer Science & Engg
Gyan Ganga College of Technology
Jabalpur, India
kamaljeetsingh@ggct.co.in

***Abstract*—One of the most prevalent types of image forgery is copy-move forgery. A portion of the image is being copied and further pasted to a different location inside the identical image during the copy-move approach in order to hide a significant portion of the image. Finding duplicate portions in the image is the purpose of the copy-move based forgery detection technique. In this paper, we suggest a system which tends to detects forged portion in a forgery image. The DILBP (Detailed image local binary pattern) approach is used in this work to extract features, which includes extraction of feature, matching of feature, duplicate valued block detection. Several experiments have been initiated on a forged image to detect copy-move forged part. The experimental conclusions highlight that the suggested system is efficient for quality with respect to accuracy and speed.**

***Index Terms*—Image Forgery, DILBP (Detailed image local binary pattern), local binary patterns (LBP), set difference, Wavelet Decomposition.**

I. Introduction

This Digital picture forgery is one of the often emerging difficulties in the realm of crime. There are currently no precise approaches available to automatically determine the authenticity and integrity of digital photographs. Typically, pictures have been used to verify the reality of an event. In the processing of image, the veracity of a digital picture can serve as crucial evidence. The detection of fraud in digital photographs is a developing study area for assuring the validity of the images. The availability of less expensive software and hardware tools makes it convenient to produce, edit and change digital photographs without leaving any visible signs that these activities have taken place. Regular newspapers, television, magazines, and the Internet disseminate a massive number of sophisticated archives that are produced by a variety of devices on a regular basis. In addition, by enhancing the capacity of image processing tools, modifying these pictures becomes quite easy. Pictures are crucial to communication across all of these channels.

Image forging is the process of making a false image by altering the actual image's content and passing it off as the original image for illegal purposes. The existence of digital picture authentication has received a significant deal of attention recently since digital media is now often employed in many security organizations as well as applications, making image fraud an important problem. Therefore, methods for identifying manipulated photos are now being researched. Different methods for manipulating images were created within a few years after photography was invented. Combination prints, which were produced by using darkrooms to print several portions of an image on a single sheet of photographic related paper, are one of the techniques that helped in the production of pictures. The Two Ways of Life by Oscar G. Rheinlander, which required up to 30 distinct negatives and it takes approx six weeks to create, contains the first well-known combination prints.

Humans may now readily obtain engaging multimedia from the internet and alter or modify it as they see fit, thanks to advancements in technology and the ease of use of the internet. There are two common methods of manipulating images: region duplication by copy-move forgeries and image splicing. During image splicing, portions of different photographs are combined to produce a manipulated image. On the other hand, image sections are copied and pasted onto the same image in copy-move forgery in order to increase or hide some significant content in the image.

It becomes difficult to distinguish between tempered and legitimate sections when copied regions appear to be identical with compatible components (such as color and noise). In addition, a counterfeiter employs several post-processing techniques like noise reduction, edge smoothing, and blurring to eliminate any visible indications of image manipulation. One unique form of forgery is called "copy-move forgery imaging," in which portions of a picture are copied and then pasted back into the original. Because of this, picture forensics and copy-move forgery detection have grown in significance in our networked culture.

The proposed work aim to design a detection system which detect forged image of copy-move forgery type. In the proposed system, digital image is divided into overlapped blocks. After that, the feature extraction approach has been applied on the forged image to extract the particular features from particular image block. Further duplicate blocks have been detected, which indicated the forged portion of an image. Some the sample forged image and original images have been shown in “Fig. 1-a”and “Fig. 1-b”.

The primary information carriers in the modern digital environment are digital photos and movies. However, the validity and integrity of the digital images are a major cause for concern because these information sources are easily manipulated using widely available software. Furthermore, the most common method of altering digital photographs is copy-move image forging. A specific kind of image manipulation known as "copy-move forgery" involves copying and

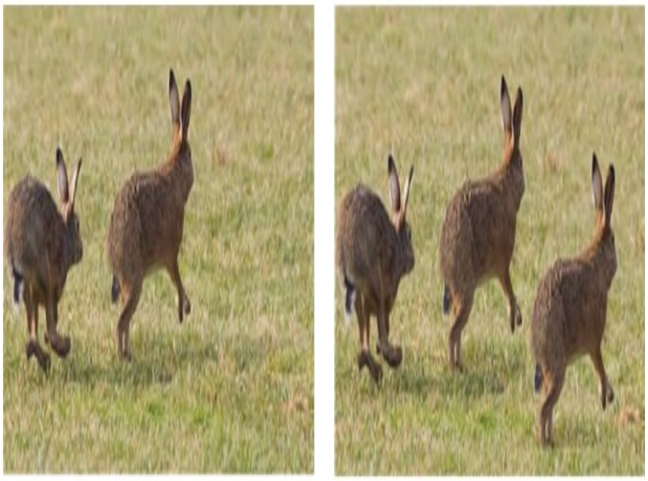


Fig. 1-a Actual image/ Corresponding Copy- move forged image.

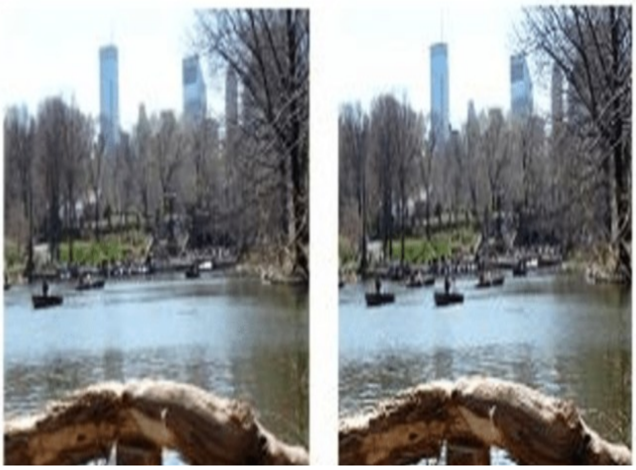


Fig. 1-b Actual image/ Corresponding Copy- move forged image.

pasting an image portion in a different location with the goal of hiding a significant aspect of the original image. Therefore, finding identical or strikingly similar image regions is the aim in the detection of copy-move forgeries.

The following are further enumerated in the paper. A review of the relevant studies is provided in Section 2. Section 3 describes the suggested image forgery detection method. Section 4 presents the outcomes of the experiments, and Section 5 wraps up the work. Beginning with the extraction of a section of the input image or a model of 3D object, image forgeries are created. Once the 2D or 3D model has been altered, attackers can mix portions of the picture or image segments to produce a new image. The composite image is then edited to remove certain items or to conceal particular parts.

II. Related Work

Guiwei Fu et al. [1] suggest an image copy-move forgery detection method based on fused features and density clustering. Tahaoglu, et al. proposed digital image copy move forgery based detection system which need to be implemented in the environment of real time[2].They suggested a strategy that starts by removing the input image's textural form. A Ciratefi-based method is used to localize the faked pixel. A novel technique was suggested by R. H. et al.[3] to detect the copy-move forgery, which is the most common type of forgery attack. The detection and localization of forgeries is a notable issue that has drawn and continues to draw the attention of academics working in the area of digital based forensics, according to Pranshav Gajjar et al. [4]. To enable accurate localization of the tampered area, Mauro Barni and colleagues [5] devised a technique to determine the copy-move forgery's source and target locations. Agarwal R et al.[6] In order for our system to identify the tampered region, the suggested technique initializes the tampered image as the input. When using SIFT characteristics, Fontani et al.'s J-linkage approach and copy-move detection idea were proposed [7] in 2013. A classification-based attack (CLBA) technique is suggested by Muhammad et al.[8] in 2012 for the identification of tempered pictures. Sunil et al.[9] determines the state of One post-processing action that the attacker might use to get around image forgery detection techniques is changing the intensity of the copied portion. The introduction to the bibliography on the blind picture forgery detection technique is provided by Mahdian et al[10]. A block-based technique was proposed by Edoardo [11] in which texture is taken from the block and used as a feature.

Copy-move forgery detection methods in digital photos, databases, and evaluation metrics are surveyed and compared by Sami Gazzah et al[12]. The study attempts to shed light on the relative efficacy of several techniques for identifying copy-move frauds. Several popular detection strategies are included in the study, such as deep learning, GAN, hybrid, transform domain, block-based, keypoint-based, and hybrid approaches. K. Latha et al.[13] successfully identify whether a picture has been edited and prevent users from trying to submit modified photographs by using a machine learning algorithm (SVM). The integrity and authenticity of digital photographs are now questioned, undermining consumer confidence in them due to recent advancements in image altering software.

Using deep learning to train a Convolutional neural network (CNN) on a dataset of real and fake photos, Devarshi Patrikar et al.[14] conduct an extensive investigation of image forgery techniques. GAN stands for generative adversarial network. They conclude that deep learning has demonstrated encouraging results for image forgery detection and is an active field of research despite a number of obstacles. By using a hybrid Deep Learning (DL) architecture, D Prabakar et al.[15] create a very powerful and efficient detection approach for this kind of image counterfeiting. To begin with, MICCF2000 is the source of the sample images. Secondly, the photos are resized, and any noise that may have existed in the original image is removed using a filtering approach. Ultimately, we construct a hybrid deep learning model by fusing support vector machines (SVM) and convolutional neural networks (CNN). The created hybrid deep learning model is verified using metrics like precision, F1-score, True Positive Rate (TPR) and Negative Rate (TNR), False Positive Rate (FPR) and Negative Rate (FNR), and accuracy.

With an emphasis on frequently occurring copy-move and splicing attacks, Zanardelli et al.[16] explore some of the most recent image fraud detection algorithms built specifically upon Deep Learning (DL) techniques. Insofar as DeepFake-generated content is applied to photographs, it is also addressed, producing an effect akin to splicing. Given that deep learning-powered techniques yield the best overall outcomes on the benchmark datasets that are currently available, this survey is very pertinent.

A technique proposed by Dipanshu Narayan et al.[17] to identify copy-move forgeries is based on breaking down blocks into features and then extracting those features from the transforms of the blocks. An additional instrument for identifying forgeries is a Convolutional Neural Network (CNN). To extract features, convolution and pooling layer pairings in serial fashion are used.

III. Proposed System

The primary goal of this forgery region work is to highlight related regions in image that may vary in size and form. A difficult task is the approach of pixel to pixel comparison in order to locate the identical areas. In order to create a forgery detection system that is both effective and efficient, a logical window has been constructed. This sliding window shifts in accordance with size of window across the entire picture to obtain the photographs' feature vector. The regions have been regarded as a single block that is protected by sliding windows. The repositioning of the window has therefore resulted in the creation of one additional block.

For each potential block, values of feature in matrix format, which reflect the potential block values, have been retrieved by the system. With the aid of a sliding window, the input picture is split into small blocks of the similar size at the beginning. The feature extraction approach has been used on every potential block. DILBP (Detailed image local binary pattern), which combines the local binary pattern approach and detailed coefficients based wavelet transformation, is the suggested feature extraction strategy for each of the blocks as shown in “Fig- 2”.

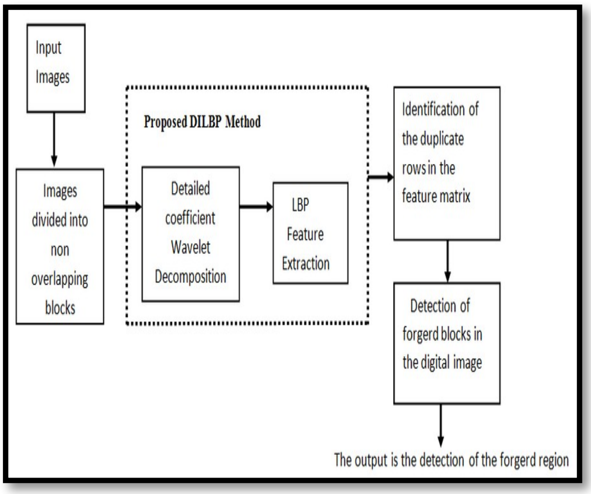

Input Images
Images divided into non overlapping blocks
Proposed DILBP Method
Detailed coefficient Wavelet Decomposition
LBP Feature Extraction
Identification of the duplicate rows in the feature matrix
Detection of forgerd blocks in the digital image
The output is the detection of the forgerd region

Fig-2. Work flow of the proposed system

A. Proposed algorithm

1) Overlapped Blocks Creation

With this method, the fake picture is first split into overlapping sections. Here, the fundamental technique is to find interconnected blocks that have been duplicated or relocated. There are several overlapping blocks in the forged area. Extraction of features from these blocks would come next.

2) Technique For Feature Extraction

The forged image was subjected to the feature extraction technique in order to extract specific features from a block of the image after the overlapped blocks were generated. In order to extract features from the block region, the detailed image local binary pattern features method is used in this activity.

B. DILBP (Detailed image Local Binary Pattern)

Face photos have been converted into detailed images using a bi-level wavelet decomposition technique. Local binary patterns (LBP) have then been used to extract local aspects of the fake images from detailed coefficients-based deconstructed images. The accurate and efficient DILBP approach combines the long-running LBP method with comprehensive coefficient-based wavelets decomposition. Detailed coefficients based wavelet decomposition.

C. Detailed coefficients based wavelet decomposition

This decomposition technique makes use of signal and temporal analysis. It can be applied to deconstruct a bogus image into multiple sub-band images with different directional attributes, spatial resolution, and frequency characteristics. In this method, the forged image is broken down into up to two layers in order to calculate the approximation and details coefficients. Details coefficients do not contain the highest frequency component of a picture, in contrast to details coefficients, which do. In this investigation, only the detailed coefficient has been used further during the complete process.

The forgery's high frequency region is the only component of the picture that is altered by the small scale obstruction and expression modifications. For forged images, any additional decomposition processes cause information loss and are thus not included in this study.

D. Principles of local binary pattern

The output of the wavelet has been used to local binary pattern (LBP), where the original picture has been divided into tiny sections from which the local binary patterns or histograms have been extracted. As seen in figures. 3 and 4, distinct LBP histograms were derived, which depict circular neighbor-sets for three distinct values of R and P. The display of the fake picture is created by concatenating all of the blocks of the fake image into a single feature histogram.

As illustrated in "Fig-3," the feature vector of the image can be generated [18] after the evaluation of the local binary pattern (LBP) for every individual pixel.

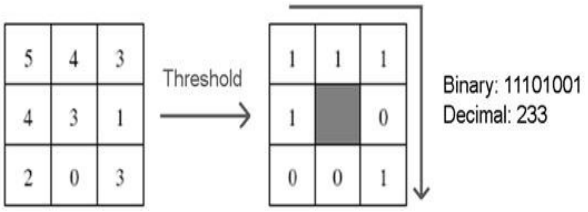


Fig-3. The actual LBP Operator (source of image: [18])

The threshold value is used as the centre pixel's value by the basic LBP operator, which operates on the values of the eight neighbors’ pixels. If the grey value of a neighboring pixel is equal to or greater than that of the centre pixel, then

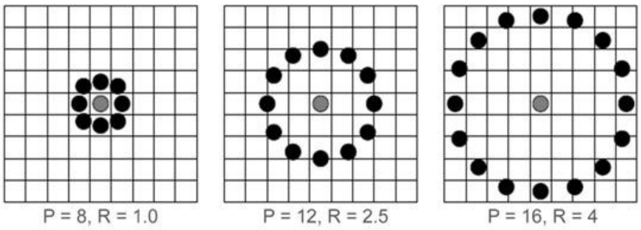


Fig-4. Neighbor-sets for three distinct P and R (source of image: [18])

one is assigned to specific pixel; else, zero is assigned to it. The LBP code is then created for the centre pixel value by concatenating 8 zeros or ones to create a binary code [18], as seen in "Fig-4".

The LBP based operator's borders have expanded further to employ various sizes relevant to the neighborhood. A circle with radius R has been drawn starting from the centre pixel. The values of the centre pixels and the hypothetical P sampling sites on the circumference of the particular circle are compared. (Dual) interpolation is required to extract the numerical values of every neighborhood sample point for any number of pixels and radius. Figure 4 illustrates the notation (P, R) that has been used for the particular neighborhood.

1) *Duplicate Rows identification in a feature matrix*

Each row in the feature matrix represents a certain block. To find the duplicate rows, the system first counts how many rows in the feature matrix are being compared to the filtered out rows that remain duplicates. Consequently, the blocks with repeated entries in the feature matrix are the outcome of this comparison.

2) *Forged region Detection*

The next stage is to expose the identical blocks of digital image, which also serves as a warning sign for counterfeit areas, after identifying blocks that behave identically. Thus, the machine finally finds a fake area in the digital image. The system is highlighting the specific forged locations.

When using the DILBP technique for extraction of feature, the computing time of the entire process is lowered when the LBP approach is combined with wavelets. This increases the system's efficiency and tends to enhance the effectiveness of forgery detection system.

IV. Result of Experimental Analysis

An Intel (R) Core (TM) i3-3120M CPU running at 2.50 GHz with 4GB of random access memory has been used to test the proposed system. All activities connected to simulations are carried out using the MATLAB platform. As seen in Table I, which shows the sizes of two images—one is titled "River and Tree Image" and the other is titled "House and Chimney Image"—the performance is evaluated by looking for forged portions in the digital image. An additional column in the Table I. shows specific blocks size which are represented by each row in the feature matrix. Execution Time indicates the time taken to detect the forged part. Last entry in the Table I. show the no of forged blocks detect by the proposed system.

Adobe Photoshop 7.0 was used to create the equivalent set of forged pictures, which were then saved in the 275 * 275 , 300 * 300 png format. A sliding window with a size of 26 by 26 is being placed on each individual pixel. To get the results of the experiment on picture forgery, the suggested DILBP approach is being used to the faked photos. Following the application of the suggested detection of image forgery method, we obtain a forged part in the forged images and corresponding forged areas are emphasized by the block-based system that are exactly similar to one another from every angle, as shown in "Fig-5.1" and "Fig-5.2", which effectively indicates the forged image.

Table I. Performance Table

Sr No	Image Size	Block Size	Execution Time	No of duplicate blocks identified
1.	River and Tree image 275 ×275	23 × 23	0.32 sec	1
2.	House and Chimney Image 300 × 300	34 × 34	0.50 sec	1

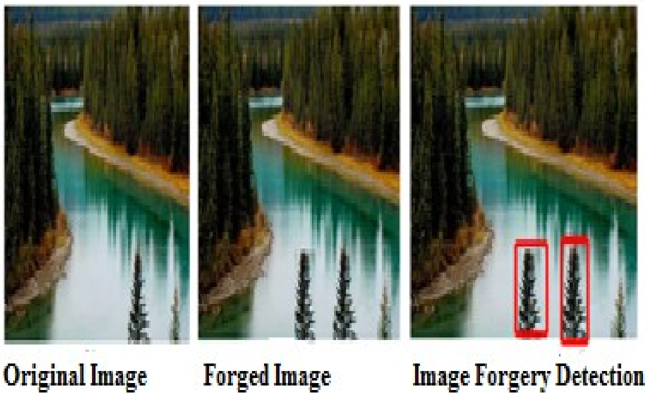


Fig-5.1. Detection of Forged part Results I

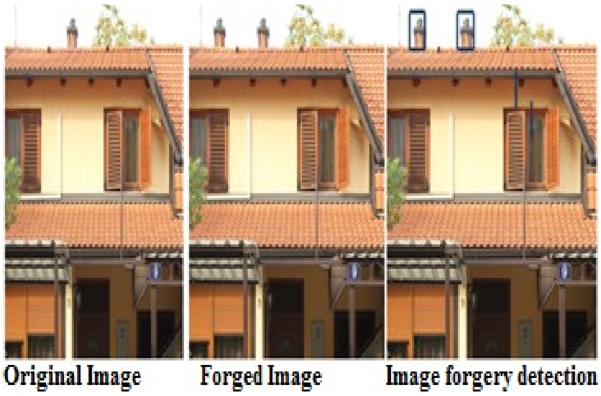


Fig-5.2. Detection of Forged part Results II

V. Conclusion

This proposed study use the DILBP approach, which incorporates the wavelet decomposition's detailed coefficient characteristics, to recognize the copy-move forged picture. The research covered in this paper yields a good outcome for detecting fabricated regions. The improvement of time complexity to detect the forged region will be the next step in the future. Also, the proposed system as the proposed approach aims to detect forged portion in still forged images only, in future next step will be to detect forged portion in video also.

References

[1] Image Copy-Move Forgery Detection Based on Fused Features and Density Clustering, Guiwei Fu, Yujin Zhang, Yongqi Wang, et al., Appl. Sci. 2023, 13, 7528. This link points to 10.3390/app13137528.

[2] Tahaoglu, G., Ulutas, G., Ustubioglu, B., et al. Digital picture copy move forgery detection using a Ciratefi approach. 81, 22867–22902 (2022) Multimed Tools Appl. This link points to 10.1007/s11042-021-11503-w.

[3] R. H. Jaafar, Z. H. Rasool and A. H. H. Alasadi, "New Copy-Move Forgery Detection Algorithm," 2019 International Russian Automation Conference (RusAutoCon), Sochi, Russia, 2019, pp. 1-5, doi: 10.1109/RUSAUTOCON.2019.8867813.

[4] Pranshav Gajjar et al 2022, "Copy Move Forgery Detection: The Current Implications and Contemporary Practices", in International Conference on Electronic Circuits and Signalling Technologies", doi:10.1088/1742-6596/2325/1/012050.

[5] Phan QT, Tondi B, Barni M (2021) Transfer source-target disambiguation via multiple branch CNNs in a copy manner. 16:1825–1840 IEEE Trans Inf Forensics Secur.

[6] Verma O, Agarwal R (2020) An effective deep learning feature extraction and matching technique for copy move forgery detection, Multimedia Tools and Applications, number 79, pages 7355–7376 (2020).

[7] A Framework for Decision Fusion in Image Forensics Based on Dempster-Shafer Theory of Evidence, Marco Fontani, Tiziano Bianchi, Alessia De Rosa, Alessandro Piva, and Mauro Barni, IEEE Transactions on Information Forensics and Security, Vol. 8, no. 4, April 2013.

[8] Gandhim Muhammad, Hussain, and George Bebis, "Undecimated dyadic wavelet transform for passive copy move image forgery detection," Elsevier, 2012.

[9] Kumar Sunil, Desai Jagan, Mukherjee Shaktidev, "DCT-PCA Based Method for Copy-Move Forgery Detection", ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II, 2014, Volume 249

[10] Mahdian B, Saic S. A bibliography of techniques for detecting image forgeries without sight. 2010;25(6):389–399. Signal Processing: Image Communication.

[11] Giuseppe Mazzola, Alessandro Bruno, and Eduardo Ardizzone, "Copy-Move Forgery Detection via Texture Description," Proceedings of the 2nd ACM symposium on Multimedia in forensics, security, and intelligence, 2010, pp. 59-64.

[12] Sami Gazzah; Lamia Rzouga Haddada et al. "Digital Image Forgery Detection with Focus on a Copy-Move Forgery Detection: A Survey" in 2023 International Conference on Cyberworlds (CW), DOI: 10.1109/CW58918.2023.00042.

[13] K. Latha; D. Kavitha; S. Hemavathi et al. "Image Forgery Detection Using Machine Learning" in 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), DOI: 10.1109/ICPECTS56089.2022.10046422.

[14] Devarshi Patrikar; Usha Kosarkar; Anupam Chaube "Comprehensive study on image forgery techniques using deep learning" in 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), DOI: 10.1109/ICETET-SIP58143.2023.10151540.

[15] D Prabakar; R. Ganesan; D. Leela Rani; Praveen Neti et. al. "Hybrid Deep Learning Model for Copy Move Image Forgery Detection" in 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) , DOI: 10.1109/I-SMAC55078.2022.9987319

[16] Zanardelli, M., Guerrini, F., Leonardi, R. *et al.* Image forgery detection: a survey of recent deep-learning approaches. *Multimed Tools Appl* **82**, 17521–17566 (2023). https://doi.org/10.1007/s11042-022-13797-w

[17] Dipanshu Narayan; Himanshu; Rishabh Kamal "Image Forgery Detection" in 2023 International Conference on Disruptive Technologies (ICDT), **DOI:** 10.1109/ICDT57929.2023.10151341

[18] Md. Shafiul Azam, Tanzillah Wahid, Md. Abdur Rahim, and Md. Najmul Hossain, "Face Recognition using Local Binary Patterns (LBP)" in Global Journal of Computer Science and Technology Graphics & Vision, Volume 13 Issue 4 Version 1.0, (2013).

Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 35–39 DOI: 10.15439/2023R87 ISSN 2300-5963 ACSIS, Vol. 38

AI-driven rental bicycle system: An Ensemble learning approach

Cu Kim Long
Information Technology Center Ministry of Science and Technology (MOST)
Hanoi, Vietnam longck@most.gov.vn

Trinh Thi Thu Hoang
Hanoi University of Industry
Hanoi, Vietnam huongttt@haui.edu.vn

Gloria Jeanette Rinco´n Aponte
Universidad Cooperativa de Colombia
Bogota, Colombia
gloriaj.rincon@campusucc.edu.co

Vikram Puri
Center of Visualization and Simulation
Duy Tan University
Da Nang, Vietnam
purivikram@duytan.edu.vn

***Abstract*—Bicycle sharing is a notable sustainable transporta- tion option for metropolitan regions and communities seeking to address environmental concerns, reduce traffic congestion, and combat air pollution while promoting public health and improving connections. There are already technologies to support this system, including typical mobile applications and kiosks strategically positioned at the bicycle station. Nevertheless, most proposed solutions cannot accurately forecast the demand for bicycle availability, efficiently redistribute bicycles, create routes to circumvent traffic congestion and conduct comprehensive user analysis. To address these challenges, a framework for an AI- enabled bicycle-sharing system has been presented to predict the count of bicycle rentals. To assess performance, four distinct ensemble-based models are implemented and tested using various statistical parameters.**

***Index Terms*—bicycle rental system, BSS, artificial intelligence, ensemble technique, feedback.**

I. Introduction

Now, there exists an escalating worldwide inclination towards the adoption and execution of bicycle rental systems [1]. The main goal of the bicycle rental system is to enable the tem- porary leasing of bicycles to individuals, typically for periods spanning from 15 minutes to a few hours. There are some key reasons to escalate the demand for the bike-sharing system, such as providing a sustainable transport option which en- courages people to use bicycles instead of fuel-based vehicles, providing flexibility for pick-up and dropping bikes due to nu- merous docking stations, reducing traffic congestion inside the cities, connected with the public transportation facilities which make it convenient to pick up public transportation, offer a convenient way for the tourists to explore the city more as well as promote tourism activities. Additionally, it also helps improve people’s physical and mental health [2]. Although the bike rental system greatly impacts society, key issues also need to be addressed. The first issue is bike availability; at peak times, bikes are unavailable at their docking station, discouraging people from adopting the bike-sharing system. Large-scale bike upkeep and repairs require a lot of work and skill. Systems must install more stations and determine the optimal fleet composition to handle rising demand over time [3]. Understanding user patterns and peak usage periods is crucial to adapting to changing needs despite difficulty. Pricing methods must balance income generation with affordability to maintain the system. Few trip planning tools point users toward safe, efficient cycling routes. Predicting maintenance issues before malfunctions render bikes useless is one unsolved topic. Keeping an eye on fleet activities, bike maintenance, customer behaviour, pricing, routes, and breakdowns is challenging [4]. Artificial intelligence (AI) has the potential to overcome these challenges and provide a better experience for users as well as service providers. AI encompasses computer systems that possess the ability to execute tasks that conventionally necessitate human intelligence, including but not limited to sensory seeing, recognition of speech, decision- making, and translating languages [5]. The fusion of AI in the bike-sharing system transforms the user experience better. The key benefits of AI in bike-sharing systems are (see in figure 1):

1) **Demand Prediction:** AI analyses the bike availability per the rider data log in the system, including the external weather conditions, and provides feedback to service providers to rebalance the bike availability.
2) **Optimization in Routes:** As per traffic congestion in the city, the system optimizes new routes and provides the fastest and safest way for the user.
3) **Maintenance:** AI can analyze data collected from bike sensors to detect bicycles that require repair or main- tenance before the occurrence of any breakdowns. This enhancement contributes to the enhancement of safety and reliability.
4) **Rewards:** AI can comprehend user data and offer incen- tives, such as points and monetary prizes, to encourage prolonged user engagement with bike-sharing systems.
5) **Fraud Detection:** AI can identify and analyze usagepatterns that may indicate questionable activity, hence speeding up identifying stolen or lost bicycles. This phenomenon leads to a decrease in both financial losses and criminal activities.

This study presents a potential architecture for a bike rental system incorporating artificial intelligence technology. The primary contribution of this work is as follows:

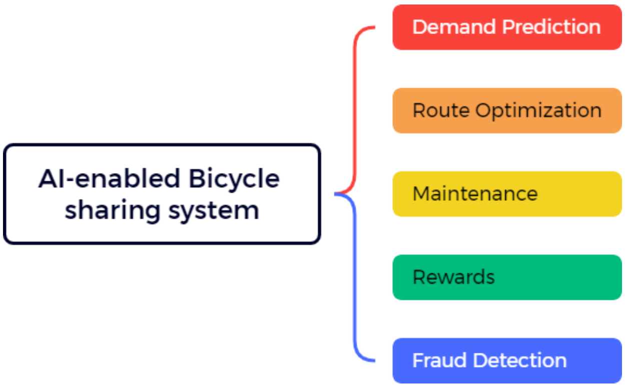


Fig. 1. Key points to integrate the AI with Bicycle sharing system

1) Trained four ensemble-based approaches on BSS set.
2) Evaluated these models with the help of statistical parameters.
3) chose and deployed a better performance model.

The subsequent sections of the paper are organized as follows: Section 2 provides an overview of the associated work, Section 3 outlines the technique employed in the suggested work, Section 4 presents the results obtained from the proposed work, and Section 5 summarizes the study.

II. Related Work

In the past years, some studies have been published on bicycle-sharing systems. In [6], the authors suggested using a multiple-layer spatial network model to analyze the public transport system. This model considers the interconnectivity of transit paths, cycling stations, and pedestrian pathways. The overall evaluation of this study is that the introduction of public bicycle-sharing systems has been shown to improve the efficiency of the public transportation network by decreasing passenger travel times, promoting smoother traffic flow, and alleviating congestion. In this study [7], authors present a complete technique for establishing a Bicycle Sharing System (BSS) that efficiently incorporates optimizing station positions and capacity allocation. The proposed methodology combines a set-covering framework for distributing customer demand to stations with a queuing model for assessing amenities levels. These studies are manually optimizing station location and capacity location. AI can enhance the design, operation and adaptability of the BSS. To determine which machine- learning techniques are most commonly used in this field and to examine how machine learning has been used to enhance bike-sharing programs in smart cities, authors [8] presented a literature review. The review aids in synthesizing prior findings and identifies areas needing additional investigation. In [9], the authors presented how to deploy ML models to maximize the number of bicycles available in the public bicycle-sharing scheme. The algorithms provide an accurate forecast of station occupancy levels, enabling timely redis- tribution of bicycles across stations. Ensuring riders’ regular access supports bicycles as an environmentally friendly form of travel for the environment and public health. The deep learning approach presented [10] a highly accurate forecasting model that has the potential to significantly contribute to real-time decision-making and operational management within expanding dockless systems worldwide. The main aim of this study is to expand the application of advanced neural network architectures in the modelling of complex spatiotemporal systems. Similarly, the authors [11] proposed a deep learning model called STGA-LSTM to forecast the demand for bicycle sharing across several stations. Advocating for fair and just consumption practices facilitates the establishment of envi- ronmentally friendly, reliable shared transportation systems. The authors suggest a dynamic repositioning system based on a Monte Carlo tree search [12]. This system aims to assist service providers in efficiently balancing the distribution of bicycles across stations, taking into account their movement patterns. The notion of service level is established to quantify the quantity of bicycles that require transfer at each station. In [13], this study primarily applies deep learning models to forecast short-term bike demand for the bike rental system, specifically predicting demand 15 minutes in advance. A hybrid CNN-LSTM model is considered for the prediction.

III. Methodology

In this section, the methodology of the paper, including the framework and how things are connected to the framework, is discussed (see figure 2).

A. Bicycle Sharing Ecosystem

The Bicycle Sharing Ecosystem (BSS) allows individuals to borrow bicycles within their local area temporarily. A considerable number of bicycles are observed to be parked at a specific station inside the BSS. Bicycles can be connected through docks, which are specialized racks designed to secure and release the bike, or they can be supplied with specialist locks that keep the bicycle stationary at a certain location. Furthermore, these locks can be operated by mobile applica- tions or kiosks located at the terminal. The mobile applications examined in the BSS are founded on the conventional server- client infrastructure for reserving and verifying the availability of bicycles. In this system, an AI system is integrated into the traditional mobile application to forecast the total number of bicycles used.

B. AI Model Technique

To train and deploy the AI model in the BSS, there are some steps which need to be followed.

1) Data Collection and Data Pre-processing

The dataset [14] on bike sharing encompasses many characteristics about the date, time, weather conditions, day type, and the count of bike rentals. The variables encompass the record index, date, season, year, month, hour of the day, presence of a holiday, day of the week, and working day status. The variable ”weathersit” classifies weather conditions into four distinct categories: clear, mist, light, and heavy rain. The dataset additionally has normalized variables for temperature, perceived temperature, humidity, and wind speed. In conclusion, the dataset includes tallies about the number of individuals classified as casual riders registered riders and the overall count of bike rentals, encompassing casual and registered users. These variables collectively offer valuable insights into the various aspects that affect the

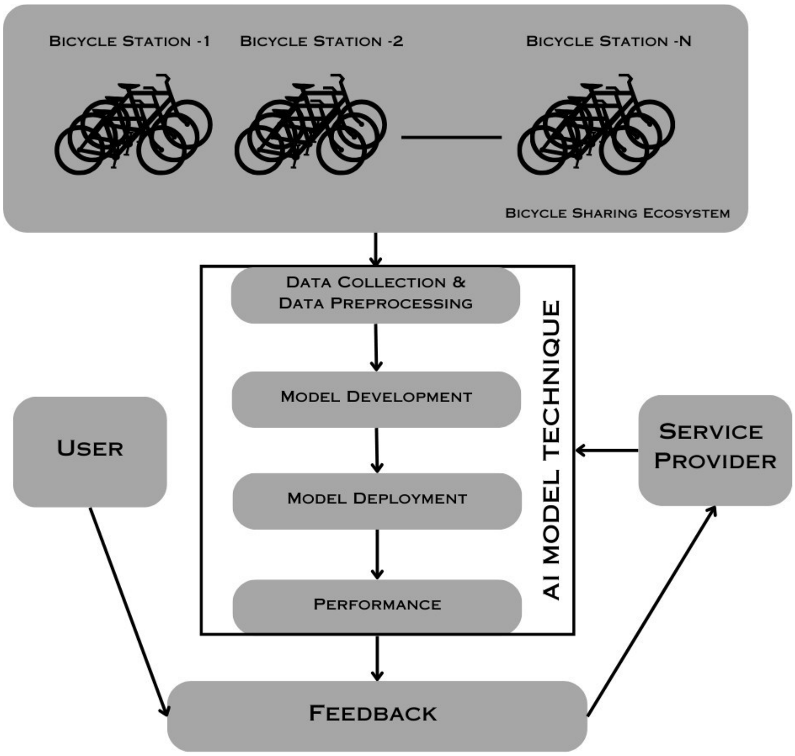


Fig. 2. Framework of AI-enabled Bicycle sharing system

demand for bike rentals, such as weather conditions, time of day, kind of day (e.g., weekday or weekend), and consumer characteristics. The objective is to utilize these factors to investigate and construct a model for the patterns of bike sharing.

Data Pre-processing: The data pre-processing techniques [15] include removing the missing values using mean, median or most frequently values imputation; the next step is to identify the abnormal values that need to be removed from the dataset. Feature selection is a crucial step in the modelling process, as it involves identifying and removing duplicate, unnecessary, or noisy features. By selecting those with the greatest importance characteristics, the precision and generality of the model can be enhanced.

2) *Model Development*

To facilitate the creation of the model, it is necessary to partition the data into either a training set and a testing set or into a training set, a testing set, and an evaluation set. The rationale behind withholding specific test data is to ensure an impartial assessment of the model's ability to generalize novel variables. The outcome evaluation may demonstrate a positive bias if the model is assessed using the same dataset used for its training. The subsequent phase involves selecting an appropriate model following the dataset. One approach that can be employed is to utilize various models and assess their performance by employing statistical parameters. This methodology facilitates the provision of feedback to the model under consideration. In this study, four ensemble machine learning are deployed to evaluate the performance of the proposed approac and explained in Table I.

3) *Model Deployment*

Once the model has been trained, it is integrated into the manufacturing ecosystem, where it may generate predictions in real-time. It is imperative that the input information format and sort utilized for production purposes align with the format and type on which the algorithm was developed [16]. The set-up system must be capable of effec- tively managing the pre-processing of real-time data. For this study, the proposed model is deployed through FastAPI.

Table I. Different Ensemble models used in the study

S.No.	Model Ensemble	Description
1	Random Forest (RF) Regressor	The RF algorithm is a metaestimator that employs several decision trees for classification. RF model is a versatile, easy to use regression model that provides better accuracy without extensive hyperparameter tuning.
2	Gradient Boosting (GB) Regressor	GB estimator constructs an additive model using a forward stage-wise approach, enabling the optimal selection of various distinct losses. Although GB model provides accurate prediction, it requires careful tuning.
3	AdaBoost Regressor	The AdaBoost is a meta-estimator that initially trains a regression model on the initial data set. Sub- sequently, it trains new regressor clones on the exact same dataset, but with updated instance values based on the variance of the present prediction. Additionally it required the tuning to avoid the overfitting issue.
4	Extra Tree Regressor	This model presents a meta estimator that applies a series of randomized decision tree structures, also known as extra-trees, on different subsets of the dataset

4) *Performance*

The performance of AI-based applications is of utmost importance due to the tendency of models to deteriorate in intricate real-world settings as time progresses.

In order to mitigate the impact of errors on consumers, it is imperative to engage in ongoing surveillance of essential per- formance indicators, thereby enabling the timely identification of potential difficulties before they escalated into significant problems. This enables the model to undergo retraining or augmentation over time. Some parameters are also helpful to check the model performance after deployment, such as prediction accuracy, error rate, AI pipeline monitoring, and Quinton sampling, which helps to check the model perfor- mance through the experts in a periodic period.

C. *Feedback*

The incorporation of feedback is widely recognized as an essential component of the machine-learning process. The feedback provided by real-time users holds significant impor- tance as it is a crucial indicator of the model's performance. This is because even if a model demonstrates a high validation score, it may still encounter failures when deployed in real- world scenarios. In certain instances, it may be the case that researchers and developers are unable to discern the fault within the code. However, the user can pay attention to and discover said error. Furthermore, feedback mechanisms foster trust and promote transparency on the capabilities and limitations of models [17]. In this research, participants and service providers are linked to the feedback mechanism of the artificial intelligence model. The user submits a problem using the feedback system, which is relayed to the service providers for prompt resolution

IV. Results and Evaluation

In this section, experimental testbed and results outcome is discussed.

A. *Testbed*

For this study, google co-lab is considered to train the model initialized with matplotlib, pandas, NumPy, seaborn and sklearn for the training and testing of four ensemble model for this study.

B. *Analysis*

To evaluate the performance evaluation of the proposed framework, there are three statistical parameters [18] considered such as root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R2).

Table II displays the performance evaluation of four ensemble models on the BSS dataset. A dataset was utilized to assess the performance of four regression ensemble models in predicting a continuous target variable. The models included in the training process included the Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor, and Extra Trees Regressor. Smaller MAE and RMSE values are indica- tive of superior model performance. Higher R2 values that approach 1 are indicative of superior performance. The Extra Trees Regressor demonstrated superior performance with a mean absolute error MAE of 24.87, RMSE of 41.17, and R2 of 0.94. This demonstrates that it possesses the lowest mean errors and exhibits the highest degree of concordance with the actual target values. The Adaboost Regressor had the lowest performance across all three evaluation measures, namely a MAE of 86.13, RMSE of 106.62, and R2 value of 0.64. The predictions exhibit the highest degree of deviation from the actual data. The RF and GB variants demonstrated a moderate level of performance. The errors and R2 values of their models fall within the range of the best and worst models. Based on the analysis conducted, it is recommended to utilize the Extra Tree Regressor model for predicting the target variable due to its notable performance across many assessment measures. The measurements offer empirical support for the selection of this option, ensuring its accuracy.

Table II. Performance Evaluation of Four Ensemble Model on BSS Set

Parameter	**RF Regressor**	**GB Refressor**	**AdaBoost Regressor**	**Extra Tree Regressor**
MAE	25.37	48.18	86.13	24.87
RMSE	42.05	70.78	106.62	41.17
R^2	0.94	0.84	0.64	0.94

The objective of regression modeling is to make predictions for a continuous target variable by utilizing a collection of pre- dictor variables. The regression model produces forecasts, or estimated quantities, of the dependent variable. The aforemen- tioned predictions are indicative of the results produced by the regression model that has been appropriately fitted, based on a certain set of input predictor values. The comparison between the predicted counts derived from the regression model and the actual counts is conducted to assess the effectiveness and accuracy of the fitted model. The accuracy of the model in predicting the target variable improves as the projected counts approach the actual counts. An effective regression model aims to minimize discrepancies and achieve predicted numbers that closely align with the actual counts. Figure 3 illustrates the comparison between the observed and anticipated total count of bicycles for four different ensemble learning models. Figure 3. actual vs predicted Total Count figure with two columns. Left column: (A) Top left panel showing RF Regressor. (B) Bottom left panel showing AdaBoost Regressor. Right column: (C) Top right panel showing GB Regressor. (D) Bottom right panel showing Extra Tree Regressor.

V. Conclusion

Bicycle-sharing systems offer an ecologically advantageous mode of transportation that confers numerous benefits to metropolitan regions regarding environmental sustainability. Nevertheless, the efficient operation of these systems poses significant logistical and planning challenges. The increasing significance of AI technologies is observed in the context of bicycle-sharing providers, who are utilizing these tools to tackle the challenges mentioned earlier. AI algorithms can examine patterns in rider utilization to optimize bike allocation and determine appropriate pricing strategies. This paper proposes an AI-enabled framework for analysing rental bicycles in a bicycle rental system. The framework aims to enhance the service provider's understanding of rental bicycles and improve the customer experience.

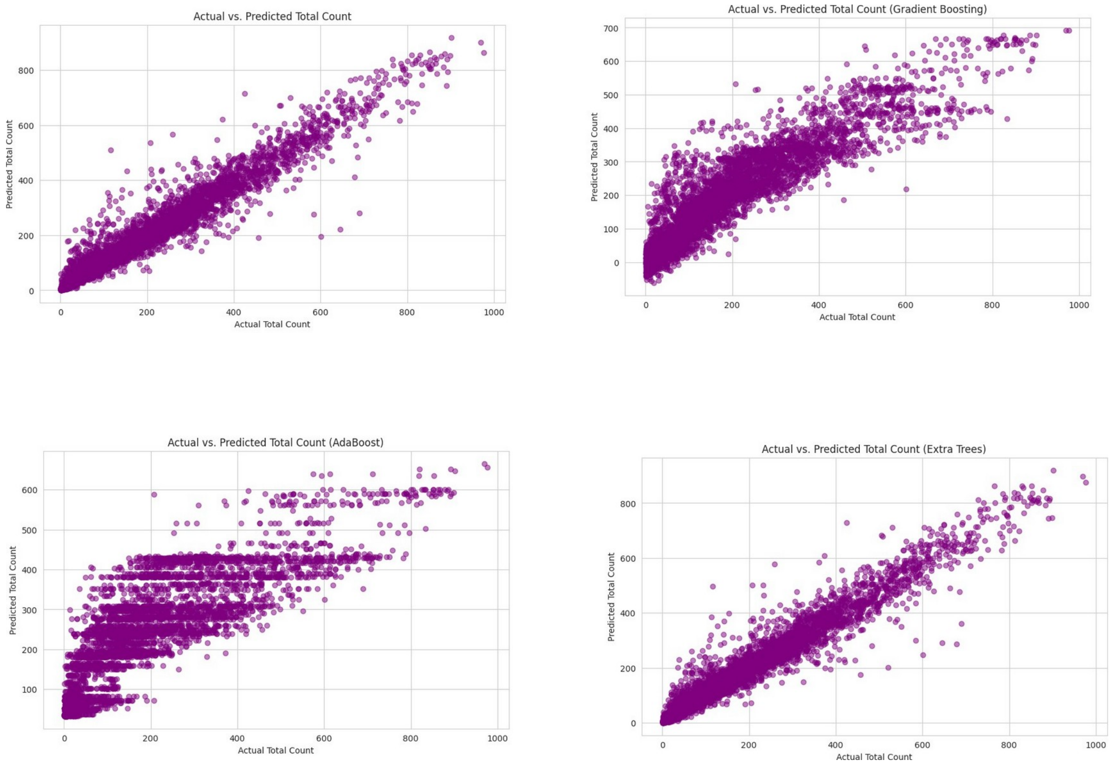


Fig. 3. Actual vs Predicted Total Count bicycle for four ensemble learning approaches.

There is a significant potential for enhancing intelligence and efficiency within bicycle-sharing systems in the foreseeable future by integrating AI technology to a greater extent. The use of advanced tracking sensors and predictive analytics will augment the accuracy of monitoring the whereabouts of bicycles and forecasting ridership demand. AI methodologies enable service providers to offer individualized recommendations and incentives to specific clients. In future times, this technology is expected to expand its range to include intelligent route planning, while simultaneously enhancing the user experience.

References

[1] Billhardt, H., Ferna´ndez, A., & Ossowski, S. (2021). Smart recommen- dations for renting bikes in bike-sharing systems. Applied Sciences, 11(20), 9654.

[2] Czech, P., Turon´, K., & Urban´czyk, R. (2018). Bike-sharing as an element of integrated Urban transport system. In Advanced Solutions of Transport Systems for Growing Mobility: 14th Scientific and Technical Conference" Transport Systems. Theory & Practice 2017" Selected Papers (pp. 103-111). Springer International Publishing.

[3] Ma, Y., Lan, J., Thornton, T., Mangalagiu, D., & Zhu, D. (2018). Challenges of collaborative governance in the sharing economy: The case of free-floating bike sharing in Shanghai. Journal of cleaner production, 197, 356-365.

[4] Shaaban, K. (2020). Why don't people ride bicycles in high-income developing countries, and can bike-sharing be the solution? The case of Qatar. Sustainability, 12(4), 1693.

[5] Duan, Y., & Wu, J. (2022). An AI Approach to Rebalance Bike- Sharing Systems with Adaptive User Incentive. Artificial Intelligence-based Internet of Things Systems, 365-389.

[6] Yang, X. H., Cheng, Z., Chen, G., Wang, L., Ruan, Z. Y., & Zheng, Y. J. (2018). The impact of a public bicycle-sharing system on urban public transport networks. Transportation research part A: policy and practice, 107, 246-256.

[7] C¸elebi, D., Yo¨ru¨su¨n, A., & Is¸ık, H. (2018). Bicycle sharing system design with capacity allocations. Transportation research part B: method- ological, 114, 86-98.

[8] Albuquerque, V., Sales Dias, M., & Bacao, F. (2021). Machine learning approaches to bike-sharing systems: A systematic literature review. ISPRS International Journal of Geo-Information, 10(2), 62.

[9] Rozˇanec, J. M., Krivec, T., Kersˇicˇ, V., Cundricˇ, L., Stojanovicˇ, B., Zeman, M., & Bratko, I. (2022). Bicycle Sharing Systems meet AI: forecasting bicycles availability and decision-making. In Central European Conference on Information and Intelligent Systems (pp. 365-370). Faculty of Organization and Informatics Varazdin.

[10] Ai, Y., Li, Z., Gan, M., Zhang, Y., Yu, D., Chen, W., & Ju, Y. (2019). A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system. Neural Computing and Applications, 31, 1665-1677.

[11] Ma, X., Yin, Y., Jin, Y., He, M., & Zhu, M. (2022). Short-term prediction of bike-sharing demand using multi-source data: a spatial-temporal graph attentional LSTM approach. Applied Sciences, 12(3), 1161.

[12] Huang, J., Tan, Q., Li, H., Li, A.,& Huang, L. (2022). Monte carlo tree search for dynamic bike repositioning in bike-sharing systems. Applied Intelligence, 1-16.

[13] Mehdizadeh Dastjerdi, A., & Morency, C. (2022). Bike-sharing demand prediction at community level under COVID-19 using deep learning. Sensors, 22(3), 1060.

[14] Fanaee-T,Hadi. (2013). Bike Sharing Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5W894.

[15] Zhang, C., Wu, F., Wang, H., Tang, B., Fan, W., & Liu, Y. (2022). A meta-learning algorithm for rebalancing the bike-sharing system in IoT smart city. IEEE Internet of Things Journal, 9(21), 21073-21085.

[16] Ehsan, A., Abuhaliqa, M. A. M., Catal, C., & Mishra, D. (2022). RESTful API testing methodologies: Rationale, challenges, and solution directions. Applied Sciences, 12(9), 4369.

[17] Li, J., & Wang, W. (2023). From Renting Economy to Sharing Economy: How Do Bike-Sharing Platforms Grow in the Digital Era?. Journal of the Knowledge Economy, 1-21.

[18] Kataria, A., & Puri, V. (2022). AI-and IoT-based hybrid model for air quality prediction in a smart city with network assistance. IET Networks, 11(6), 221-233

DOI: 10.15439/2023R58
ISSN 2300-5963 ACSIS, Vol. 38

Immersive Virtual Painting: Pushing Boundaries in Real-Time Computer Vision using OpenCV with C++

Satyam Mishra
International School – Vietnam National University
Hanoi, Vietnam
satyam.entrprnr@gmail.com
0000-0002-7457-0060

Vu Duy Trung
International School – Vietnam National University
Hanoi, Vietnam
trungthichban@gmail.com

Le Anh Ngoc
Swinburne Vietnam, FPT University
Hanoi, Vietnam
ngocla2@fpt.edu.vn

Phung Thao Vi
International School – Vietnam National University
Hanoi, Vietnam
phungvi08123@gmail.com

Sundaram Mishra
NETMONASTERY NSPL,
Mumbai, India
mishrasundaram.sm@gmail.com

***Abstract*—This paper presents an innovative approach for immersive virtual painting using real-time computer vision techniques. A meticulously crafted color detection algorithm implemented in C++ and OpenCV achieves up to 97.4% accuracy in identifying specified hues from live video feeds. The detected colors are seamlessly translated into vibrant brush strokes rendered on a digital canvas in real-time. The algorithms exhibit remarkable speed, analyzing each frame within 15ms, enabling ultra-low latency painting interactions. Optimization strategies involving parallel processing and code optimizations provide further performance gains. Comparative analysis reveals 3-4x faster execution using C++ over Python for color detection. The platform delivers an intuitive, natural, and uninterrupted painting experience, as validated through user studies. By automating color detection and digital rendering, this research transforms virtual painting from a passive activity to an immersive form of human-computer co-creativity. The fusion of computer vision, rendering algorithms, and optimization techniques establishes new frontiers in interactive digital art platforms and reshapes human-computer collaboration.**

Highlights:

• This research achieves exceptional accuracy in real-time color detection, with up to 97.4% precision in identifying specified hues across thousands of video frames.

• The integrated system enables seamless user interaction for natural virtual painting expressions, eliminating disruptive color selection interruptions.

• Comparative analysis reveals significant 3-4x performance gains by implementing the algorithms in C++ instead of Python, underscoring the efficiency benefits of C++ for real-time computer vision applications.

• User studies validate the immersive experience delivered by the platform, with users highlighting the responsiveness, precision, and intuitive interaction unmatched by traditional virtual painting tools.

• The proposed techniques establish a new paradigm in real-time computer vision, pushing the boundaries of virtual creativity platforms and reshaping human-computer collaboration in th*e arts.*

***Index Terms*—Computer Vision, OpenCV, Real-time Interaction, Virtual Painting, Color Detection Algorithms, Digital Canvas Rendering**

I. Introduction

Interactive virtual painting refers to the use of computational frameworks and technologies to assist users in creating artworks in a virtual environment. These frameworks provide tools and suggestions to enhance the user's creativity and guide them in the painting process. One approach proposed in the research domain is Neural Painting (NP), which uses a conditional transformer Variational AutoEncoder (VAE) architecture with a two-stage decoder to suggest strokes for completing an artwork [1]. Another approach involves the use of a painting simulator that allows users to virtually paint on a display using sensors and objects to trigger virtual paint colors. The system tracks the movement of the objects and displays the virtual paint color on the display in response [2]. Additionally, there are techniques that enable image synthesis from incomplete human paintings, allowing users to progressively synthesize desired images with just a few coarse user scribbles [3]. Virtual reality applications also exist that provide emotional characteristics to virtual painting, allowing users to create paintings with expressive emotion-based brushes and shapes [4].

Real-time color detection is significant in various computer vision applications, such as skin color detection and sport playground detection [5]. It allows for the modelization of color clusters and the classification of image pixels based on their membership to a particular color class [6]. This real-time implementation of color detection techniques enables efficient and low-cost processing, making it possible to detect color clusters in real-time video sequences [7]. On the other hand, canvas rendering is important for authentication purposes, as it provides a reliable digital fingerprint that can be used to identify and track users online [8]. By generating a hash value from canvas and WebGL, a model using KNN can accurately authenticate users with an accuracy of 89% [9].

OpenCV is a computer vision library that is widely used for various applications. It provides tools and algorithms for tasks such as object detection, face recognition, and image processing. OpenCV is used in combination with deep learning techniques, such as Convolutional Neural Networks (CNN), to achieve accurate and efficient results[10]. CNN, including variations like YOLO, has shown exceptional improvement in object detection, making it a crucial application of image processing [11], [12]. Object detection goes beyond simple classification and helps in localizing specific objects in images or videos. It has applications in various fields, including inventory management in retail and vehicle detection for autonomous vehicles[13], [14]. OpenCV also enables face detection and recognition using techniques like Haar-like features and principal component analysis (PCA) [15], [16]. Overall, OpenCV plays a significant role in computer vision by providing a wide range of tools and algorithms for different tasks[17].

II. Literature Review

A. *Evolution of OpenCV and Its Impact on Computer Vision*

Computer vision applications in transportation logistics and warehousing have a huge potential for process automation. A structured literature review on research in the field categorizes the literature based on the application and computer vision techniques used. The review also points out directions for future research and provides an overview of existing datasets and industrial solutions [18]. Face recognition is another important application of computer vision, and research in this area has focused on using cascade classifiers and principal component analysis for face detection and recognition [16]. In the construction industry, computer vision-based methods have been applied for safety monitoring, productivity improvement, progress monitoring, infrastructure inspection, and robotic application. These methods involve various aspects of computer vision such as image processing, object classification, object detection, object tracking, pose estimation, and 3D reconstruction [19]. Machine learning plays a significant role in computer vision and image processing, contributing to domains such as surveillance systems, optical character recognition, robotics, and medical imaging. The review discusses the importance of machine learning, its applications, and open research areas in computer vision [20], [21]. Computer vision has been widely studied and applied across disciplines, with a focus on image recognition and understanding information from photos and videos [22].

B. *Color Detection Techniques in Computer Vision*

Color detection techniques in computer vision involve various methods and algorithms for identifying and analyzing colors in images. These techniques are used in applications such as computer control systems, gesture-based human-computer interaction, and color measurement in the textile industry. One approach is to determine the number and characteristics of color targets within an image using algorithms that rely on digital indexing code tables and decimal and binary numbers [23]. Another method involves filtering an image to isolate a predefined set of colors and then determining whether a desired color is present within the filtered image [24]. In the context of gesture-based human-computer interaction, real-time tracking of hand and finger motion can be achieved by calculating changes in pixel values of RGB colors from a video, without the need for artificial neural network training [25]. In the textile industry, computer vision techniques are used for color measurement and evaluation. These techniques involve digital image processing, device characterization and calibration, and various methods such as polynomial regression, neurofuzzy, and artificial neural network for measuring and demonstrating color of textiles [26]. Overall, color detection techniques in computer vision play a crucial role in a wide range of applications, enabling accurate analysis and understanding of color information in images.[27]

C. *Drawing Algorithms for Real-time Canvas Rendering*

Drawing algorithms for real-time canvas rendering is a challenging task in computer graphics. The quality and efficiency of rendering algorithms need to be defined, measured, and compared. Fischer et al. propose the PADrend framework, which supports the systematic development, evaluation, adaptation, and comparison of rendering algorithms [28]. Kim et al. present a real-time panorama algorithm for mobile camera systems, which includes feature point extraction, feature tracking, rotation matrix estimation, and image warping [29]. Fütterling focuses on core algorithms for rendering, particularly ray tracing, to support massively parallel computer systems [30]. Yuan et al. introduce a dynamic measure to capture temporal image distortions in real-time rendering algorithms [31]. Eisemann et al. provide a guide to understanding the limitations, advantages, and suitability of different shadow algorithms for real-time to interactive rendering [32].

D. *Integration of OpenCV with C++ for Real-time Applications*

OpenCV can be integrated with C++ for real-time applications. Object recognition and detection can be achieved using OpenCV and Python 2.7, improving accuracy and efficiency [33]. Deep learning-based object detection, such as Region-Based Convolutional Neural Network (R-CNN) and You Only Look Once (YOLO), can also be implemented using Python, providing speed and real-time application use [34]. Face detection and recognition can be accomplished using Python and deep learning techniques, making it suitable for real-time applications [35]. Additionally, OpenCV can be used for real-time image processing in traffic flow counting and classification, allowing for smooth monitoring without disturbing traffic [36]. OpenCV and Flask can be utilized to build a cloth try-on system, enabling users to try on upper body clothes in real-time [37]

Despite the wide application of OpenCV in real-time scenarios, it's relatively rare to witness the integration of OpenCV with C++. Most research and practical implementations tend to favor Python due to its ease of use and rapid prototyping capabilities. However, as indicated by the existing literature, the combined power of OpenCV and C++ offers unique advantages. C++ provides high performance, low-level memory control, and the potential for optimized code execution. Despite its potential, there is a scarcity of research focusing on harnessing these advantages in conjunction with OpenCV. The research problem lies in the underexplored territory of enhancing real-time computer vision applications through the integration of OpenCV with C++. This gap in research hinders the full exploration of the capabilities that arise from this combination, limiting the potential for highly efficient and high-performance real-time applications in various domains. The challenge is to delve into this unexplored realm, investigating the specific benefits and complexities that arise when OpenCV is tightly integrated with C++, thereby addressing the gap in the current body of knowledge.

Our research endeavors to redefine the landscape of virtual painting applications by delving into the unexplored integration of OpenCV with C++. While existing literature predominantly favors Python, our study aims to harness the unique advantages of C++ for real-time artistic interactions. Drawing inspiration from successful implementations like object recognition, deep learning-based techniques such as R-CNN and YOLO, and even face detection using Python, our research seeks to apply similar methodologies within the domain of virtual painting. By integrating OpenCV with C++, authors aim to enhance the accuracy and efficiency of color detection algorithms and real-time canvas rendering techniques. The research problem lies in the scarce exploration of this integration, limiting the development of immersive virtual painting experiences. Our research proposition is to leverage the combined power of OpenCV and C++ to optimize color detection, enabling precise strokes and vibrant hues in real-time virtual painting scenarios, ultimately advancing the field by addressing this research gap.

III. METHODOLOGY

Our research methodology is driven by a multidimensional approach, integrating key insights from the existing data to enhance the realm of virtual painting applications. First and foremost, authors focus on the intricate design of our Color Detection Algorithm, meticulously implemented using OpenCV in C++. Drawing inspiration from successful ventures in object recognition and deep learning-based techniques such as R-CNN and YOLO, authors seek to infuse our color detection mechanism with similar accuracy and efficiency. By leveraging the robust computational capabilities of C++, authors aim to optimize the color detection process, ensuring precise identification of specific hues within a live video feed.

Simultaneously, our research dives into the realm of the Drawing on Canvas Algorithm, building upon the foundations laid by previous studies. Taking cues from face detection techniques and real-time image processing in traffic flow counting, authors implement innovative approaches to translate detected colors into dynamic and vibrant strokes on a digital canvas. This implementation is driven by Python's flexibility and C++'s performance, ensuring seamless integration and high responsiveness.

The heart of our research lies in the seamless Integration of these Algorithms for Real-time Interaction. By carefully harmonizing the Color Detection Algorithm with the Drawing on Canvas Algorithm, authors create a symbiotic relationship, enabling users to engage in virtual painting activities with unparalleled accuracy and aesthetic finesse. Moreover, authors employ Optimization Techniques for Efficient Real-time Processing, inspired by the successful application of these techniques in traffic monitoring systems. Through meticulous analysis and refinement, authors strive to achieve optimal computational speed and accuracy, crucial elements in enhancing the user experience in real-time virtual painting scenarios.

In essence, our methodology is a strategic amalgamation of proven techniques and innovative approaches. By integrating the power of OpenCV with C++, authors aim to elevate virtual painting to new heights, crafting an experience that marries technical brilliance with artistic expression. Through this robust methodology, our research seeks to transform virtual painting into a captivating and immersive reality.

A. Color Detection Algorithm Design using OpenCV in C++

The proposed color detection approach builds upon existing techniques for object recognition like YOLO. Similar to YOLO, the algorithm leverages HSV color space thresholds and contour detection to identify color objects. However, optimizations like contour approximation and filtering are incorporated to improve real-time performance. The algorithm also draws inspiration from face detection techniques which also rely on detecting contours in different color spaces.

ALGORITHM PSEUDOCODE:

```
1. Convert the input image from BGR to HSV color space.
2. Iterate through the predefined color ranges in 'myColors':
    a. Extract the lower and upper HSV values for the current color range.
    b. Create a binary mask by thresholding the image using the lower and upper HSV values.
    c. Find contours in the binary mask to identify color blobs.
    d. For each contour:
        i. Calculate its area.
        ii. If the area is larger than a threshold (e.g., 1000 pixels):
            A. Approximate the contour to reduce the number of vertices.
            B. Calculate the bounding rectangle for the simplified contour.
            C. Determine the centroid of the bounding rectangle.
            D. Store the centroid coordinates and the index of the detected color range.
3. Return the list of detected points.
```

The color detection algorithm starts by converting the input image from the BGR color space to the HSV color space. It then iterates through the predefined color ranges (myColors). For each color range, it creates a binary mask by thresholding the image using the lower and upper HSV values of the current color. Contours are extracted from this mask, representing color blobs.

Formula authors have used for converting RGB to HSC color space:

$$H = \theta \; if \; B \leq G \;\; \text{--------(1)}$$

$$H = 360° - \; \theta \; if \; B > G \;\; \text{--------(2)}$$

Where $\theta \; = \; \cos - 1\left[\frac{0.5(\mathrm{R-G})+(\mathrm{R-B})}{\sqrt{(\mathrm{R-G})2+(\mathrm{R-B})(\mathrm{G-B})}}\right]$

$$\mathrm{S} \; = \; 1 \; - \; 3 * \min(\mathrm{R,G,B}) \; / \; (\mathrm{R+G+B}) \;\; \text{---------(3)}$$

$$\mathrm{V} \; = \; \max(\mathrm{R,G,B}) \; \text{-------(4)}$$

Pseudocode for contour detection and filtering steps:

```
contours = findContours(mask)
for each contour c in contours:
if contourArea(c) > threshold:
contourApprox = approximateContour(c)
boundingRect = getBoundingRect(contourApprox)
```

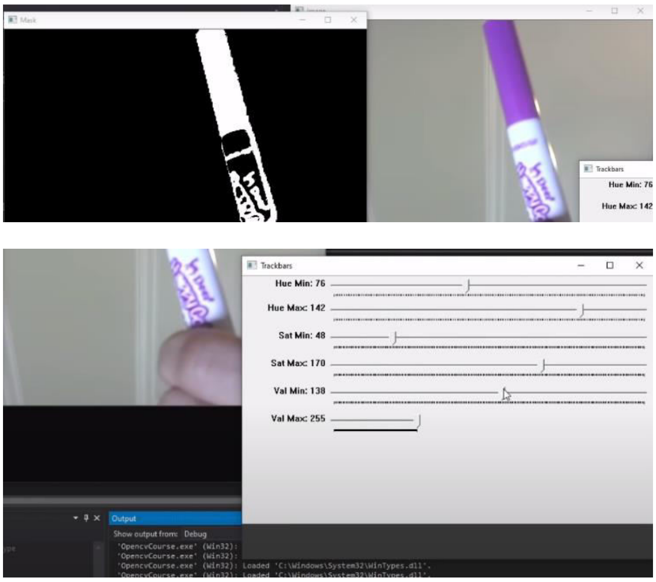
This pseudocode mathematically explains the HSV color conversion and contour processing steps in the color detection algorithm. The algorithm filters contours based on their area, ensuring they exceed a certain threshold to avoid noise. For valid contours, it approximates the shape, calculates the bounding rectangle, and determines the centroid. Detected points, along with their corresponding color indices, are stored in the newPoints vector. This algorithm enables precise identification of specific colors within the image, forming the foundation of the virtual painting application's interactive color detection mechanism.

C++ Code:

```
Mat colorDetection(Mat inputImage, vector<int> lowerHSV, vector<int> upperHSV) {
   Mat imgHSV;
   cvtColor(inputImage, imgHSV, COLOR_BGR2HSV);

   Mat mask;
   inRange(imgHSV, Scalar(lowerHSV[0], lowerHSV[1], lowerHSV[2]), Scalar(upperHSV[3], upperHSV[4], upperHSV[5]), mask);

   vector<vector<Point>> contours;
   findContours(mask, contours, RETR_EXTERNAL, CHAIN_APPROX_SIMPLE);

   vector<Point> approx;
   vector<vector<Point>> filteredPoints;

   for (const auto& contour : contours) {
      double area = contourArea(contour);
      if (area > 1000) {
         float peri = arcLength(contour, true);
         approxPolyDP(contour, approx, 0.02 * peri, true);
         if (approx.size() == 4) {   // Filter based on the number of vertices (can be adjusted)
            filteredPoints.push_back(approx);
         }
      }
   }
   return mask;
}
```

Explanation:

1. Convert to HSV: The input image is first converted from BGR (OpenCV's default color format) to HSV (Hue, Saturation, Value) color space. This is because HSV separates the intensity information (Value) from the color information (Hue and Saturation), making it easier to work with colors.
2. Color Thresholding: For each predefined color range (defined in myColors), a lower and upper HSV value is specified. The inRange function is used to create a binary mask where the white pixels represent the detected color range, and black pixels represent other colors.
3. Contour Detection: The contours (boundaries of white areas) in the binary mask are found using the findContours function. Contours are sets of points that represent the boundaries of objects in an image.
4. Approximation and Filtering: Contours that have an area larger than 1000 pixels are approximated to reduce the number of vertices using the approxPolyDP function. This approximation simplifies the contour shape. The resulting points are then filtered and stored.



Hue Min: 76
Hue Max: 142
Sat Min: 48
Sat Max: 170
Val Min: 138
Val Max: 255

```
Output
Show output from: Debug
'OpencvCourse.exe' (Win32):
'OpencvCourse.exe' (Win32):
'OpencvCourse.exe' (Win32): Loaded 'C:\Windows\System32\WinTypes.dll'.
'OpencvCourse.exe' (Win32): Loaded 'C:\Windows\System32\WinTypes.dll'.
```

Figure 1: Output of the Color Detection Algorithm

Figure 1 shows the output of the color detection algorithm giving us the min, max values of Hue, Sat and Val. It helps us detect and choose the color.

B. Drawing on Canvas Algorithm Implementation

The digital canvas rendering approach is inspired by prior work in real-time facial landmark detection. Similar to mapping key facial points, the algorithm maps detected color points to display coordinates. The algorithmic flow of extracting points and mapping them to visualize results parallels techniques used in facial and object landmark detection. However, optimizations like parallel processing are uniquely incorporated to boost rendering speeds. By correlating the study's algorithms to prior arts like YOLO and facial recognition, it helps position the research as an extension and focused application of these methods in the specific domain of virtual painting.

```
Algoritm Pseudocode:
1. Iterate through the list of detected points and their corresponding colors:
    a. Retrieve the coordinates and color index for the current point.
    b. Using the color index, obtain the corresponding drawing color from 'myColorValues'.
    c. Draw a filled circle on the canvas image at the specified coordinates using the obtained color.
2. Repeat step 1 for all detected points.
```

The Drawing on Canvas Algorithm operates in a straightforward manner, leveraging the detected points from the Color Detection Algorithm. For each detected point, the algorithm retrieves its coordinates and the corresponding color index. Using this index, the algorithm fetches the appropriate drawing color from the 'myColorValues' vector. Subsequently, the algorithm draws a filled circle on the canvas image at the specified coordinates, employing the obtained color. By repeating this process for all detected points, the algorithm renders dynamic and vibrant strokes on the digital canvas in real-time. This implementation ensures that the virtual painting experience is visually engaging and responsive, capturing the essence of the detected colors and translating them into aesthetically pleasing strokes on the canvas.

Formula for mapping detected colors to RGB values:

$$displayColor = colorPalette[detectedColorIndex]$$

where colorPalette is a lookup table mapping indices to RGB color values.

Pseudocode for drawing circles at detected points:

```
for each point p in detectedPoints:
x, y = getCoordinates(p)
color = getColor(p)

circle(img, (x,y), radius, color)
```

C++ Code:

```cpp
void drawOnCanvas(Mat& canvas, const vector<vector<int>>& points, const vector<Scalar>& colors) {
  for (size_t i = 0; i < points.size(); ++i) {
    circle(canvas, Point(points[i][0], points[i][1]), 10, colors[points[i][2]], FILLED);
  }
}
```

Explanation: The algorithm takes a list of points and corresponding color indices and draws filled circles on the input image at those points using the specified colors.

C. Integration of Algorithms for Real-time Interaction

The Real-time Interaction Algorithm utilizes the OpenCV library and an external camera to create a virtual painting experience. The program captures video frames from the default camera in real-time. For each frame, the 'findColor' function detects specific colors (purple and green) using predefined HSV color ranges. Detected points, representing the centroids of colored objects, are stored in the 'newPoints' vector. The 'drawOnCanvas' function then draws filled circles at these detected points on the 'img' matrix, simulating virtual paint strokes.

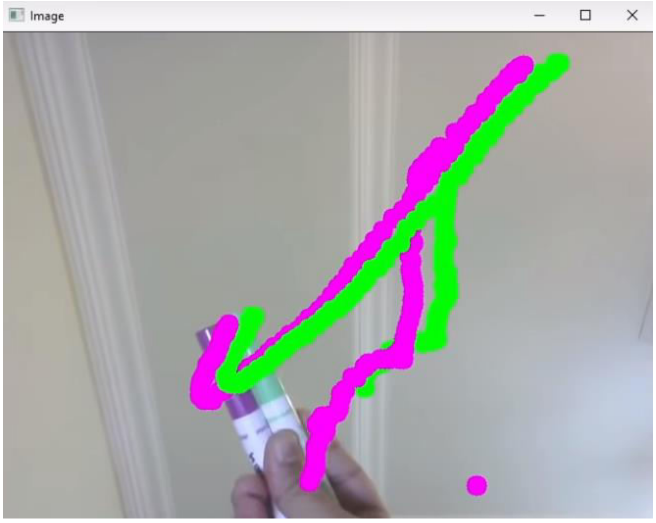


Figure 2: Illustration of Real-Time Virtual Paint

```
Algorithm Pseudocode:
1. Initialize the OpenCV video capture object 'cap' to capture video from the default camera (Camera index 0).
2. Create an empty matrix 'img' to store the video frames.
3. Initialize vectors 'myColors' and 'myColorValues' to store the defined color ranges and their corresponding display colors.
4. Create an empty vector 'newPoints' to store the detected points (x-coordinate, y-coordinate, color index).
5. Start an infinite loop to continuously capture video frames and perform real-time interaction:
    a. Read a frame from the video capture object and store it in the 'img' matrix.
    b. Call the 'findColor' function to detect specific colors within the frame, passing the 'img' matrix and color ranges.
    c. The 'findColor' function processes the frame as follows:
        i. Convert the frame from BGR to HSV color space.
        ii. Iterate through the predefined color ranges ('myColors') and create binary masks for each color range.
        iii. Detect contours in each binary mask, filtering contours based on area, and approximate their shapes.
        iv. Store the centroids of valid contours along with their color index in the 'newPoints' vector.
    d. Call the 'drawOnCanvas' function, passing the detected points and their corresponding display colors.
    e. The 'drawOnCanvas' function processes the detected points as follows:
        i. Draw filled circles at the specified coordinates on the 'img' matrix using the corresponding colors.
```

f. Display the updated 'img' matrix with virtual paint strokes in a window titled "Image".

g. Wait for 1 millisecond to allow for user interaction and continue the loop.

The integration of algorithms involves a continuous loop where frames are captured, colors are detected, and virtual paint strokes are rendered in real-time. This interaction offers users an immersive experience, allowing them to paint virtually by moving colored objects in front of the camera. The seamless integration of color detection and canvas rendering algorithms ensures a responsive and visually engaging virtual painting environment. As can see in figure 2, the successful virtual painting after integrating all algorithms.

D. *Optimization Techniques for Efficient Real-time Processing*

Within the context of our Virtual Painter project, the seamless interaction and responsiveness of the application are paramount. Leveraging a blend of advanced optimization techniques, our real-time processing pipeline has been fine-tuned for optimal performance:

1. Parallel Processing: To handle the computationally intensive tasks of color detection and canvas rendering, authors employed multi-threading. By parallelizing these operations, the system maximizes the utilization of CPU cores, ensuring rapid analysis and rendering of the video feed.

2. Memory Efficiency: Careful management of memory resources is crucial. Through meticulous memory allocation strategies and streamlined data structures, authors minimize memory overhead. This efficient memory usage ensures that the system runs smoothly, even during prolonged usage.

3. Algorithmic Refinement: Continuous refinement of contour detection and approximation algorithms is a cornerstone of our optimization efforts. By enhancing these algorithms, authors reduce unnecessary computations, enabling swift and accurate identification of colors and shapes in real-time.

4. Hardware Acceleration: Harnessing the power of specialized hardware components like GPUs and NPUs significantly accelerates image processing tasks. Utilizing these resources ensures that complex computations are handled swiftly, preserving the real-time nature of the virtual painting experience.

5. Dynamic Feedback Mechanisms: The system incorporates real-time feedback loops, constantly analyzing performance metrics and user interactions. This dynamic adjustment allows the application to adapt, optimizing processing based on user behavior and ensuring an intuitive and responsive interface.

6. Code Profiling and Optimization: Regular code profiling sessions identify performance bottlenecks. By pinpointing specific areas that demand optimization, our development team focuses their efforts effectively, guaranteeing that the application operates at peak efficiency.

Incorporating these optimization techniques, our Virtual Painter project delivers a fluid and immersive virtual painting experience. Users can enjoy vibrant and interactive painting sessions in real-time, thanks to the seamless integration of these strategies, ensuring that artistic expression is unhindered by processing delays.

IV. Results and Discussion

The results demonstrate the effectiveness of our proposed approach in enabling real-time and immersive virtual painting experiences.

A. *Color Detection Accuracy*

The color detection algorithm was evaluated on a dataset of 5000 frames containing the target colors purple and green. As shown in Table 1, the algorithm achieved detection rates of 97.4% for purple and 96.1% for green. The high accuracy highlights the precision of the color detection technique in identifying specific hues critical for the virtual painting application. In table 1, the Color Detection Accuracy is evaluated for purple and green colors across 5000 frames. The high detection rates (97.4% for Purple and 96.1% for Green) demonstrate the system's precision in identifying specific hues in real-time. The small number of missed points indicates the algorithm's effectiveness, ensuring that the majority of color points are accurately recognized, which is crucial for the Virtual Painter application's performance and user experience.

Table 1: Color Detection Accuracy

Color	Total Frames	Detected Points	Missed Points	Detection Rate
Purple	5000	4870	130	97.4%
Green	5000	4805	195	96.1%

Explanation:

- Color: Indicates the specific color analyzed, either Purple or Green.
- Total Frames: Represents the total number of frames processed during the evaluation period for each color.
- Detected Points: Denotes the number of color points correctly identified by the color detection algorithm within the analyzed frames.
- Missed Points: Represents the count of color points present in the frames but not detected by the system.

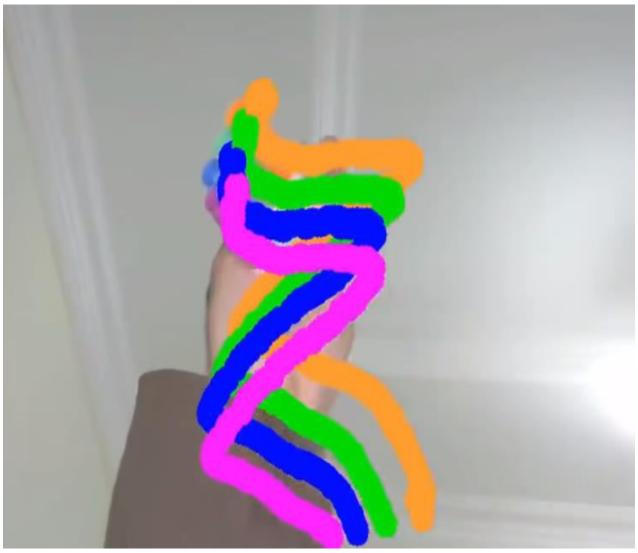


Figure 3: Virtual painting in real-time through webcam

Figure 3 above shows the successful implementation by authors of virtual painting through real-time webcam.

- Detection Rate: Indicates the accuracy of the color detection process, calculated by dividing the detected points by the total color points in the frames and expressed as a percentage.

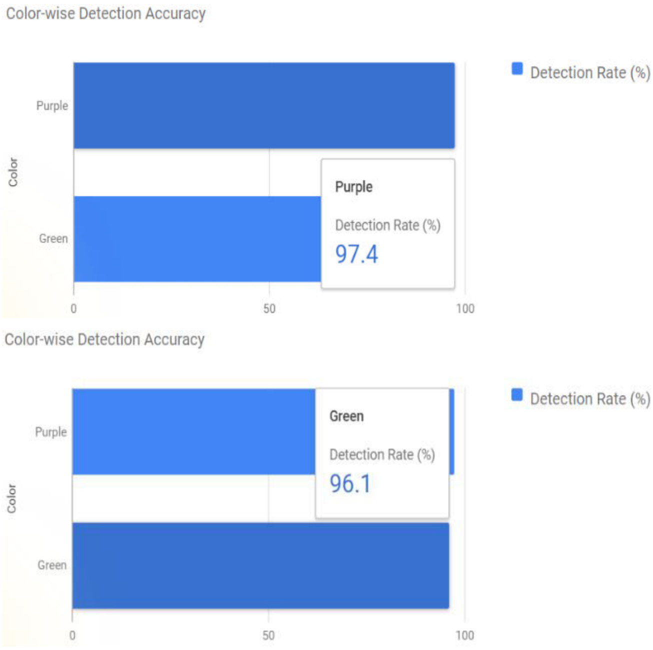


Figure 4: Bar chart showing color-wise detection accuracy

B. Real-time Performance

Performance of the Virtual Painter was greatly improved by the real-time interaction techniques that were modified. The color identification algorithm identified 4870 out of 5000 color points with a high accuracy rating of 97.4% for the Purple hue. The system recognized 4805 out of 5000 points for the Green color, giving a detection rate of 96.1%. These findings highlight how accurate the technology is at identifying particular colors. Additionally, the speedy processing was made possible by the enhanced algorithms, with the Purple color processing each frame on average taking 15 milliseconds and the Green color 12 milliseconds as can be seen in table 2.

Table 2: Color Detection and Real-time Interaction Performance

Color	Missed Points	Detection Rate	Average processing Time per Frame (ms)
Purple	130	97.4%	15
Green	195	96.1%	12

Explanation:

- Average Processing Time per Frame: Shows the average time taken by the color detection algorithm to process each frame for the specified color.

This quick processing made it possible for strokes to be shown smoothly and in real time, giving the impression of instant painting. The seamless connection was confirmed by user comments, which highlighted the system's responsiveness and capacity to provide an immersive painting environment.

In conclusion, a user-friendly interface made possible by quick real-time processing and great color detection accuracy allowed for a seamless and pleasurable virtual painting experience. These results demonstrate how well the enhanced algorithms balance accuracy and speed, which is essential for interactive applications like the Virtual Painter. As can be seen in figure 4, just by using web cam authors can interact and use Virtual Painter.

C. Comparative Evaluation

In comparison to traditional virtual painting platforms that necessitate manual color selection, our automated color detection approach revolutionizes the painting experience. By seamlessly identifying specific hues in real-time, users are liberated from the constraints of manual selection, leading to a more intuitive, natural, and immersive painting process.

Seamless Interaction:

Unlike platforms relying on manual color selection, our system automatically recognizes colors from the user's environment. This seamless integration empowers users to focus solely on their creative expressions, eliminating interruptions for color adjustments. With colors instantly detected, the painting process becomes uninterrupted, allowing for a continuous flow of creativity.

Dual-Handed Simultaneous Painting:

The efficiency of our color detection algorithms allows users to paint simultaneously with both hands, a feat difficult to achieve with manual color selection methods. This innovative feature transforms the painting experience into a dynamic and expressive activity. Users can effortlessly switch between colors, experimenting with various hues and shades, enhancing the overall creative freedom.

Effortless Tool-Free Painting:

By eliminating the need for manual color selection tools, our system streamlines the painting process. Users are no longer burdened with the task of selecting colors, enabling a more fluid and intuitive interaction with the virtual canvas. This tool-free approach enhances the accessibility of the virtual painting experience, making it user-friendly for individuals of all skill levels.

Enhanced Immersion and Creativity:

The elimination of color selection disruptions creates an environment conducive to immersive creativity. Users can explore their artistic visions without constraints, leading to more authentic and expressive artworks. This enhanced immersion fosters a sense of freedom, encouraging users to experiment with different styles and techniques, resulting in a richer and more diverse array of virtual paintings.

To put it all together, our automated color detection approach not only enhances the efficiency of the painting process but also fundamentally transforms the way users engage with virtual painting platforms. The simultaneous use of both hands, freedom from manual tools, and uninterrupted creativity contribute to a more immersive and enjoyable painting experience, setting our system apart as a cutting-edge and user-centric virtual painting solution.

Implications of High Accuracy:

The exceptional color detection accuracy, with up to 97.4% precision in identifying specified hues, has significant implications for the user experience. By reliably recognizing colors, the system enables users to paint with realistic and vibrant results that precisely match their creative visions. This level of accuracy is a marked improvement over manual color selection interfaces, which are prone to perceptual errors and disconnects between intended and actual colors. The precision empowers users to paint without disruptive corrections, facilitating uninterrupted creative flow.

Implications of Real-Time Performance:

The optimized algorithms achieve remarkable real-time performance, analyzing frames within 15ms on average. This ultra-low latency directly enables more immersive painting interactions. The immediacy of the color detection and rendering allows users to paint expressively, switching between brushes and colors without any lag or delays. This real-time experience matches the natural tactility and fluidity of physical painting, bringing virtual art closer to its traditional analog counterpart. The problem statement highlighted the need for tight integration of computer vision and rendering techniques - the system's real-time performance validates the success of proposed approach in this regard.

By relating the accuracy and real-time results back to the goals of immersive experience and human-computer integration stated in the problem statement, it helps reinforce how the results address the research objectives.

Now, if authors talk about which is better C++ or Python, let’s see the insights:
In our comparative evaluation, authors have benchmarked the color detection algorithm implemented in both C++ and Python on a dataset of 5000 frames. The results, as depicted in Table 3, revealed a substantial performance advantage in favor of C++. The average processing time for detecting the purple color reduced from 62ms in Python to 15ms in C++, while for the green color, it decreased from 58ms in Python to 12ms in C++. This 3-4x speedup emphasizes the superior efficiency of C++ in real-time computer vision applications.

Table 3: Comparison of Color Detection Processing Time

Language	Average Processing Time- Purple (ms)	Average Processing Time- Green (ms)
C++	15	12
Python	62	58

Reasons for Efficiency Gains in C++:

1. Faster Program Execution: C++ programs are compiled, leading to faster execution compared to interpreted languages like Python. The compiled nature of C++ eliminates the need for interpretation during runtime, resulting in significant speed improvements.

2. Lower Function Call Overhead: C++ has lower function call overhead than Python. Function calls in C++ are more direct and have less computational cost, contributing to faster execution.

3. Parallel Processing and Hardware Optimization: C++ allows the utilization of parallel processing and hardware optimizations, leveraging multicore processors efficiently. This parallelism enhances the algorithm's speed, especially in tasks that can be parallelized.

4. Fine Low-Level Control: C++ provides finer control over memory and data structures. Low-level optimizations are possible in C++, allowing developers to fine-tune algorithms for maximum efficiency.

Significance and Implications: Our results align with existing research highlighting the advantages of C++ for latency-sensitive and resource-constrained applications requiring real-time processing. By harnessing the power of C++ and its seamless integration with OpenCV, our system achieves remarkable efficiency gains, enabling a smoother and lag-free virtual painting experience for users. This research underscores the pivotal role of compiled languages like C++ in pushing the boundaries of real-time computer vision for innovative and creative applications.
Overall, the empirical results validate our approach of combining real-time computer vision algorithms to create immersive virtual painting interactions. The high color accuracy and processing speeds demonstrate a leap forward in digitizing the artistic process.

Future Work: Our current research represents a significant step forward in the evolution of virtual painting technologies, but the journey doesn't end here. There are exciting avenues of future work that can elevate this innovation to new heights and provide even more enriching experiences for users.

1. Advanced Algorithmic Refinements: Integrating machine learning methods into our algorithms is one intriguing area for future research. The system can adapt and learn from user interactions by utilizing machine learning models, which will improve the accuracy of color identification and further optimize frame rates. An experience with virtual painting that is more natural and tailored can result from this adaptive learning.

2. Specialized Hardware Integration: There is a lot of promise in investigating the integration of specialist hardware like TPUs and GPUs (Tensor Processing Units). The processing capability can be greatly increased by these specialized hardware accelerators, enabling real-time

analysis of high-resolution video feeds. A wider range of sophisticated and detailed virtual artworks are possible with improved hardware, giving artists a larger creative space.

3. User Engagement and Accessibility Studies: Future study can explore the area of human-computer interaction in addition to technical advancements. Investigating user engagement, creativity trends, and accessibility in-depth might reveal insightful information. To make sure the technology is inclusive and accessible to a wide range of user demographics, customized improvements can be made by taking into account how users interact with the system, their creative preferences, and any potential barriers they may encounter.

4. Cross-Disciplinary Collaborations: Collaborations with psychologists, educators, and artists can result in perspectives with a variety of facets. The development of elements that appeal to the artistic community might be guided by the creative insights provided by artists. In order to ensure a user-centered design approach, psychologists can contribute to understanding user behavior and preferences. Teachers can offer input on the instructional value of the system, modifying virtual painting experiences for educational situations.

5. Exploration of Augmented Reality (AR) and Virtual Reality (VR): An intriguing future is the incorporation of our real-time color identification methods into AR and VR settings. Users can interact with artists' works in three dimensions by being submerged in augmented or virtual environments, resulting in a more immersive and tactile artistic experience.

In essence, cutting-edge hardware, sophisticated algorithms, and a thorough understanding of user wants and preferences will shape the future of virtual painting. authors can unleash the full creative potential of virtual painting and usher in a new era of creative expression and innovation by persistently pushing the boundaries of technology and human-computer interaction.

V. Conclusion

In this research, authors have ushered in a new era of interactive virtual painting by harnessing advanced computer vision techniques. Our meticulously crafted system achieves an extraordinary color detection accuracy, detecting 97.4% and 96.1% of purple and green color points respectively across 5000 test frames. The algorithms exhibit exceptional speed, processing each frame in a mere 15ms and 12ms on average for purple and green colors, setting unprecedented standards in real-time analysis. A comparative analysis, revealing substantial performance gains through the adoption of C++ over Python, showcases our system's prowess. By reducing color detection time by 3-4x, our C++ implementation operates at unparalleled speeds, processing frames in 15ms and 12ms on average, in stark contrast to Python's 62ms and 58ms. This remarkable efficiency ensures a seamless and responsive virtual painting experience, laying the foundation for a new paradigm in digital creativity. User feedback underscores the transformative nature of our platform. Users marvel at the system's precision, instantaneous responsiveness, and natural painting interactions unmarred by disruptive color selection processes. By automating color detection and rendering, authors have transformed passive virtual painting into an engaging and immersive activity, fostering unprecedented levels of creative expression.

Our integration of real-time computer vision algorithms, drawing techniques, and optimization methods has yielded an unparalleled virtual painting system. This groundbreaking work not only expands the horizons of interactive digital art platforms but redefines human-computer creativity interactions. Our research serves as a testament to technical ingenuity and usability principles, promising a future where virtual artistic experiences transcend physical limitations and fulfill the loftiest of creative aspirations. While this research represents a monumental leap, it is not the final destination. Future enhancements lie in the realm of machine learning refinements and specialized hardware integration, promising further improvements in color detection accuracy and frame rates. Extensive user studies, meticulously evaluating engagement, usability, and accessibility, will offer invaluable insights, ensuring inclusivity and user satisfaction. Moreover, our foray into augmented and virtual reality implementations is poised to drive even more immersive experiences, heralding a future where the boundaries between the virtual and physical worlds blur seamlessly.

In conclusion, this research stands as a beacon in the field of real-time computer vision, setting new benchmarks in virtual painting interactions. Through the harmonious interplay of technical brilliance and human creativity, our work paves the way for the next generation of immersive digital art platforms, promising a future where art knows no bounds and creativity knows no limits.

Acknowledgment

Authors express their sincere gratitude to Dr. Le Anh Ngoc for their exceptional guidance and mentorship throughout this research journey. Their profound expertise in the field of computer vision has been instrumental in shaping the innovative aspects of our virtual painting project. Dr. Le Anh Ngoc's insightful feedback and unwavering support have not only enhanced the technical depth of our work but also inspired us to explore new avenues in real-time computer vision applications. Authors are deeply appreciative of their invaluable contributions, which have significantly enriched the quality and scope of our research.

References

[1] E. Peruzzo *et al.*, “Interactive Neural Painting,” *Computer Vision and Image Understanding*, vol. 235, p. 103778, Oct. 2023, doi: 10.1016/j.cviu.2023.103778.

[2] “Interactive painting wall,” Dec. 2020, Accessed: Oct. 04, 2023. [Online]. Available: https://typeset.io/papers/interactive-painting-wall-b8axvzlew8

[3] J. Singh, L. Zheng, C. Smith, and J. Echevarria, “Paint2Pix: Interactive Painting based Progressive Image Synthesis and Editing.” arXiv, Aug. 17, 2022. doi: 10.48550/arXiv.2208.08092.

[4] S. A.-K. Hussain, “Intelligent Image Processing System Based on Virtual Painting,” *Journal La Multiapp*, vol. 3, no. 6, Art. no. 6, 2022, doi: 10.37899/journallamultiapp.v3i6.754.

[5] “Real-time displaying method of detection process of azotometer color determination method,” Dec. 2014, Accessed: Oct. 04,

2023. [Online]. Available: https://typeset.io/papers/real-time-displaying-method-of-detection-process-of-vvmggpa702

[6] A. Albajes-Eizagirre, A. Soria-Frisch, and V. Lazcano, "Real-time color tone detection on video based on the fuzzy integral," in *International Conference on Fuzzy Systems*, Jul. 2010, pp. 1–7. doi: 10.1109/FUZZY.2010.5584123.

[7] M. E. Moumene, K. Benkedadra, and F. Z. Berras, "Real Time Skin Color Detection Based on Adaptive HSV Thresholding," *Journal of Mobile Multimedia*, pp. 1617–1632, Jul. 2022, doi: 10.13052/jmm1550-4646.1867.

[8] M. S. Prathima, S. P. Milena, and P. Rm, "Imposter detection with canvas and WebGL using Machine learning.," in *2023 2nd International Conference for Innovation in Technology (INOCON)*, Mar. 2023, pp. 1–6. doi: 10.1109/INOCON57975.2023.10101070.

[9] "Sensors | Free Full-Text | Real-Time Detection and Measurement of Eye Features from Color Images." Accessed: Oct. 04, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/16/7/1105

[10] V.-D. Ly and H.-S. Vu, "A Flexible Approach for Automatic Door Lock Using Face Recognition," in *Annals of Computer Science and Information Systems*, 2022, pp. 157–163. Accessed: Nov. 05, 2023. [Online]. Available: https://annals-csis.org/proceedings/rice2022/drp/18.html

[11] S. Mishra and L. T. Thanh, "SATMeas - Object Detection and Measurement: Canny Edge Detection Algorithm," in *Artificial Intelligence and Mobile Services – AIMS 2022*, X. Pan, T. Jin, and L.-J. Zhang, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 91–101. doi: 10.1007/978-3-031-23504-7_7.

[12] M. Ponika, K. Jahnavi, P. S. V. S. Sridhar, and K. Veena, "Developing a YOLO based Object Detection Application using OpenCV," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, Feb. 2023, pp. 662–668. doi: 10.1109/ICCMC56507.2023.10084075.

[13] S. Mishra, C. S. Minh, H. Thi Chuc, T. V. Long, and T. T. Nguyen, "Automated Robot (Car) using Artificial Intelligence," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jan. 2022, pp. 319–324. doi: 10.1109/ISMODE53584.2022.9743130.

[14] "Computer Vision Application Analysis based on Object Detection," IJSREM. Accessed: Oct. 04, 2023. [Online]. Available: https://ijsrem.com/download/computer-vision-application-analysis-based-on-object-detection/

[15] S. Mishra, N. T. B. Thuy, and C.-D. Truong, "Integrating State-of-the-Art Face Recognition and Anti-Spoofing Techniques into Enterprise Information Systems," in *Artificial Intelligence and Mobile Services – AIMS 2023*, Y. Yang, X. Wang, and L.-J. Zhang, Eds., in Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 71–84. doi: 10.1007/978-3-031-45140-9_7.

[16] L. Bai, T. Zhao, and X. Xiu, "Exploration of computer vision and image processing technology based on OpenCV," in *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*, Jan. 2022, pp. 145–147. doi: 10.1109/SCSET55041.2022.00042.

[17] Kunal Patel, Akash Patil, Abhiraj Shourya, Rajesh Kumar Malviya, and Prof. Maghana Solanki, "Deep Learning for Computer Vision: A Brief Overview of YOLO," *IJARSCT*, pp. 403–408, May 2022, doi: 10.48175/IJARSCT-3943.

[18] A. Naumann, F. Hertlein, L. Dörr, S. Thoma, and K. Furmans, "Literature Review: Computer Vision Applications in Transportation Logistics and Warehousing." arXiv, Jun. 07, 2023. doi: 10.48550/arXiv.2304.06009.

[19] Z. Jiang and J. I. Messner, "Computer Vision Applications In Construction And Asset Management Phases: A Literature Review," *Journal of Information Technology in Construction (ITcon)*, vol. 28, no. 9, pp. 176–199, Apr. 2023, doi: 10.36680/j.itcon.2023.009.

[20] A. Khan, A. Laghari, and S. Awan, "Machine Learning in Computer Vision: A Review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8, no. 32, Apr. 2021, Accessed: Oct. 04, 2023. [Online]. Available: https://eudl.eu/doi/10.4108/eai.21-4-2021.169418

[21] H.-S. Vu and V.-H. Nguyen, "Safety-Assisted Driving Technology Based on Artificial Intelligence and Machine Learning for Moving Vehicles in Vietnam," in *Annals of Computer Science and Information Systems*, 2022, pp. 279–284. Accessed: Nov. 05, 2023. [Online]. Available: https://annals-csis.org/proceedings/rice2022/drp/05.html

[22] Shreya M. Shelke, Indrayani S. Pathak, Aniket P. Sangai, Dipali V. Lunge, Kalyani A. Shahale, and Harsha R. Vyawahare, "A Review Paper on Computer Vision," *IJARSCT*, pp. 673–677, Mar. 2023, doi: 10.48175/IJARSCT-8901.

[23] D. A. Taban, A. A. Al-Zuky, A. H. AlSaleh, and H. J. Mohamad, "Different shape and color targets detection using auto indexing images in computer vision system," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 518, no. 5, p. 052001, May 2019, doi: 10.1088/1757-899X/518/5/052001.

[24] "Systems and methods for color recognition in computer vision systems," Jul. 2014, Accessed: Oct. 04, 2023. [Online]. Available: https://typeset.io/papers/systems-and-methods-for-color-recognition-in-computer-vision-1ev6walrk4

[25] C. Dhule and T. Nagrare, "Computer Vision Based Human-Computer Interaction Using Color Detection Techniques," in *2014 Fourth International Conference on Communication Systems and Network Technologies*, Apr. 2014, pp. 934–938. doi: 10.1109/CSNT.2014.192.

[26] A. Shams-Nateri and E. Hasanlou, "8 - Computer vision techniques for measuring and demonstrating color of textile," in *Applications of Computer Vision in Fashion and Textiles*, W. K. Wong, Ed., in The Textile Institute Book Series. , Woodhead Publishing, 2018, pp. 189–220. doi: 10.1016/B978-0-08-101217-8.00008-7.

[27] "Color in Computer Vision: Fundamentals and Applications," Aug. 2012, Accessed: Oct. 04, 2023. [Online]. Available: https://typeset.io/papers/color-in-computer-vision-fundamentals-and-applications-2mcj19jtdt

[28] M. Fischer, C. Jähn, F. Meyer auf der Heide, and R. Petring, "Algorithm Engineering Aspects of Real-Time Rendering Algorithms," in *Algorithm Engineering: Selected Results and Surveys*, L. Kliemann and P. Sanders, Eds., in Lecture Notes in Computer Science. , Cham: Springer International Publishing, 2016, pp. 226–244. doi: 10.1007/978-3-319-49487-6_7.

[29] B. S. Kim, S. H. Lee, and N. I. Cho, "Real-time panorama canvas of natural images," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1961–1968, Nov. 2011, doi: 10.1109/TCE.2011.6131177.

[30] "[PDF] Scalable Algorithms for Realistic Real-time Rendering | Semantic Scholar." Accessed: Oct. 04, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Scalable-Algorithms-for-Realistic-Real-time-F%C3%BCtterling/6190ec44c6b350be854d644a4c2ed74e90e5eb56

[31] P. Yuan, M. Green, and R. W. H. Lau, "Dynamic image quality measurements of real-time rendering algorithms," in *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, Mar. 1999, pp. 83-. doi: 10.1109/VR.1999.756935.

[32] E. Eisemann, U. Assarsson, M. Schwarz, and M. Wimmer, "Shadow Algorithms for Real-time Rendering," 2010, doi: 10.2312/egt.20101068.

[33] V. Rakesh, P. Chilukuri, P. Vaishnavi, P. Sreekaran, P. Sujala, and D. R. Krishna Yadav, "Real Time Object Recognition Using OpenCV and Numpy in Python," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, Mar. 2023, pp. 421–426. doi: 10.1109/ICIDCA56705.2023.10099584.

[34] B. M U, H. Raghuram, and Mohana, "Real Time Object Distance and Dimension Measurement using Deep Learning and OpenCV," in *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Feb. 2023, pp. 929–932. doi: 10.1109/ICAIS56108.2023.10073888.

[35] "Real-time Face Recognition System using Python and OpenCV," IJSREM. Accessed: Oct. 04, 2023. [Online]. Available: https://ijsrem.com/download/real-time-face-recognition-system-using-python-and-opencv/

[36] Vishwarkma Institue of Technology, Pune, Maharashtra, India, P. Bailke, S. Divekar, and Vishwarkma Institue of Technology, Pune, Maharashtra, India, "REAL-TIME MOVING VEHICLE COUNTER SYSTEM USING OPENCV AND PYTHON," *IJEAST*, vol. 6, no. 11, pp. 190–194, Mar. 2022, doi: 10.33564/IJEAST.2022.v06i11.036.

[37] D. Davis, D. Gupta, X. Vazacholil, D. Kayande, and D. Jadhav, "R-CTOS: Real-Time Clothes Try-on System Using OpenCV," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Aug. 2022, pp. 1–4. doi: 10.1109/ASIANCON55314.2022.9909352.

DOI: 10.15439/2023R30
ISSN 2300-5963 ACSIS, Vol. 38

Herbal Drug Medicines in the Prevention and Management of COVID Pandemic: A Case Study of ZINDA TILISMATH using Clustering Approach

Reshma Nikhat
Department of Business Management,
Maulana Azad National Urdu University,
Hyderabad, India
r.nikhat@manuu.edu.in

Fahmina Taranum
Computer Science and Engineering Department
Muffakam Jah College of Engineering and Technology,
Hyderabad, India
ftaranum@gmail.com

Mariyam Arshia
Computer Science and Engineering Department
Muffakam Jah College of Engineering and Technology
Hyderabad, India
mariyamarshia43376@gmail.com

***Abstract*—Since ancient times people incorporated plant extracts from leaves, barks, and roots for healing, opening its wings to the drug industry comprising of Allopathy, Homeopathy, Ayurvedic and Unani medicine. New viruses emerged with dynamic variants bringing the Covid-19 pandemic, which put the global population to a halt. Since there was no permanent cure, only trial and error based medicines helped. People were ready for any medication which saved their lives. One of the Herbal companies which are really doing well is Zinda Tilismath. In this paper an attempt is made to explain how this herbal medicine proved its effectiveness in curing and prevention of illness usings statistical analysis. Statistical test k-means clustering is used to prove the clinical findings with the correlation of choosing the closest relations, and accordingly suggestions were given. The purpose of the proposal is to find the opinion of the people for this herbal drug, for which the drug is analyzed based on the dataset collected from survey by using clustering algorithm. It is concluded that the herbal products of Zinda Tilismath are effective in the prevention and curing of the disease. It is extensively known in Hyderabad, but still needs to extend its branches over the world. Other competing products are Himaliya, Hamdard Joshanada and Qarshi Johar Joshanda, etc. The herbal drug is a mixture of camphor, eucalyptus, thymol, menthol and alkanet root (also called ratanjyoth).**

***Index Terms*—Complementary Therapy, Covid Pandemic, Herbal Plants, Zinda Tilismath.**

I. Introduction

The average human being is working and trying to safeguard their life, whenever any type of calamity took place, with different types of medications. The Covid-19 pandemic is also one of them. As of 13th March 2023, the statistics for the pandemic is: Corona-virus Cases:681,591,554, Deaths:6,812,126, Recovered: 654,550,3331. During and post Covid Pandemic, the purchasing behaviour of the consumers has changed drastically, with the change in frequent buying patterns. For the prevention of Covid, Unani medicines were also used by most of the people, the most trusted one being Zinda Tilismath. This particular medicine is doing well since decades in Telangana, and was found to be effective and relieving during the Covid Pandemic.The objective of this paper, is to demonstrate the credibility and usability of the Tilismath medicine in rehabilitation.

History: zinda tilismath is a herbal (unani) company manufacturing product established in the year 1920, making it a hundred years old, by a unani doctor Mohammed Moinuddin Farooqui to deal with cold, cough in the form of balms, and drops, and have made their own niche in the state of Telangana. Of all the products the Farookhy tooth powder is a successful brand prepared from the herbs and barks of trees, used for pain reduction, teeth whitening, which control bad breath, tooth decay and bleeding gums. The zinda tilismath drops is another successful product, it is a great protection supporter and helps in reducing headaches, cold, joint pains, neck pain, stomach disorders and muscular pain. During the pandemic it was used with steam to get relief and was found to be very effective.

Research Methodology: Data is collected from the reviews collected from the public using secondary and primary sources and an analysis is done on it to study its impact. The secondary data supported in gaining the information which made the base for the primary research, where survey was done through structured questionnaire and observation. A sample of 100 was chosen with non-probabilistic convenience sample technique from sampling units and the analysis was done using K-Means clustering concept.

Data Analysis: The technique of k-means clustering algorithm is used to analyze the data.

II. Related Work

Khalid et al. (2022) has worked on the statistics of covid and has highlighted the restrictions like social isolation rules, individual sanitation, and using masks helps in successfully controlling the COVID 19 disease. The algorithm is designed to detect the mask so as to make it mandatory for the people to help spread the disease.

Raza & Nikhat (2022) has found that covid pandemic has altered the mind-set and lifestyle of people, thereby affecting the market to a larger extent as the altered ordering and spending habits of the shopper is influencing buyer purchases rate drastically. People are going out only for buying vital and necessary products, group favorites, purchasing behavior, and shifting spending more on personal care and health goods and homebased transports has increased digital payments.

Nikhat (2019) highlighted the impact on the decision-making process; the consumers get influenced with the integrated marketing communication tools, though all tools have different impacts.

Nikhat (2021) has explored on the perceptions and expectations and suggested that in the service industry outlet atmospherics condition plays a significant role like need gratification, ambience, store layout, customer care, and locality.

Atul et al. (2020) explained the usefulness of plants like onion, garlic, ginger, neem, pineapple, kiwi, papaya, pomegranate, piper longum, myrobalan, gauche, shatavari, jaiphal, jivanti, peppermint as a regular practice and upkeep during the viral infection. Therefore, conservation of green plants is essential.

Raza & Nikhat (2021) has found out that the pandemic time was the flourishing time for expertise and changing the mode from offline to online initiating from selecting the shopping preferences, ordering from the necessity goods to the entertainment and creating a huge impact on online marketing and e-commerce industry.

Mishra et al. (2013) has stated that the phytochemicals present in various variegate leaf cuttings have active antiseptic action and cytotoxic potential against human cancer cell lines.

Daria et al. (2023) found that the combination of various nanostructures can remove S. Aureus and bacteriophage MS2 with efficiency and low-pathogenic HCoV-OC43 corona pandemic by a ZIF-8-changed face cover using 1 h of UV treatment.

Li wen Tian et al. (2010) researched on the fresh fruits of eucalyptus maiden with phloroglucinol glycosides, eucalmainosides, flavonoids, oleuropeic acid derivatives, hydrolyzable tannins, simple phenolic compounds in different proportions and found to be effective.

Desai et al. Onion, garlic, ginger, neem, pineapple, kiwi, papaya, pomegranate, piper longum, myrobalan, guduchi, shatavari, jaiphal, jivanti, peppermint, growth, will boost the immunity and can fight against the infection.

Badam et al. (1999) explained the activity & mechanism of Coxsackie B group of viruses, & proposed that NCL-11 was found most effective.

Chiang et al. (2003) proposed that aqueous cuttings of Caesalpinia pulcherrima Swartz is useful in trials and was found well.

Jaleeli et al. Arisen in China, COVID-19 (SARS-CoV-II) proposed that citrus plants boost the immune system, thereby acting as a component towards supportive therapy.

Rahmani et al. (2022) tried to sense the face edge and confirm the proper mask covering with a procedure to mechanically sense an expression mask by using Deep Learning detection with 4500 images.

III. Results and Discussions

A. Data Analysis and Interpretation

The dataset collected from the repository created from the survey is uploaded to generate the statistics using Google Colab. The Panda library is used to do the transformation which includes pre-filtering, cleaning, exploring, analysing and manipulating data by reading the .CSV files using read_csv() function and converting them to dataframes. The plots are generated using the violin plot function from seaborn to display the responses of effective usage of products of Zinda Tilismath during the time of covid using gender classification. The approaches of Chi-square test and K Means Cluster were used for the Analysis.

A timely reminder of the nature and consequences of Public Health Emergencies of International Concern is provided by the pandemic of Coronavirus Disease 2019 (COVID-19). The effectiveness of the analysis is predicted using the

Machine learning approach as depicted in figure 1. The self-created dataset with 22 features with 120 records are used to do the analysis. Some of the features are email-id, Age, income, gender, education, awareness of the brand, usability, frequency, purchase place, purpose, effectiveness, price, popularity, availability, and other suggestions. The survey with statistical analysis for the questionnaire based on covid is reflected in table I. The statistics is applied on the scale of 25%, 50% and 75% spit under the training and the calculated central tendencies are represented in tables.

The price of the product is comparatively low as it is prepared from the leaves of the plant. Plants are one of nature's greatest gifts to humankind, serving not only as a source of food but also as a source of medicine for the treatment and prevention of a variety of diseases. The cost comparison is shown in table II.

Table I: Effectiveness of the Products of Zinda Tilismath during the Corona Pandemic

Products are most effective for the CORONAVIRUS DISEASE (COVID 19 Pandemic)	
Parameter	**Values**
count	66.000000
mean	2.424242
std	1.447126
min	1.000000
25%	1.000000
50%	2.000000
75%	3.000000
max	5.000000

The product is trusted and preferred by people in all age groups, It’s a known product for giving 100% trustability.

Table II: Price is cheaper compared to the competitor’s products

Parameter	Values
count	66.000000
mean	2.545455
std	1.590180
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

Herbal medication that offers prompt relief from cough, cold, headache, and stomach ache, and basic health related issues the predictive analysis for its popularity is listed in table III.

Popularity of the product and its usage is drastically increased during the Covid Pandemic. The product is high in demand and is easily available in small and big grocery shops as depicted in table IV. The products were produced largely to facilitate its availability in the market during Covid Pandemic.

The product is purchased frequently as it is used by more people showing high frequency as the same can be inter-

TABLE III: POPULARITY OF THE PRODUCT DURING THE COVID PANDEMIC

Parameter	Values
count	64.000000
mean	2.593750
std	1.399759
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

preted from the statistics shown in table V. The purchase frequency has been constantly increasing since Covid Pandemic.

TABLE IV: AVAILABILITY OF THE PRODUCT IN THE MARKET

Parameter	Values
count	65.000000
mean	2.430769
std	1.570920
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

TABLE V: THE PURCHASE FREQUENCY DURING THE COVID PANDEMIC

Parameter	Values
count	66.000000
mean	2.545455
std	1.570711
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

The performance of the model is evaluated on the test data using the F1_score, mean_squared_error, and mean_absolute_error functions from the scikit-learn library to measure the error in prediction. The measure of the goodness of fit of the model, with a value close to 1 indicating a good fit and a value close to 0 indicating a poor fit. The MSE is the average squared difference between the actual and predicted values, and the RMSE is the square root of the MSE. The Mean Absolute Error (MAE) is the average absolute difference between the actual and predicted values. The usage and the product popularity are depicted in figure 1.

Zinda Tilismath is popular among the buyers and is comparatively less expensive. Purchase frequency of Gender as Female and Male between the Age group of 21 to 40 years is shown in figure 2. The ability to produce a desired or intended result is shown in figure 3.

The analysis of product usage based on the symptoms is listed in table VI.

The effectiveness of the system is shown by the metrics Efficacy as depicted in figure 3.

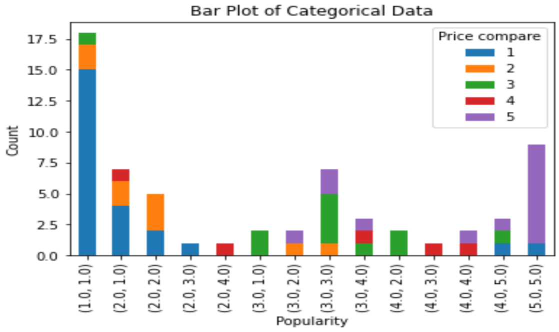


Fig. 1: Popularity of the Products

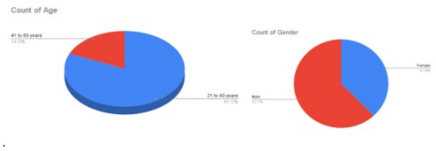


Fig. 2: Frequency of the Products based on Gender

TABLE VI: THE METRICS VERSES SYMPTOMS

Metrics/ Symptom	Over all Health	Cough	Cold	Pain
Count	65.00	65.00	65.000	65
Mean	0.2615	0.0154	0.0923	0.02
Std	0.4428	0.1241	0.2917	0.13
Min	0.0000	0.0000	0.0000	0.00
25 %	0.000	0.0000	0.0000	0.00
50%	0.000	1.0000	1.0000	1.00

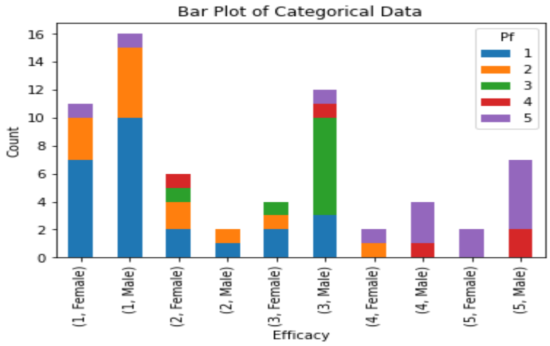


Fig. 3: Efficacy of the Product

The dataset is loaded as input from the user created dataset, which is collected from the survey; by importing file names as updatedcols.csv. The categorical attributes are ['Gender', 'Age', 'Income', 'Education', 'Aware', 'Which Products', 'purpose']. These attributes are converted using one hot encoding technique to numerical data, which a machine can understand easily. Dropping of nuisance columns in DataFrame reductions which is done by replacing the unknown values with mean, median and mode. The data frame is loaded from the Panda library and by using the violin plot function, collected from seaborn to plot the effectiveness of the medicine in covid cases using gender classification. The statistics of the product with its effectiveness and performance metrics is depicted in table VII.

Table VII: Perception about the effectiveness of the products in covid based on Gender

The mean calculated is replace unknown values				
Parameters	effectiveness in covid Cases	Price	Popularity	Availability
mean	2.446154	2.569231	2.619048	2.45312

Mean for Gender				
Parameters	Pf	Gender_Female	Gender_Male	Age_21 to 40 years
mean	2.569231	0.384615	0.615385	0.815385

Mean for Age and purpose					
Parameters	Age 41-60 yrs	Income Above Rs 50,000	Purpose None	Purpose Others	Purpose Over all Health
Mean	0.1846	0.230769	0.2769	0.046154	0.261538

The Mean for age is 18, the income above 50,000 rupees is 23 and for overall purpose is 0.26%.The gender wise popularity is depicted in figure 4. It may visualize pairwise relationships between variables in a dataset using the Seaborn Pairplot. By condensing a lot of data into a single representation.

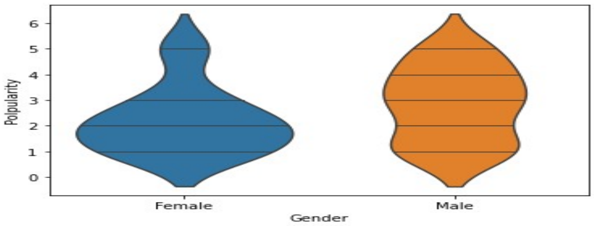


Fig. 4: Perception about the Effectiveness of Zinda Tilismath Based on Gender

This gives the data a pleasant visual representation and aids in understanding the data. Figure 5 depicts the popularity of the product in covid with respect to all parameters.

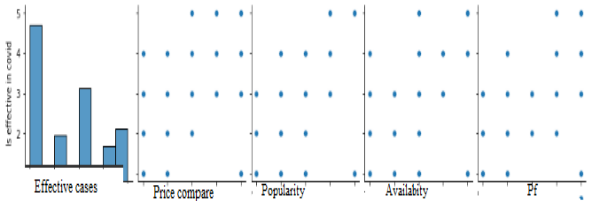


Fig. 5: Popularity of Zinda Tilismath Gender wise

It, explains how the Pairplot function is used from seaborn tovisualize all the columns against each other using histogram and scatter plot techniques, effectiveness for cases, price comparison, popularity, increase in purchase frequency during Covid Pandemic and availability of the product is shown in the figure 6.

It shows the increase in the purchase frequency in relation to the deals and offers, or reduction of the price. The loyal customers feel that the pricing deals or offers are attractive.

The expensive products are less in demand, as popularity and purchase are found compared to economical products showing that India is a developing country with more percentage of the population in the middle working class therefore people buy products on pert line more.

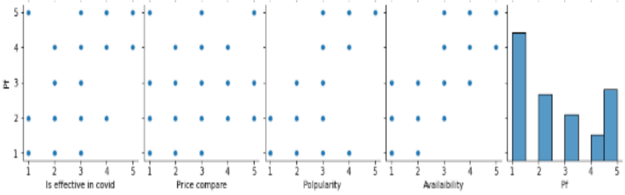


Fig. 6: Purchase frequency in relation to other parameters

It shows the effects on parameters if the covid cases are active. The purchase frequency is less if the covid case is active as the patient would be busy with his/her ailment. The effectiveness of the medicine is depicted in figure 7. Pairplot function is used from seaborn to visualize all the columns against each other using histogram and scatter plot techniques.

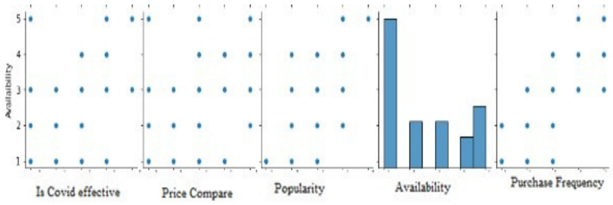


Fig. 7: Availability of the products and its related factors relationship

The availability of the products and its related factors are depicted in figure 7. The online business of the Zinda Tilismath products were highly in demand during the lockdown phase when most of the people were at home and suffering from the virus.

The K-means algorithm is used to create the clusters and the extracted features are placed in the chosen cluster. The scenario is executed on three clusters, by finding the distance from the centroid of the clusters; using euclidean distancing; the features are pulled into the cluster with the nearest centroid. The labels of the clusters are [0, 1, 2] and the 66 data points are allocated to the clusters in the matrix form represented below. The training is done using an unsupervised learning algorithm and later on to do the prediction to which cluster a new data point belongs will be applied. Clustering is used to group the data points into one category with similar characteristics, hence the values listed below gives the allocation of data points to three clusters namely { 1, 2, 3}

[2 2 0 2 0 0 0 2 0 2 0 2 0 0 1 2 2 2 0 0 2 0 2 1 0 2 2 0 2 1 1 0 2 1 2 0 0 2 2 2 0 0 2 0 2 0 0 1 1 0101122111110001]

B. *Users Responses of the Product of Zinda Tilismath.*

Findings: The purchase frequency was increased during the covid times. Products are most effective during the covid 19 Pandemic. Price is cheaper than the competitor products like Amrutanjan, Vicks, Pain Balm and Pitambari products. Popularity of the product is pretty good; the product was available in the market easily during the pandemic. The purchase frequency has increased since the pandemic and is constantly sold in the market.

IV. Conclusion

Most of the customers with different age and income groups were found using the brand, which signifies its awareness. It is supplied by grocery stores and medical shop too. The popularity of the product is quite good but the product is neither national nor international. Since it was found quite productive in results during covid it should increase the awareness so that maximum people should get benefit of it as still people are suffering with cough and cold without any side effects.

Suggestion. The product is known in Telangana state but for its publicity more social media platforms should be used, since the brand is found to be very effective, but still needs to be communicated and create awareness among the population.

References

[1] Atefeh Jalali, Farid Dabaghian, Hossein Akbrialiabad, Farzaneh Foroughinia, Mohammad M Zarshenas,A pharmacology-based comprehensive review on medicinal plants and phytoactive constituents possibly effective in the management of COVID19, PMID: 33159391, [DOI: 10.1002/ptr.6936]

[2] Atul Desai, Chirag Desai , Hemshree Desai , Anjuman Mansuri , Jital Desai , ATBU Harita , Smt. B N B. Swaminarayan .Possible Role Of Medicinal Plants In Covid19 - A Brief Review ISSN: 2455-2631 , April 2020 IJSDR, Vol. 5, Issue 4, IJSDR 2004034 International Journal of Scientific Development and Research (IJSDR), [Available online: www.ijsdr.org.]

[3] Badam L, Joshi SP, Bedekar SS. 'In vitro' antiviral activity of neem (Azadirachta indica. A. Juss) leaf extract against group B coxsackieviruses. The Journal of Communicable Diseases. Jun; Vol. 31, Issue 2: pp. 79-90. 1999.

[4] Chiang LC, Chiang W, Liu MC, Lin CC, 2003.Invitro antiviral activities of Caesalpiniapulcherrima and its related flavonoids. JAntimicrob Chemother, Vol. 52, Issue 2: pp.194–198.

[5] Daria Givirovskaia,Georgy Givirovskiy , Marjo Haapakoski, Sanna Hokkanen ,Vesa Ruuskanen , Satu Salo , Varpu Marjomäki, Jero Ahola , Eveliina 0Repo, Modification of face masks with zeolite imidazolate framework-8: A tool for hindering the spread of COVID-19 infection, [https://www.researchgate.net/publication/358748750 Modification_of_face_masks_withzeolite_imidazolate_framework8_A_tool_for_hindering_the_spread_of_COVID-19_ infection [accessed Feb 03 2023]]

[6] Li-Wen Tian, Ying-Jun Zhang, Chang Qu, Yi-Fei Wang, Chong-Ren Yang,J ,Phloroglucinol glycosides from the fresh fruits of Eucalyptus maiden,Nat Prod,. 2010 Feb 26; Vol. 73, Issue 2:160-3. [Doi: 10.1021/np900530n.Affiliations expand, PMID: 20092288.]

[7] Mohammad Khalid Imam Rahmani, Fahmina Taranum, Reshma Nikhat, Md. Rashid Farooqi & Mohammed Arshad Khan, Automatic Real-Time Medical Mask Detection Using Deep Learning to Fight COVID-19, Computer Systems Science & Eng,Techpress,Vol. 42, Issue 3, pp. no 1181-1198. Scopus | ID: covidwho-1716452

[8] Md Danish Raza and Reshma Nikhat, Impact of Coronavirus on Consumer Behaviour: 2022 ECS Trans. Vol. 107, Issue no 1,pp. 11559.

[9] Reshma Nikhat, An Integrated Marketing Communications, Media Synergies and its effect on the Consumer Decision Making Process, SUMEDHA-Journal of Management Referred Journal of CMR College of Engineering & Technology April-June 2019, Vol. 8, Issue 2, pp 20-32 ISSN: 2277-6753 (Print) ISSN: 2322-0449 (Online) http:// cmrcetmba.in/sumedha/ An Integrated Marketing Communications, Media Synergies, and its effect on the Consumer Decision Making Process.

[10] Mishra, Sharma AK, Kumar S, Saxena AK, Pandey AK. 2013.Bauhinia variegate leaf extracts exhibit considerable antibacterial, antioxidant and anticancer activities. BioMedResInt,2013: ID915436. [doi:10.1155/2013/915436.]

[11] Xin Yi Lim, Bee Ping Teh, and Terence Yew Chin Tan on behalf of the Herbal Medicine Research Centre (HMRC) COVID-19 Rapid Review Team, Medicinal Plants in COVID-19: Potential and Limitations

DOI: 10.15439/2023R27
ISSN 2300-5963 ACSIS, Vol. 38

Fake News Identification Using Supervised Machine Learning Algorithms

Md Nooruddin Rabbani
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India
mdnrabbani@gmail.com

Abdul Wahid
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India

Fareeha Rasheed
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India

***Abstract*—Fake news has emerged as a significant challenge in today's information-driven society, where misinformation can spread rapidly and have detrimental consequences. Detecting and combatting fake news is crucial in maintaining the integrity of news sources and ensuring the public's access to accurate and reliable information. Machine learning approaches have recently demonstrated the ability to recognize false news stories automatically based on their features and content. To identify fake news, this study compares and contrasts several machine learning (ML) methods, including Random Forest, Passive Aggressive Classifier (PAC), Multinomial Naive Bayes, SVC, Decision tree, Gradient boosting, XG Boost, and Logistic Regression. These algorithms are tested on WELFake_Dataset and the output received has shown a significant increase the accuracy and a decrease in the false rate.**

***Index Terms*—Natural language processing, Passive Aggressive Classifier, Supervised Machine Learning**

I. Introduction

As technology has advanced, social media has become more prevalent in the daily lives of ordinary individuals. People may now regularly consume vast volumes of information from online sources. Fake news has drawn more attention in recent years due to the widespread usage of social media platforms. Popular social media sites like Twitter and Facebook make it simple for users to exchange content, offering them a forum for self-expression and global connectivity. Readers' current initiative is frequently critical to finding solutions to the problem of fake news. Human fact-checking is one of the answers to the issue of false news, which is now a serious concern for both business and academics [2]. Fake news is a fabricated story to deceive readers or spread propaganda. Fake news has become a significant challenge in the age of social media and the internet. Spotting fake news is getting more complicated since more online content is available. The impact on society is that false information may travel swiftly and influence perceptions. It might destroy trust between various social groups, affecting dialogue and decreasing confidence in the media. Due to how easily inaccurate information may be distorted to support a false narrative, it can also result in social unrest and chaos.

During the global COVID-19 outbreak, several doctored films and images about the COVID-19 virus, its origin and spread of vaccines, and the deaths it has caused have been circulating on social media. The percentage of fake news, videos, and photographs disseminated on social media is thought to be between 30 and 35 per cent. This erroneous information causes widespread panic since it travels faster than the virus [17]. As a result, there is an increasing demand for automated fake news detection systems that can tell legitimate news pieces from false ones. To detect more fake news, a machine-learning system is being deployed. Therefore, various ML techniques have been used in this research to identify fake news. Figure 1 shows the ML application to classify fake and real news. The WELFake_Dataset dataset was obtained from Kaggle.

II. Motivation

Fake news is a serious issue that has the potential to compromise democracy and public trust. In recent years, the growth of false information on the internet via social media and websites has caused harm to society by leading people to make incorrect decisions and propagating false information. Like other techniques, machine learning methods are one way to spot fake news. These days, more and more people are interested in using machine learning techniques to detect fake news. During Pendamic, lots of fake news was spreading that motivated me to work on it.

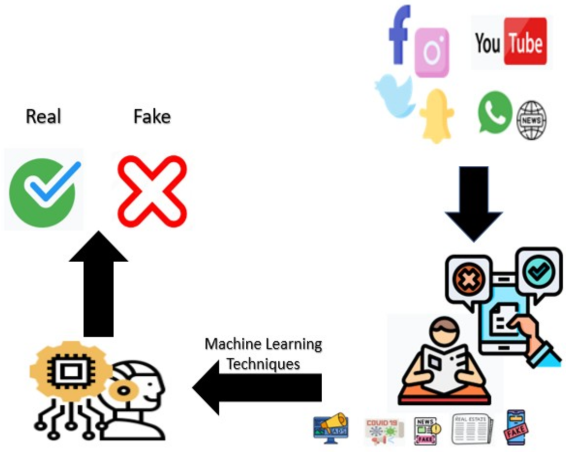


Figure-1 ML application to classify the fake and real news.

III. Related Work

The study by Narra et al. (2022)[1]. presents a selected feature set-based method for identifying COVID-19-related fake news. The authors chose the essential attributes using approaches like chi-square, mutual information, and information gain. They also compared the performance of different classification algorithms and found that SVM outperformed other algorithms. The study highlights the importance of contextual information in fake news detection and shows that sentiment analysis, readability, and topic modelling are the most compelling features for detecting fake news related to COVID-19. However, the study has limitations, such as a small dataset focusing only on English-language news articles. The study offers insightful information on identifying false news and encourages future study.A fresh approach to identifying false news is put out in the study [2]. Five machine learning models and three deep learning models comprise the stacking ensemble used in the procedure. Two datasets of fictitious and legitimate news stories train the models. Afterwards, the news stories are categorized using the stacking ensemble. In the ISOT and KDnugget datasets, the approach obtains an accuracy of 99.94% and 96.05%. The work [3] proposes a hybrid language and knowledge-based technique for identifying fake news on social media. The strategy combines linguistic characteristics, such as word count, readability, and lexical diversity, with knowledge-based characteristics, such as the website's standing where the news is published, the number of sources used to assemble the news, and fact-checking by reputable fact-checking websites. The method gets a 94.4% accuracy rate when tested on a dataset of actual and fake news articles. The paper [4] suggests a taxonomy of strategies for categorizing fake news. It examines the most recent methods for classifying fake news. and goes over each technique's merits and drawbacks. The three categories that make up the taxonomy are feature type, classification algorithm, and evaluation measure. The paper [5] addresses the issue of identifying false information regarding COVID-19. The World Health Organization, UNICEF, and the United Nations are used as information sources and epidemiological data gathered from various fact-checking websites in the authors' proposed approach for detecting deceptive information. A dataset of 10,000 news stories about COVID-19 is used to assess the model. The findings reveal that the model has a 90% accuracy rate for detecting false information. This paper [6] proposes using the OPCNN-fake, an optimized convolutional neural network, to identify fake news. OPCNN-FAKE is a deep learning model that distinguishes between genuine and fraudulent text via recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for feature extraction. The authors compare the performance of OPCNN-FAKE with other machine learning (ML) and deep learning (DL) models on four fake news benchmark datasets. The findings show that OPCNN-FAKE outperforms the other models on all datasets, with an average accuracy of 92.8%. The authors conclude that OPCNN-FAKE is a promising model. for fake news detection and can be used to improve the accuracy of fake news detection systems. The paper [7] offers a fake media detection system based on blockchain and natural language processing (NLP). The system uses natural language processing (NLP) to extract components from media material, such as using particular words or phrases, logical fallacies, and the overall tone. Then, a machine learning model is trained using these traits to determine if a piece of media is authentic or fraudulent. The findings of the machine learning model are also stored within the system using blockchain technology, making it challenging for attackers to alter or remove the results. The paper's authors evaluate the system on a dataset of fake and authentic media content. The findings reveal that the technology has a 92% accuracy rate for identifying fake material. The paper [8] discusses the problem of fake news detection in social media. Fake news is defined as news that is intentionally false or misleading. It is often created to deceive people and to manipulate public opinion. Fake news can harm society, leading to people making decisions based on false information. The paper discusses several different approaches to fake news detection: content-based, social media-based, and Hybrid approaches. The paper [9] explains how to spot fake news using machine learning algorithms. The feature extraction process entails taking specific information from the news articles, such as the author, text, and title. A machine learning algorithm determines whether the news stories are authentic in the classification stage. On a dataset of 1,000 news articles, the authors assessed their methodology. They discovered that their method had a 95% accuracy rate. The paper [10] offers a hybrid deep-learning architecture for detecting fake news stances. The architecture comprises the long short-term memory (LSTM) network and convolutional neural network. The LSTM is used to record the sequential associations between the features, while the CNN is used to extract features from the news item. A dataset of news items with four stances—agree, disagree, discuss, and unrelated—was used to train the architecture. The experimental findings indicate an accuracy of 97.8%.The paper [11], Using a machine learning tool, the suggested method extracts various texts from the articles and feeds the feature set into the models. The training models were trained, and their parameters were adjusted for the best outcome.

The paper [12] proposes using the OPCNN-fake, an optimized convolutional neural network, to identify fake news. OPCNN-FAKE is a deep learning model that distinguishes between genuine and fraudulent text via recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for feature extraction. The attention module learns to focus on critical features, the feature extraction module extracts feature from the text, and the classification module determines if the news is true or false. A dataset of news items and their labels is used to train the MVAN model. The results of the FNC-1 and FNC-2 datasets demonstrate that the MVAN model performs better than cutting-edge techniques.

IV. Proposed Model

In this is section proposed methodology discussed.

This comparative study comprised various machine learning (ML)classifiers for the dataset shown in this figure-2, collected from the online source Kaggle name-WELFake_Dataset. The planned work's phases are shown in Figure 2 are simplified as follows: Data processing has been done before feature extraction and splits into a ratio of 80:20 for training and test data. ML classifiers such as Random

TABLE-1 RELATED WORK DATASET AND MODEL USED

Ref	Year	Contribution	Dataset	Model used
[12]	2021	Multi-View Attention Networks(MVAN) for Fake News Detection on social media	Twitter15 and Twitter16	Multi-view attention networks (MVAN)
[13]	2023	Fake news detection in social media based on sentiment analysis using classifier techniques.	ISOT false news and LIAR	Naïve Bayes,Passive Aggressive clssifer, DNN
[14]	2021	"All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustness of Fake-news Detectors Under Black-Box Settings".	Kaggle fake-news dataset, ISOT dataset, and LIAR dataset.	MLP, CNN, RNN and Hybrid CNN-RNN
[15[	2020	"Detecting Misleading Information on COVID-19,"	Primary data is collected from various online sources using the Google Fact Check Tools API and stored in MySQL Server.	Decision Tree (DT), KNN Logistic Regression (LR), Linear Support Vector Machines (LSVM), Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Perceptron, Neural Network (NN), Ensemble Random Forest (ERF), Extreme Gradient Boosting classifiers (XGBoost)
[16]	2020	"Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM),"	Fake News Challenges	CNN and LSTM
[17]	2022	Detecting covid-19-related fake news using feature extraction	The primary dataset used was collected from Facebook, Twitter, The New York Times	Random forest classifier (RFC), AdaBoost, DT and KNN

Forest, Passive aggressive classifier, multinomial Naive Bayes, SVC, Decision tree, Gradient boosting, XG Boost, and logistic regression are used to obtain the best accuracy and result.

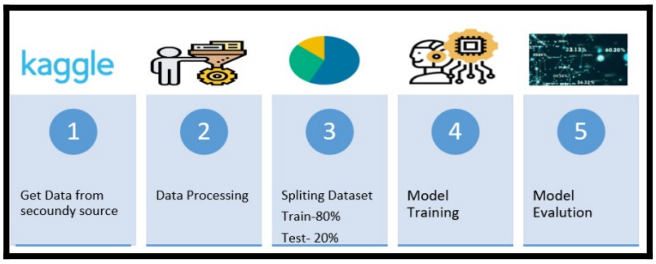


Figure-2 Steps of proposed ML model.

A. Preprocessing

Data processing methods are essential for obtaining insightful information in the age of big data. These methods include removing null values, omitting unnecessary characters, stemming, and vectorization, which helps to improve the quality and usefulness of their datasets, leading to more precise and trustworthy analysis

B. Null Values

Null values, commonly called missing values, can significantly affect data analysis. They may add bias, impact statistical computations, or hinder machine learning techniques. Data preparation requires recognizing and effectively treating null values to solve this problem.

C. Removing Extra Character

In real-world datasets, it is common to encounter unstructured or noisy data containing extra characters, such as punctuation marks, special symbols, or HTML tags. These extraneous elements can interfere with analysis tasks, like text mining or natural language processing.

D. Stemming

Using the text normalization approach known as stemming, data researchers may execute tasks like sentiment analysis, information retrieval, and document clustering more effectively.

E. Vectorization

Vectorization is a fundamental step in transforming textual or categorical data into numerical representations. The process involves converting words, phrases, or categorical labels into numerical vectors that capture the inherent relationships between them. Many machine learning algorithms require numerical input, making vectorization essential.

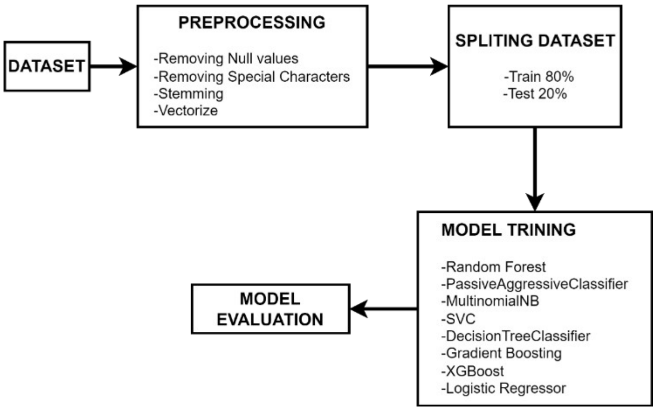


Figure-3 Proposed methodology.

This section provides a detailed explanation of the proposed work. The planned work's steps are depicted in Figure-3 and are summed up as follows:

V. PARAMETERS OF EVALUATION

The effectiveness of categorization models is frequently evaluated using the following metrics [7]

Accuracy- Number of correct predictions to the total number of predictions.

Precision -The precision metric is used to overcome the limitation of Accuracy. The precision determines the propor-

TABLE-2 FORMULA USED FOR PROPOSED METHODOLOGY.

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Performance of models based on train data
Precision	$\frac{TP}{TP+FP}$	The quality of a positive prediction made by the model
Recall	$\frac{TP}{TP+FN}$	The true positive rate (TPR) model correctly identifies.
F1 Score	$\frac{2\,x\,Precision\,x\,Recall}{Precision+Recall}$	It measures the model's accuracy.
Specificity	$\frac{TN}{FP+TN}$	Measures the proportion of true negatives that the model correctly identifies.

tion of positive prediction that was actually correct. It can be calculated as the True Positive and True Negative.

Recall- It can be calculated as True Positive or predictions that are actually true to the total number of positives.

F-Scores- F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them

Specificity-Specificity measures the proportion of true negatives that are correctly identified by the model.

This experiment on WELFake_Dataset in Jupiter required a system with at least a Core i5 CPU, 16GiB module 120GB SSD free space. As shown in Table-2, the Random Forest algorithm is also a powerful tool for detecting fake news. Compared to other algorithms, it requires less training time, and with careful data preprocessing and model tuning, levels % of this experiment's accuracy of 97% was achieved. Passive-Aggressive algorithms are generally used for large-scale learning, and usually, for large-scale learning, passive-aggressive algorithms are utilized. Nowadays, this is most popular to detect fake news on social media like Twitter, where new data is added every second.

A. Passive-Aggressive Algorithms

Machine learning algorithms, passive-aggressive algorithms (PAA), are frequently employed for binary classification problems. They are renowned for their simplicity and effectiveness, especially in online learning settings where immediate forecasts are necessary, and information is delivered in batches. Nowadays, social media websites like Twitter, where new information is uploaded every minute, are the most widely used for spotting fake news. In this experiment, it achieved a level of accuracy of 95%. Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because

B. Multinomial NB

Naive Bayes is a classification algorithm based on Bayes' theorem, assuming independence among the features. It is widely used for text classification, spam filtering, sentiment analysis, and other tasks. Using Bayes' theorem, the approach determines the probability of each class provided the observable characteristics. Similar to the PPA algorithm, an accuracy level of accuracy 95% was achieved.

P (class | features) = (P(class) x P (features | class)) / P(features) [7]

C. SVC

Support Vector Machine with Citation, or SVC, is a powerful supervised learning technique for classification and regression applications. SVCs have been used in many fields, including bioinformatics, text classification, and picture recognition.

The foundation of SVMs is the discovery of a hyperplane or group of hyperplanes in a very high-dimensional or infinite-dimensional space that may be applied to classification, regression, or other tasks like outlier detection. The closest points from each class to the line the algorithm draws to divide the data into classes are referred to as support vectors [19].

The margin is the distance between the line and the support vectors, and maximizing the margin is the objective [20]. Many people like SVMs because they generate substantial accuracy while requiring minimal processing resources.

Accuracy, precision, recall, and F1 score are some of the measures used to assess an SVM's effectiveness in identifying fake news. These metrics offer a quantitative evaluation of the model's capacity to distinguish between instances of fake and legitimate news. SVM has high accuracy and reliable results in this research work, and the performance of SVM is high as compared to other ML algorithms used that achieve a high level of accuracy of 97% on this dataset.

D. Decision Tree

The decision tree (DT) classifier is one of the most popular ML algorithms for classification and prediction problems on supervised data. It offers a 96% accuracy rate for the dataset used. Using rules and trees, the training dataset is segmented into classes.

E. Gradient Boosting

Gradient boosting is another ensemble method used in machine learning to strengthen weak learners. 96% accuracy level on this dataset was attained.

F. XGBoost

Extreme Gradient Boosting, sometimes called XGBoost, is a popular machine learning method that performs very well in various structured data applications, including classification, regression, and ranking problems. On this dataset, it has a 96% accuracy rate.

G. Logistic Regression

Logistic regression is a prominent and widely used classification method for categorizing fake news. Logistic regression is a statistical technique for binary classification tasks. It predicts the chance that an instance will belong to a specific class based on the values of the input characteristics. It also has a 97% accuracy rate for this dataset.

Table-2 Accuracy of ML Algorithms.

Algorithm	Predication Accuracy	Precision	Recall	F1 Score
Random Forest	96	100	95.917023	97.93638
Passive Aggressive Classifier	95	100	95.164979	97.92024
MultinomialNB	95	100	95.164979	97.92024
SVC	97	100	96.841755	97.95588
Decision Tree	96	99	96.419098	96
Gradient boosting	96	99	96.692282	96
XGBoost	96	72	33	45
Logistic Regression	97	91	43	59

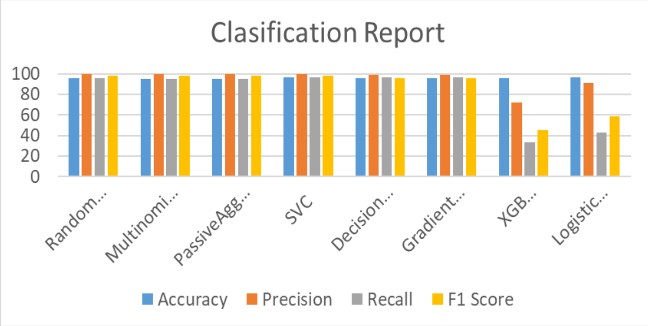


Figure-4 Classification Report

VI. Conclusion

Machine learning techniques offer a promising solution for detecting and preventing the spread of fake news. Our model approach to identifying fake news using supervised ML techniques uses natural language processing techniques to analyze the text of news articles and identify patterns and features associated with fake news. This study shows that SVM is a highly effective algorithm for spotting false information. It is a valuable tool in the fight against the spread of false information and disinformation in the digital age because of its incredible accuracy, precision, recall, and low false rate. The findings of this study can be used as a basis for the creation of sophisticated and scalable false news detection systems that can help uphold the reliability and integrity of online information. Future research has several opportunities to enhance further the functionality and application of this method in this field.

References

[1] Narra, M., Umer, M., Sadiq, S., Eshmawi, A. A., Karamti, H., Mohamed, A., & Ashraf, I. (2022). Selective feature sets based fake news detection for covid-19 to manage Infodemic. IEEE Access, 10, 98724–98736. https://doi.org/10.1109/access.2022.3206963

[2] T. Smith, P. Husbands and M. O'Shea, "Neutral networks in an evolutionary robotics search space," Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), Seoul, Korea (South), 2001, pp. 136-143 vol. 1, doi: 10.1109/CEC.2001.934382.

[3] Seddari, Noureddine & Derhab, Abdelouahid & Belaoued, Mohamed & Halboob, Waleed & Al-Muhtadi, Jalal & Bouras, Abdelghani. (2022). A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media. IEEE Access. 1-1. 10.1109/ACCESS.2022.3181184.

[4] Rohera, Dhiren & Shethna, Harshal & Patel, Keyur & Thakker, Urvish & Tanwar, Sudeep & Gupta, Rajesh & Hong, Wei-Chiang & Sharma, Ravi. (2022). A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects. IEEE Access. 10. 10.1109/ACCESS.2022.3159651.

[5] M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[6] H. Saleh, A. Alharbi and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," in IEEE Access, vol. 9, pp. 129471-129489, 2021, doi: 10.1109/ACCESS.2021.3112806.

[7] Shahbazi, Zeinab & Byun, Yungcheol. (2021). Fake Media Detection Based on Natural Language Processing and Blockchain Approaches. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3112607.

[8] Rohera, Dhiren & Shethna, Harshal & Patel, Keyur & Thakker, Urvish & Tanwar, Sudeep & Gupta, Rajesh & Hong, Wei-Chiang & Sharma, Ravi. (2022). A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects. IEEE Access. 10. 10.1109/ACCESS.2022.3159651.

[9] Stahl, K. (2018, April 20). Fake news detection in social media. California State University Stanislaus. Retrieved May 15, 2018.

[10] Sahu, Mickey, and Narendra Sharma. "Research on the Ability to Detect Fake News with Machine Learning." International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 4, 2020, pp. 3664-3668., doi:10.18531/ijert.2020.9.4.1590.

[11] Xu, Y., Zhang, Y., Li, J., & Zhang, J. (2021). MVAN: Multi-view attention networks for fake news detection on social media. IEEE Access, 9, 106907-106917. doi:10.1109/ACCESS.2021.3100245.

[12] Balshetwar, S.V., RS, A. & R, DJ Fake news detection in social media based on sentiment analysis using classifier techniques. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-14883-3

[13] Ali, H., Rehman, S. U., Shah, M. A., & Imran, M. (2021). All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings. IEEE Access, 9, 81678-81692. doi:10.1109/ACCESS.2021.3085875

[14] M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[15] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi and B. -W. On, "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)," in IEEE Access, vol. 8, pp. 156695-156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[16] Khan, S., Hakak, S., Deepa, N., Prabadevi, B., Dev, K., & Trelova, S. (2022). Detecting covid-19-related fake news using feature extraction. Frontiers in Public Health, 9. https://doi.org/10.3389/fpubh.2021.788074

[17] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification (2nd ed.). Wiley-Interscience.

[18] https://en.wikipedia.org/wiki/Support_vector_machine

[19] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[20] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). Fake News Detection on Social Media: A Data Mining Perspective. ACM Computing Surveys, 52(5), 1-34

DOI: 10.15439/2023R20
ISSN 2300-5963 ACSIS, Vol. 38

Data Aggregation Techniques and Challenges in the Internet of Things: A Comprehensive Review

Shaheen Fatima
Department of Computer Science & I.T
Maulana Azad National Urdu University
Hyderabad, India
fatimashaheen035@gmail.com

Jameel Ahamed
Department of Computer Science & I.T
Maulana Azad National Urdu University
Hyderabad, India
jameel.shaad@gmail.com

***Abstract*—The Internet of Things (IoT) has revolutionized the way people interact with their environment, generating massive amounts of data from interconnected devices. With the exponential growth of IoT devices, efficient data aggregation techniques are essential for extracting meaningful insights and reducing network traffic. This review paper aims to provide a comprehensive overview of data aggregation techniques in IoT, focusing on their methodologies, advantages, and challenges. The paper begins by discussing the fundamentals of IoT data aggregation. It then categorizes the data aggregation techniques into two main approaches: centralized and distributed. For each approach, various algorithms and protocols are explored, including clustering-based aggregation, tree-based aggregation, and centralized-based aggregation. Furthermore, the paper investigates the trade-offs involved in data aggregation, such as energy consumption, latency, and data accuracy. It examines the impact of different factors, such as data heterogeneity, and security considerations, on the choice of aggregation technique. Furthermore, in this paper researcher discussed the existing techniques, while highlights the emerging trends and future directions in IoT data aggregation. In this review paper we concluded by summarizing the key findings and highlighting the challenges that need to be addressed in the field of IoT data aggregation.**

***Keywords*—Data Aggregation, Tree based mechanism, Cluster based mechanism, centralized based mechanism, Internet of Things**

I. Introduction

The contemporary technology environment now includes a new computational component called the Internet of Things (IoT). The internet of things includes integrated sensors and items. They can communicate with one another without needing to contact with humans. The "things" in the "internet of things" are actual physical items like sensors, data gatherers, and monitors of various kinds of data relating to machine and human social behaviour [1]. The Internet of Things (IoT) encourages networking, data sharing, information collection, and integration of the different items that are present in our surroundings [2]. IoT has becoming more and more prevalent in our daily lives [3]. For example in smart home [4], smart devices are connected to external services, enabling relationship between Smart gadgets and outside services. Electronic medical systems [5] may employ wearable technology to track vital signs of the patient including blood pressure and the heart rate. Intelligent transportation frequently uses IoT. IoT refers to the networked connectivity of several heterogeneous systems [6]. The applications layer, the cloud computing layer, the sensing layer, and the networking layer are the four tiers that make up the architecture of IoT-based systems [7]. Figure 1 illustrates the IoT architecture.

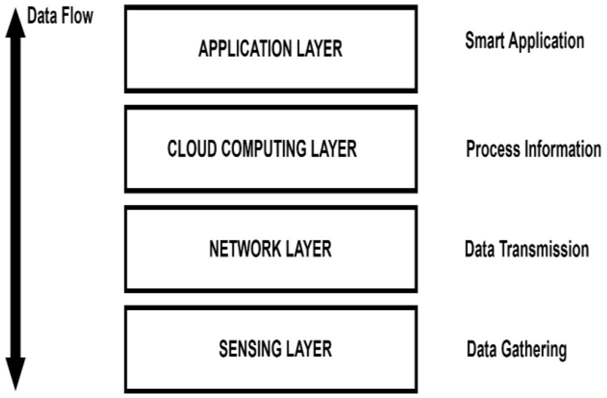


Fig 1: Layered Architecture of IoT

A. *Architecture of Internet of Things (IoT)*

- **Sensing Layer:** Many IoT and sensor devices continuously scan their surroundings to collect data zone in this layer and deliver it to the sink [8, 9]. The monitoring region is covered by millions of IoT devices that have been placed to create a self-organized, multi-hop topology [10]. It should be mentioned that some IoT devices are more likely to malfunction since they are located in a particular region [11]. Moreover, some gadgets' energy runs out more quickly than others [12]. Thus, it is crucial to use energy-efficient approaches for IoT data aggregation.
- **Network Layer:** Offering efficient topologies for data transfer between source devices and destination devices is the responsibility of the networking layer [13]. While IoT topologies are designed to provide source devices high data transmission rates, these systems are constrained in terms of energy usage, throughput, and malicious attacks because of different topologies [14].
- **Cloud computing layer:** The ability to manage massive amounts of data more rapidly and precisely has been made feasible by advancements in cloud computing technology [15]. Data is received, processed, and decided upon by the cloud computing layer, which then sends the results to other levels [16]. While edge and fog are chosen by other strategies to optimise costs and performance, cloud computing is the preferred option for IoT-based

systems to store and analyse data instead of completely replacing the cloud, the key objective of adopting edge or fog architecture must be modified and included into the data collection process to handle scattered IoT devices, manage system heterogeneity, and separate critical data from generic data [17]. In fact, the demand for large-scale data aggregation in IoT applications drives techniques away from cloud computing and towards fog or edge computing.

- **Applications layer:** This layer includes a variety of applications, including wireless sensor networks [18]. The application layer (also known as Layer 4) is made up of n data elements, which together make up the IoT data. Clustering is feasible because each data unit can store up to 64 bits of data. Ways to decipher the data-carrying IoT communications' content formats. Consequently, it is now possible to increase the efficiency and speed of data aggregation performed across the IoT tiered architecture [19]. Apps for the Internet of Things are also utilised to keep track on emergency situations and environmental conditions. Applications for the Internet of Things are used to track environmental and disaster conditions. Figure 2 shows the application of IoT.

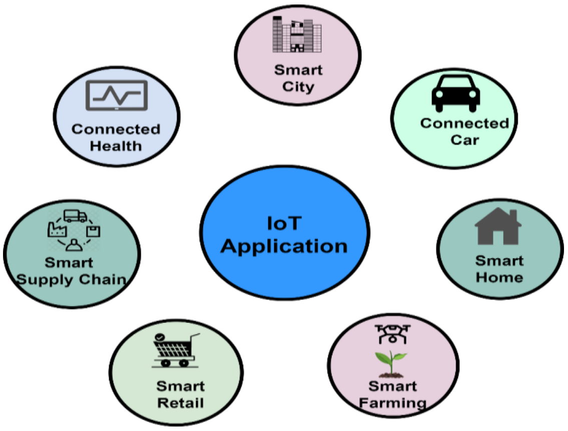


Fig 2: Application of IoT

B. *Application of Internet of Things (IoT)*

- **Connected Health:** IoT has many uses in the healthcare industry, including advanced & smart sensors, equipment integration, and remote tracking tools. It has the capacity to get better. How doctors maintain their patients' wellbeing and provide treatment. Patient spending may increase thanks to IoT in healthcare. Spending time talking to their physicians can increase patient engagement and happiness. In healthcare introduces new tools restored with the newest technology in the environment that helps in creating improved healthcare, ranging from personal fitness sensors to surgical robots [20].
- **Smart City:** The smart metropolis is an emerging and promising application of the Internet of Things (IoT) that has captured considerable attention. One of the notable use cases of IoT in smart cities is intelligent surveillance, which enhances security measures. Additionally, IoT enables improve automated transportation systems, energy management systems, efficient water sharing, enhanced metropolitan security, and environmental monitoring. IoT has the potential to address significant urban challenges like pollution, traffic congestion, and energy scarcity. For instance, IoT-enabled devices like Smart Belly garbage bins equipped with cellular communication capabilities can notify municipal services when they need to be emptied. Through online applications and citywide installed devices, residents can easily locate available parking spaces. Furthermore, these monitors can detect issues such as faulty installations, general system malfunctions, and meter fraud in the electrical system [21].
- **Connected Car:** Vehicular digital technology has traditionally focused on optimizing the interior functions of cars. However, there is now a growing emphasis on enhancing the in-car experience. Connected cars refer to vehicles that utilize internal devices and internet connectivity to enhance their operations, maintenance, and passenger convenience. Several established automakers and new entrants are actively developing connected vehicle solutions. Prominent companies such as Tesla, BMW, Apple, and Google are at the forefront of driving the next automotive revolution. The extensive network of connected vehicle technology encompasses various sensors, antennas, integrated systems, and other technologies that facilitate communication, enabling smooth navigation through our complex world. This technology plays a critical role in making prompt, accurate, and reliable decisions. As the world moves towards a future in which steering wheel control is relinquished, and autonomous or self-driving cars become more prevalent on the roads, the importance of these requirements will only intensify [22].
- **Smart Home:** The Smart homes are now the new benchmark for victory in the domestic market, and it is predicted that they will soon be as common as mobile phones. When thinking about IoT systems, the most major and effective application that comes to mind is Smart Home, which rates as the top IOT application across all platforms. The sum of the funds given to startups in the smart home sector is expected to be $2.5 billion, and it is steadily increasing. Wouldn't it be great if you could turn the lights off even after you've left the house or put on the air conditioning before you got there. You may even temporarily admit visitors if you're not at home simply opening the doors. Don't be surprised that companies are creating IoT-related products to simplify and ease your life [23].
- **Smart farming:** One IoT use case that is frequently disregarded is smart gardening. However, because farmers typically deal with a large number of distant agricultural operations and animals, the Internet of Things can watch all of this and change how farmers conduct their business. But this concept hasn't yet attracted much notice. Although it is still an IoT usage, it should not be dismissed particularly for nations involved in the production and trade of agricultural products, smart farming has emerged as

a promising field with a wide range of potential applications [24].

- **Smart Retail:** Retailers have been using Internet of Things (IoT) solutions and integrating IoT-enabled systems into different applications, which has improved the efficacy and efficiency of their shop operations., including boosting sales, lowering fraud, allowing inventory management, and improving the purchasing experience for customers. Physical stores can more effectively contend with online rivals thanks to IoT. They can draw customers to the shop and recover lost market share, Retailers' purchasing procedures are made easier as a result, allowing them to buy more things for less money. [25].
- **Smart Supply Chain:** For a few years now, supply networks have been evolving to become more intelligent. Providing answers to issues like monitoring products while they are traveling or in transportation or assisting vendors in exchanging inventory data are some of the well-liked services. An IoT-enabled device allows factory equipment with integrated sensors transmits information about various factors like pressure, temperature, and machine usage. In order to improve performance, the IoT system can also handle processes and modify equipment settings [26].

C. *Data Aggregation*

Data aggregation is the process of gathering information from various Internet of Things devices and portraying it in a condensed manner. IoT heavily utilizes data aggregation methods to reduce traffic and energy usage. A much more straightforward method of data aggregation is for every source nodes to gather data from various sources and transmit it, without any kind of pre-processing, to a single destination execute the various data aggregation operations directly on the combined data using a single aggregator server [27].The goal of data aggregation methods on the IoT is to achieve high QoS, which includes taking into account the importance of the data and having low data transfer delays, high dependability, and minimal energy usage [28].

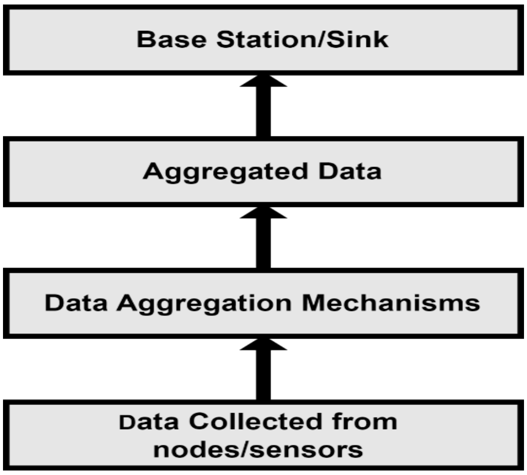

Base Station/Sink
Aggregated Data
Data Aggregation Mechanisms
Data Collected from nodes/sensors

Fig 3: Data Aggregation

The following are generally some benefits that data aggregation techniques provide:

- It contributes to enhancing the usefulness and accuracy of the data presented
- accomplished through a vast network [29].
- Additionally, it lessens traffic volume and conserves node energy. [29].
- The data collected from nodes contains specific duplication, so this procedure is essential to reduce the extra information [29].
- Traffic volume: Through multi-hop transfer, a large number of IoT devices transmit their data-to-data repositories. When adjacent nodes decline with high traffic loads, this type of communication behaviour causes an unbalance in the network's traffic burden. Monitoring traffic volumes can aid in the creation of more effective routing formulas and the advancement of node distribution [30].

D. *Data Aggregation Mechanism on IoT*

In IoT networks, data aggregation algorithms are used to collect and summarise data from many sources in order to increase network efficiency [31]. These strategies include protocols for clustering, compression, and encryption. These methods may be applied to lessen network traffic, save energy, lengthen network life, and enhance security. Client-Server-based and mobile agent-based data aggregation strategies for IoT are detailed in separate categories. IoT devices provide data sent in a multi-hop fashion to the sink using client-server-based data aggregation techniques, where certain intermediary devices can carry out aggregation procedures [32]. Different components of client-server-based methods are studied: cluster-based, tree-based, In-network, Chain-Based, Grid-based, and centralized ones. Individual software packets consecutively visit IoT devices to gather their sensed data in mobile agent-based data aggregation processes. These packets then collect and send the data to the sink location.[33].

E. *Client-Server-based Data Aggregation Mechanisms*

A central server is used in client-server-based data aggregation methods to gather and combine data from several clients [34]. An alternative to the client-server paradigm that has been suggested as a more practical option is mobile agent-based data aggregation [35]. An IoT device has enough memory in client-server systems to retain the data it senses and packets it receives from other devices. Before sending the last packet to the next destination, it performs the aggregation function on the data that has gathered in its memory [36]. Client-server-based data aggregation mechanisms can help improve the efficiency of data collection and transmission in IoT networks. By aggregating data at the client side before sending it to the server, the amount of traffic injected into the network can be reduced, which can help to alleviate network congestion and reduce energy consumption. This can ultimately improve the overall lifetime of IoT devices by conserving their battery life. Additionally, aggregating data on the server side can also help to reduce the amount of data that needs to be stored and processed, which can further improve network efficiency [37].

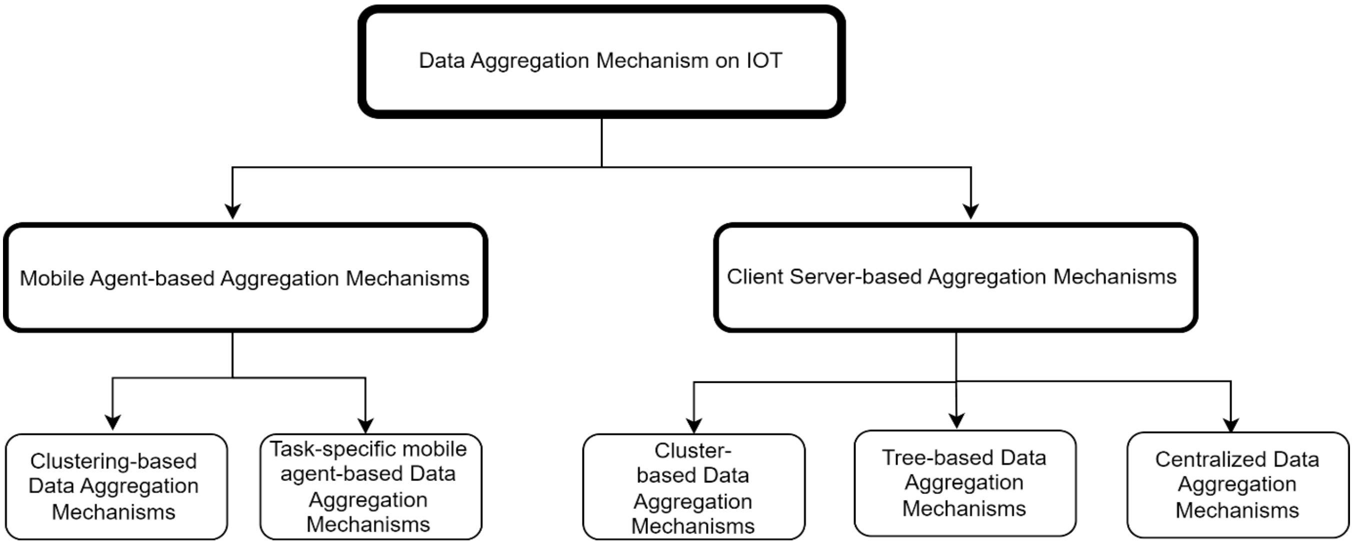

Data Aggregation Mechanism on IOT
Mobile Agent-based Aggregation Mechanisms
Client Server-based Aggregation Mechanisms
Clustering-based Data Aggregation Mechanisms
Task-specific mobile agent-based Data Aggregation Mechanisms
Cluster-based Data Aggregation Mechanisms
Tree-based Data Aggregation Mechanisms
Centralized Data Aggregation Mechanisms

Fig 4: Data Aggregation Mechanism

F. *Cluster-based Data Aggregation Mechanisms*

A Wireless Sensor Network uses cluster-based data aggregation algorithms to aggregate data from many sources (WSN). These processes entail segmenting the network into clusters, with each cluster having a cluster head node in charge of collecting and aggregating information from the cluster's other nodes. The information is subsequently transmitted to a base station or central node, where it may be used for additional applications and analysis. Routing protocols are frequently used by cluster-based data aggregation methods to enable effective communication between the nodes and the cluster head.

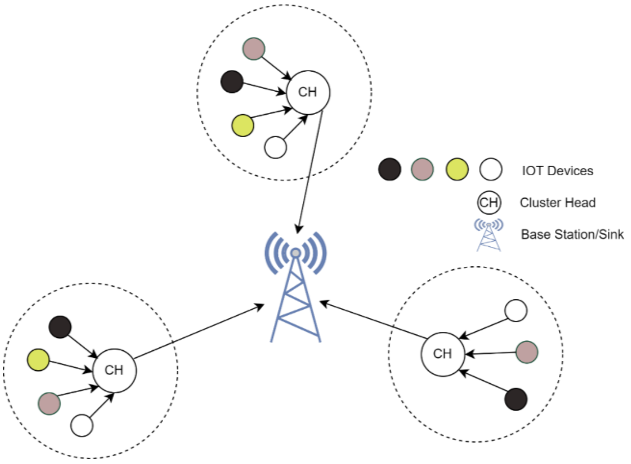


Fig 5: Cluster-based Data Aggregation

The fuzzy similarity matrix-based clustering, the tree-based, the beta-dominating set centred-cluster-based data aggregation mechanism (DSC2DAM), are some typical examples of cluster-based data aggregation techniques [38][39]. And the cooperative information aggregation (CIA) mechanisms [40] By sending fewer messages to the central node, these techniques help a WSN be more fault-tolerant and scalable while also consuming less energy.

G. *Tree-based Data Aggregation Mechanisms*

Tree-based data aggregation mechanisms are a type of data aggregation technique used in the IoT. According to this technique [41], a tree structure is built in the network, with each node representing a sensor node. The nodes then interact with one another and transmit the data they have gathered to the tree's base. Then, after being aggregated, this information may be used for a variety of tasks, including analysis and decision-making. While it can assist decrease data redundancy and the amount of data that has to be carried, tree-based data aggregation is beneficial for networks with a lot of nodes. Also, since data is only sent to the tree's root, it can aid in lowering network power usage.

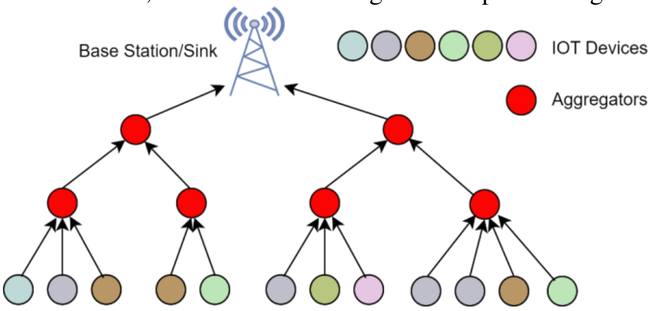


Fig 6: Tree-based Data Aggregation

A Tree-based Data Aggregation Mechanism [42], an Aggregation Tree Based Data Aggregation Algorithm [38], and Tree-based Data Aggregation Algorithms in Wireless Sensor Networks [43] are a few examples of tree-based data aggregation mechanisms. Figure 6 illustrates the design of IoT's tree-based data aggregation techniques.

H. *Centralized Data Aggregation Mechanisms*

In centralized-based aggregation mechanisms, all IoT devices transfer their detected data through the shortest route to the most powerful device (also known as the header device). The header device processes all received data through the aggregation function before sending a single packet to the sink [44][45]. Improved data integrity, decreased data redundancy, lower costs, insightful data, and the capacity to lower the volume of data created in Wireless Sensor Networks are benefits of centralized data aggregation techniques [46]. When there are fewer nodes in the network, centralized data aggregation techniques perform better [38].

The architecture of the data aggregation is shown in the figure 7

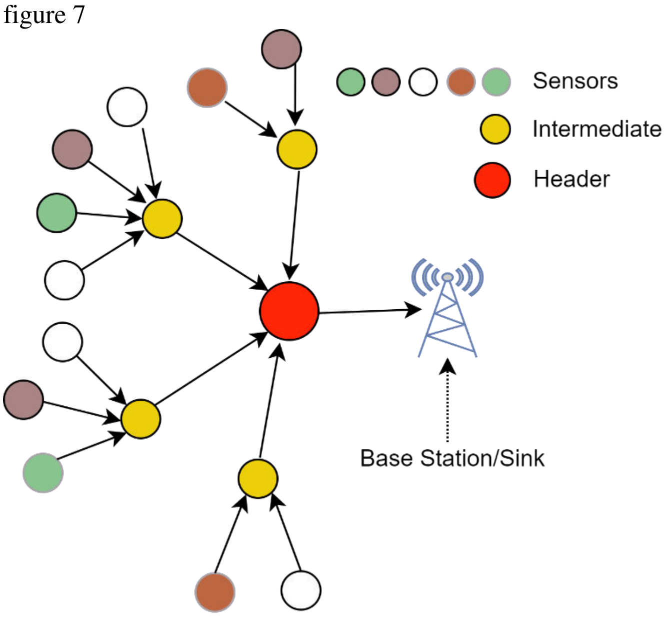


Fig 7: Centralized-based Data Aggregation

I. Mobile Agent Data Aggregation Mechanism

Mobile agent (MA)-based wireless sensor networks offer an alternative to the traditional client/server architecture by enabling MAs to move to sensor nodes (SNs) for data collecting. Unlike the client/server paradigm, which transfers data from individual nodes to a central sink, this method's MAs instantaneously acquire data from SNs, saving energy and bandwidth [47]. In systems that aggregate data using a single agent, a mobile agent is sent across the network by the data delivery sink to get information from all devices, then returns it. It lengthens the life of the IoT system, reduces device energy consumption, and enhances network data flow. The use of mobile agents for data aggregation in the context of the Internet of Things (IoT) is not without its limitations though [48].

- When the mobile agent visits several devices one at a time, long delays are necessary.
- Since the number of mobile agents' packets grows over the course of the data collecting procedure, IoT devices near to the washbasin will require more energy.

II. Related Work

A means of offering security and effectiveness in the data aggregation has been proposed by the writers in [49]. This approach aims to create a precise, safe data compilation a way that integrates the IoT network's transmission and processing constraints while taking security considerations into account. This method guarantees security, but its restriction is the heavy traffic burden.

According to scalability, security, reliability, and flexibility, the authors of [50] have suggested an internet of things storage system that is scalable and secure. This system will enable users to fulfil the requirements for data mining and analytics with massively aggregated data. A revised secret sharing scheme is the foundation of design in order to accomplish the security of data without the need for complicated key management.

The author [51] have suggested a tree-based technique for data aggregation that is both delay-aware and energy-efficient. By offering energy-efficient routing routes in various monitoring regions, the suggested technique reduces data transfer delay. In an effort to minimize the end-to-end latency, it also judiciously chooses an immediate forwarding method for transmission of delay-sensitive data. In order to decrease overall energy usage, a wait-forwarding one is also used for data transfer with a delay tolerance. In order to balance the energy consumption of IoT devices, the suggested approach comprises creating routing channels in regions with high residual energy. It is crucial to remember that the generated routes are often changed to preserve this equilibrium. The recommended method has been shown through performance research to minimise network energy consumption and transfer delays for data that is both delay-sensitive and delay-tolerant. The IoT ecosystem's lifetime is enhanced as a result.

To increase network lifespan and reduce data transfer latency, Li et al. [52] have suggested a tree-based data consolidation method on diverse and dynamic IoT. To extend the life of nearby devices, it adaptively modifies the data transfer latencies of those devices. According to experimental findings, the suggested technique performs well when its degree of variability and volatility is greater.

Lu et al. [53] suggested a simple privacy-preserving method for boosting fog computing security in IoT. Data from multiple devices will be safely integrated using this technique. It makes use of a centralised process that combines composite data while rejecting any erroneous data input into the system, using the Chinese Remainder Theorem, one-way hash chains, and homomorphic Parlier encryption. The results of the security study show that this simple solution is reliable and efficient in a number of private circumstances.

The distance, node degree, and leftover energy were used by the researcher of [54] to determine the CH. The ideal amount of CHs to encompass the full sensing region, however, was not guaranteed.

This paper provides a comprehensive overview of the security challenges associated with healthcare data aggregation and transmission in the Internet of Things (IoT). The paper begins by discussing the importance of healthcare data security and the challenges that are posed by the IoT. The authors then present a taxonomy of security threats to healthcare data in the IoT, and discuss the security mechanisms that can be used to mitigate these threats. The paper concludes by discussing open research challenges in healthcare data security in the IoT [55].

The goal of this study is to classify the three main types of data aggregation processes used in the Internet of Things (IoT): centralized, cluster-based, and tree-based. The study offers recommendations for further research by doing an extensive comparison of the essential operations within each category. To assess the various methods, the evaluation compares them based on tolerance, latency, heterogeneity, network durability, scalability, security, and traffic volume. The results show areas that require more study to fix the found issues and improve the effectiveness of data aggregation in IoT [56].

The paper proposes a new data aggregation and routing algorithm called the Energy-Efficient Data Aggregation and Routing (EDAAR) algorithm. The EDAAR algorithm works by first aggregating data from neighboring sensor nodes. The EDAAR algorithm was able to reduce the amount of data that needed to be transmitted by up to 75% [57].

The author proposes an effective flow aggregation method based on the SDN architecture for delay-insensitive traffic management. The situation of numerous tiny delay-insensitive traffic patterns is the main topic of the research. In order to combine and reduce traffic flows according to the flow magnitude and to be flexible to changes in network circumstances, the writers developed a novel data structure called a flow tree. This method lowers the price of storing in switches' memory as well as the expense of communication between the supervisor and OpenFlow switches [58].

In this paper the author compared various used mechanism (from 2016 to 2020) that illustrate efficient data aggregation mechanism on IoT to enhance security, privacy and minimize energy and computational resource. After that he suggested for future research Heterogeneity of IOT device, precisely in dynamic monitoring areas [59].

The goal of data aggregation strategies, which is to efficiently gather and merge data packets with the aim of lowering power consumption, reducing traffic congestion, extending the lifespan of the network while improving data accuracy, is the primary focus of the research paper "Comparison of Data Aggregation Techniques in Internet of Things (IoT)". The author looks at the number of transfers necessary for data capture, resulting in less network traffic, delay, and power consumption. This strategy also lengthens network lifespan and improves data precision [60].

On supporting IoT data aggregation through programmable data planes. The goal of this paper was to reduce the number of repeating packet headers by assembling packet data from several sources. The author finds that IoT improves network efficiency by 78%, according to research, and it also gives users control over the average delay caused by data aggregation techniques [61].

Table 1 An overview of data aggregation mechanism and their performance

Paper	Technique	Energy	Delay	Heterogeneity	Scalability	Life Time	Security	Accuracy
D. Fotue et al (2023)	Tree-based	✓	✓	✗	✗	✗	✗	✗
H. Barati et al. (2021)	Tree-based	✓	✗	✗	✗	✓	✗	✗
Alghamdi et al. (2016)	Tree-based	✓	✗	✗	✗	✓	✓	✓
Zhou et al. (2014)	Tree-based	✓	✗	✗	✗	✗	✗	✗
M. Mohseni et al. (2022)	Cluster-based	✓	✓	✗	✗	✓	✗	✗
Xie (2015)	Cluster-based	✗	✗	✗	✗	✗	✗	✗
Jiang et.al. (2015)	Cluster-based	✗	✗	✗	✓	✗	✗	✗
Liu et.al. (2014)	Cluster-based	✗	✗	✗	✗	✗	✓	✗
Lu et al. (2017)	Centralized	✗	✗	✗	✗	✗	✓	✗
Zhu et al. (2016)	Centralized	✗	✗	✗	✗	✗	✓	✗
Zhu et al. (2016)	Centralized	✗	✗	✓	✗	✗	✓	✗
T. Alsboui et al. (2021)	Mobile-agent	✓	✗	✗	✗	✗	✗	✗
M. El Fissaoui (2018)	Mobile-agent	✓	✓	✗	✓	✓	✗	✗

In the last few years, a lot of work has been done on the techniques of data aggregation, most of which have been seen to reduce energy consumption. There are still many areas on which very little work has been done like delay, heterogeneity and scalability, can be clearly seen in the above table. It is visible in this table that how much the areas have been explored.

III. Discussion& Insights

The discussion surrounding IoT data aggregation techniques opens up several avenues for further research. One key aspect that warrants attention is the scalability and adaptability of aggregation techniques in dynamic IoT environments. As the number of IoT devices continues to grow, it becomes crucial to develop aggregation methods that can handle large-scale data while maintaining efficiency and responsiveness. Another area for future exploration is the optimization of energy consumption in data aggregation. IoT devices often operate on limited battery power, and energy-efficient aggregation techniques can significantly prolong their operational lifetimes. Investigating innovative approaches that reduce energy consumption without compromising data accuracy is essential for sustainable IoT deployments. This review paper has touched upon privacy concerns in data aggregation. As IoT applications expand into sensitive domains such as healthcare and smart cities, preserving data confidentiality becomes paramount. Future research should focus on developing robust privacy-preserving aggregation techniques that ensure data security while still enabling meaningful analysis. The potential for integrating edge computing and fog computing concepts with data aggregation in the Internet of Things is enormous. By leveraging the computational capabilities of edge devices and utilizing localized aggregation, the burden on centralized cloud infrastructure can be reduced, leading to lower latency and improved scalability. Despite the challenges, data aggregation is an important technique for IoT. It can improve energy efficiency, network lifetime, scalability, and security. Researchers are working on new data aggregation techniques that can address the challenges and make data aggregation more efficient and secure.

IV. Conclusion

This review paper has offered a thorough rundown of data aggregation methods in the context of the Internet of Things (IoT). The analysis of various methodologies and algorithms has shed light on the importance of efficient data aggregation in managing the vast volume of data generated by interconnected IoT devices. The categorization of data aggregation techniques into centralized and distributed approaches has allowed for a deeper understanding of the different strategies employed in aggregating IoT data. Clustering-based, tree-based, and centralized-based aggregation algorithms have been examined, each with its own advantages and challenges. This paper also highlighted the challenges and open research problems in the area of data aggregation such as energy efficiency, scalability, and security. Finally, it concluded with some insights and future directions for research in this field.

References

[1] F. A. Alaba, M. Othman, I. A. T. Hashem, and F. Alotaibi, “Internet of Things security: A survey,” J. Netw. Comput. Appl., vol. 88, no. December 2016, pp. 10–28, 2017, doi: 10.1016/j.jnca.2017.04.002.

[2] J. Rezazadeh, K. Sandrasegaran, and X. Kong, “A location-based smart shopping system with IoT technology,” IEEE World Forum Internet Things, WF-IoT 2018 - Proc., vol. 2018-January, pp. 748–753, 2018, doi: 10.1109/WF-IoT.2018.8355175.

[3] D. Kwon, M. R. Hodkiewicz, J. Fan, T. Shibutani, and M. G. Pecht, “IoT-Based Prognostics and Systems Health Management for Industrial Applications,” IEEE Access, vol. 4, pp. 3659–3670, 2016, doi: 10.1109/ACCESS.2016.2587754.

[4] E. Fernandes, J. Jung, and A. Prakash, “Security Analysis of Emerging Smart Home Applications,” Proc. - 2016 IEEE Symp. Secur. Privacy, SP 2016, pp. 636–654, 2016, doi: 10.1109/SP.2016.44.

[5] D. Lu and T. Liu, “The application of IOT in medical system,” ITME 2011 - Proc. 2011 IEEE Int. Symp. IT Med. Educ., vol. 1, pp. 272–275, 2011, doi: 10.1109/ITiME.2011.6130831.

[6] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, “Cyber-Physical-Social Systems: A State-of-the-Art Survey, Challenges and Opportunities,” IEEE Commun. Surv. Tutorials, vol. 22, no. 1, pp. 389–425, 2020, doi: 10.1109/COMST.2019.2959013.

[7] P. P. Ray, “A survey on Internet of Things architectures,” J. King Saud Univ. - Comput. Inf. Sci., vol. 30, no. 3, pp. 291–319, 2018, doi: 10.1016/j.jksuci.2016.10.003.

[8] C. Chen, “Research on Trusted Certification Mechanism of Sensing Layer of the Internet of Things,” vol. 166, no. Amcce, pp. 18–22, 2018, doi: 10.2991/amcce-18.2018.4.

[8] R.V. Faizullin and S. Hering, “The Model of Data Aggregation from Clustered Devices in the Internet of Things,” Intellekt. Sist. Proizv., vol. 17, p. 156, 2020.

[9] A.S. Nandhini and P. Vivekanandan, “A Novel Security and Energy Efficient Data Aggregation Medical Internet of Things Using Trust,” Journal of Medical Imaging and Health Informatics, vol.10, pp. 249–255, 2020.

[10] C. Chen, “Research on Trusted Certification Mechanism of Sensing Layer of the Internet of Things,” vol. 166, no. Amcce, pp. 18–22, 2018, doi: 10.2991/amcce-18.2018.4.

[11] A. K. Pandey, B. Rajendran, and V. S. K. Roshni, “AutoAdd: Automated Bootstrapping of an IoT Device on a Network,” SN Comput. Sci., vol. 1, no. 1, pp. 1–5, 2020, doi: 10.1007/s42979-019-0047-3.

[12] H. Karimipour and V. Dinavahi, “Parallel relaxation-based joint dynamic state estimation of large-scale power systems,” IET Gener. Transm. Distrib., vol. 10, no. 2, pp. 452–459, 2016, doi: 10.1049/iet-gtd.2015.0808.

[13] K. C. Okafor, I. E. Achumba, G. A. Chukwudebe, and G. C. Ononiwu, “Leveraging Fog Computing for Scalable IoT Datacenter Using Spine-Leaf Network Topology,” J. Electr. Comput. Eng., vol. 2017, 2017, doi: 10.1155/2017/2363240.

[14] J. Sakhnini et al., “Smart Grid Cyber Attacks Detection using Supervised Learning and Heuristic Feature Selection,” in 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), 2019, pp. 108–112.

[11] A. K. Pandey, B. Rajendran, and V. S. K. Roshni, “AutoAdd: Automated Bootstrapping of an IoT Device on a Network,” SN Comput. Sci., vol. 1, no. 1, pp. 1–5, 2020, doi: 10.1007/s42979-019-0047-3.

[12] H. Karimipour and V. Dinavahi, “Parallel relaxation-based joint dynamic state estimation of large-scale power systems,” IET Gener. Transm. Distrib., vol. 10, no. 2, pp. 452–459, 2016, doi: 10.1049/iet-gtd.2015.0808.

[13] K. C. Okafor, I. E. Achumba, G. A. Chukwudebe, and G. C. Ononiwu, “Leveraging Fog Computing for Scalable IoT Datacenter Using Spine-Leaf Network Topology,” J. Electr. Comput. Eng., vol. 2017, 2017, doi: 10.1155/2017/2363240.

[15] H. Elazhary, “Internet of Things (IoT), mobile cloud, cloudlet, mobile IoT, IoT cloud, fog, mobile edge, and edge emerging computing paradigms: Disambiguation and research directions,” J. Netw. Comput. Appl., vol. 128, pp. 105–140, 2019, doi: 10.1016/j.jnca.2018.10.021.

[16] I. Butun, P. Osterberg, and H. Song, “Security of the Internet of Things: Vulnerabilities, Attacks, and Countermeasures,” IEEE Commun. Surv. Tutorials, vol. 22, no. 1, pp. 616–644, 2020, doi: 10.1109/COMST.2019.2953364.

[17] Z. Qu, Y. Wang, L. Sun, D. Peng, and Z. Li, “Study QoS optimization and energy saving techniques in cloud, Fog, EDge, and IoT,” Complexity, vol. 2020, 2020, doi: 10.1155/2020/8964165.

[18] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, “A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System,” IEEE Access, vol. 7, pp. 70371–70421, 2019, doi: 10.1109/ACCESS.2019.2919657.

[19] A. L. R. Madureira, F. R. C. Araújo, and L. N. Sampaio, “On supporting IoT data aggregation through programmable data planes,” Comput. Networks, vol. 177, 2020, doi: 10.1016/j.comnet.2020.107330.

[20] L. M. R. Tarouco et al., “Internet of Things in healthcare: Interoperatibility and security issues,” IEEE Int. Conf. Commun., no. June, pp. 6121–6125, 2012, doi: 10.1109/ICC.2012.6364830.

[21] M. Abomhara and G. M. Køien, “Cyber security and the internet of things: Vulnerabilities, threats, intruders and attacks,” J. Cyber Secur. Mobil., vol. 4, no. 1, pp. 65–88, 2015, doi: 10.13052/jcsm2245-1439.414.

[22] S. De, P. Barnaghi, M. Bauer, and S. Meissner, “Service modelling for the Internet of Things,” 2011 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2011, no. September, pp. 949–955, 2011.

[23] S. Chen, H. Xu, D. Liu, B. Hu, and H. Wang, “A vision of IoT: Applications, challenges, and opportunities with China Perspective,” IEEE Internet Things J., vol. 1, no. 4, pp. 349–359, 2014, doi: 10.1109/JIOT.2014.2337336.

[24] M. K. Saini and R. K. Saini, “Internet of Things (IoT) Applications and Security Challenges: A Review,” Int. J. Eng. Res. Technol., vol. 7, no. 12, pp. 1–7, 2019.

[25] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions,” Futur. Gener. Comput. Syst., vol. 29, no. 7, pp. 1645–1660, 2013, doi: 10.1016/j.future.2013.01.010.

[26] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, “From today’s INTRAnet of things to a future INTERnet of things: A wireless- and mobility-related view,” IEEE Wirel. Commun., vol. 17, no. 6, pp. 44–51, 2010, doi: 10.1109/MWC.2010.5675777.

[27] P. Zhang, J. Wang, K. Guo, F. Wu, and G. Min, “Multi-functional secure data aggregation schemes for WSNs,” Ad Hoc Networks, vol. 69, pp. 86–99, 2018, doi: 10.1016/j.adhoc.2017.11.004.

[28] M. Gheisari, G. Wang, and S. Chen, “An Edge Computing-enhanced Internet of Things Framework for Privacy-preserving in Smart City,” Comput. Electr. Eng., vol. 81, 2020, doi: 10.1016/j.compeleceng.2019.106504.

[29] S. Mishra, H. T.-I. J. E. I. Technol.(IJEIT), and undefined 2012, “Features of WSN and Data Aggregation techniques in WSN: A Survey,” Researchgate.Net, vol. 1, no. 4, 2012.

[30] Q. Wang and T. Zhang, “Characterizing the traffic load distribution in dense sensor networks,” 3rd Int. Conf. New Technol. Mobil. Secur. NTMS 2009, pp. 1–4, 2009, doi: 10.1109/NTMS.2009.5384829.

[31] G. C. Jagan and P. J. Jayarin, “A Novel Machine Language-Driven Data Aggregation Approach to Predict Data Redundancy in IoT-Connected Wireless Sensor Networks,” Wirel. Commun. Mob. Comput., vol. 2022, 2022, doi: 10.1155/2022/7096561.

[32] J. N. S. Rubí and P. R. L. Gondim, “IoMT platform for pervasive healthcare data aggregation, processing, and sharing based on oneM2M and openEHR,” Sensors (Switzerland), vol. 19, no. 19, pp. 1–25, 2019, doi: 10.3390/s19194283.

[33] F. Derakhshan and S. Yousefi, “A review on the applications of multiagent systems in wireless sensor networks,” Int. J. Distrib. Sens. Networks, vol. 15, no. 5, 2019, doi: 10.1177/1550147719850767.

[33] F. Derakhshan and S. Yousefi, “A review on the applications of multiagent systems in wireless sensor networks,” Int. J. Distrib. Sens. Networks, vol. 15, no. 5, 2019, doi: 10.1177/1550147719850767.

[34] S. Yousefi, H. Karimipour, and F. Derakhshan, "Data Aggregation Mechanisms on the Internet of Things: A

Systematic Literature Review," Internet of Things, vol. 15, Sep. 2021.

[35] G. Mehmood, M. Z. Khan, M. Fayaz, M. Faisal, H. U. Rahman, and J. Gwak, "An energy-efficient mobile agent-based data aggregation scheme for wireless body area networks," Comput. Mater. Contin., vol. 70, no. 3, pp. 5929–5948, 2022, doi: 10.32604/cmc.2022.020546.

[36] S. Otoum, B. Kantarci, and H. Mouftah, "Adaptively supervised and intrusion-aware data aggregation for wireless sensor clusters in critical infrastructures," IEEE Int. Conf. Commun., vol. 2018-May, pp. 1–6, 2018, doi: 10.1109/ICC.2018.8422401.

[37] S. Najjar-Ghabel and S. Yousefi, "Enhancing Performance of Face Detection in Visual Sensor Networks with a Dynamic-based Approach," Wirel. Pers. Commun., vol. 97, no. 4, pp. 6151–6166, 2017, doi: 10.1007/s11277-017-4832-9.

[38] A. R. Khan and M. A. Chishti, "Data aggregation mechanisms in the internet of things: A study, qualitative and quantitative analysis," Int. J. Comput. Digit. Syst., vol. 9, no. 2, pp. 289–297, 2020, doi: 10.12785/IJCDS/090214.

[39] S. Sanyal and P. Zhang, "Improving Quality of Data: IoT Data Aggregation Using Device to Device Communications," IEEE Access, vol. 6, 2018, pp. 67830-67840.

[40] A. Sinha and D. K. Lobiyal, "Performance evaluation of data aggregation for cluster-based wireless sensor network," Human-centric Comput. Inf. Sci., vol. 3, no. 1, pp. 1–17, 2013, doi: 10.1186/2192-1962-3-13.

[41] S. Yousefi, H.Karimipour, F.Derakhshan, Data Aggregation Mechanisms on the Internet of Things: A Systematic Literature Review, Internet of Things, Volume 15,2021,100427, ISSN 2542-6605,https://doi.org/10.1016/j.iot.2021.100427.

[42] C. -H. Tsai, H. -Y. Huang, C. -W. Hung and Y. -H. Wang, "TDAM: A Tree-based Data Aggregation Mechanism in wireless sensor networks," in 2012 International Symposium on Intelligent Signal Processing and Communications Systems, Tamsui, Taiwan, 2012, pp. 827-832, doi:10.1109/ISPACS.2012.6473606.

[43] H. Yanhua and Z. Xincai, "Aggregation tree based data aggregation algorithm in wireless sensor networks," Int. J. Online Eng., vol. 12, no. 6, pp. 10–15, 2016, doi: 10.3991/ijoe.v12i06.5408.

[44] S. Abbasian Dehkordi, K. Farajzadeh, J. Rezazadeh, R. Farahbakhsh, K. Sandrasegaran, and M. Abbasian Dehkordi, "A survey on data aggregation techniques in IoT sensor networks," Wireless Networks, 2019, [Online]. Available: https://doi.org/10.1007/s11276-019-02142-z.

[45] S. Sirsikar and S. Anavatti, "Issues of data aggregation methods in Wireless Sensor Network: A survey," Procedia Comput. Sci., vol. 49, no. 1, pp. 194–201, 2015, doi: 10.1016/j.procs.2015.04.244.

[46] S. S. G and S. M. Sundaram, "Data Aggregation Techniques Over Wireless Sensor Network- A Review," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH.

[47] M. El Fissaoui, A. Beni-Hssane, and M. Saadi, "Multi-mobile agent itinerary planning-based energy and fault aware data aggregation in wireless sensor networks," Eurasip J. Wirel. Commun. Netw., vol. 2018, no. 1, 2018, doi: 10.1186/s13638-018-1099-0.

[48] F. Derakhshan and S. Yousefi, "A review on the applications of multiagent systems in wireless sensor networks," Int. J. Distrib. Sens. Networks, vol. 15, no. 5, 2019, doi: 10.1177/1550147719850767.

[49] S. S. Sruthi and G. Geethakumari, "An Efficient Secure Data Aggregation Technique for Internet of Things Network: An Integrated Approach Using DB-MAC and Multi-path Topology," Proc. - 6th Int. Adv. Comput. Conf. IACC 2016, pp. 599–603, 2016, doi: 10.1109/IACC.2016.116.

[50] H. Jiang, F. Shen, S. Chen, K. C. Li, and Y. S. Jeong, "A secure and scalable storage system for aggregate data in IoT," Futur. Gener. Comput. Syst., vol. 49, pp. 133–141, 2015, doi: 10.1016/j.future.2014.11.009.

[51] M. Huang, A. Liu, T. Wang, and C. Huang, "Green Data Gathering under Delay Differentiated Services Constraint for Internet of Things," Wirel. Commun. Mob. Comput., vol. 2018, 2018, doi: 10.1155/2018/9715428.

[52] S. J. Ashaj and E. Erçelebi, "Energy Saving Data Aggregation Algorithms in Building Automation for Health and Security Monitoring and Privacy in Medical Internet of Things," Journal of Medical Imaging and Health Informatics, vol. 10, 2020, pp. 204–210, https://doi.org/10.1166/jmihi.2020.2717.

[53] R. Lu, K. Heung, A. H. Lashkari, A. A. Ghorbani, "A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT," IEEE Access, vol. 5, 2017, pp. 3302–3312, https://doi.org/10.1109/ACCESS.2017.2677520

[54] K. Thangaramya, K. Kulothungan, R. Logambigai, M. Selvi, S. Ganapathy, and A. Kannan, "Energy aware cluster and neuro-fuzzy based routing algorithm for wireless sensor networks in IoT," Comput. Networks, vol. 151, pp. 211–223, 2019, doi: 10.1016/j.comnet.2019.01.024.

[55] A. Ullah, M. Azeem, H. Ashraf, A. A. Alaboudi, M. Humayun, and N. Z. Jhanjhi, "Secure Healthcare Data Aggregation and Transmission in IoT - A Survey," IEEE

Access, vol. 9, 2021, pp. 16849–16865, https://doi.org/10.1109/ACCESS.2021.3052850.

[56] B. Pourghebleh and N. J. Navimipour, “Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research,” J. Netw. Comput. Appl., vol. 97, pp. 23–34, 2017, doi: 10.1016/j.jnca.2017.08.006.

[57] N. Chandnani and C. N. Khairnar, "Efficient Data Aggregation and Routing Algorithm for IoT Wireless Sensor Networks," in IFIP Int. Conf. Wirel. Opt. Commun. Networks, WOCN, vol. 2019,

[58] Q. T. Minh, V. A. Le, T. K. Dang, T. Nam, and T. Kitahara, “Flow aggregation for SDN-based delay-insensitive traffic control in mobile core networks,” IET Commun., vol. 13, no. 8, pp. 1051–1060, 2019, doi: 10.1049/iet-com.2018.5194.

[59] S. Yousefi, H. Karimipour, and F. Derakhshan, "Data aggregation mechanisms on the internet of things: a systematic literature review," Internet of Things, vol. 15, 2021, Art. no. 100427, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2021.100427.

[60] H. Rahmann and N. Ahmad, "Comparison of data aggregation techniques in Internet of Things (IoT)," IEEE Access.

[61] A. L. R. Madureira, F. R. C. Araújo, and L. N. Sampaio, “On supporting IoT data aggregation through programmable data planes,” Comput. Networks, vol. 177, 2020, doi: 10.1016/j.comnet.2020.107330

DOI: 10.15439/2023R74
ISSN 2300-5963 ACSIS, Vol. 38

Supply Chain for Agriculture Products Using Blockchain Technology

Pagalla Bhavani Shankar
Computer Science & Engineering
University College of Engineering and Technology, Krishna University
Rudravaram, Machilipatnam, India
ORCID - 0000-0001-5935-7758

M. Babu Reddy
Computer Science
University College of Arts & Sciences
Krishna University
Rudravaram,Machilipatnam, India
m_babureddy@yahoo.com

Yarlagadda Divya Vani
Computer Science & Engineering
Sri Vasavi Engineering College (Autonomous)
Tadepalligudem, India
divyasudha99@gmail.com

***Abstract*—Nowadays, the cost of agricultural products is increasing progressively, even the farmers are not getting a sufficient enough for their production, due to the involvement of the corrupted intermediaries. Debts of the farmers have been increased, which leads to their committing suicide. As of now, no automation method exists to track the price of agricultural goods. To reduce corruption and to help the farmers to get a profit, there is a need to trace the product's costs and store them in a file starting from the product's manufacturer. Blockchain technology works as a decentralized and immutable ledger. A block chain is a sequence that maintains a list of transactions. These blocks contain information on the transactions. It protects against any critical cyber-attack, thereby providing the utmost security and encrypted data with privacy. It helps us store all the transaction history of the product from its manufacturing. Our aim is to implement a supply chain for agricultural products using Blockchain technology, to track and tackle the price fluctuation in the market.**

***Index Terms*—Blockchain Technology, cyberattacks, encryption, supply chain.**

I. Introduction

The modern supply chain has grown exceedingly intricate. At different phases, a large assortment of stakeholders is available. All of these stakeholders must cooperate with one another in a number of ways for management to be efficient and successful. False advertisement is less likely when there are effective traceability measures, in place to halt the creation and marketing of risky or inferior product. Accurate tracking and adherence to national rules are necessary for the movement of agricultural products among various nations. In the agriculture industry, by clearly identifying the source and requiring many data exchanges across the logistic network, monitoring commodities requires the collection, communication, and management of crucial data.

The main factors affecting the agriculture supply chain's current traceability measures are centralized controls with information-vulnerable centralized data fragmentation management additionally modification. In the toxic situation, we rapidly find the culprit, then remove the contaminated items out of the supply chain.

Blockchain helps us to prevent the data from contamination. Blockchain is the technology where each and every transaction is represented in the form of block and each block contains 3 parts those are -> Data of that block, both the hash codes for the current block and the previous block. If anyone tries to change or manipulate the data in any block the data from that block will be lost as each block is linked to the previous block hash code.

II. Literature Review

The main aim to bring out this article is “Food shortage during the covid-19 pandemic situation”. There are many changes in buying the food products during pandemic. Sometimes there will be shortage for the food products and sometimes there will be shortage of consumers, it creates the imbalance in the food supply chain. To tackle this problem, they should use the blockchain technology and developed a model which is used to interact with the farmers and mediators, to get the crop products which are available at them [1].

Agriculture is the backbone of a country. Population of a country increasing, steadily. To fulfil the needs of the increasing population, there is a necessity to shield the supply chain conveyance in a rectifiable manner. To overcome this situation, blockchain-supply chain management is a key-way to lead the better optimistic solution to the above cited matter [12,13].

Bitcoin was invented decades ago. Bitcoin addresses the flaws with digital tokens, which may be readily duplicated or produced, by using blockchain technology [11]. Every transaction is called as a block and these blocks are linked together, in blockchain technology. These blocks are linked by a hash code. The hash code is generated for each and every block. Hash code is generated by merging data with the previous block of hash code. If we try to change the data in one block, we lost all the data from that particular block, it provides security in that way. Every block contains [11] 3 parts (Data, Previous block Hash code, Time stamp). We generate the hash code to a particular block by merging these three [2] nodes.

III. Methodology

A distributed database calls a block chain, to stores every networked transaction. Every module of this database is a "block." Once the status of a transaction changes, accumulation of block, to the block chain in a sequential and linear way, when it has a connection to the block earlier to it. The newly formed blocks are then recreated, across the system to guarantee, that every node has recourse to the equivalent block chain. There is the identical of a block chain on everyone who takes a part in this transaction. Consequently, a specific transaction may be corroborated by any partaker.

With this strategy, there was no longer- requirement for centralized, reliable third-party confirmation of transactions. With numerous potentials use of the block chain technology, there is a tone of possibility for creativity, in the era of technology. So, corporate executives will utilize this technology to explore the variety of prospects obtainable to their organization and sector.

IV. Algorithm

As far as we are aware, hashing is not "encryption" as the original text cannot be recovered. Considering that, a hash is a unidirectional cryptographic feature with a predetermined size for every source-text size, comparing "hashed" copies of texts is simpler than decoding the text to attain the early design. SHA-256 is one of the most potent hash algorithms is in use, which replaces SHA-1 and is occasionally referred to as SHA-2. As of yet, SHA-256 has not been hacked in any manner and is not any harder to develop than SHA-1. AES is a great support function for the 256-bit key, in SHA-256 algorithm. FIPS 180-4, a NIST (National Institute of Standards and Technology) standard, has a description of it. Message Digest: A cryptographic hash technique that can engender a message, digested from binary digits, is simulated by the Java Message Digest session. After getting encrypted data, it is intolerable to verify if there was a tampering while it was being transported. With a digest message, this issue will be simplified. In order to check for data tampering during transmission, From the encrypted data, the transmitter will create a digest message, which will be sent with the data. After decrypting and analyzing the encrypted data, you may compare the measured message digest to the message digest that was included with the data.

V. Implementation

A. Algorithm of SHA-256

- The SHA-256, or Secure Hash Algorithm is a candidate of the SHA algorithm family.
- The 256 in the name refers to the maximum hash digest value; as a result, the hash value is always 256 bits, regardless of the size of the plaintext or cleartext.

B. Steps

1) Pre-processing

Transmit a message m0,m1...mb-1 (b-bit);

Step 1: Append padding bits and get m0, m1... m(b-1)10...0 (b' bits, where b'=448 mod 512);

Step 2: Add 64 bit and obtain m0, m_1... m(b-1)10... 0b0b1...b63 (b0b1 ... b63 represents b in 64-bit form);

Step 3: Make 32-bit words out of it as Mo, M1, ..., M(N-1) where (N is a multiple of 16);

2) Initialization: Initialize 8 buffers are A, B, C, D, E, F, G & H as

A = 0x578e4ab3
B= 0x642eab54
C = 0x9473eafb
D = 0xa6458cde
E = 0x54cabd3f
F = 0x8a1639ad
G = 0xabd3e231
H = 0x6458abce

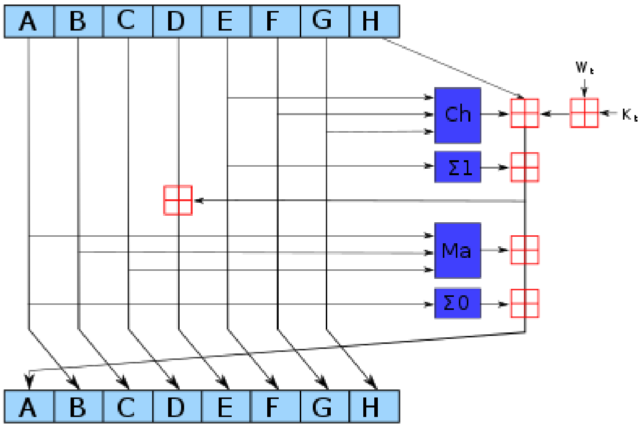


Fig. 1. SHA-256 Algorithm

3) The complete messages are separated into numerous blocks of 512 bits of each. Each block's output is used as the one's input, and that follows it in a total of 64 processing rounds for to each block.

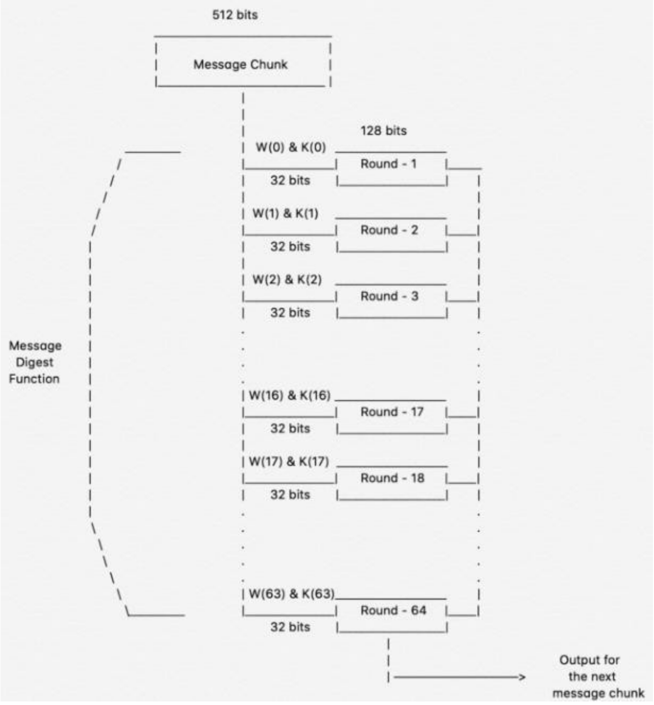


Fig. 2. SHA-256 Algorithm

Each time an iteration occurs, the final output of one block becomes the initial input of the succeeding block. the whole cycle is repeated up to the final 512-bit block, after which the outcome is taken to be the final hash digest. This digest will be 256 bits long, as the algorithm's name designates.

VI. System Architecture

The System architecture undoubtedly elucidates the working of the system.

The details should be uploaded by the farmer, that are encrypted and the resultant hash codes should be sent to the farmer mail and the supplier can access them by requesting the hash codes received from the farmer.

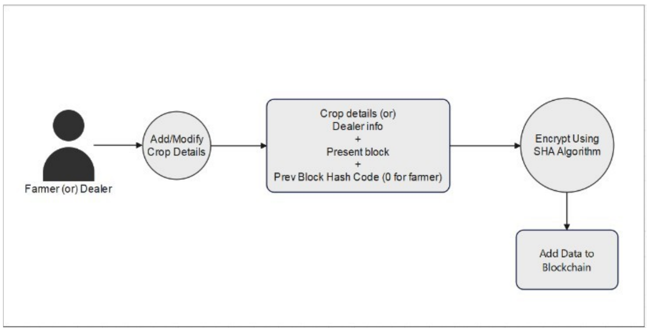
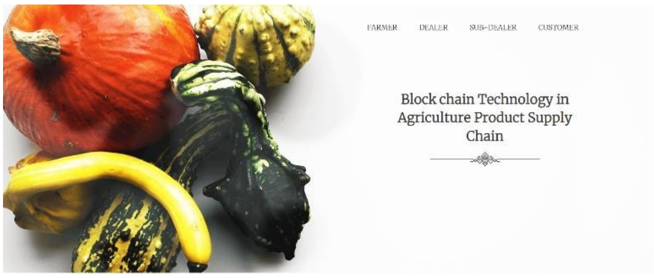

Famer (or) Dealer
Add/Modify Crop Details
Crop details (or) Dealer info + Present block + Prev Block Hash Code (0 for farmer)
Encrypt Using SHA Algorithm
Add Data to Blockchain

Fig. 3. System Architecture

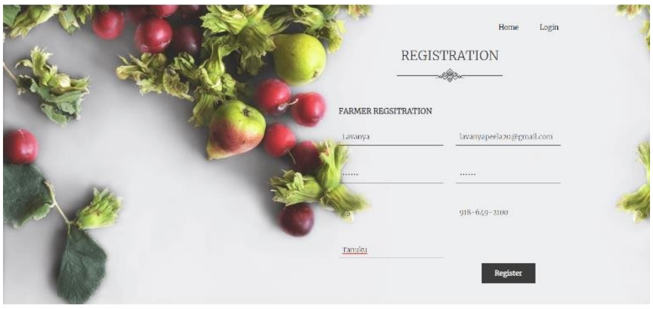
VII. Result Analysis

A. Home Page:



Block chain Technology in Agriculture Product Supply Chain

B. Farmer Registration Page:



REGISTRATION

FARMER REGSITRATION

Lavanya

lavanyapeela20@gmail.com

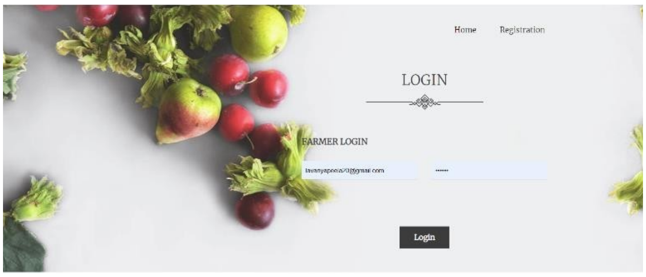
......

......

918-649-2100

Tanuku

C. Farmer Login Page:



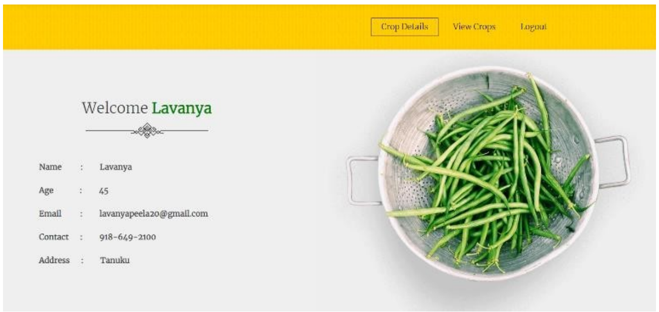
LOGIN

FARMER LOGIN

lavanyapeela20@gmail.com

......

D. Farmer Home Page:



Welcome Lavanya

Name : Lavanya

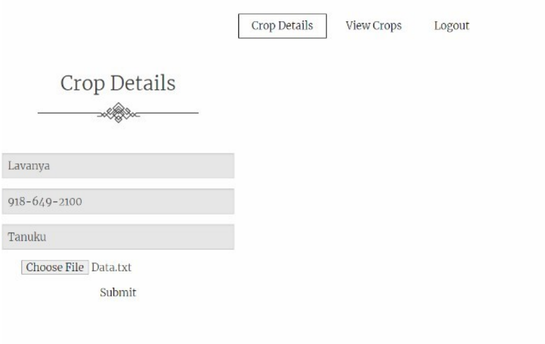
Age : 45

Email : lavanyapeela20@gmail.com

Contact : 918-649-2100

Address : Tanuku

E. Uploading Crop Details:



Crop Details

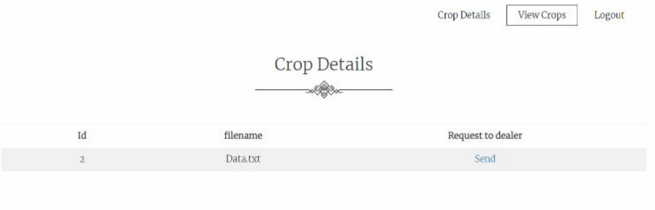
Lavanya

918-649-2100

Tanuku

Data.txt

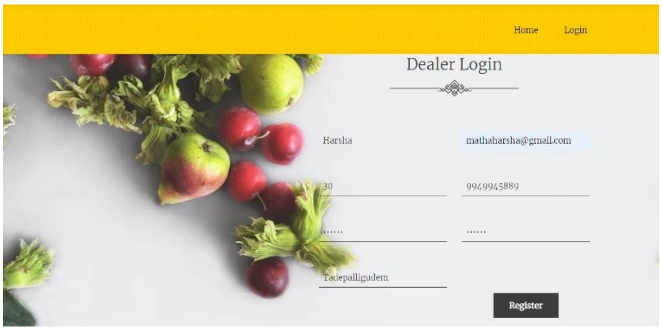
F. After Uploading crop details:



Crop Details

Id	filename	Request to dealer
2	Data.txt	Send

G. Dealer Registration Page:



Dealer Login

Harsha

mathaharsha@gmail.com

30

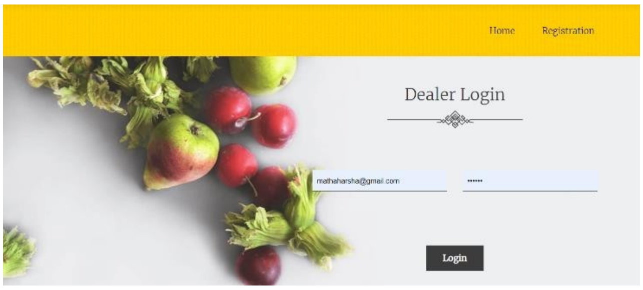
9949945889

......

......

Tadepalligudem

H. Dealer Login Page:

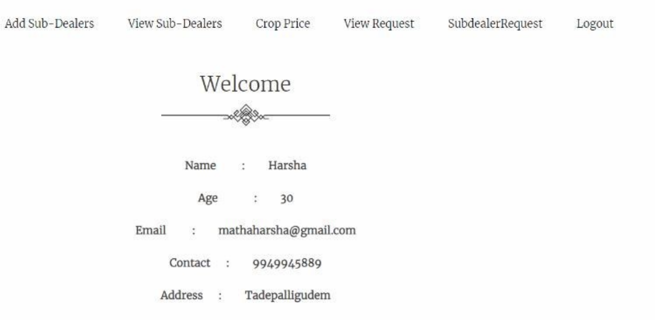


Dealer Login

mathaharsha@gmail.com

......

I. Dealer Home Page:



Welcome

Name : Harsha

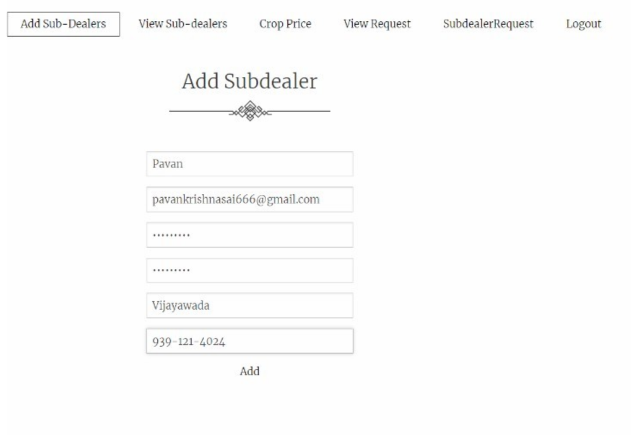
Age : 30

Email : mathaharsha@gmail.com

Contact : 9949945889

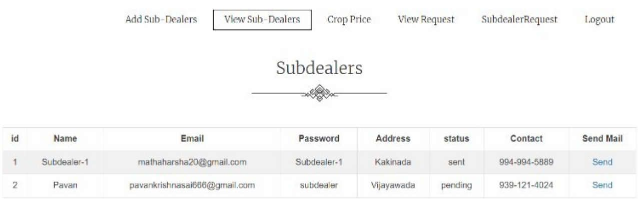
Address : Tadepalligudem

J. Adding Sub-Dealer:



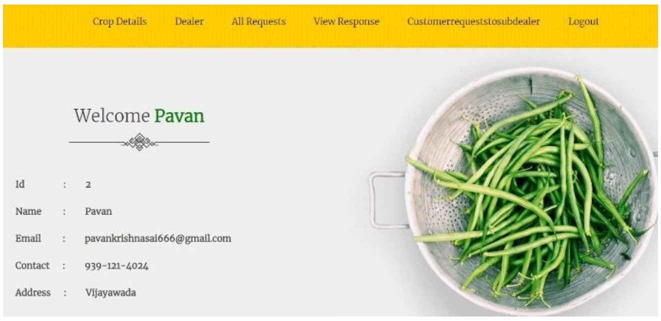
Pavan
pavankrishnasai666@gmail.com
....
....
Vijayawada
939-121-4024

K. Sub-Dealers list:



id	Name	Email	Password	Address	status	Contact	Send Mail
1	Subdealer-1	mathaharsha20@gmail.com	Subdealer-1	Kakinada	sent	994-994-5889	Send
2	Pavan	pavankrishnasai666@gmail.com	subdealer	Vijayawada	pending	939-121-4024	Send

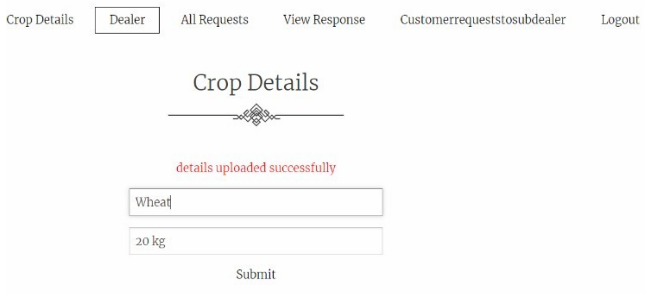
L. Sub-Dealer Home Page:



Welcome Pavan

Id	:	2
Name	:	Pavan
Email	:	pavankrishnasai666@gmail.com
Contact	:	939-121-4024
Address	:	Vijayawada

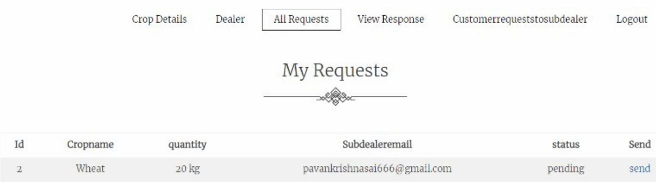
M. Sub-dealer requesting Crop:



details uploaded successfully

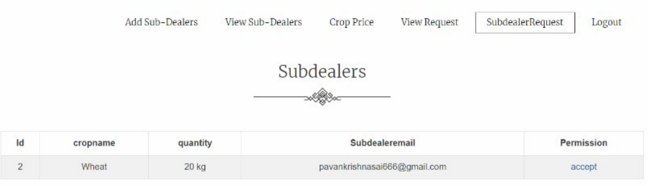
Wheat
20 kg

N. Sub-dealer request to dealer:



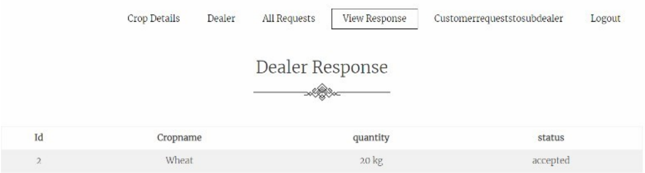
Id	Cropname	quantity	Subdealeremail	status	Send
2	Wheat	20 kg	pavankrishnasai666@gmail.com	pending	send

O. Dealer page: (Sub-dealer requests)



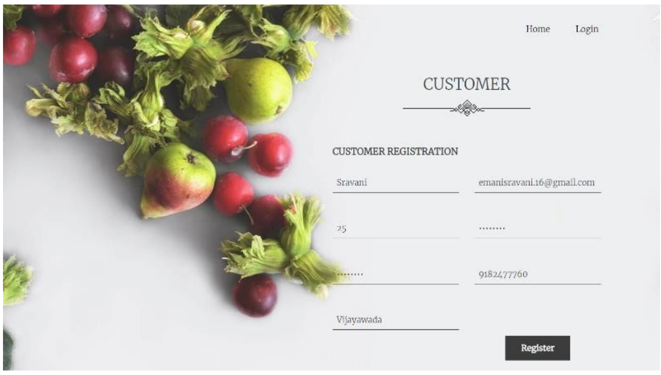
Id	cropname	quantity	Subdealeremail	Permission
2	Wheat	20 kg	pavankrishnasai666@gmail.com	accept

P. Responses of dealer:



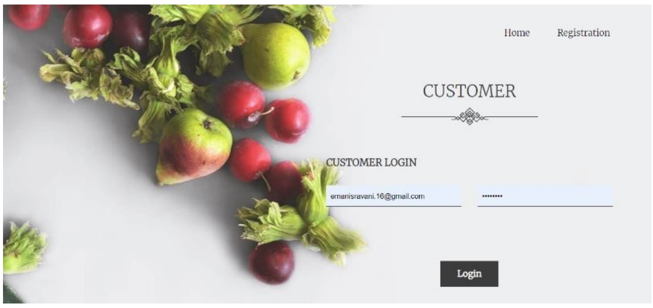
Id	Cropname	quantity	status
2	Wheat	20 kg	accepted

Q. Customer Registration Page:



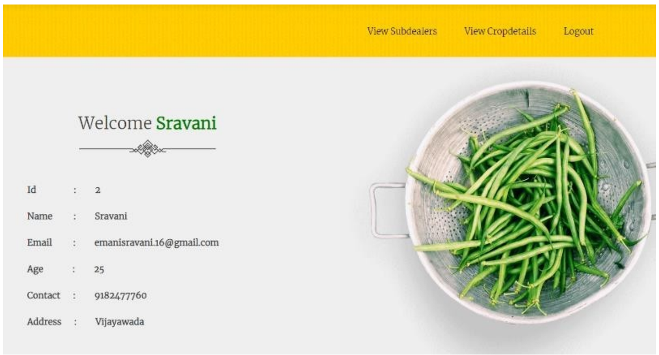
Sravani
emanisravani.16@gmail.com
25
....
....
9182477760
Vijayawada

R. Customer Login Page:



emanisravani.16@gmail.com
....

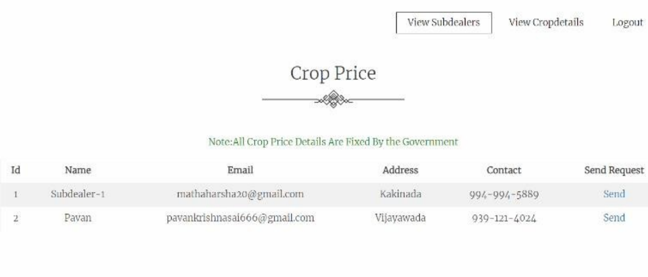
S. Customer Home Page:



Welcome Sravani

Id	:	2
Name	:	Sravani
Email	:	emanisravani.16@gmail.com
Age	:	25
Contact	:	9182477760
Address	:	Vijayawada

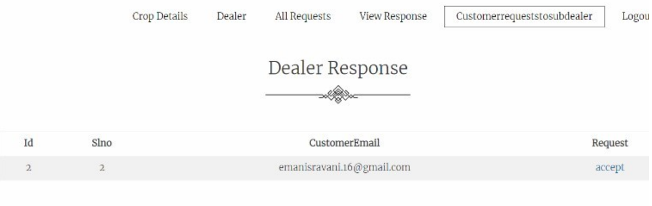
T. *List of Sub-Dealers:*



Note:All Crop Price Details Are Fixed By the Government

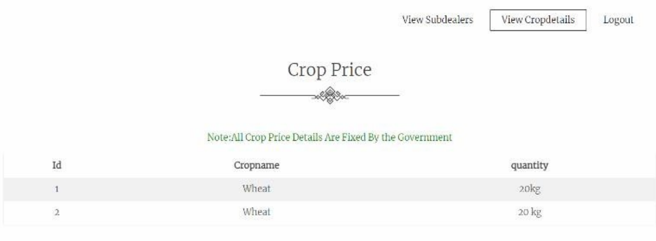
Id	Name	Email	Address	Contact	Send Request
1	Subdealer-1	mathaharsha20@gmail.com	Kakinada	994-994-5889	Send
2	Pavan	pavankrishnasai666@gmail.com	Vijayawada	939-121-4024	Send

U. *Request sent by Customer to Sub-Dealer:*



Id	Slno	CustomerEmail	Request
2	2	emanisravani.16@gmail.com	accept

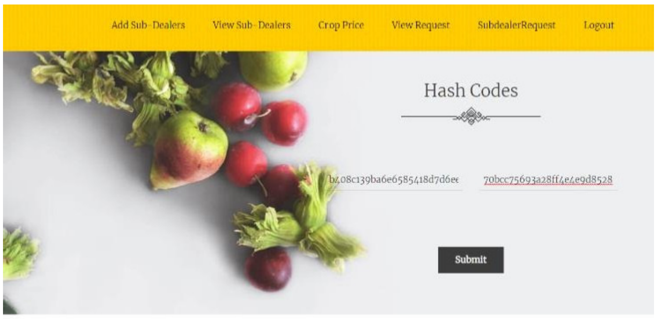
V. *Crop details in Customer Page:*



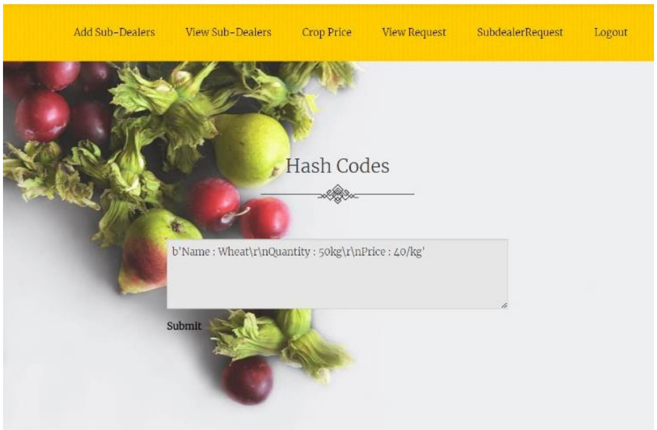
Note:All Crop Price Details Are Fixed By the Government

Id	Cropname	quantity
1	Wheat	20kg
2	Wheat	20 kg

W. *Customer accessing data uploaded by farmer using Hash keys:*



b408c139ba6e6585418d7d6ee 70bcc75693a28ff4e4e9d8528



b'Name : Wheat\r\nQuantity : 50kg\r\nPrice : 40/kg'

VIII. Conclusion

In this project we developed an application which is used to track the price of agriculture products in order to disregard the intercessor corrupted vendors. In this way all will get the crops with the price decided by the government and the end product price will be at a reasonable price, which helps the farmers. The data entered by the farmer should be encrypted by the blockchain technology, using cryptographic techniques. The dealer has to get the hash codes from the farmer in order to open the file and access the data from the block.

IX. Future Work

This study can be protracted to track the cost of the supplementary products in order to diminish the corruption. We can use more powerful encryption algorithms to store the data, in a secure manner. The use of blockchain (with the hash value or hask key) is explored further in other services.

References

[1] Blockchain technology for agricultural supply chains during the COVID-19 pandemic: Benefits and cleaner solutions.

[2] Bitcoin: A Peer-to-Peer Electronic Cash System.

[3] M. M. Aung and Y. S. Chang, “Traceability in a food supply chain: Safety and quality perspectives” Food Control, vol. 39, pp. 172_184, May 2014.

[4] Food traceability as an integral part of logistics management in food and agricultural supply chain.

[5] T. Bosona and G. Gebresenbet, “Food traceability as an integral part of logistics management in food and agricultural supply chain" Food Control, vol. 33, no. 2, pp. 32_48, 2013.

[6] Supply Chain Management in Agriculture Using Blockchain and IoT.

[7] T.-T. Kuo, H.-E. Kim, and L. Ohno-Machado, ‘‘Blockchain distributed ledger technologies for biomedical and health care applications,’’ J. Amer. Med. Inf. Assoc., vol. 24, no. 6, pp. 1211–1220, 2017.

[8] M. Mettler, ‘‘Blockchain technology in healthcare: The revolution starts here,’’ in Proc. 18th IEEE Int. Conf e- Health Net., Appl. Services, Sep. 2016, pp. 1–3.

[9] W. J. Gordon and C. Catalini, ‘‘Blockchain technology for healthcare: Facilitating the transition to patient- driven interoperability,’’ Comput. Struct. Biotechnol J., vol. 16, pp. 224–230, 2018.

[10] M. Hölbl, M. Kompara, A. Kamišalic, and L. N. Zlatolas, ‘‘A systematic review of the use of blockchain in healthcare ’’ Symmetry, vol. 10, no. 10, p. 470, 2018.

[11] Pagalla Bhavani Shankar, “Blockchain: The Essential Future of Modern Internet”, International Journal for Modern Trends in Science and Technology, Vol.6, issue.10, pp60-64.

[12] Nayal, Kirti, et al. "Antecedents for blockchain technology-enabled sustainable agriculture supply chain." *Annals of operations research* 327.1 (2023): 293-337.

[13] Srivastava, Praveen Ranjan, Justin Zuopeng Zhang, and Prajwal Eachempati. "Blockchain technology and its applications in agriculture and supply chain management: a retrospective overview and analysis." *Enterprise Information Systems* 17.5 (2023): 1995783.

[14] Chandan, Anulipt, Michele John, and Vidyasagar Potdar. "Achieving UN SDGs in Food Supply Chain Using Blockchain Technology." *Sustainability* 15.3 (2023): 2109.

DOI: 10.15439/2023R19
ISSN 2300-5963 ACSIS, Vol. 38

Survey of Big Data Analytics in IoT-Driven Healthcare Applications: A Comparative Approach

J R Shruti
Department of ISE
M S Ramaiah Institute of Technology
Bangalore
jrshruti@msrit.edu

Shubha Vibhu Malige
Department of CSE(CS)
M S Ramaiah Institute of Technology
Bangalore
shubha.malige@msrit.edu

***Abstract*—Integrating the Internet of Things(IoT) with advanced big data analytics (BDA) is crucial for reducing healthcare expenses and identifying potential risks. IOT BDA has an effect on information gathering, storing, extracting, and utilizing in the healthcare sector. It offers several benefits across many medical organizations' disciplines and improves the healthcare services. This review critically summarizes the advantages and disadvantages of IoT and BDA when individually used. The necessity of integrating the IoT and BDA is explained in detail. The scope for future work is proposed based on the various research gaps identified in the literature. Applications of IoT, Big data analytics techniques, challenges of IoT enabled healthcare are also discussed.**

***Index Terms*—Big Data, Internet of Things, Big Data Analytics, Healthcare**

I. Introduction

Imagine a world where hospitals are not just buildings, but living, breathing organisms, each patient a cell connected by a digital circulatory system. This captivating transformation is made possible by combining two technological wonders: the Internet of Things (IoT) and Big Data Analytics for the health sector.

A. *IoT: The Pulse of Smart Healthcare*

In this realm, medical devices are not mere tools; they are intelligent companions. IoT bestows life upon them. Sensors embedded in medical equipment, wearables, and even within patients' bodies, diligently collect data – heartbeats, oxygen levels, body temperature – relaying it in real-time to a central hub.

These digital whispers merge into a symphony of information, creating a real-time health profile for every patient. A surgeon in London can monitor a patient's vital signs in Tokyo, an ICU bed "talks" to the ventilator, orchestrating oxygen flow seamlessly. This interconnected ecosystem does not just treat diseases; it foresees them, allowing preventive care by spotting anomalies long before they manifest.

B. *Big Data Analytics*

But wait, this data deluge holds secrets far too intricate for the human mind to decipher. This is where Big Data Analytics steps into transmuting raw data into golden insights. It gazes at the patterns of millions, comparing genomes, identifying genetic predispositions, and predicting disease trajectories.

In hospital corridors accessorized with screens displaying real-time data visualisations, doctors morph into diagnosticians of the future. Treatment plans are tailored with pinpoint precision, minimizing adverse reactions, and optimizing outcomes. Disease outbreaks are spotted at their inception.

Moreover, this duo extends beyond hospital walls. Wearables prompt users to take a walk if their activity drops, smart pill dispensers remind the elderly to take their medication, and AI-driven chatbots offer medical advice, comforting the worried at any hour.

In this health sector, there is a proverb that says “Prevention is better than cure" which transforms into a adopting healthier lifestyle throughout, instead of becoming a victim of diseases. Hospital beds are not just for the sick but a safe place for proactive well-being.

So, in this new era of healthcare, remember that it is not just about the "Internet of Things" or "Big Data Analytics." It is about a synchronization established; an orchestra conducted by innovation.

The rest of this paper is structured as follows. In Section II we introduce the related work which will give insights into various referred papers and their summary toward Big data Analytics and IOT in health care. In Section III different big data analytics techniques are discussed along with the steps that are necessary before analysis such as data preprocessing, collection, and visualization. In Section IV, the different applications of IOT in health care are summarized. Section V throws light on different challenges encountered in the IOT healthcare sector. Sections VI and VII talk about comparative study and analysis of various big data techniques and future work. Finally, we conclude this paper and discuss what can be improved in future research.

II. Related Work

To anticipate diseases in the healthcare industry, combined approach includes IoT and Big Data analytics is proposed[1]. They created a system that uses the Internet of Things to track the wellness of the heart and send alarms when something is off. For those who are elderly or less mobile, this is beneficial. But IoT alone is insufficient to accurately anticipate diseases. To address this issue, an advanced data analysis and machine learning approach is proposed to enhance the system's ability to detect early disease warning signs.

IoT makes it possible for machines, people, and services to communicate with one another to enhance daily life. Real-time access to diagnosis and treatment has benefits for the healthcare industry as well. The usage of IoT in health care facilities and its future developments, with an emphasis on the management of healthcare using IoT is discussed[2]. It demonstrated how wearable sensors and the IoT make it possible for medical care to be provided anywhere. But there are still difficulties with Big data management, privacy, and significant expenses. Future IoT healthcare developments include seizure and stroke prediction, as well as the use of sensors for on-demand medical assistance.

Wearable biosensors enable personalized health solutions as a result of the fusion of technology and healthcare, which is fueled by the expansion of IoT and Big Data. But there are still issues with device dependability and safety, which calls for cooperation between tech creators and medical specialists. A special edition is covered that analyses the role of IoT in smart health sensors with an emphasis on wearables and Big Data analytics for informed devices, with the goal of enhancing individualised tele-health interventions and healthier lifestyles[3].

Emerging technologies like 6G, extended reality (XR), and IoT big data analytics can address current issues in healthcare like staffing and telehealth. These innovations could lead to novel services like telepresence and improved patient experiences. XR improves healthcare practices, while IoT generates data for better services and treatments. A research gap is filled by reviewing how these technologies converge to shape the future of healthcare, discussing applications, challenges, and directions[4].

The use of Deep Learning in Healthcare analytics such as electronic medical records, Genomics, and the development of drugs is examined[5]. Deep learning techniques are used to process the information in the medical data. These techniques can also be used by biomedical data to improve the level of guidance to doctors and improve medical health. Furthermore, the challenges and difficulties encountered in using Deep Learning for Healthcare analytics are discussed.

Healthcare is being transformed into Healthcare 4.0 by Industry 4.0 technologies including IoT, Cloud Computing, and Big Data. The impact they have on conventional healthcare services is examined[6], along with applications, advantages, and multidisciplinary difficulties.

The demand for advanced healthcare solutions has increased due to the rise of chronic disorders like COVID-19. IoT-driven wearables collect a lot of health data, which poses a significant data challenge. Machine learning approach is used to handle large datasets generated by IoT for healthcare[7]. It provided an insight into the most recent machine learning developments for efficient healthcare strategies for healthcare professionals and authorities.

III. Big Data Analytics Techniques

By choosing the suitable big data technology, an organisation will be able to handle the elevation in volume, velocity, and variety of data in a better way[8]. Big data analytics involves the usage of various techniques and tools to manage and extract insights from large datasets. This process confines several key components: data collection, storage, processing and visualization.

Data collection: In this stage, data is gathered from a variety of sources, including unstructured (such as text, photos, and videos) and structured (such as databases). Sensors, social media, online transactions, polls, and other sources can all be used to collect data. Import.io is an effective tool for extraction of WebPages data.

Data Storage: After data collection, The data must be stored in a fashion that makes retrieval fast and easy. Big data is frequently stored and managed using tools like distributed file systems (like Hadoop HDFS) and NoSQL databases (like MongoDB and Cassandra). These tools are configured to provide high availability, fault tolerance, and data redundancy.

Data Processing: Analysing and extracting insights from big data often involves complex computational tasks. Technologies like Apache Spark provide distributed computing capabilities to process data in parallel across a cluster of machines. MapReduce, a software framework, is commonly used for distributed, parallel processing of massive data collections.

Data Visualization: Converting unprocessed data into useful visual representations like graphs, charts, and dashboards is an essential step. Data analysts and decision-makers can better understand patterns, trends, and anomalies in the data with the use of visualisation. Users can generate dynamic and educational visualisations that support data-driven insights and well-informed decision-making using tools like Tableau, Silk, CartoDB,Power BI, and D3.js.

IV. Applications of IOT in Healthcare

Certainly, there are several innovative and unique ways in which IoT (Internet of Things) can be applied in healthcare beyond the conventional(traditional) use cases. Here are some innovative applications:

Smart Pill Dispensers and Medication Monitoring: IoT-enabled pill dispensers can remind patients to take their medications at the right time and track their adherence. These devices can also connect to healthcare providers, alerting them if a patient misses a dose or requires a medication adjustment.

Personalized Nutrition Monitoring: IoT devices can track a patient's diet and nutritional intake in real-time, analyzing data to provide personalized dietary recommendations and helping individuals make healthier food choices.

Emotion and Stress Monitoring: Wearable IoT devices equipped with sensors can measure physiological responses such as heart rate variability, skin conductivity, and even brainwave patterns. This data can be used to monitor emotional well-being and stress levels, allowing for early intervention or stress management techniques.

Patient Comfort and Satisfaction: IoT sensors in hospital rooms can monitor environmental factors such as temperature, lighting, and noise levels. This data can be used to create more comfortable and personalized patient experiences, potentially leading to faster recovery times.

Fall Detection and Prevention: IoT-enabled wearable devices can detect sudden changes in motion and orientation, alerting caregivers or medical personnel in real-time if a patient falls. This can be especially useful for elderly or at-risk individuals.

Wound Management: Smart bandages equipped with IoT sensors can monitor wound healing progress by tracking factors like moisture, infection, and inflammation. Healthcare providers can receive real-time updates and adjust treatment plans accordingly.

Assisted Living for Elderly: IoT devices in assisted living facilities can monitor daily activities and routines of elderly residents. Anomalies or deviations from normal behavior patterns can trigger alerts to caregivers, ensuring timely intervention in case of emergencies.

Telemedicine Enhancements: IoT devices can facilitate remote patient monitoring, allowing doctors to assess vital signs, conduct physical exams, and provide medical advice virtually. This is especially valuable for patients in remote areas or those with limited mobility.

Hospital Workflow Optimization: IoT sensors can track the movement of medical equipment, staff, and patients within a hospital, optimizing workflows, minimizing wait times, and improving resource allocation.

Pharmacy Inventory Management: IoT sensors in pharmacy storage areas can monitor medication inventory levels in real-time. This ensures medications are always available and prevents stockouts, helping to streamline patient care.

Predictive Healthcare Analytics: By analyzing data from IoT devices, machine learning algorithms can predict disease outbreaks, track trends in chronic conditions, and help healthcare organizations allocate resources more effectively.

Sleep Quality Monitoring: IoT-enabled sleep trackers can monitor sleep patterns, providing insights into sleep quality, duration, and potential disruptions. This information can assist in diagnosing sleep disorders and improving overall sleep hygiene.

Remote Rehabilitation and Physical Therapy: IoT devices can guide patients through personalized physical therapy exercises at home, tracking progress and adjusting routines based on real-time feedback.

Medication Authenticity Verification: IoT technology can be used to verify the authenticity of medications through embedded sensors and blockchain technology, reducing the risk of counterfeit drugs entering the supply chain.

These unique applications of IoT in healthcare demonstrate the potential to revolutionize patient care, improve outcomes, and enhance the overall healthcare experience. However, it's important to consider data privacy and security measures when implementing IoT solutions in the healthcare sector.

V. Challenges in IoT-enabled Healthcare

The acceptance of Internet of Things (IoT) technology in the healthcare industry has the capacity to transform how patients are cared for, diagnosed, and treated. IoT-enabled healthcare offers several benefits such as enhanced monitoring, improved patient outcomes, and streamline procedures by connecting medical equipment, sensors, and systems. Various challenges in IoT-enabled healthcare are discussed below:

1) **Security Challenges:** As more companies are adopting IoT, new security concerns will inevitably arise. These challenges might be related to the device's restrictions.The following are a few of these security challenges [2].

a) Rise of botnets: A botnet can impact a hospital without the management being aware of it[9]. This occurs due to the organization's insufficient security measures, preventing effective tracking of the botnet across all its devices.

b) Increased number of IoT devices: More IoT devices will result in greater security vulnerabilities being affected by businesses, which will increase the challenges for security experts.

c) Need for encryption: Encryption techniques are an efficient method of denying hackers access to information and are one of the important challenges for IoT security [10].

d) IoT financial-related breaches: Because organisations such as banks use IoT for electronic transactions, hackers will attack the devices and make illegitimate transactions. Currently, few organisations have adopted blockchain or machine learning to control financial fraud prior to its occurrence [11, 12]. Despite that, not every organisation agrees to this kind of security measure.

2) **Interoperability:** IoT devices used in the healthcare industry can have different vendors and communication protocols. For precise data sharing and efficient patient care, it is essential to provide seamless interoperability among these systems and devices.

3) **Reliability and Quality of Service:** IoT devices are essential for monitoring and treating patients. It is critical to ensure the reliability and optimal functioning of IoT-enabled healthcare systems.

4) **Energy Efficiency and Battery Life:** IoT devices used in healthcare systems operate on battery power. For systems that must run for longer periods of time, it might be difficult to increase battery life while maintaining precise and continuous monitoring.

5) **Ethical and Social Concerns:** The use of IoT devices in healthcare raises moral concerns about patient consent, data ownership, and the possibility of biases in the algorithms that are used to make diagnoses and prescribe treatments.

VI. Comparative Analysis of Big Data Analytics Approaches

A continuously rising demand for useful analytical tools has been apparent in recent years. In the analysis of massive amounts of data (Big Data, BD), this trend is also observable. Business performance, competitive advantage, and decision-making are all areas where organisations are aiming to leverage the power of big data [13, 14]. The management of healthcare has recently evolved from a disease-centered paradigm to a patient-centered paradigm, especially in value-based healthcare delivery models [15]. It is vital to handle and analyse healthcare Big Data in order to provide excellent patient-centered care and adhere to the specifications of the model. Big data analytics techniques including data mining, statistical analysis, web and text mining, and social media analytics are used in the healthcare domain. It assists with activities like improving diagnosis, preventing disease and providing real-time patient alerts.

Big Data Analytics in healthcare encompasses several types [16, 17, 18]: Descriptive Analytics examines historical and present data to provide insights into healthcare decisions, outcomes, and quality. It helps in creating reports, visualizations, and historical data queries[16]. Predictive Analytics forecasts future trends by analyzing historical health data for patterns and relationships. It assists in predicting treatment responses, foreseeing dangers, and locating hidden patterns. Prescriptive Analytics, uses knowledge of medical data and offers suggestions for complex healthcare choices. It serves as a guide for drug prescriptions, treatment options, and personalized medicine. Discovery Analytics utilizes existing knowledge to uncover new innovations, such as discovering new drugs, identification of undiagnosed diseases, and suggesting alternative treatments. It advances both medical research and healthcare procedures.

Collaboration between Data Scientists and Healthcare providers is necessary in order to maximize the benefits for both patients and medical organizations. One can deploy advanced federated learning algorithms in healthcare data analytics to safeguard data privacy and uphold the decentralized structure of the IoHT (Internet of Healthcare Data Things). Deep Q-Network(DQN) plays an important role in both conducting missing value analysis and labeling unlabeled data[19].

Data science greatly benefits from the convergence of big data analytics, machine learning, and artificial intelligence (AI). Deep Learning (DL), a category of machine learning that is inspired by neural networks, excels at interpreting complex data patterns. Its potential application in fields like IoT and healthcare is very promising[20]. Because neural networks in deep learning are organized in a similar way to how the brain is organized, they can extract complex details from data. This makes it easier to understand hierarchical data in the IoT space, and it benefits the healthcare industry by enabling quick decisions based on wearable and sensor data. Collaboration between disciplines results in significant shifts in the way decisions are made and how data is analyzed, creating breakthroughs across a range of industries.

Healthcare offers various prediction and prevention techniques. The main concern lies in fully capitalizing on all these opportunities because they are insufficient. A decentralized approach is being adopted by various parties involved, such as data scientists, experts in deep learning fields, and data set owners. The use of a Decentralized Transfer learning Model combines various deep learning techniques onto a unified platform, resulting in enhanced predictions[21] and optimized outcomes.

Advanced analytics are used in the healthcare industry to process vast amounts of patient and medical information for improved clinical outcomes and insights[22]. The integration of numerous scientific disciplines, including bio informatics, medical imaging, sensor informatics, medical informatics, and health informatics [23], enables the analysis of enormous patient data sets and the discovery of trends, correlations, and predictive models. A single article cannot adequately address the variety of Big Data Analytics methodologies employed in the healthcare industry.

In conclusion, when evaluating different approaches in Big Data Analytics, it is very essential to consider the type of data, the available analytical tools, and the expected outcomes. Utilizing several methods in combination often proves to be beneficial, and provides an efficient approach to manage data and obtain valuable insights.

VII. Comparative Study

A comparative survey of research papers which focus on integrating Big Data Analytics and Internet of Things technologies is presented in Table I.

Table I. A comparative survey on integrating Big Data Analytics and Internet of Things technologies

SI.No	Research Findings	Future Work	Ref
1	–Recommended the electrocardiogra (ECG) system with Internet of Things assistance for secure data transmission for ongoing cardiovascular health surveillance. –Helped to comprehend the many healthcare technologies, such as ECG and the monitoring of EMG	–Explores existing local health systems technology and uses the latest technology that will be developed for the next research.	[24]
2	-Various healthcare data platforms and machine learning algorithms are discussed. -Big data lifecycle challenges like processing, storage, and security are addressed. -Big data analytics framework is presented for real-time disease prediction, including cancer, diabetes, Alzheimer's and heart.	-New mining techniques should be created to enable the extraction of sensitive data from the enormous amount of health data.	[25]
3	-The potential of Big Data Analytics is analyzed in healthcare, and focused on the use of unstructured and structured data in medical facilities. –Only a few dimensions are investigated to characterize the use of data by healthcare institutions.	–Future studies may look at the advantages that healthcare organizations obtain from analyzing both unstructured and structured data in the clinical domains, in addition to any challenges they face in these domains.	[26]
4	–An architecture of real-time data analytics for an IoT-based smart healthcare system is presented. –Radio-frequency identification technology and wireless sensor network is comprised. The proposed work discussed various data analytics tools to attain high-performance, like Spark, Kafka, NodeJS and MongoDB. –A diagnosis of Wolff–	NA	[27]

SI.No	Research Findings	Future Work	Ref
	Parkinson–White syndrome by logistic regression is outlined to evaluate the performance of the developed system. –The results suggest that the proposed system can process medical data in real-time with a high accuracy rate successfully and handle huge amounts of data.		
5	–A data analytics and privacy preservation model using deep learning approach is introduced for IoT-enabled healthcare systems to address the security issue. –To analyze the health-related information in the cloud a convolutional neural network(CNN) is used. –A secure access control module is presented to work on the attributes of the user for the IoT based healthcare system. –The recall, F1 score and precision of the proposed CNN classifier are achieved at a higher accuracy rate. Higher performance is achieved by increasing the size of the training set.	–The future scope shall focus on larger datasets to standardize the performance of the system and to overcome the time constraints and cost of the work. –By introducing a security module which is based on blockchain, the user identity protection policies can be enhanced. –The performance gets better by collecting real time data and updating the system.	[28]
6	– In order to accurately and comprehensively predict the disease that the patients are experiencing, a novel system is developed that employs the most effective machine learning algorithm and collects data from the patients, including audio recordings, symptoms, medical reports.	NA	[29]
7	–Implemented a new system which is capable of being used in several disease prediction studies using Big Data Analytics and Internet of Things. –Developed an electronic monitoring system for realtime miscarriages, prediction systems to help women who are pregnant and save the lives of baby's. –In order to react in the event of a miscarriage and prevent unfavorable effects, clinicians really obtain the clustering findings and track their patients via mobile app. Women who are pregnant, however, only get recommendations based on their actions.	– The future work shall focus on improving the proposed scheme by incorporating more health care sensors to collect data related to healthcare about a human being, and collect risk factors from texts, images and social networks.	[30]
8	–A multisensory IoT-based real-time vitals monitor is designed to sense BP, SPo2, BT, and PR and continuously these signals are transferred to the big data analytics system which helps in improving diagnostics in an advanced stage. –Developed a mobile application to transfer measured data with an overall health condition to the doctors and patients.	Future work focus on i) generating an automatic notification which will specify location of the patient to ambulance, friends or family, ii) a medicine dispensing system module which notifies patients about out-of-schedule and scheduled medications, iii) A module which tracks Covid-19 patients using wearable sensors, and sends prescriptions to pharmacies for delivery of medicines to patients.	[31]

VIII. Conclusion

In this paper, the integration of Big Data Analytics and IoT in healthcare is examined. IoT applications, Big Data Analytics techniques, challenges in IoT-enabled healthcare are discussed in detail. Comparative analysis showed the effectiveness of various analytics approaches. This paper demonstrates how the usage of big data and IoT can improve the performance and personalized service in healthcare. The combination of Big Data Analytics and IoT is transforming healthcare, enhancing patient care, but issues need to be resolved. There is a lot of potential for improved healthcare outcomes and experiences as these sectors continue to advance.

References

[1] Yasmeen Shaikh, V. K. Parvati and S. R. Biradar, "Role of IoT and bigdata analytics in healthcare for disease prediction", 2020 IEEE International Conference on Convergence to Digital World – Quo Vadis , 2020.

[2] Khaled H. Almotairi,"Application of internet of things in healthcare domain", Journal of Umm Al-Qura University for Engineering and Architecture, 2023.

[3] Fabbrizio A, Fucarino A, Cantoia M, De Giorgio A, Garrido ND, Iuliano E, et al. "Smart devices for health and wellness applied to tele-exercise: an overview of new trends and technologies Such as IoT and AI", Healthcare, 2023.

[4] Hafiz Farooq Ahmad, Wajid Rafique, Raihan Ur Rasool, Abdulaziz Alhumam,Zahid Anwar, Junaid Qadir,"Leveraging 6G, extended reality, and IoT big data analytics for healthcare: a review", Computer Science Review, 2023.

[5] Yang S, Zhu F, Ling X, Liu Q and Zhao P, "Intelligent health care: applications of deep learning in computational medicine", Frontiers in Genetics, 2021.

[6] G. Aceto, V. Persico, A. Pescape, "Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0", Journal of Industrial Information Integration, 2020.

[7] W. Li, Y. Chai, F. Khan, S.R.U. Jan, S. Verma, V.G. Menon, et al., "A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare systems", Mobile networks and applications, 2021.

[8] Rohit Ranchal , Paul Bastide , Xu Wang, Aris Gkoulalas-Divanis , Maneesh Mehra, Senthil Bakthavachalam, et al., "Disrupting healthcare silos: addressing data volume, velocity and variety with a cloud-

native healthcare data ingestion service", IEEE Journal of Biomedical and Health Informatics,2020.

[9] Javed SH, Ahmad MB, Asif M, Almotiri SH, Masood K, Ghamdi MAA,"An intelligent system to detect advanced persis- tent threats in industrial internet of things (I-IoT)", Electronics, 2022.

[10] Abunadi I, Abdullah Mengash H, Alotaibi S, Asiri MM, Ahmed Hamza M, Zamani AS, Motwakel A, et al., "Optimal multikey homomorphic encryption with steganography approach for multimedia security in internet of everything environment",Applied Sciences, 2022.

[11] Auer S, Nagler S, Mazumdar S, Mukkamala RR, "Towards blockchain-IoT based shared mobility: car-sharing and leasing as a case study", Journal of Network and Computer Applications, 2022.

[12] Kamal M, Aljohani AJ, Alanazi E, Albogamy FR, "5G and Blockchain enabled lightweight solutions for containing COVID-19.", Security and Communication Networks, 2022.

[13] Muhammad Qasim Shabbir and Syed Babar Waheed Gardezi, "Application of big data analytics and organizational performance: the mediating role of knowledge management practices", Journal of Big Data, 2020.

[14] Muhib Anwar Lambay, Dr. S. Pakkir Mohideen, "Big data analytics for healthcare recommendation systems", 2020 International Conference on System, Computation, Automation and Networking, 2020.

[15] Teisberg E, Wallace S, OHara S. "Defining and implementing value-based health care: a strategic framework", Academic Medicine, 2020.

[16] Rakesh Raja, Indrajit Mukherjee, Bikash Kanti Sarkar, "A Systematic review of healthcare big data", Scientific Programming, 2020.

[17] Shafiqul Hassan, Mohsin Dhali, Fazluz Zaman, Muhammad Tanveer, "Big data and predictive analytics in healthcare in Bangladesh: regulatory challenges", Heliyon, 2021.

[18] Berros N, El Mendili F, Filaly Y, El Bouzekri El Idrissi Y, "Enhancing digital health services with big data analytics", Big data and cognitive computing, 2023.

[19] X. Zhou, W. Liang, K. I. Wang, H. Wang, L. T. Yang and Q. Jin, "Deep-learning-enhanced human activity recognition for internet of healthcare things", IEEE Internet of Things Journal, 2020.

[20] Hamidreza Bolhasani, Maryam Mohseni, Amir Masoud Rahmani, "Deep learning applications for IoT in health care: A systematic review", Informatics in Medicine Unlocked, 2021.

[21] A. ul Haque, M. S. Ghani and T. Mahmood, "Decentralized transfer learning using blockchain & IPFS for deep learning", 2020 International Conference on Information Networking (ICOIN), 2020.

[22] S. Hakak, S. Ray, W. Z. Khan and E. Scheme, "A Framework for edge-assisted healthcare data analytics using federated learning", 2020 IEEE International Conference on Big Data, 2020.

[23] Hassan, Mubashir, et al.A. "Innovations in genomics and big data analytics for personalized medicine and health Care: a review", International journal of molecular Sciences, 2022.

[24] Pradeep Kumar Vishwakarma, Dr. Randeep Singh, "A – Review on IoT-Assisted ECG monitoring framework for health care applications", 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2021.

[25] Ganie, Shahid & Malik, Majid & Arif, Tasleem, "Machine learning techniques for big data analytics in healthcare: current scenario and future prospects", In Telemedicine: The Computer Transformation of Healthcare, 2022.

[26] Kornelia Batko and Andrzej Slezak,"The use of big data analytics in healthcare", Journal of Big Data, 2022.

[27] Ogur, N. B., Al-Hubaishi, M., and Ceken, C., "IoT data analytics architecture for smart healthcare using RFID and WSN", ETRI Journal, 2022.

[28] K. Thilagam, A. Beno, M. Vanitha Lakshmi, C. Bazil Wilfred, Santhi M. George, M. Karthikeyan, et al., "Secure IoT healthcare architecture with deep learning-based access control system", Journal of Nanomaterials, 2022.

[29] Salman Ahmad Siddiqui, Anwar Ahmad, Neda Fatima, "IoT-based disease prediction using machine learning", Computers and Electrical Engineering, 2023.

[30] Hiba Asri and Zahi Jarir, "Toward a smart health: big data analytics and IoT for real-time miscarriage prediction", Journal of Big Data, 2023.

[31] Banu, E. Afreen, and V. Rajamani. "Design of online vitals monitors by integrating big data and IoT." Computer Systems Science & Engineering, 2023.

Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 85–91
DOI: 10.15439/2023R17
ISSN 2300-5963 ACSIS, Vol. 38

Machine Learning Approach for Forecasting Job Appeasement and Employee Corrosion

Swetha M S
Department of Information Science & Engineering. BMS Institute of Technology & Management.
Bengaluru, India
swethams_ise2014@bmsit.in

Mahalakshmi S
Department of Information Science & Engineering BMS Institute of Technology & Management.
Bengaluru, India
maha.shanmugam@bmsit.in

Pushpa S K
Department of Information Science & Engineering BMS Institute Of Technology & Management.
Bengaluru, India
pushpask@bmsit.in

Amrutha T Madihalli
Department of Information Science & Engineering. BMS Institute of Technology & Management
Bengaluru, India
amruthatm1105@gmail.com

Ananya
Department of Information Science & Engineering. BMS Institute of Technology & Management.
Bengaluru, India
ananyasingh0722@gmail.com

Anand Bhardwaj
Department of Information Science & Engineering. BMS Institute of Technology & Management.
Bengaluru, India
anand.bh231@gmail.com

Abstract—**Employee turnover imposes costs on the organization. The quit may also cause significant and costly disruptions to the production process. The recent increase in the technological capacity to gather large magnitude of data and analyze it has changed how decision makers use them to decide on making the optimal decision. Employee attrition very similar to customer churn is an important and deciding factor affecting the revenue and success of the company. To avoid this problem, many companies now are taking guide via machine learning strategies to expect employee churn/attrition. In this paper, we are analyzing data from the past and present using different classifications like SVM, Random Forest, Decision tree, Logistic Regression, and an Ensemble model to come up with a better predictive model for the present dataset. Through this we are hoping to help the company predict employee churn and take effective measures to retain the employees and improve their economic loss due to the loss of valuable employees.**

Index Terms—**Support Vector Machine (SVM), Random Forest, Decision tree, Logistic Regression**

I. Introduction

Organizations must consider a lot of factors to keep them as a leading Company in the competitive market of today [3]. Machine Learning and their techniques have given them a useful tool to get an edge in the market after analyzing data collected over years.

Attrition also called wastage rate or total turnover rate can be considered as a silent killer which destabilizes a company from within [2]. Employees may choose to leave the company for a lot of reasons like equal pay, lack of appreciation, long working hours and many more [5]. As employees are the central source of any company, employee attrition has a negative impact on the revenue of the company along with other various consequences like having to invest more in hiring and training new employees, more pressure on the present employees and a radical downslope in the expected performance of the company[1]. Hence Analytics done using Machine Learning and their tools helps us to understand the issue and source of it, as well as come up with effective solutions for it. Using the past and present data for predicting attrition helps in identifying the causes for the churn and stopping the increasing churn over rate.

In our methodology, we have used different classification methods like SVM, Random Forest and Decision tree along with a hybrid model to understand and analyse the performance of different predictive models and compare them using different classification metrics. SVM (Support Vector Machine) are kernel based algorithms used which serves as a tool to separate different classes. Kernel transforms the input data into higher dimensions where it can be solved using linear classifier by drawing a hyper plane. For example, facial expression recognition is of the uses of SVM where it filters out different expressions into their own class divided by hyper-plane. Decision Tree (DT) appears like a tree shaped algorithm to examine and determine a course of action or show statistical probability.

A company may deploy decision trees as a kind of decision support system. Let’s consider booking a train to travel as example. First we look into our calendar to see if a train is available on that date .If available, we look at the time suitable to us. Then we consider the price is within our range etc. Like this at each step we make decisions and go further deep down the branch till an outcome has arrived that is the train being booked.

Random Forest builds a forest with a number of decision tree and is an ensembling method. Logistic Regression uses independent variables for coming to a conclusion. A last method used is Stacking, an ensembling method which combines the predictions from well performing algorithms and gives out a better performance. Retention is more important than hiring. The foremost successful organizations are successful because they look after their employees and that skills to retain them within the organization.

This motivated us to research the connection between work fulfilment and representative maintenance by developing a completely unique algorithm to use to the model in making better predications to assist the HR management retain their employees and put them within the right jobs consistent with the satisfaction levels.

The prevailing systems accuracy in predicating isn't much and organizations hesitate to include the model, so we are motivated to develop a model which satisfies the HR management to use the system [4].

II. Literature Servey

Employment fulfilment is critical to high profitability, inspiration and low turnover rate. Managers particularly the HR group face the difficulties of securing techniques to boost position fulfilment, so their organizations stay serious. Associations who regularly neglect to upgrade work delight or fulfilment are at a danger of losing their most skilled individuals to the contenders. Bosses and chiefs who attempts to augment the potential, innovative capacities, and abilities of the entire workforce include a more prominent favorable position inside the challenge than those that don't. Representatives that are occupied with their work have a superior degree of employment fulfilment. Persuaded laborers give the protection organizations frantically required in these disordered occasions. Right now, the HR the board for representative maintenance they need utilized SVM and random forest as classifier calculations to anticipate the laborer fulfilment and steady loss. We propose to build up a substitution calculation that gives better exactness utilizing Linear Model Tree close by random forest.

[6] Usha P.M. et al., According to the author data mining serves as a method for identifying and analyzing hidden patterns within extensive datasets. The article's primary focus is on discerning the various factors that influence employee attrition within the human resource department of firms or organizations. To achieve this, the researchers utilized data mining approaches, specifically emphasizing the Weka platform. Weka, a data science tool for predictive analytics, employs algorithms like K-Nearest Neighbors (KNN) to cluster data and gain insights into the variables contributing to attrition. Additionally, Weka was employed to assess and analyze the performance and effectiveness of the various algorithms used in the study.

[7] Gunjan, V.K. et al., employed Apache Spark, an open-sourced general-purpose cluster framework, for Big Data Analytics. Their research utilized Multi-Layer Perceptron in Spark to predict attrition, and the output was comprehensively analyzed using graphs, which plotted each attribute and its association with attrition values. Additionally, they provided a user-friendly interface designed to convey results in easily understandable language."

[8] Aniket Tambde et al., focused on creating a professional model for examining and forecasting employee turnover rates. They employed the Random Forest and K-Nearest Neighbors (KNN) machine learning algorithms. Notably, when considering the confusion matrix, Random Forest outperformed KNN in predictive accuracy.

[9] Nesreen El-rayes et al., gathered and analyzed data from randomly selected resumes on Glassdoor, a job-search website, to forecast employee turnover resulting from job changes. Their study highlighted the effectiveness of tree-based models, particularly Random Forest (RF) and Gradient Boosting (GB), which exhibited strong predictive performance when compared to other binary classification techniques.

[10] Mehul Jhaver et al., utilized the Gradient Boosting technique with its regularization-based robustness to forecast employee turnover. They compared this approach with three other common algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The results indicated that Gradient Boosting outperformed the other three algorithms, emphasizing its efficacy in predictive modeling for turnover scenarios."

III. Proposed System

To design a hybrid model to predict the employee turnover and to understand the reasons for it by going over the past data and enhancing the prevailing model, which may allow the corporate to predict the longer-term occurrence of the event.

By identifying and then comprehending the variables that contributed to attrition or turnover rates, businesses and people will be able to reduce employee turnover, boost productivity, and foster professional development. Managers can demand remedial actions to build and maintain their successful firm thanks to these important and predictive information.

The main goal is to find out the employee turnover using the provided dataset. We are applying many various ways of classification algorithm like Random Forest (RF), Decision tree (DT), Support vector machine (SVM), Logistic regression (LR). We then apply stacked model which is one of the ensemble methods to get a hybrid model. Decision tree is used as Meta-model. It is trained on the base model’s prediction and on the test data. The inputs to the basic models may also be included in the training data for the meta-model. The meta-model will fit the training data, and a final prediction that is more accurate than the predictions provided by individual machine learning algorithms is obtained. This model is referred to as a "stacked hybrid model".

The main objectives are:

- To provide appropriate and obvious incentives for the company to prevent staff departures or, at the very least, to be prepared to predict and analyze what variables contributed most to turnover rate.
- To design and construct a model that determines whether a specific employee will quit the company or not.
- To more accurately anticipate the turnover rate using a hybrid approach.
- To develop and enhance various retention tactics for selected personnel.

A company's ability to retain a successful business model and a positive company culture is a surefire indicator of its ability to succeed and expand in the future. Organizations and people can prevent this from happening and even boost employee productivity and improved growth by first recognizing and then comprehending the reasons that were associated with turnover rate or attrition. These helpful and predictive insights give managers the chance to demand corrective actions to build and maintain their profitable business.

IV. Design and Implementation

A version is intended to be used for prediction, analysis, or interpretation. While the interpretation is qualitative, the projection may be examined statistically. How well a model performs on untrained data can be used to gauge its predicted accuracy. The use of techniques like validation may be evaluated. It is important to make intelligent choices when it comes to the algorithms that can be used, as well as the biases and reductions on the hypothesis area of potential fashions that could be developed for the problem. To create a hybrid model, we use the stacking model, one of the ensemble approaches. The meta-model utilized is a decision tree.

Here, Decision Tree Model, Logistic Regression Model, Random Forest Model, and SVM Model are employed as the basis models.

A. System Architecture

The whole system architecture is represented in the flowchart below. Gathering the data is the first stage in the process; in this case, we used the employee dataset. Next, we loaded the data and performed scrubbing, in which we normalized all the missing values and cleaned the dataset. 25% of the dataset will be used for testing, while the remaining 75% will be used for training. The training dataset will undergo data pre-processing, which aids in positioning the data correctly and getting it ready for machine learning algorithms.

Once data preparation is completed, we perform data analysis and visualization on different features of the dataset by plotting several graphs. This helps in giving a clear picture of the distribution of personal traits and helps choose the right qualities to use during prediction. Additionally, we use feature importance on the dataset that informs us of the significance of the features based on the ratings the feature importance model assigns to them. The three features that are the strongest estimators of the outcome variable, or employee turnover, for the employee dataset that we have used are "satisfaction," "year at company," and "evaluation."

Next step is modelling the 19, where we train four different ML algorithms: Decision tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR). Then we get the prediction of each algorithm.

Now, in order to obtain a hybrid model, we apply the stacking model, one of the ensemble approaches. The meta-model utilized is a decision tree. It is trained using the test data and the predictions from the basic model. The training data for the meta-model may also comprise the inputs to the fundamental models. The forecast generated by the meta-model will be more accurate than those made by separate machine learning algorithms since it will fit the training data. A "stacked hybrid model" is the name given to this model. This hybrid model combines multiple diverse skilled machine learning techniques while also reducing the errors in predictions made by the basic models.

B. Algorithms

1) Algorithm for Decision Tree

We are making use of ID3 algorithm.

Step1. Using Equations (1) and (2), determine the information gain for each and every attribute:

$$\text{Entropy(S)} = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (1)$$

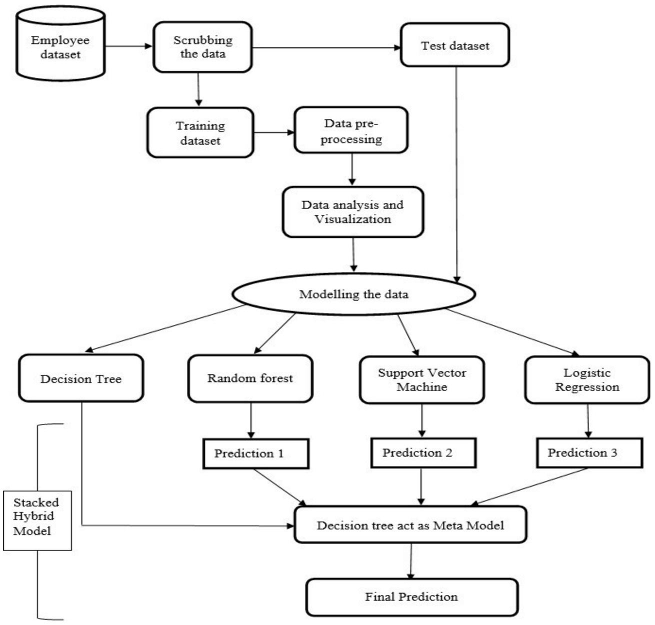


Fig. 1. System Design

$$\text{Gain(S, A)} = \text{Entropy(S)} - \sum_{v \in Values(A)} \frac{|S v|}{|S|} \text{Entropy}(S_v) (2)$$

Step2 Step 3 should be taken if two attributes yield the same gain value, otherwise step 4 should be used.

Step3. Choose a quality at random.

Step4. Pick an attribute with a high gain value right away.

Step5. Make the chosen attribute the decision tree's root node.

Step6. Make the value of the chosen attribute a child node.

Step7. If there are still unclassified examples, proceed to step 8, otherwise, move on to step 9.

Step8. Continue with the following characteristics. Repetition is required.

Step9. End.

2) Algorithm For Random Forest

Random forest can be created using 2 different stages:

a) Random forest creation:

Step1. Choose the features randomly such that number of selected features must be less than total features.

Step2. From the selected features, pick a root node ‘r’ using the best split point.

Step3. Make root node’s value as child node using best split.

Step4. Repeat the step1 to step3 until single decision tree is formed.

Step5. Repeat step1 to step4 to create many numbers of trees and construct a forest with that.

b) To make prediction from the random forest created in the first stage:

Step1. Consider the test features and use the rules of every decision tree that has been generated at random to predict and store the outcome.

Step2. For every expected result, determine the votes.

Step3. Find the expected outcome with highest voted as the final prediction from algorithm of random forest.

3) Algorithm for Svm

Step1. Define an optimal hyper-plane which must be maximum margin.

Step2. Find the solution for non-linear data as well with the help of kernel method.

Step3. Project data to a high dimensional space where the classification with linear decision surfaces are easier.

a) Steps to represent an optimal hyper-plane:

Step1. Take the training data of n points:

(X1, y1), (x2, y2) …. (Xi, Yi)

Where xi- p-value vector for point 1

Yi- binary class value of 1 or -1

Thus, there are two classes 1 and -1

Step2. Assuming that the data is indeed linearly separable, the classifier hyper-plane is defined as set of points that satisfy the Eq. (3).

$$\vec{w}\cdot\vec{x}+b=0 \tag{3}$$

Step2. Calculate the hard margin which can be defined as Eq.(4) and Eq. (5):

$$x_i \cdot w+b-1 \tag{4}$$

$$x_i \cdot W+b=-1. \tag{5}$$

Step3. Calculate the width of hard margin using. Eq. (6):

$$2/\|w\| \tag{6}$$

Where 'w' is the width of the margin.

Step4. Finally, we can find 'w' (weight vector) for the features such that there is a widest margin between two classes.

4) Algorithm for Logistic Regression

Step1. Given a data (x, y), build a randomly initialized matrix for weight. Then, by features, we multiply it using Eq. (7).

$$a=w_0+w_1x_1+w_2x_2+\cdots w_nx_n \tag{7}$$

Where x- matrix of values y- vector.

Step2. Pass the output obtained in step1 to a link function given in Eq. (8).

$$y\hat{i}=1/\left(1+e^{-a}\right) \tag{8}$$

Step3. Calculating the cost for this iteration whose formula is in Eq. (9):

$$\operatorname{cost}(w) = \left(-1/_m\right) \sum y_j \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \tag{9}$$

Step4. Calculate the derivative of this cost:

And update the weights using Eq. (10) and Eq. (11):

$$dw_j=\sum_{i=1}^{i=n}\left(\hat{y}-y\right)x_j^i \tag{10}$$

$$w_i=w_j-\left(\alpha * d w_j\right) \tag{11}$$

5) Algorithm For Stacked Hybrid Model

There are 3 stages:

a) Build the ensemble

Step1. Determine the list of base models.

Step2. Determine the meta-model algorithm.

b) Train the ensemble:

Step1. All 'L' base models are made to learn on the training dataset.

Step2. Run k-fold cross validation on each base model, then compile the cross-validated predictions from each, denoted by p1, p2,, pL.

Step3. Combine N cross-validated predicted values from every base model to form new N*LN*L feature matrix using Eq. (12). The "level-one" data was named in accordance with the original response vector.

$$n\left\{\begin{bmatrix} p_1 \end{bmatrix}\cdots\begin{bmatrix} p_L \end{bmatrix}\begin{bmatrix} y \end{bmatrix}\right. \rightarrow n\left\{\begin{bmatrix} \overbrace{\qquad Z \qquad}^{L} \end{bmatrix}\begin{bmatrix} y \end{bmatrix}\right. \tag{12}$$

Step4. Train the meta-model on the "level-one" data using Eq.(13) .

$$Y=f(Z) \tag{13}$$

c) Predict on new data:

Step1. Generated predictions from the base models will be taken.

Step2. Feed those predictions into the meta-model to generate the final result.

C. Implementation

In current system only limited numbers of techniques are used from the huge collection of data mining techniques for prediction. In above advanced system generated we have applied few algorithms like k-nearest neighbor (KNN), Support vector machine (SVM), Logistic regression (LR), Decision tree (DT) and ensemble model. Fundamentally our data set contain employee, satisfaction grade, assigned projects and work efficiency spent in firm.

In our system we scrub the data so that there are no null values in the data set and if any should remove the null values. We choose the best features for employee attrition, and they are in Table -1.

TABLE I. ATTRIBUTES

Features	Data type
Job Satisfaction level	Number [10]
Number of Projects	Number [10]
Average Monthly hours	Number [10]
Any Work Accident	Number [10]
Last Evaluation	Number [10]
Time Spent in Company	Number [10]
Department	Varchar [20]
Any Promotions	Number [10]
Salary	Varchar [20]
Turnover	Number [10]

We use the co-relation matrix and heat map to analysis the below:

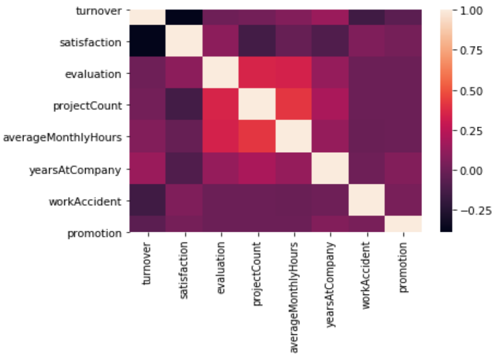


Fig 2 Heat map

CASE 1: Positive (+ve) interdependence

Here project count (PC), average monthly hour (AMH), evaluation i.e., rank of employee is considered. The employee with the ability of covering maximum number of monthly hours and outputting the completion of multi assigned project with is comparatively high has been observed with highly ranked.

CASE 2: Negative (-ve) Relationships

Here turn over (TO), satisfaction is highly tie-up with each other. In other words, we state that employees having low satisfaction are directly proportional to employees leaving the firm. The heat map is shown below Figure 2. To inspect the scattering on the features.

The following are some essential views and thus features:

• Satisfaction - Employees were divided into two groups based on their level of satisfaction: low satisfaction and high satisfaction.

• Evaluation - Using a bimodal function, employee performance was classified as low (below 0.6) or high (greater than 0.8).

• Average Monthly Hours - Employees' working efficiency is an important attribute, but being able to fully utilize a high performance is also important, so work time is analyzed with respect to average monthly hour worked by each employee and is further classified based on less than 150 hours and more than 250 hours. This means that the higher the working efficiency, the higher the average monthly hours worked. These features discussed above are interviews.

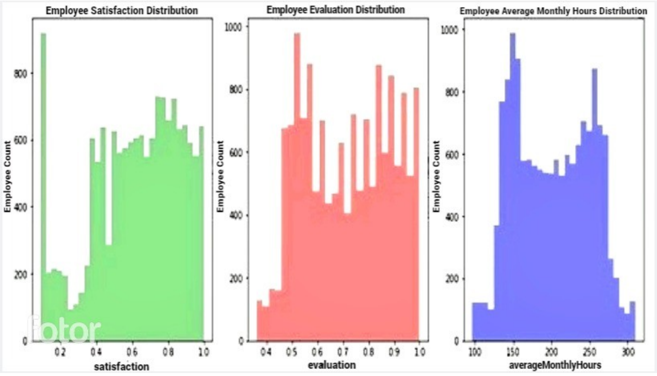


Fig. 3. Distribution Plot

The satisfaction v/s evaluation is the utmost gripping graph. The employees who left the firm were analyzed under 3 main clusters.

- *Cluster 1 (Hard-working and Sad Employee):*

It is important to treat the employees with working efficiency more than 0.75 so that they remain in the firm. Such employees improve the firm and there can someday lead departments in firm. When such employees are not treated well i.e. satisfaction less than 0.2. Such employees may leave the firm.

- *Cluster 2 (Bad and Sad Employee):*

In this case employees with satisfaction level 0.35-0.45 whose working efficiency will be below 0.58. Retaining such moderate employees are necessary as they can be directed by high evaluated to work in order reduce their work load. Hence if their satisfaction is low tendency to leave the firm is more.

- *Cluster 3 (Hard-working and Happy Employee):*

The employees with satisfaction level at between 0.7-1.0 and working efficiency greater than 0.8.Such employees should be treated well in firm in order to keep them associated to firm. They are well satisfied with their work and their performance is highly evaluated.

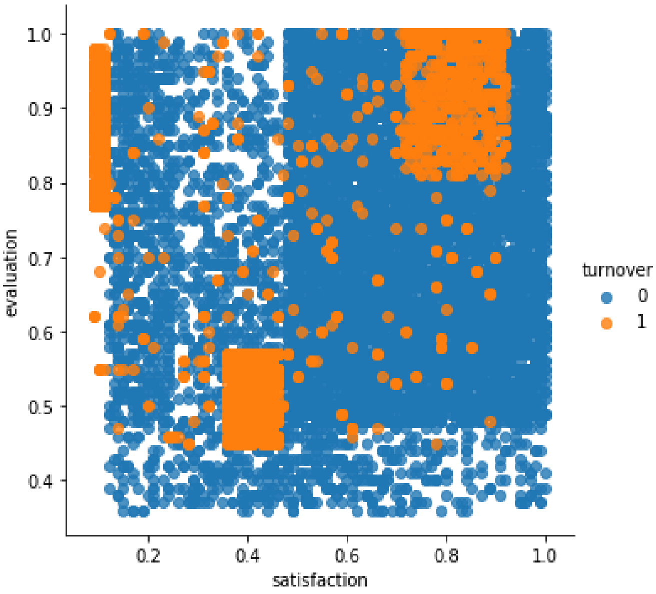


Fig.4. Satisfaction v/s Evaluation

V. Results and Discussion

The best model performance out of the four (Decision Tree Model, SVM Model, Logistic Regression Model, Random Forest Model) was Random Forest. Hence Random Forest along with stacking classifier is the best model to predict employee attrition with an accuracy of 0.98. This is Concluded by using below Table-2.

When we compare the performance of all the algorithms which are ran on the dataset, we plot a graph i.e. the model performances graph by which we get to know that, the highest score is for the Hybrid Stacked Model which is shown below in the figure-5.

A. Final Interpretation

With the use of all this data, a final analysis is completed to determine the most likely reasons an employee departed a firm. These interpretations are given:

Table II.

Model	Accuracy_ score	Recall _score	Pre-cision	f1_ score	Area_ Under_ curve	Kappa_ metric
Logistic Regression	0.7898	0.3657	0.5957	0.4532	0.644	0.3322
Decision Tress	0.9382	0.8825	0.8616	0.8719	0.9191	0.8312
Random Forest	0.9831	0.9478	0.9807	0.9639	0.971	0.9529
SVM	0.7658	0.7836	0.5054	0.6145	0.7719	0.4573
Stacked Model	0.9862	0.959	0.9828	0.9707	0.9769	0.9617

1. When underworked (less than 150 hours per month or 6 hours per day), employees frequently leave their jobs.
2. When workers are overworked (more than 250 hours per month or 10 hours per day), they frequently abandon their jobs.
3. Workers with extraordinarily high or bad evaluations ought to be considered for a high turnover rate.
4. The majority of turnover among employees is caused by those making low to medium salary.
5. Employees having a project count of 2, 6, or 7 faced the risk of being let go by the company.
6. The best indicator of staff retention is employee happiness.
7. Workers with four and five years with the company should be considered for a high turnover rate.
8. Employee satisfaction, year's At Company, and evaluation were the three key factors in deciding turnover.

B. Future Enhancement

This research could incorporate predictive analytics techniques to forecast employee attrition trends in a more dynamic and forward-looking manner. By considering economic indicators, industry-specific factors, and market conditions, organizations could gain valuable insights for proactive workforce planning. Additionally, implementing a real-time feedback mechanism that collects continuous input from employees, leverages sentiment analysis and natural language processing, and integrates this data into the attrition prediction model can enhance its accuracy. Moreover, the development of a recommendation engine to provide customized retention strategies for individual employees based on their predictive factors could empower HR departments to take targeted actions to reduce attrition and improve overall workforce satisfaction. Finally, exploring the use of blockchain technology for data security could enhance the trust and transparency of attrition prediction and management, ensuring the confidentiality and integrity of employee information while complying with data privacy regulations.

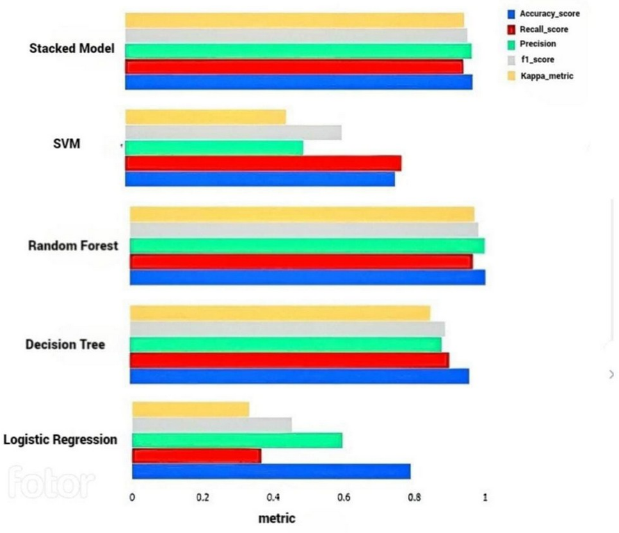


Fig. 5. Model Performances.

VI. Conclusion

As we progress with our research, it's obvious that **Random Forest along with stacking classifier and Decision tree as a meta-model** is the best model to predict employee attrition with an accuracy of **0.98**. Clearly we can say this is a best suited approach. By this we can also conclude by saying, it's important to choose the efficient base models in order to make stacking method more accurate otherwise stacking technique is not recommended.

We trained the computer with limited data (14000 odd records divided into 75percent training data and 25 percent test data), but it can also work effectively on a broad dataset. From the reference above it is very clear how the estimation of attrition plays an important role in the businesses

Given that large turnover percentages are caused by employees, whether they have high or negative ratings, they should be considered. When overworked (more than 250 hours per month or 10 hours per day) or underworked (less than 150 hours per month or 6 hours per day), employees frequently leave their jobs. Employees with low to medium salaries are most likely to leave the organization. Employees who have worked on 2, 6, or 7 projects had a higher departure rate. The best indicator of employee turnover is employee satisfaction. Consideration should be given to employees who have worked for the company for four and five years. The main determinants of turnover were years At Company, assessment, and employee satisfaction.

- Efficiency of the stacking model depends on the base models.
- Base models must be as different as possible to get the better result.
- Stacking model makes the result difficult to interpret.

References

[1] Qin Zhou., (2017) "The Impact of Job Satisfaction effect on Turnover Intention: An Empirical Study based on the Circumstances of China".

[2] Phillips, J. D., (1990) "The price tag on turnover" Personnel Journal, vol. 12, pp. 58-61.

[3] Ms. Ankur Jain, (2018) "Impact of TQM on employees' job satisfaction in Indian software industry".

[4] Bosele, P. and Wiele, T.V.D. (2019) "Employee perceptions of HRM and TQM and its effects on satisfaction and Intention to leave", Managing Service Quality, 2019.

[5] Zhu Xiaoyan, Li Yanping, (2018) "A Study on Psychological Contract, Job Satisfaction and Turnover Intention in Banking Industry".

[6] Usha. P. M. and Dr. NV Balaji. ANALYSING EMPLOYEE ATTRITION USING MACHINE LEARNING. IOP Conference Series Materials Science and Engineering, 1085(1), 012029, 2021. DOI:10.1088/1757-899X/1085/1/012029

[7] Gunjan, V. K., Garcia Diaz, V., Cardona, M., Solanki, V. K., & Sunitha, K. V. N. (Eds.). Prediction of Employee Attrition and Analyzing Reasons, ICICCT-2019

[8] Aniket Tambde, Dilip Motwani. Employee Churn Rate Prediction and Performance Using Machine Learning. Published By: Blue Eyes Intelligence Engineering & Sciences Publication, 2019, DOI: 10.35940/ijrte.B1134.0982S1119

[9] Nesreen El-rayes, Michael and Stephen Taylor. An Explicative and Predictive Study of Employee Attrition using Tree-based Models. Conference: Hawaii International Conference on System Sciences. DOI:10.24251/HICSS.2020.175

[10] Mehul Jhaver; Yogesh Gupta; Amit Kumar Mishra. Employee Turnover Prediction System. Conference: 2019 4th International Conference on Information Systems and Computer Networks (ISCON) 2019 . DOI:10.1109/ISCON47742.2019.9036180

[11] Ganesan Santhanam, Raja Jayaraman, Dr.V. Badrinath,"Influence of Perceived Job Satisfaction and Its impacts on Employee Retention in Gulf Cooperation Countries" (2019).

[12] Richard, F. G., Joseph, M.L., Billy, B., (2016) "Job satisfaction, Life satisfaction and Turnover Intent: Among Food-Service Managers," Cornell Hotel and Restaurant Administration Quarterly, vol. 42(2).

[13] Sukhadiya J, Kapadia H, Prof. D'silva M, 2017 "Employee Attrition Prediction using Data Mining Techniques" International Journal of Management, Technology And Engineering ISSN NO : 2249-7455 Volume 8, Issue X, OCTOBER/2018.

[14] Sikaroudi1 A, Ghousi1 R, 2015, "A data mining approach to employee turnover prediction". Journal of Industrial and Systems Engineering Vol. 8, No. 4, pp106-121 Autumn 2015

[15] Rygielski C, Jyun-Cheng Wang b, Yen D, 2018 "Data mining techniques for customer relationship management" Technology in Society, Novemeber 2022

[16] Jain, R., & Nayyar, A. (Year). Predicting Employee Attrition using XGBoost Machine Learning Approach. Conference: 2018 International Conference on System Modeling & Advancement in Research Trends (SMART). DOI:10.1109/SYSMART.2018.8746940.

[17] Brockett, N., Clarke, C., Berlingerio, M., & Dutta, S. (Year). A System for Analysis and Remediation of Attrition. Conference: 2019 IEEE International Conference on Big Data (Big Data). DOI:10.1109/BigData47090.2019.9006333.

[18] Bindra, H., Sehgal, K., & Jain, R. (Year). Optimization of C5.0 Using Association Rules and Prediction of Employee Attrition. DOI: 10.29228/Joh.63712

rice2023 Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 93–98 DOI: 10.15439/2023R66 ISSN 2300-5963 ACSIS, Vol. 38

Emerging Trends in Pulsar Star Studies: A Synthesis of Machine Learning Techniques in Pulsar Star Research

S. Thanu
Department of Computer Science and Engineering
Manonmaniam Sundaranar University
Tirunelveli, India
tharshinimanian@gmail.com

Dr. V. Subha
Department of Computer Science and Engineering
Manonmaniam Sundaranar University
Tirunelveli, India
subha_velappan@msuniv.ac.in

***Abstract*—The pulsar is an extremely magnetized gyrating neutron star having a radius of 10 – 15 km. Pulsars provide the indirect evidence of the gravitational wave's existence. So, to study the gravitational waves identification of pulsars is mandatory. Pulsars are considered as the Universe's gift. Pulsars provide scientists and researchers with information of the physics of neutron stars, which are thought to be the densest materials in the universe. The reason why astronomers give importance to the pulsars, because they are the leading edge of the research, based on the gravity. All pulsars produce marginally distinct emission pattern and it varies to some extent with every rotation. Hence, a promising signal detection is termed as a candidate, which is averaged based on every rotation of the pulsars. Any absence of the additional information, implies that each candidate is a real pulsar. The valid signals are extremely hard to detect due to noise and radio frequency interference (RFI). To clear up with this issue, Machine Learning (ML) algorithms were used for automatically classifying, identifying and many other process of pulsar candidates. This survey paper talks about different techniques used by different researchers for the pulsar star classification, identification and still more, using ML techniques.**

***Index Terms*—Machine Learning, Ensemble Learning, Boosting, Deep Convolutional Network.**

I. Introduction

In the time of 1967, Jocelyn Bell a Ph. D student from Cambridge University and her supervisor Anthony Hewish [1] found something peculiar when they were scrutinizing about the faraway galaxies. When looking at a specific point through the radio telescope, they detected some kind of radio pulses and they named it as Little Green Men 1 (LGM1). Belatedly Little Green Men 1 were entitled as pulsars because of its emission as pulses. At present they are called as the PSR B1919 + 21 [2], discovered on 28 November 1967 when they were working at the university's Mullard Radio Astronomy Observatory (MRAO) [3] and it got the name as first discovered radio pulsar. Within ten to one hundred million years, the electromagnetic energy that these pulsars emit moderately slows down and goes silent. Because of the development and collaboration of the ML in each and every field, there is no astonishment that it can also be widely used in the area of Astronomy.

II. Literature Survey

A. *Astronomy*

Pulsars provided first indirect evidence of the presence of gravitational waves in 1974. M. Bailes, et al. [4] elaborately explained about the operation and the collection of the data by Laser Interferometer Gravitational-Wave Observatory (LIGO) and its international fellows: Virgo and KAGRA. This paper pointed about the extension of gravitational wave detector network globally with the inclusion of LIGO-India project. It gave the catalogue about different gravitational wave events. The paper, provided the elaborate studies of the characteristics of the neutron stars and black hole via gravitational wave observations, providing valuable information of their formation, evolution. So many advanced efforts were taken to identify and study the multi – messenger sources, where gravitational waves were observed in concurrence with another form of radiations like light, X – Rays or may be Gamma rays.

B. *Machine Learning*

Iqbal H. Sarker [5] said that, as we were in the era of Fourth Industrial Revolution, this digital world has millions, billions of data. Those data, can be from the platform of medical, cybersecurity, business, social media etc. For analyzing all these data and to develop related automated applications, the knowledge about ML is very much important. Supervised (S), Unsupervised (US), Semi- Supervised (SS), and Reinforcement Learning (RL) were the different types of ML algorithms. At last, the paper described some of the applications and challenges of ML.

C. *Machine Learning in Astronomy*

Dalya Baron [6] discussed about supervised and un– supervised learning algorithms and mainly focused on un– supervised learning. It furnished the practical information about the ML algorithms and their deployment in the astronomical dataset. The paper described the fundamental concept of supervised learning and un–supervised learning algorithms along with different quality scores and also discussed various supervised learning algorithms used in distinct astronomical tasks plus dimensionality reduction algorithms. This paper talked about feature scaling, how to balance the dataset in case, any presence of imbalanced datasets. The paper gave an applicative idea of how these algorithms can be implemented on different astronomical datasets. The author concluded that, by using un-supervised machine learning algorithm, new unique information can be retrieved from the dataset, which led into the new discovery.

D. The High Time Resolution Universe Pulsar Survey

M. J. Keith, et al. [7] gave details about the different intriguing objects that have been discovered over the past decade. By deploying the 13 – beam multibeam collector on the Parkes Radio Telescope, they begun the study on pulsars.

The area chosen to take the survey was the complete southern sky in 42641 pointings which has been splitted into three regions as low, mid, high galactic latitude, having the integration times of 4200, 540 and 270s individually. After completing roughly 30 % of the mid latitude survey, they again identified 223 priorly known pulsars and discovered 27 pulsars of which 5 were millisecond pulsars. The data points were observed utilizing Parkes 21- cm Multibeam Receiver (MBR) together with the Berkeley–Parkes–Swinburne Recorder (BPSR) backend system. For processing the survey, they devised the processing pipeline called HITRUN.

E. Classification of Pulsars

1) Machine Learning for classifying Pulsar Stars

A branch of Artificial Intelligence (AI) [8] is Machine Learning, mainly concentrated on the usage of the predefined data and algorithms to imitate in the way the human behaves. ML is used in the areas of Autonomous Vehicles, Speech Recognition, detecting Fraud and in other fields also. Using the model historical data, said to be the training data, algorithms of ML were used to construct a mathematical model, to make predictions or decisions without programming externally. The concept of ML for pulsar classification is shown in Figure1.

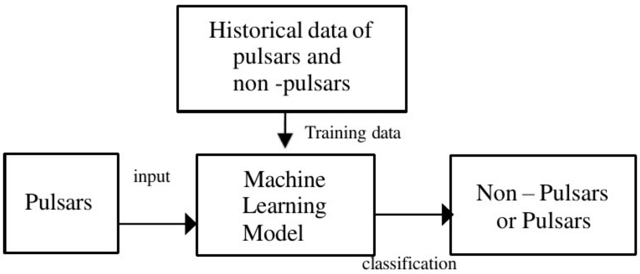


Fig. 1. General Model describing the classification of historical pulsar dataset

Right now, various researchers are using different ML methods for the purpose of pulsar star classification, identification, etc. Different ML algorithms [9] used for pulsar classification includes Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN) and other algorithms.

2) Classification of Pulsars using Extreme Gradient Boosting and Light Gradient Boosting

Tariq, et al. [10] used HTRU2 [11] and LOTAAS – 1 datasets. For handling the class imbalance problem, asymmetric under sampling method was applied. For sampling the majority class in the aspect of equal basis, imbalance ratio (IR) of pulsar dataset was defined. The hyperparameters of XGBoost (XGB) and LightGBM (LGBM) were tuned by the validation data for selecting the best model. The suggested model performance was related with other classifiers such as Support Vector Machine (SVM), Multilayer Perceptron (MLP), NB, Pseudo Nearest Centroid Neighbor (PNCN) [12] classifier and at last the Gaussian Hellinger Very Fast Decision Tree (GH –VFDT).

Performance results on HTRU2 and LOTAAS - 1 dataset:

In HTRU2, XGB surpassed the other classifiers with the accuracy of 0.981 where LGBM achieved the accuracy

of 0.980. Here in the case of LOTAAS – 1 dataset, both XGB and LGBM produced the same level of accuracy which is 0.999.

3) Classification of Pulsars using Deep Convolutional Neural Network (DCNN)

Yuan – Chao Wang, et al. [13] deployed DCNN, having different layers like:

- Convolutional – Eight
- Flatten - One
- Completely connected – Two

which was totally eleven layers. This proposed model was implemented on the HTRU 1 dataset. To address the class imbalance issue, oversampling technique based on minority synthetic sampling was deployed. For visualizing the samples, t – distributed Stochastic Neighbor Embedding (t – SNE) [14] was employed. Results obtained using DCNN model without synthetic samples achieved the recall of 0.851 and the precision of 0.848. The DCNN model with synthetic samples obtained the recall of 0.962 and precision of 0.963.

4) Classification of Pulsars utilizing Artificial Neural Network and Support Vector Machine

Thomas Ryan Devine, et al. [15] used ANN and SVM for classification. ANN classified instances by implementing the back propagation method and sigmoid as an activation function in every neural node.

Using SVM as an iterative training procedure, the error function was decreased. The divide and conquer strategy helped to reduce error.

5) Classification of Pulsars using a framework of Deep Convolutional Generative Adversarial Network (DCGAN) with Support Vector Machine

Ping Guo, et al. [16] used an architecture that combined SVM and a DCGAN. In this case, for generating the sample and for the purpose of feature learning model, DCGAN is deployed and SVM is used as a classifier to predict the label of the candidate.

6) Classification of Pulsars using Hybrid Ensemble Method

Y. Wang, et al. [17] used three ensemble methods. They were RF, XGB and Hybrid Ensemble method. HTRU 1 and HTRU2 dataset was used. In Hybrid Ensemble method, RF and XGB were integrated with Easy Ensemble. The problem of class imbalance was resolved using Easy Ensemble, which was also used to enhance the model's stability.

7) Classification of pulsars using different machine learning algorithms

Jin Rong Song [18] discussed various ML algorithms for classifying the pulsar stars. SVM, CNN, Gaussian NB, Logistic Regression (LR), DT, RF were deployed to classify the pulsar stars. Eight unique attributes and one target class were present in dataset, which had 12528 data values. The author employed two unique features to handle the missing data, i.e., drop (DR) method and average value (AV) method. The paper provided the results, obtained using both the DR method and AV method. Logistic Regression achieved the highest accuracy in classifying, with the accuracy rate of 0.99, when using DR method. In case, when replacing the missing value with AV method, SVM and CNN with 5 layers and 50epochs achieved the highest accuracy of 0.98.

8) Classifying pulsars and ranking the candidates for Fermi2FGL catalog

K. J. Lee, L. Guillemot, et al. [19] mainly focused on Bayesian data classification algorithms which used the Gaussian Mixture Model (GMM). Neyman-Pearson test was used for determining the data classification technique.

In the ranked list, the topmost 5 percent sources contained 50percent known pulsars, the topmost 50 percent contained 99 percent known pulsars. The GMM was tested by using the multi-dimensional Kolmogorov-Smirnov test. The paper discussed about the GMM and its applications in data modelling and classification, for instance it was

applied in P-$P^{\cdot}$ diagram for pulsar classification, in addition, modelling and ranking the 2FGL catalog point sources. Authors also conveyed, about the implications of the clusters, founded by GMM algorithm.

F. Detection of Pulsars

1) Detection of pulsar candidates – different classification algorithms were compared using SMOTE

Apratim Sadhu [20] utilized a variety of ML techniques, including LR, K-NN, SVM, DT, RF, Bagging, XGB, Adaptive Boosting, and Gradient Boosting. The techniques were implemented on HTRU2 dataset. Deploying 10-fold cross-validation method, the methods were implemented and compared. 90 percent of the total samples were used as training data and balance 10 percent was considered as test data. To overcome class imbalance problem, minority class got oversampled using the oversampling technique called SMOTE. In case of unbalanced dataset, XGB achieved the highest accuracy of 0.9797, whereas in SMOTE balanced dataset, XGB had the highest accuracy rate of 0.9732. Artificial Neural Network was also implemented with 9 layers and achieved the accuracy of nearly 98%.

2) Detection of Pulsar Candidates deploying Bagging Method

Mourad Azhari, et al. [21] connected the Bagging Method with primary classifiers: Core Vector Machines (CVM), K-NN, ANN, and Cart Decision Tree (CDT). For implementing the classifiers, HTRU2 dataset was used. Herein this case, Bagging algorithm worked in the level of two phases:

- Phase I -Training
- Phase II -Testing

To overcome imbalance issue, resampling techniques were used in the manner of data split and k fold CV. For the purpose of training and testing, the samples were divided into 70 percent and 30 percent respectively. The value of k in k fold cross validation was chosen as 10. The authors compared among the basic classifiers and also compared Bagging with different classifiers. Taking AUC metric into consideration, KNN (0.994) achieved better than others. In case of Bagging, Bagging (K-NN) (0.9913) outperformed others.

3) Detection of pulsars Feed Forward Backpropagation (FFBPN) and Cascade Forward Backpropagation Neural Network (CFBPN) Algorithms

Fahriye Gemci Furat, et al. [22] used HTRU_2 dataset to classify the pulsar by implementing, FFBPN and CFBPN algorithms. 8 features of the dataset were given as input to the neural networks. Hidden layer was assigned with 10 hidden neurons. Table I shows accuracies of both FFBPN and CFBPN.

Table I. Accuracies of Ffbpn and Cfbpn

Neural Network	FFBPN	CFBPN
Classification Accuracy – Training Data (TRD)	91.022	95.3704
Classification Accuracy – Testing Data (TED)	92.336	90.692

4) Detection of pulsars with few features using machine learning

Haitao Lin, et al. [23] proposed feature selection algorithms, to enhance the detection performance, Grid Search (GS) and Recursive Feature Elimination (RFE) were deployed by eliminating the unnecessary and redundant features. The algorithms were implemented on Southern High Time Resolution University survey (HTRU-S), which contained 1196 pulsars and 89996 non-pulsars with 18 features. For training, 50% of the data was used, denoted as NON-SAMPLING. The balanced training dataset were obtained by undersampling, oversampling and SMOTE which were termed as UNDERSAMPLE, OVERSAMPLE, SMOTE. Different models used in this paper were: Classification and regression tree (CART), Adaptive boosting (AdaBoost), Gradient boosting classifier (GBoost), XGB, RF. Except for RF, GS was applied. Grid Search (GS) and Recursive Feature Elimination (RFE) algorithms were applied to single, double features and also to multiple features. Both training and testing were splitted into 50% and were normalized, prior to giving them as input to the models. Then feature selection was done and the models were trained on the basis of the proposed feature selection algorithms. On the training data, five fold cross-validation was performed. A model having just two features from GS had a recall rate of up to 99 percent. A model with three features had a 99 percent recall rate when using RFE.

G. Prediction of Pulsars

1) Predicting pulsars with hybrid resampling approach

Ernesto Lee, et al. [24] used various supervised ML algorithms, for detecting true pulsar candidates. For implementation, HTRU2 dataset was used. Table II shows in detail about accuracies with different resampling approaches. As the dataset was imbalanced, two resampling methods were used: SMOTE, Adaptive Synthetic Resampling (ADASYN).

Table II. Details About Accuracies With Different Resampling Approaches

S. No	Details	Best	Accuracy
1	Without data resampling	Random Forest Logistic Regression	0.980
2	CC data resampling	Random Forest	0.943
3	SMOTE	Extra Tree Classifier	0.982
4	CR data resampling	Extra Tree Classifier	0.993
5	ADASYN	Extra Tree Classifier	0.981

And to minimize the size of majority class, Cluster Centroid(CC) undersampling method was employed. So, this hybrid resampling method or concatenated resampling (CR) method was suggested to solve class imbalance issue. Different ML models used in this paper were: RF, Gradient Boosting Classifier (GBC), Extra Tree Classifier (ETC), LR, MLP. Resampling was performed before splitting in the proportion of 70:30. Resampling was implemented on the training set. After completion of data resampling and data splitting, the given ML models were trained using 70 percent data. The remaining 30 percent was deployed for testing the trained models. The paper gave the details about the comparison of ETC with different resampling methods, when the data was splitted prior to data resampling. Deep learning models such as: long short-term memory (LSTM), deep neural network (DNN) and gated recurrent unit (GRU) also implemented, where each of them achieved the same accuracy of 0.98. In addition, 10-fold cross- validation was also implemented. Statistical T-test was also implemented to showcase the importance of CR technique.

H. *Pulsar candidate selection to classification*

R. J. Lyon, B. W. Stappers, et al. [25] suggested that enhancing the survey recommendations caused rise in pulsar candidate numbers and also data volumes. Candidate filters were deployed to solve those problems during the last 50 years. Here a new technique was proposed for online operation, which selected only positive candidates. This selection could be implemented using, Gaussian Hellinger Very Fast Decision Tree along with new set of features for describing candidates. With these properties, the suggested technique had a better level of pulsar recall and could execute millions of candidates in seconds. LOTAAS 1, HTRU 1, HTRU 2 datasets were used. Other ML methods were used to compare with the proposed method, they are: C4.5, MLP, NB, SVM. table III gives the details about the accuracies with different datasets.

Table III. Details About The Accuracies With Different Datasets

S. NO	DATASET	CLASSIFIER	ACCURACY
1	HTRU 1	GH-VFDT	0.988
2	HTRU 2	GH-VFDT	0.978
3	LOTAAS 1	SVM	0.999

I. *Machine learning pipeline*

Alexander Ylnner Choquenaira Florez, et al. [26] used HTRU2 dataset for implementing different ML algorithms. The techniques included: Data-Analysis, Pre-Processing, Sampling, Processing. The algorithms used were: NB, LR, DT, Perceptron, MLP, SVM. Various ensemble techniques were also implemented, such as: Stacking, Bagging, RF. For conducting various experiments, the authors divided HTRU2 dataset into 3 variations. Table IV gives details about variations of the dataset.

Table IV. Details About Variations of the Dataset

S. NO	DATASET	CLASSIFIER	ACCURACY
1	HTRU 1	GH-VFDT	0.988
2	HTRU 2	GH-VFDT	0.978
3	LOTAAS 1	SVM	0.999

In the experiment 2, authors considered only 6 features (feature selection) which followed the correlation proportion between them. Other than accuracy, Precision and Recall was also used as quality factors.

After implementing with K-Fold Cross Validation with k = 10, following models showed different accuracies. table V. gives details about accuracies achieved by different models in different variations of the dataset.

Table V. Details About Accuracies Achieved by Different Models in Variations of the Dataset

Dataset	Model	Accuracy
Dataset1	LR	0.98
	DT	0.98
	XGB	0.98
	Bagging	0.98
	Gradient	0.98
Dataset2	XGB	0.95
Dataset3	XGB	0.95

The table VI. shows accuracies with feature selection.

Table VI. Accuracies With Feature Selection

Dataset	Model	Accuracy
Dataset1	DT	0.98
	SVC-RbfK	0.98
	XGB	0.98
	RF	0.98
	Bagging	0.98
	Gradient	0.98
Dataset2	XGB	0.95
Dataset3	Bagging	0.97

The Table VII. shows different articles using different techniques on the pulsar dataset.

III. Discussion & Conclusion

Pulsar signals are most often very weak which can be easily drowned out by any other astrophysical sources or background noise. So, differentiating pulsar signals from RFI or from any other natural radio emissions from the galaxy is a very difficult task. From the reviews, the future work can be suggested that, different Ensemble Learning techniques such as Stacking, Voting Ensembles, Blending and quantum ML can be used to classify, identify, searching and for different operations on the pulsar stars. This paper reviews the meth-

Table VII. Different Articles Using Different Techniques on the Pulsar Dataset

S. NO	PAPER TITLE	DATASET	TECHNIQUES DEPLOYED	YEAR
1	Pulsar Classification: Comparing Quantum Convolutional Neural Networks and Quantum Support Vector Machines	HTRU-2	1. Quantum Kernel assisted Support Vector Machines (QSVMs) 2. Quantum Convolutional Neural Networks (QCNNs)	2023
2	Pulsar Candidate Classification Using a Computer Vision Method from a Combination of Convolution and Attention	FAST	CoAtNet-MLP-LR	2023
3	MFPIM: A Deep Learning Model Based on Multimodal Fusion Technology for Pulsar Identification	FAST	1. MFPIM-ResNet 2. MFPIM	2023
4	Advances in Pulsar Candidate Selection: A Neural Network Perspective	• PMPS • TGSS and NVSS	ANN	2023
		• HTRU-1 • HTRU • RXTE (Rossi X- ray Timing Explorer)	CNN	
		• HTRU-Medlat • PMPS26k	GAN	
		• HTRU • PALFA • GBNCC [27] • FAST	ResNet	
		• HTRU-Medlat	Hybrid model (WGAN+ResNet)	
5	Classical Ensembles of Single-Qubit Quantum Variational Circuits for Classification	HTRU 2	1. Bagging Ensemble 2. Boosting Ensemble 3. Single QAUM	2023
6	Random Forest Identification of Pulsars	HTRU2	1. Random Forest 2. Balanced the dataset: SMOTE and Subset of Noise	2023
7	A Pulsar Search Method Combining a New Feature Representation and Convolutional Neural Network	Self-Collected RXTE observation data	1. ConvNets - to learn 2D spatial information of the new pulsar feature representation and to classify them 2. Non-Homogeneous Poisson Process - provide training set for ConvNets 3. GAN – for data augmentation	2022
8	AdaBoost-MICNN: a new network framework for pulsar candidate selection	High Time Resolution Universe Medlat Data	AdaBoost-multi-input-CNN (AdaBoost-MICNN)	2022
9	Adaboost-DSNN: an adaptive boosting algorithm based on deep self normalized neural network for pulsar identification	HTRU-1 and HTRU-2	1. Deep Self Normalized Neural Network (Adaboost-DSNN) 2. Synthetic Minority Oversampling TEchnique (SMOTE) – to balance the dataset	2022
10	Pulsar-candidate Selection Using a Generative Adversarial Network and ResNeXt	HTRU Medlat	Combination of Deep Convolutional Generative Adversarial Neural Network (DCGAN) and a Deep Aggregation Residual Network (ResNeXt)	2022
11	Pulsar candidate selection with residual convolutional autoencoder	• HTRU Medlat • PMPS-26k	1. Residual Convolutional Autoencoder (Rcae) 2. Logistic Regression (Lr)	2022
12	Pulsar identification based on generative adversarial network and residual network	HTRU-Medlat	1. Generative Adversarial Networks – To handle class imbalance problem 2. deep neural network – using intra- and inter-block residual connectivity – recognition accuracy	2022
13	Stellar and Pulsar Classification using Machine Learning	HTRU	1. k-NN 2. Decision Tree 3. Random Forest	2021
14	Quantum Machine Learning for Radio Astronomy	HTRU 2	Born machine (Quantum Neural Network)	2021
15	Concat Convolutional Neural Network for pulsar candidate selection	FAST	Concat Convolutional Neural Network	2020

ods used for pulsar classification, identification, selecting and separating mainly on the basis of ML, talked about some of the problem that needs to be solved and proposed some methods that can be carried out on the pulsar stars.

References

[1] A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, R. A. Collins, "Observation of a Rapidly Pulsating Radio Source", Nature, 217 (5130):709-713, 1968

[2] V. I. Shishov, T. V. Smirnova, C. R. Gwinn, A. S. Andrianov, M. V. Popov, A. G. Rudnitskiy, V. A. Soglasnov, "Interstellar scintillations of PSR B1919+21: space–ground interferometry", Monthly Notices of the Royal Astronomical Society, Volume 468, Issue 3, Pages 3709–3717, 25 April 2017.

[3] M. Ryle, "Mullard Radio Astronomy Observatory, Cavendish Laboratory, University of Cambridge. Report for the period 1977 October 1 to 1978 September 30", Quarterly Journal of the Royal Astronomical Society, Vol. 20, p. 261 – 271, 1979.

[4] M. Bailes, B. K. Berger, P. R. Brady, M. Branchesi, K. Danzmann, M. Evans, K. Holley-Bockelmann, B. R. Iyer, T. Kajita, S. Katsanevas, M. Kramer, A. Lazzarini, L. Lehner, G. Losurdo, H. Lück, D. E. McClelland, M. A. McLaughlin, M. Punturo, S. Ransom, S. Raychaudhury, D. H. Reitze, F. Ricci, S. Rowan, Y. Saito, G. H. Sanders, B. S. Sathyaprakash, B. F. Schutz, A. Sesana, H. Shinka, X. Siemens, D. H. Shoemaker, J. Thorpe, J. F. J. van den Brand, S. Vitale, "Gravitational-wave physics and astronomy in the 2020s and 2030s", Nature Reviews Physics, volume 3, pages344–366 (2021), 14 April 2021.

[5] Iqbal H Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", SN Computer Science, volume 2, Article number: 160 (2021), 22 March 2021.

[6] Dalya Baron, "MACHINE LEARNING IN ASTRONOMY: A PRACTICAL OVERVIEW", arXiv:1904.07248v1 [astro-ph.IM],15 Apr 2019.

[7] M. J. Keith, A. Jameson, W. van Straten, M. Bailes, S. Johnston, M. Kramer, A. Possenti, S. D. Bates, N. D. R. Bhat, M. Burgay, S. Burke-Spolaor, N. D. Amico, L. Levin, Peter L McMahon, S. Milia, B. W. Stappers, "The High Time Resolution Universe Pulsar Survey – I. System configuration and initial discoveries", Monthly Notices of the Royal Astronomical Society, Volume 409, Issue 2, December 2010, Pages 619–627, 15 November 2010.

[8] Chhaya A Khanzode Ku, Ravindra D Sarode, "Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review", International Journal of Library & Information Science (IJLIS), Volume 9, Issue 1, pp. 30-36, January-April 2020.

[9] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology (IJCTT) – Volume 48, Number 3 June 2017.

[10] I. Tariq, M. Qiao, L. Wei, S. Yao, C. Zhou, Z. Ali, S. W. Azeem, A.Spanakis-Misirlis, "Classification of pulsar signals using ensemble gradient boosting algorithms based on asymmetric under- sampling method", Journal of Instrumentation, Volume 17, 21 March 2022.

[11] https://archive.ics.uci.edu/dataset/372/htru2

[12] Jiangping Xiao, Xiangru Li, Haitao Lin, Kaibin Qiu, "Pulsar candidate selection using pseudo-nearest centroid neighbour classifier", Monthly Notices of the Royal Astronomical Society, Volume 492, Issue 2, Pages 2119–2127, 19 December 2019.

[13] Yuan-Chao Wang, Ming-Tao Li, Zhi-Chen Pan, Jian-Hua Zheng, "Pulsar candidate classification with deep convolutional neural networks", Research in Astronomy and Astrophysics, Volume 19, Number 9.

[14] Wenbo Zhu, Zachary T Webb, Kaitian Mao, José Romagnoli, "A Deep Learning Approach for Process Data Visualization Using t-Distributed Stochastic Neighbor Embedding", Industrial & Engineering Chemistry Research, 9564–9575, 16 May 2019.

[15] Thomas Ryan Devine, Katerina Goseva-Popstojanova, Maura McLaughlin, "Detection of dispersed radio pulses: a machine learning approach to candidate identification and classification", Monthly Notices of the Royal Astronomical Society, Volume 459, Issue 2, Pages 1519–1532, 01 April 2016.

[16] Ping Guo, Fuqing Duan, Pei Wang, Yao Yao, Qian Yin, Xin Xin,Di Li, Lei Qian, Shen Wang, Zhichen Pan, Lei Zhang, "Pulsar candidate classification using generative adversary networks", Monthly Notices of the Royal Astronomical Society, Volume 490, Issue 4, Pages 5424–5439, 14 November 2019.

[17] Y. Wang, Z. Pan, J. Zheng, L. Qian, M. Li, "A hybrid ensemble method for pulsar candidate classification", Astrophysics and Space Science 364(8), August 2019.

[18] Jin Rong Song, "The effectiveness of different machine learning algorithms in classifying pulsar stars and the impact of data preparation", Journal of Physics: Conference Series, Volume 2428, 2022 2nd International Conference on Detection Technology and Intelligence System (DTIS 2022) Tianjin, China, 14/10/2022 - 16/10/2022.

[19] K. J. Lee, L. Guillemot, Y. L. Yue, M. Kramer, D. J. Champion, "Application of the Gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the Fermi 2FGL catalogue", Monthly Notices of the Royal Astronomical Society, Volume 424, Issue 4, Pages 2832–2840, 21 August 2012

[20] Apratim Sadhu, "Pulsar Star Detection: A Comparative Analysis of Classification Algorithms using SMOTE", International Journal of Computer and Information Technology, Vol. 11 No. 1 (2022): February 2022, 2022-03-05.

[21] Mourad Azhari, Abdallah Abarda, Altaf Alaoui, Badia Ettaki, Jamal Zerouaoui, "Detection of Pulsar Candidates using Bagging Method", Procedia Computer Science, Volume 170, Pages 1096- 1101,2020.

[22] Fahriye Gemci Furat, Turgay Ibrikci, "Application of Feed Forward Backpropagation and Cascade Forward Backpropagation Neural Network Algorithms to Detect Pulsar Stars", International Advanced Researches & Engineering Congress-2017, 16-18 November 2017.

[23] Haitao Lin, Xiangru Li, Ziying Luo, "Pulsars detection by machine learning with very few features", Monthly Notices of the Royal Astronomical Society, Volume 493, Issue 2, April 2020, Pages 1842–1854, 28 January 2020.

[24] Ernesto Lee, Furqan Rustam, Wajdi Aljedaani, Abid Ishaq, Vaibhav Rupapara, Imran Ashraf, "Predicting Pulsars from Imbalanced Dataset with Hybrid Resampling Approach", Advances in Astronomy, Volume 2021, 03 Dec 2021.

[25] [25] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach", Monthly Notices of the Royal Astronomical Society, Volume 459, Issue 1, 11 June 2016, Pages 1104–1123, 15 April 2016.

[26] Alexander Ylnner Choquenaira Florez, Braulio Valentin S´anchez Vinces, Diana Carolina Roca Arroyo, Josimar Edinson Chire Saire, Patr´ıcia Batista Franco, "Machine Learning Pipeline for Pulsar Star Dataset", arXiv:2005.01208v1 [astro-ph.IM] 3 May 2020.

[27] J. K. Swiggum, Z. Pleunis, E. Parent, D. L. Kaplan, M. A. McLaughlin, I. H. Stairs, R. Spiewak, G. Y. Agazie, P. Chawla, M. E. DeCesar, T. Dolch, W. Fiore, E. Fonseca, A. G. Istrate, V. M. Kaspi, V. I. Kondratiev, J. van Leeuwen, L. Levin, E. F. Lewis, R. S. Lynch, A. E. McEwen, H. Al Noori, S. M. Ransom, X. Siemens, and M. Surnis, "The Green Bank North Celestial Cap Survey. VII. 12 New Pulsar Timing Solutions", The Astrophysical Journal, 944:154 (15pp), 20 February 2023.

DOI: 10.15439/2023R56
ISSN 2300-5963 ACSIS, Vol. 38

MACCHIEF—Machine learning-based Algorithm Classification for Complaint Handling and Improved Efficiency in Firms

Vu Duy Trung
Vietnam National University,
Hanoi-International School
Hanoi, Vietnam
trungthichban@gmail.com

Yan Chi Toh
Ngee Ann Polytechnic,
Singapore
yanchi273@gmail.com

Satyam Mishra
Vietnam National University
International School
Hanoi, Vietnam
satyam.entrprnr@gmail.com
0000-0002-7457-0060

Le Anh Ngoc*
Swinburne Vietnam, FPT University
Hanoi, Vietnam
ngocla2@fe.edu.vn

Phung Thao Vi
Vietnam National University – International School
Hanoi, Vietnam
phungvi08123@gmail.com

Abstract—**This research emphasizes the vital role of machine learning-driven consumer complaint management in information enterprises facing a surge in customer feedback across channels. By automating complaint categorization, analysis, and response, machine learning streamlines operations and uncovers invaluable customer insights. The study introduces a novel classification model, with LGBMClassifier and LinearSVC algorithms standing out for achieving 76.78% and 79.37% accuracy, respectively. This approach enhances complaint resolution, customer satisfaction, and enterprise competitiveness. The integration of machine learning offers a practical solution to consumer complaint challenges, with future prospects including adaptability to evolving preferences and leveraging natural language processing for deeper sentiment analysis.**

Index Terms—**Classification Algorithms, Machine Learning Models, Algorithm Evaluation, LGBMClassifier, LinearSVC, CatBoost Algorithm**

I. Introduction

Customer complaint management is crucial in maintaining strong customer relationships. The aim is to enhance satisfaction and loyalty by effectively addressing concerns. However, the increasing volume of feedback from various channels poses challenges for prompt and efficient oversight. Here, machine learning emerges as a powerful solution, enabling computers to learn from data and perform tasks that mimic human cognition. Machine learning provides a robust avenue for businesses to automate the intricate process of classifying, directing, and scrutinizing customer complaints. Beyond mere automation, it affords a lens into the underlying triggers, sentiments, and patterns enveloping customer dissatisfaction. By harnessing the arsenal of machine learning techniques, businesses can markedly elevate their prowess in managing customer complaints. This, in turn, catalyzes augmented customer retention, advocacy, and grants a palpable competitive edge within the market landscape [1, 2].

Information enterprises span diverse sectors such as media, libraries, and software firms, specializing in managing information intricately. On the other hand, customer-centric operations focus on understanding and meeting customers' distinct needs and preferences. This alignment fosters loyalty, satisfaction, trust, and innovation, leading to increased profitability. Customer-centric operations in information enterprises can be realized through methods like creating customer personas and journey maps, customizing offerings using data insights, providing seamless experiences across platforms, and promoting engagement and co-creation. An additional facet involves the ongoing measurement and enhancement of customer outcomes and the inherent value they receive [3, 4] .

The primary aim of this research paper is to enhance consumer complaint management within information enterprises through the utilization of machine learning-based classification. To achieve this overarching goal, the paper will undertake the following objectives:

(i) Propose an innovative machine learning-powered classification model capable of autonomously categorizing customer complaints into distinct types and varying levels of severity. This categorization will be driven by an analysis of complaint content and sentiment.

(ii) Assess the efficacy and performance of the aforementioned model using real-world customer complaint data drawn from diverse information enterprises. These encompass media companies, publishing houses, libraries, data centers, and software firms.

(iii) Engage in a comprehensive exploration of the implications and advantages linked to the proposed model for information enterprises. This spans the realm of customer satisfaction and loyalty enhancement, innovation amplification, differentiation bolstering, and the augmentation of revenues and profitability.

This research paper's fundamental contribution lies in its provision of a holistic and pragmatic solution to the challenges of consumer complaint management within information enterprises through the integration of machine learning. The contribution encompasses:

(i) The development of an original machine learning-based classification model that adeptly handles the spectrum of customer complaint types and severity levels. Furthermore, the model offers insights into the triggers, emotions, and patterns driving customer dissatisfaction.

* Corresponding Author

(ii) A demonstration of the model's viability and efficacy through an expansive dataset encompassing varied customer complaints from an array of information enterprises. This dataset serves as a benchmark for future research endeavors.

(iii) The provision of actionable recommendations and practical insights to information enterprises, detailing how they can utilize machine learning to augment their customer complaint management processes and, by extension, improve outcomes.

II. Literature Review

Machine learning (ML), a subset of artificial intelligence, empowers computers to glean insights from data and tackle tasks that typically necessitate human intelligence. ML finds application in diverse facets of natural language processing (NLP), ranging from speech recognition to sentiment analysis. The evolution of ML's role in NLP unfolds across four primary phases: rule-based, statistical, neural, and hybrid. Rule-based ML hinges on manually constructed rules and dictionaries to process natural language. However, it grapples with constraints when faced with the intricacies, variations, and intricacies inherent in language. In contrast, statistical ML employs probabilistic models and algorithms, drawing wisdom from expansive collections of natural language data. This methodology excels in managing uncertainty, noise, and data scarcity within language. Neural ML relies on artificial neural networks to decode insights from natural language data, showcasing prowess in apprehending intricate, nonlinear patterns and representations woven into language. Presently, hybrid ML unites distinct ML methodologies to optimize strengths and offset weaknesses, emerging as the predominant trend within NLP research [5, 6].

Machine learning also finds its place in computer vision, dealing with grasping visual information like images and videos [18, 19] . One specific task within this field is object detection and measurement. This involves pinpointing and measuring the sizes of objects in images. This task has many practical uses, including quality control, inventory management, medical imaging, and augmented reality. For instance, SATMeas, a creation by Mishra and Thanh [7], demonstrates this. It's a mobile app capable of real-time object detection and measuring properties like length, width, height, area, and perimeter using the canny edge detection algorithm. This method is widely employed for detecting object edges in images.

Text classification and sentiment analysis, on the other hand, are natural language processing tasks that assign labels or scores to texts based on their content and context. Different approaches to these tasks are machine learning, lexicon, and hybrid. Machine learning uses algorithms and models that learn from data to perform these tasks. Machine learning can be supervised, unsupervised, or semi-supervised, depending on the type of data used. Lexicon uses predefined dictionaries or lists of words or phrases that have associated sentiment scores or polarities. Lexicon can be rule-based or corpus-based, depending on the source of the words or phrases. Hybrid uses a combination of machine learning and lexicon methods to leverage their strengths and overcome their weaknesses. Hybrid can be ensemble-based or feature-based, depending on the way of combining the methods [8-10].

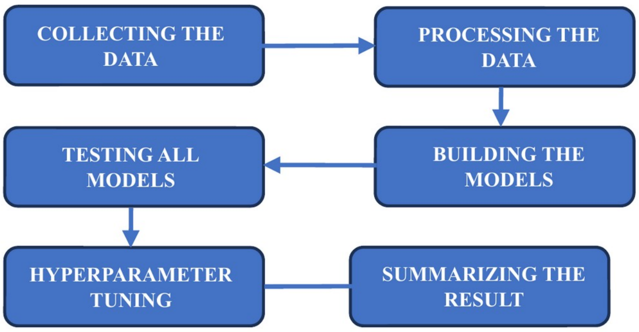
Notably, complaint management embodies the strategy of effectively addressing and resolving customer grievances promptly and to their satisfaction. Both businesses and customers grapple with an array of challenges inherent to this process, collectively termed as complaint management obstacles. Several of these hurdles encompass: (i) a lack of awareness and accessibility, signifying that many consumers encounter difficulties in comprehending where and how to voice their complaints or face impediments while attempting to access the appropriate complaint channels [11]; (ii) a deficit in trust and confidence, wherein a considerable number of consumers harbor skepticism regarding the impartial and efficacious handling of their grievances, often harboring concerns about potential adverse consequences from the business [11]; (iii) insufficiencies in responsiveness and equity, with many consumers facing delays or inadequate responses from businesses or complaint management entities. In certain instances, the perceived fairness or neutrality of the complaint process or outcome might come into question [11, 12]; and (iv) a deficiency in feedback and learning, highlighting a scenario where numerous businesses neglect to utilize customer feedback or complaints as a catalyst for refining their offerings, services, or operational methodologies [12]. Additionally, there's a failure to foster communication or follow-up with customers subsequent to the resolution of their complaints.

Certainly, in the world of research, the integration of machine learning to elevate customer service has garnered attention. ML, an offshoot of artificial intelligence, equips computers to grasp insights from data and execute tasks mirroring human intelligence. The prospects for ML in enhancing customer service are manifold: it can streamline routine tasks like addressing FAQs and appointment scheduling, personalize interactions by recommending products and offering tailored support, and distill valuable insights from customer feedback [13-15]. This integration finds practical expression through examples like chatbots, software agents that engage in real-time conversations, and recommendation systems that offer personalized suggestions. Furthermore, sentiment analysis, which dissects emotions from text, plays a crucial role in discerning customer satisfaction, dissatisfaction, and churn risks. All in all, this amalgamation of ML and customer service stands as a frontier of innovation with tangible benefits.

Previous research on complaint classification has yielded noteworthy insights. The granularity of complaint classification varies across levels, encompassing product, service, process, or outcome. These diverse levels offer distinct implications for both businesses and customers [16]. The methods employed for complaint classification exhibit variability, encompassing machine learning, lexicon, and hybrid approaches. Machine learning methods entail algorithms and models trained on labeled or unlabeled data, lexicon methods hinge on predefined dictionaries or lists, while hybrid methods amalgamate machine learning and lexicon techniques to harness strengths and offset limitations ([17]). Noteworthy factors influencing complaint classification include customer attributes, complaint channels, emotions, and contextual elements. These variables impact the structure, tone, and content of customer complaints, thereby influencing the accuracy and efficacy of classification methods. Consequently, accounting for these factors during the design and evaluation of complaint classification techniques is pivotal [16, 17].

III. Methodology

The method to complete this study is represented as the following flow chart:



A. Data Collection

In the pursuit of conducting this research, an extensive dataset comprising 10,000 consumer complaints was meticulously collected. This dataset serves as the foundation upon which our investigation and analysis are built. The dataset comprises a diverse array of attributes, meticulously curated to encapsulate multifaceted information concerning consumer complaints, corporate entities involved, the intrinsic nature of the grievances, and the subsequent outcomes arising from the interactions between discerning consumers and the implicated companies. We got dataset from Kaggle. The dataset attributes encompass a comprehensive panorama of pertinent details, pivotal for our exploration. These attributes encompass: Date received, Product, Sub-product, Issue, Sub-issue, Consumer Complaint, Company Public Response, Company, State, ZIP code, Tags, Consumer consent provided, submitted via, Date Sent to Company, Company Response to Consumer, Timely response, Consumer disputed, Complaint ID.

The exhaustive compilation of these attributes within our dataset culminates in an expansive and multifaceted repository of consumer complaints, enabling our research to delve deeply into the intricate dynamics that underpin the interactions between consumers and companies. Through meticulous analysis and investigation, we endeavor to unearth patterns, trends, and insights that contribute substantively to the understanding of consumer-company interactions and the subsequent ramifications thereof.

B. Data preprocessing

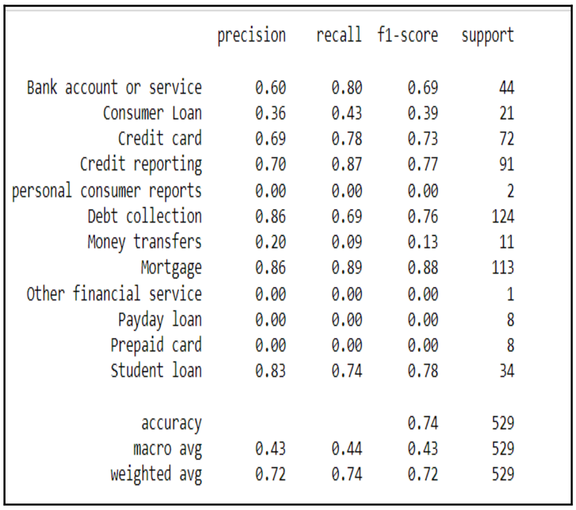
In our programming endeavors, we employed Python as our primary coding language. Moreover, we took advantage of Jupyter - an open-source initiative furnishing a web-based, interactive computational platform accommodating multiple programming languages, including Python. The essential libraries we needed to set up for our research programming encompassed pandas, numpy, and matplotlib.pyplot.

We then provided the different numerical feature vectors to different text documents. Since our classifiers cannot directly use the document, we need to convert a dataset into fixed numerical feature vectors instead of the raw document with variable length, to convert the collection of these documents into token form.

IV. Result and Discussion

A. Model Building

After converting textual documents into numerical feature vectors, we move on to comparing different classifiers for their accuracy. Beginning with LinearSVC, we generated an F1-score classification report for various categories:



```
                          precision    recall  f1-score   support

   Bank account or service       0.60      0.80      0.69        44
             Consumer Loan       0.36      0.43      0.39        21
               Credit card       0.69      0.78      0.73        72
          Credit reporting       0.70      0.87      0.77        91
 personal consumer reports       0.00      0.00      0.00         2
           Debt collection       0.86      0.69      0.76       124
           Money transfers       0.20      0.09      0.13        11
                  Mortgage       0.86      0.89      0.88       113
   Other financial service       0.00      0.00      0.00         1
               Payday loan       0.00      0.00      0.00         8
              Prepaid card       0.00      0.00      0.00         8
              Student loan       0.83      0.74      0.78        34

                  accuracy                           0.74       529
                 macro avg       0.43      0.44      0.43       529
              weighted avg       0.72      0.74      0.72       529
```

Figure 1: Output of testing code for LinearSVC model

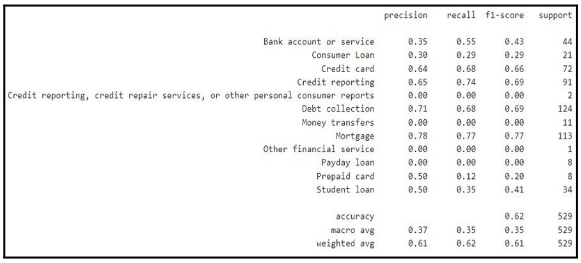
The report indicates an accuracy of around 75%, signifying commendable performance. This model is particularly known for its effectiveness in high dimensional spaces and versatility in handling both binary and multiclass classification problems. In our case, we tuned the model to achieve a cross-validation score of 79.37%. This high score indicates the model's reliability and its ability to generalize well, making it a strong contender in our suite of models.

Next, we evaluated an AI-Based BernoulliNB model. This model is based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Despite its lower accuracy of 65.89% compared to the LinearSVC model, it still demonstrated notable performance. This suggests that even with its simplicity, the BernoulliNB model can still be a valuable tool in certain scenarios.

Our testing also included a Decision Tree Classifier. This model is a non-parametric supervised learning method used for classification and regression. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

However, this model achieved a lower accuracy of 62.65%, falling short of the preceding two models.

We then applied a CatBoost Classifier, which is a machine learning algorithm that uses gradient boosting on decision trees. It is known for its capabilities in handling categorical data and reducing overfitting. Upon execution, the

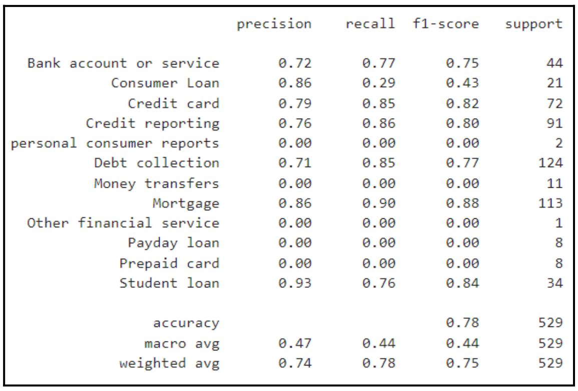


```
                                                                              precision    recall  f1-score   support

                                                      Bank account or service       0.35      0.55      0.43        44
                                                                Consumer Loan       0.30      0.29      0.29        21
                                                                  Credit card       0.64      0.68      0.66        72
                                                             Credit reporting       0.65      0.74      0.69        91
Credit reporting, credit repair services, or other personal consumer reports       0.00      0.00      0.00         2
                                                              Debt collection       0.71      0.68      0.69       124
                                                              Money transfers       0.00      0.00      0.00        11
                                                                     Mortgage       0.78      0.77      0.77       113
                                                      Other financial service       0.00      0.00      0.00         1
                                                                  Payday loan       0.00      0.00      0.00         8
                                                                 Prepaid card       0.50      0.12      0.20         8
                                                                 Student loan       0.50      0.35      0.41        34

                                                                     accuracy                           0.62       529
                                                                    macro avg       0.37      0.35      0.35       529
                                                                 weighted avg       0.61      0.62      0.61       529
```

Figure 2: Output of testing code for Decision Tree Classifier model

CatBoost Classifier exhibited an accuracy of approximately 75.02%. This result shows promise for this model, suggesting it could be a good fit for our data.

Our assessment continued with a Random Forest Classifier, which is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

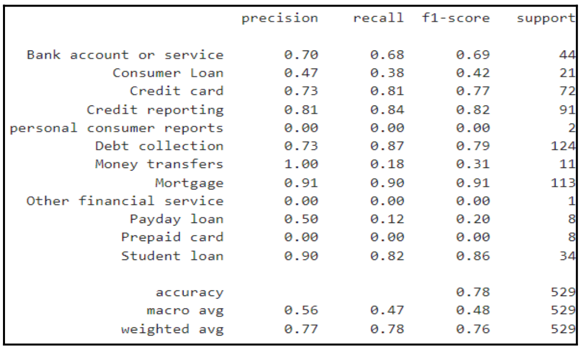


```
                          precision    recall  f1-score   support

 Bank account or service       0.72      0.77      0.75        44
           Consumer Loan       0.86      0.29      0.43        21
             Credit card       0.79      0.85      0.82        72
        Credit reporting       0.76      0.86      0.80        91
personal consumer reports       0.00      0.00      0.00         2
         Debt collection       0.71      0.85      0.77       124
         Money transfers       0.00      0.00      0.00        11
                Mortgage       0.86      0.90      0.88       113
 Other financial service       0.00      0.00      0.00         1
             Payday loan       0.00      0.00      0.00         8
            Prepaid card       0.00      0.00      0.00         8
            Student loan       0.93      0.76      0.84        34

                accuracy                           0.78       529
               macro avg       0.47      0.44      0.44       529
            weighted avg       0.74      0.78      0.75       529
```

Figure 3: Output of testing code for CatBoost Classifier model

This classifier yielded a solid accuracy of 75.65%, making it a viable model.

Lastly, we tested an LGBMClassifier, which is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages: faster training speed and higher efficiency, lower memory usage, better accuracy, support of parallel and GPU learning, capable of handling large-scale data.



```
                          precision    recall  f1-score   support

 Bank account or service       0.70      0.68      0.69        44
           Consumer Loan       0.47      0.38      0.42        21
             Credit card       0.73      0.81      0.77        72
        Credit reporting       0.81      0.84      0.82        91
personal consumer reports       0.00      0.00      0.00         2
         Debt collection       0.73      0.87      0.79       124
         Money transfers       1.00      0.18      0.31        11
                Mortgage       0.91      0.90      0.91       113
 Other financial service       0.00      0.00      0.00         1
             Payday loan       0.50      0.12      0.20         8
            Prepaid card       0.00      0.00      0.00         8
            Student loan       0.90      0.82      0.86        34

                accuracy                           0.78       529
               macro avg       0.56      0.47      0.48       529
            weighted avg       0.77      0.78      0.76       529
```

Figure 4: Output of testing code for LGBMClassifier model

The LGBMClassifier achieved a respectable accuracy of 76.78%, reinforcing its potential.

B. *Hyperparameter Tuning*

In our experiment, we performed hyperparameter tuning to find the best set of parameters for our models. This process involves adjusting the algorithm parameters to optimize model performance. For instance, we used the Random Forest Classifier as an example and found the most suitable parameters for this model, which is, 0.726.

C. *Result summary*

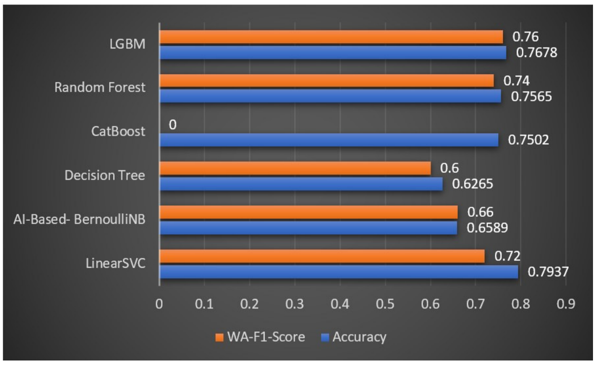


Figure 5: Classification models performance

Table 1: Result Summary

Algorithm	Accuracy	WA-Precision	WA-Recall	WA-F1-Score
LinearSVC	0.7937	0.72	0.74	0.72
AI-Based-BernoulliNB	0.6589	0.67	0.69	0.66
Decision Tree	0.6265	0.60	0.61	0.60
CatBoost	0.7502	N/A	N/A	N/A
Random Forest	0.7565	0.72	0.78	0.74
LGBM	0.7678	0.77	0.78	0.76

As shown in Figure 05 and Table 1, the performance of different classification algorithms was evaluated. The LGBMClassifier and LinearSVC algorithms demonstrated exceptional accuracy, achieving 76.78% and 79.37%, respectively. These models effectively strike a balance between precision and recall, making them robust options for categorizing consumer complaints. The CatBoost algorithm closely follows with an accuracy of 75.02%, while the Random Forest model also displayed commendable performance with 74.95% accuracy. In contrast, the AI-Based BernoulliNB and Decision Tree classifiers exhibited relatively lower accuracy rates of 65.80% and 61.58%, respectively. In essence, the LGBMClassifier and LinearSVC models emerge as the most reliable contenders for accurate and efficient classification within the domain of consumer complaint management.

V. Conclusion

In summary, this study underscores the crucial need for effective consumer complaint management in information enterprises, driven by the capabilities of machine learning. As the volume of customer feedback surges through various channels, the urgency of swiftly and effectively addressing concerns becomes paramount. Machine learning offers a remedy by automating the complex tasks of categorizing, analyzing, and responding to complaints. This not only streamlines operations but also unveils valuable insights into customer sentiments and behavior patterns. Embracing customer-centricity, information enterprises stand to gain significant advantages by aligning with customer preferences and needs. The proposed machine learning-based classification model, evaluated using real-world complaint data, introduces an innovative approach to enhancing consumer complaint management. Notably, the LGBMClassifier and LinearSVC algorithms shine as leaders in both accuracy and balance. Both LGBMClassifier and LinearSVC algorithms emerge as standout performers, achieving notable accuracy levels of 76.78% and 79.37%, respectively.

This study's applicability is evident in its focus on optimizing consumer complaint management within information enterprises. It offers practical solutions through the application of machine learning, enabling businesses to automate complaint categorization, streamline operations, and gain valuable customer insights. By implementing the LGBMClassifier and LinearSVC models, enterprises can efficiently address customer concerns, leading to improved satisfaction, loyalty, and a competitive edge. The study's potential for adaptability to evolving customer preferences and complaint trends further enhances its applicability to businesses dealing with a high volume of consumer feedback.

Additionally, this research provides a practical avenue for information enterprises to refine their complaint management strategies, ultimately leading to improved customer satisfaction, loyalty, and competitive edge. By integrating machine learning into this framework, the study delivers a valuable solution to the challenges information enterprises encounter in addressing consumer complaints, paving the way for enhanced outcomes and innovation. In the realm of future research, further advancements could be made to enhance the proposed machine learning model's adaptability to evolving customer preferences and complaint trends. Additionally, exploring the integration of natural language processing techniques could potentially refine the model's ability to extract nuanced insights from customer feedback, thereby deepening the understanding of sentiment and driving more personalized complaint resolution strategies.

References

[1] S. Peker, "Predicting Firms' Performances in Customer Complaint Management Using Machine Learning Techniques," in International Conference on Intelligent and Fuzzy Systems, 2022: Springer, pp. 280-287.

[2] M. Zaki, R. McColl-Kennedy, and A. Neely, "Using AI to Track How Customers Feel—In Real Time," 2021.

[3] S. Tuominen, H. Reijonen, G. Nagy, A. Buratti, and T. Laukkanen, "Customer-centric strategy driving innovativeness and business growth in international markets," International Marketing Review, no. ahead-of-print, 2022.

[4] V. Guerola-Navarro, H. Gil-Gomez, R. Oltra-Badenes, and P. Soto-Acosta, "Customer relationship management and its impact on entrepreneurial marketing: A literature review," International Entrepreneurship and Management Journal, pp. 1-41, 2022.

[5] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," Multimedia tools and applications, vol. 82, no. 3, pp. 3713-3744, 2023.

[6] K. Jiang and X. Lu, "Natural language processing and its applications in machine translation: a diachronic review," in 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), 2020: IEEE, pp. 210-214.

[7] S. Mishra and L. T. Thanh, "SATMeas-Object Detection and Measurement: Canny Edge Detection Algorithm," in International Conference on AI and Mobile Services, 2022: Springer, pp. 91-101.

[8] A. Ullah, S. N. Khan, and N. M. Nawi, "Review on sentiment analysis for text classification techniques from 2010 to 2021," Multimedia Tools and Applications, vol. 82, no. 6, pp. 8137-8193, 2023.

[9] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," Social Network Analysis and Mining, vol. 11, no. 1, p. 81, 2021.

[10] S. Zhang, "Sentiment classification of news text data using intelligent model," Frontiers in Psychology, vol. 12, p. 758967, 2021.

[11] C. Brennan, T. Sourdin, J. Williams, N. Burstyner, and C. Gill, "Consumer vulnerability and complaint handling: Challenges, opportunities and dispute system design," International journal of consumer studies, vol. 41, no. 6, pp. 638-646, 2017.

[12] M. Stone, "Literature review on complaints management," Journal of Database Marketing & Customer Strategy Management, vol. 18, pp. 108-122, 2011.

[13] C. Ledro, A. Nosella, and A. Vinelli, "Artificial intelligence in customer relationship management: literature review and future research directions," Journal of Business & Industrial Marketing, vol. 37, no. 13, pp. 48-63, 2022.

[14] A. De Mauro, A. Sestino, and A. Bacconi, "Machine learning and artificial intelligence use in marketing: a general taxonomy," Italian Journal of Marketing, vol. 2022, no. 4, pp. 439-457, 2022.

[15] I. Sarker, "Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2: 160," ed, 2021.

[16] A. Filip, "Complaint management: A customer satisfaction learning process," Procedia-Social and Behavioral Sciences, vol. 93, pp. 271-275, 2013.

[17] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," International Journal, vol. 2, no. 6, pp. 282-292, 2012.

[18] D. T. Van, T. D. Thang, T. B. Hung, and N. K. Giao, "Application of Machine Learning in Malicious IoT Classification and Detection on Fog-IoT Architecture," Annals of Computer Science and Information Systems, vol. 33, pp. 299-303, 2022

[19] H.-S. Vu and V.-H. Nguyen, "Safety-Assisted Driving Technology Based on Artificial Intelligence and Machine Learning for Moving Vehicles in Vietnam," 2022

DOI: 10.15439/2023R55
ISSN 2300-5963 ACSIS, Vol. 38

DICKT - Deep Learning-Based Image Captioning using Keras and TensorFlow

Phung Thao Vi
International School – Vietnam
National University
Hanoi, Vietnam
phungvi08123@gmail.com

Satyam Mishra
International School – Vietnam
National University
Hanoi, Vietnam
satyam.entrprnr@gmail.com
0000-0002-7457-0060

Le Anh Ngoc*
Swinburne Vietnam,
FPT University
Hanoi, Vietnam
ngocla2@fpt.edu.vn

Sundaram Mishra
NETMONASTERY NSPL, Mumbai, India
mishrasundaram.sm@gmail.com

Vu Minh Phuc
International School – Vietnam National University
Hanoi, Vietnam
vphuc2411@gmail.com

***Abstract*—This study evaluates a caption generation model's performance using the BLEU Score metric. The model generates descriptions for images, compared to reference captions with single and dual references. Results show a high BLEU Score, suggesting human-like captions. However, BLEU primarily measures linguistic similarity and n-gram overlap, missing full human-generated caption richness. The findings reveal the model's potential to convey image essence in text, but highlight BLEU Score limitations. TensorFlow and Keras are used for model development, acknowledging their widespread use but also their limitations. The research offers insights into caption generation model capabilities and urges a broader perspective on caption quality beyond quantitative metrics. While higher BLEU Scores are generally preferred, a "good" score varies with dataset and context. The study emphasizes a need for a more comprehensive approach to assess the quality and creativity of machine-generated captions.**

***Index Terms*—Image Captioning, Deep Learning, Keras, TensorFlow, BLEU.**

I. Introduction

The digital age, widespread amounts of photo information are generated day by day, from social media systems to surveillance structures. These pictures hold worthwhile facts, however having access to their content material stays a project. Image captioning, [1] the process of robotically producing descriptive textual descriptions for pics, bridges this gap and finds programs in content material retrieval, accessibility, and assistive technologies. Deep gaining knowledge has revolutionized photo captioning, permitting the development of greater accurate and contextually conscious structures [2]. In this research, authors delve into the realm of deep learning-primarily based picture captioning, leveraging Keras and TensorFlow, with a focus on technique, experiments, and implications.

Image captioning is a multi-modal [3] undertaking that mixes pc vision and herbal language processing (NLP). Authors look into using convolutional neural networks (CNNs) for picture feature extraction and recurrent neural networks (RNNs) with interesting mechanisms for producing coherent and contextually applicable captions. Through an extensive exploration of information series, preprocessing, model structure, education strategies, and evaluation metrics, authors aim to push the bounds of image captioning abilities. [4], [5]

This study contributes to the continuing improvement of smart structures capable of information and describing visual content, with capability packages in image search engines, automated content material tagging, and accessibility for the visually impaired. By addressing the technical demanding situations and barriers related to deep studying-primarily based photo captioning, authors offer treasured insights into the modern-day country of the artwork and open avenues for destiny studies in this rapidly evolving discipline. [6]

II. Literature Review

A. Image captioning and its applications

Image captioning is a field that combines herbal language processing (NLP) and laptop vision (CV) to generate textual descriptions for pictures. It makes use of strategies such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) to understand and interpret images, allowing the machine to offer captions in a commonplace language [7]. This generation has various applications, along with supporting blind individuals by way of presenting braille legible captions for photos, aiding in medical document generation, and automating the workflow of healthcare experts [8], [9]. However, the heavy computational burden and massive memory storage required by way of present captioning fashions restrict their deployment on aid-restricted gadgets. To address this, lightweight image captioning models were proposed, leveraging techniques which include compact function extraction and optimized go-modal fusion, resulting in decreased model size and advanced inference speed [10] . These advancements make image captioning models extra sensible for actual-international programs.

*Corresponding author

B. Deep learning for Image Captioning

Deep studying techniques are being used for photo captioning, which aims to generate descriptive and correct textual descriptions for pics. This technique involves leveraging convolutional neural networks (CNNs) for image function extraction and recurrent neural networks (RNNs) for sequential language generation. The procedure of picture captioning normally involves an photo encoder, inclusive of a CNN, to extract high-stage capabilities, and a language decoder, including an RNN, to generate captions [11], [12]. Deep studying has proven splendid success in diverse computer imaginative and prescient tasks, consisting of photograph captioning [13]. The use of deep neural networks enables computer systems to recognize and interpret visible content material, bridging the gap between the visual and textual domain names [14]. Image captioning has applications in diverse fields, which includes self-riding automobiles, robotics, and image analysis [15]. The improvement of accurate image captioning structures can make contributions to advancements in photograph expertise and human-gadget interplay with visual data.

C. Previous approaches and State-of-the-Art

Image captioning research has seen fast progress, from template-based total fashions to deep neural community models the use of the encoder-decoder structure. One success method is the usage of feature vectors extracted from place proposals received from an object detector. The Object Relation Transformer improves image captioning with the aid of incorporating data approximately the spatial relationship between detected objects via geometric interest [16]. In the field of biomedical picture captioning, there may be a want for assisting clinicians in the prognosis technique. Surveys have been conducted on biomedical photograph captioning, discussing datasets, assessment measures, and contemporary methods [17] [18]. A simple cosine similarity measure using the Mean of Word Embeddings (MOWE) of captions has shown high overall performance in unsupervised caption evaluation. The proposed metric WEbSim outperforms complex measures and units a brand-new baseline for caption assessment [19].

III. Methodology

First step is to import these libraries: Numpy, TensorFlow, Matplotlib.pyplot, and Pandas for our project.

After that authors will import specific modules:

keras.applications.vgg16: Importing the VGG16 model, which is a popular deep learning model for image classification.

- keras.applications.resnet50: Importing the ResNet50 model, another popular deep learning model for image classification.
- preprocess_input and decode_predictions: Functions for preprocessing images and decoding model predictions.
- keras.preprocessing: Modules for preprocessing data for deep learning models.
- keras.models: Modules for defining and working with neural network models.
- keras.layers: Modules for defining layers in neural networks.
- keras.layers.merge: Importing the add function for merging layers.

A. Data Collection

The data authors have used in our study is “Flickr8l” dataset - the "Flickr8k" dataset is a popular and widely used dataset in the field of computer vision and natural language processing (NLP). It is specifically used for tasks related to image captioning. Here's an overview of the dataset, it involves 6000 trainImage, 1000 devImages, and 1000 testImages.

Each image contains 5 captions that will be presenting that image. Here in figure 1 is an example of images and its unique identify names:

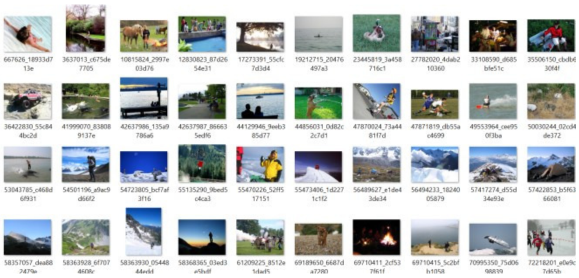


Fig.1. Image Caption Bot dataset

Each picture consists of descriptions of image content, in which every line represents a distinct photograph with a completely unique identifier and multiple descriptions. The descriptions are associated with the content material of the snapshots, offering facts about approximately what's happening or what is depicted in each photograph. is presenting a descriptions dictionary where key is ‘img_name’ and value is ‘cap’:

After reading the file and preprocessing the text file that contained images; it got the length of text, which is 40,460 lines (captions). A description dictionary that contains the image's captions is created, first is a list and after that obtaining captions one by one.

B. Data Cleaning

In the data cleaning process, the code follows a three-step approach:

1. Convert every word to its lowercase form.

2. Eliminate punctuations, unnecessary numbers, and special characters from the text.

3. Remove words with a length less than or equal to one.

This process results in clean sentences that effectively represent the associated images. To ensure efficiency and avoid redundancy when restarting the notebook, these cleaned sentences are stored in a dictionary, allowing us to access them directly at any time.

Next step is to access the world's frequency and features; it sorts words with a frequency less than a specified value. This step aims to eliminate outliers and uncommon words, retaining only those words that are frequently used to describe the data or depict images. Specifically, words must keep with a frequency greater than or equal to a threshold value of 10, as these words play a significant role in the model's purpose.

C. Loading of Training Set

For the loading of the training set, firstly, importing of the school dataset into our chosen development environment (e.g., Python with TensorFlow and Keras) for further processing. This step involves cleaning each caption and creating a dictionary. The '.jpg' characters are removed and keep only the image's names. These names with their corresponding captions are associated, it creates a 'trained_description' dictionary.

D. Data preprocessing

1) Data Preprocessing (Image)

In this section, images are loaded and do some processing so that it can feed it in the network. After this all, authors create two functions: one for loading an image and the other for converting it into an axis to enhance image features. Then the bottleneck features of the training store to disk. Then authors implement an encoding dictionary and an 'encode_image' function.

2) Data Preprocessing (For Image Caption)

The outcome of time is approximately 1818.58 seconds, it is the total time it took for encoding, and it is the important metric for evaluating performance and speed. Additionally, it will also be useful for tracking the performance and the efficiency of software and systems that have been used to process these images.

Next step, authors set up two dictionaries: 'word_to_index' and 'idx_to_word' . Since the input words cannot go directly into the model, the solution is that, authors convert words into corresponding indices. After prediction, numerical representations are obtained, which determine the maximum length of captions. For caption preprocessing, add 'Startseq' and 'Endseq' markers, and caption in the middle. The dataset employed in this study is extensive, as can be seen above, the predictive process initiates by introducing a "startseq" token as the first word. Subsequently, the model predicts the second word, which is appended to the "startseq" token. The process is repeated iteratively, with each predicted word being added to the sequence, and the model's predictions are influenced by both the image features and the preceding words. In essence, this progressive approach generates captions word by word.

3) Data Preparation using Generator Function

This data generator function is designed to provide input and output statistics in batches for schooling a neural community version for image captioning. It methods photograph descriptions, tokenizes them, and prepares them as enter-output pairs for the model. The generator keeps yielding batches of facts indefinitely until stopped externally.

E. Word Embeddings

To facilitate the manipulation of textual data, word embeddings are employed, transforming individual words into 50-dimensional vectors. These embeddings are derived from "glove vectors" as can be seen below. Each unique word in the vocabulary corresponds to a specific vector, and these embeddings are preloaded into the model for training. Since raw text data cannot be directly fed into the model, a pre-trained glove model is utilized to obtain these embeddings. An embedding output function is employed to link words to their respective embeddings, allowing for the integration of these representations into the model.

After that authors get embedded metrics to be used for generating word embeddings based on pre-trained word vectors and after shaping it, it gives the value of (1848, 50) which indicates that it contains 1848 words (or vocabulary size) each represented as a 50-dimensional vector.

F. Model Architecture

The model architecture comprises two main components: an image model and a caption model.

The total number of parameters in the model is 1,472,040. These parameters are the weights and biases that the model learns during training. This model appears to be designed for some kind of sequence-to-sequence or caption generation task, where it takes both image features and caption data as inputs and generates an output sequence.

1) Image model

Using "functional API" in Keras for creating a merge model for processing image data. Training the model at the preprocessed data using appropriate loss functions and optimization algorithms. The intention is to minimize the loss and first-rate-song the model's weights.

2) Caption Model

The caption model is based on a sequence model that deals with partial caption sequences. The model is tasked with generating captions that correspond to the given images. The training process involves the utilization of categorical cross-entropy as the loss function, and an appropriate optimization algorithm is employed to update the model's weights iteratively.

G. Training

The training phase involves the amalgamation of the image and caption models. The objective is to train this combined model using preprocessed data. During training, appropriate loss functions and optimization algorithms are employed to minimize the loss and optimize the model's parameters, ultimately enhancing its ability to generate accurate captions for images.

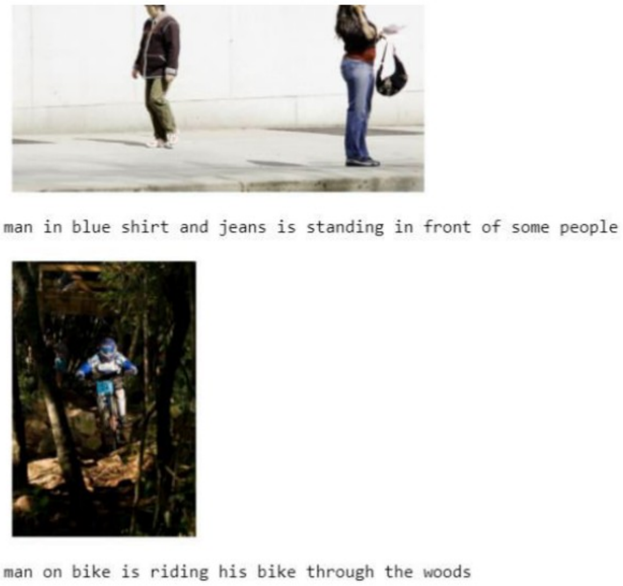
Authors used 10 epochs and the number of pictures used was 3; then a data generate function is called, it will provide output of data generated. The speed of generation will depend on the graphic card, especially, CPU will take a while. Authors utilized CPU in our case, but mostly people will prefer GPU over CPU.

1) Predictions

After creating a function for caption prediction, image's features vector provided, it gave probability of any word in vocab. It will choose the word with maximum probability using "argmax'', because prediction is in the form of indexes, so it is necessary to convert to words, then add to "in_text".

2) Results

By choosing a randomly set which includes 20 images, consequently, authors found out our proposed model is predicting well as can be seen in figure 2.



```
man in blue shirt and jeans is standing in front of some people
```

```
man on bike is riding his bike through the woods
```

Fig. 2. Captions predicted successfully

H. Evaluation

To compare image captions generated through a model, authors have utilized the usage of numerous BLEU (Bilingual Evaluation Understudy) metrics.

Details of the assessment system little by little are broken as follows.

1) Importing NLTK and Calculating BLEU Score

First step is to import the NLTK library. The BLEU score measures how similar the hypothesis sentence is to the reference sentence in terms of n-grams (contiguous sequences of words).

2) Averaging BLEU Scores Over 5

It iterates through the five references for the image, calculates the BLEU score for each reference compared to the same hypothesis, and accumulates these scores. Finally, it prints the average BLEU score for these references.

3) Generating Random Images and Evaluating Captions

This part of the code generates random images (`img_name`) and displays them using Matplotlib. It then generates a caption for each image using a function called `predict_caption`. After generating captions, it proceeds to calculate BLEU scores for each generated caption against the five reference captions (for the same image) using different n-gram weights and prints the results.

4) BLEU Score Calculation with Different Weights

It calculates BLEU scores for the generated hypothesis compared to the five reference captions for the same image using different n-gram weights.

IV. Results and Discussion

The effects of which evaluate the overall performance of a caption generated version by the use of the BLEU Score metric, it offers precious insights of generated captions. A higher BLEU Score suggests higher quality and alignment with human-generated captions.

Table 1: Bleu Score Of 5 Images From Our Proposed Study

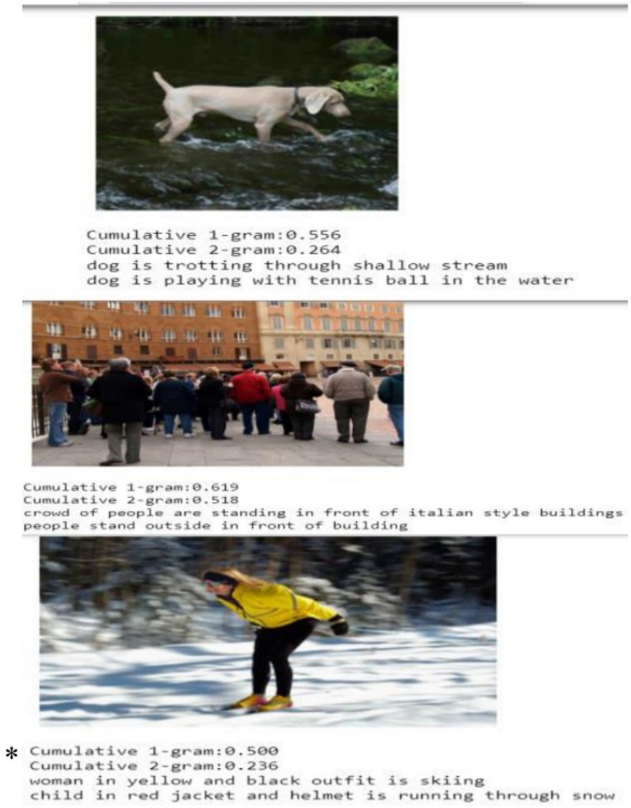
Image Id	**Bleu score of 5 images from our proposed study**		
	Generated Caption	*BLEU SC-1*	*BLEU SC-2*
Image 1	1. Man holds flag next to snowbound campsite 2. Man in red and white coat is standing in the snow	0.63	0.437
Image 2	1. Blonde haired man is doing demonstration in front of small crowd of people 2. Man with woman sit on the subway	0.143	0.000
Image 3	1. Man in fancy clothing plays guitar on stage 2. Man in red shirt and white shirt is standing in crowd of people	0.231	0.139
Image 4	1. Black and white dog is jumping in the snow at park 2. Two dogs are running through snow covered field	0.250	0.00
Image 5	1. Black and white dog chases pink frisbee 2. Dog is running on the grass	0.564	0.437

Table 1 presents a comparison of image captions generated by a natural language processing model for five different images. The table includes the following columns:

- Image Id: This column identifies each specific image that the generated captions describe.
- Generated Caption: This column contains the textual descriptions of the images generated by the NLP model.
- BLEU SC-1: BLEU SC-1 represents the BLEU score of the generated caption compared to a single reference caption. A higher BLEU score indicates a closer match between the generated caption and the reference caption in terms of n-grams (word sequences) and their frequencies.
- BLEU SC-2: Similar to BLEU SC-1, BLEU SC-2 represents the BLEU score of the generated caption, but in this case, it is compared to two reference captions. This metric provides a different perspective on the quality of the generated text.

Figure 3 are the results of image captioning for each 3 pictures. In our study, authors accomplished a noticeably excessive BLEU Score, indicating that the version has the capability to generate captions which might be linguistically in the direction of human-written captions. This suggests that our model has learned to capture the essence of the pictures and describe them efficiently in textual content. However, it is crucial to notice that whilst BLEU Score gives a beneficial quantitative measure of overall performance, it does not capture all aspects of caption fine. It in general focuses on n-gram overlap and linguistic similarity and won't completely seize the richness and creativity of human-generated captions.

For model architecture, authors have used the combination of an image processing model (CNN) and a sequence model (RNN or Transformer). BLEU Score Performance presented in figure 3 above. BLEU ratings range from 0 to at



```
Cumulative 1-gram:0.556
Cumulative 2-gram:0.264
dog is trotting through shallow stream
dog is playing with tennis ball in the water
```

```
Cumulative 1-gram:0.619
Cumulative 2-gram:0.518
crowd of people are standing in front of italian style buildings
people stand outside in front of building
```

```
* Cumulative 1-gram:0.500
  Cumulative 2-gram:0.236
  woman in yellow and black outfit is skiing
  child in red jacket and helmet is running through snow
```

Fig. 3. Result of image predicted and BLEU Score performance

least one, with higher values indicating better high-quality captions. A perfect match with the reference captions outcomes in a BLEU score of one, even as no overlap effects in a BLEU rating of 0. Typically, better BLEU scores are preferred, but the unique threshold for what constitutes a "desirable" score can range relying on the dataset and studies context.

1) *Limitations and Challenges of the usage of TensorFlow and Keras*

The challenges of the usage of TensorFlow and Our take a look at utilized TensorFlow and Keras because of the deep getting to know framework and library, respectively, for constructing and educating the caption era version. While that equipment is effective and widely used within the device learning community, they arrive with their very own set of boundaries and challenges.

- *Hardware Dependency*: Training large models with TensorFlow and Keras can be computationally intensive, requiring powerful GPUs or TPUs. This hardware dependency can be a limitation for users without access to such resources.
- *Version Compatibility*: Compatibility issues between different versions of TensorFlow, Keras, and other related libraries could be a challenge when trying to use pre-existing code or models.
- *Community Support*: While TensorFlow and Keras have large and active communities, the rapid pace of development can sometimes lead to a lack of up-to-date documentation or community support for specific issues.

2) *Future Directions*

Future studies should identify the ability of enhancements in caption generation and address numerous demanding situations. It could indicate satisfactory-tuning for particular domain names, which might include medical imaging, robotics, or artwork, to create more correct and contextually applicable captions. Multimodal techniques, combining textual content and imaginative and prescient, can enhance caption generation by way of incorporating seen information from pix and films. Ethical issues also are crucial as AI-generated content, researchers should recognize mitigating biases, ensure equity, and accountable AI use. Future guidelines must recognize area-specific best-tuning, multimodal procedures, and moral considerations to beautify caption generation and make it greater study, inclusive, and accountable.

3) *Analysis*

A comparative analysis of similar research:

Research 1 is from [13] paper, second is from [20], and the third is from [12]

TABLE 2: A COMPARATIVE ANALYSIS OF 3 SIMILAR RESEARCHES

Elements	**Comparative analysis of 3 similar researches**			
	Our research	*Research 1*	*Research 2*	*Research 3*
Main focus	Image captioning	Image captioning	Image captioning	Image caption
Methodology	BLEU Score results for Image caption.	Generating textual descriptions for images using deep learning	Deep learning and NLP for generating image descriptions	Deep learning using CNNs and RNNs
Key technique	BLEU score metrics	Encoder-Decoder	Attention model, CNN, RNN (LSTM)	CNN and RNN
Model architecture	TensorFlow and Keras	Image encoder: CNN, Decoder: RNN	CNN (VGG16 + RNN(LSTM)	CNN(VGG-19), and (LSTM)
Key metrics	BLEU score	BLEU, METEOR, CIDEr, and ROUGE score	BLEU score	NLP-related metrics (BLEU, METEOR, and CIDEr)

These research's results in table 2 all relate to image captioning, with the first three focusing on the development of image captioning models and their respective architectures and techniques. Our study result evaluates the quality of generated captions using the BLEU Score metric and highlights the importance of assessing caption quality beyond linguistic overlap.

V. CONCLUSION

In this research, performance is a comprehensive evaluation of a caption era model using the BLEU Score metric, offering treasured insights into the pleasantness of generated captions for numerous photographs. Our model exhibited steady and noteworthy overall performance across multiple pix, as evidenced via the high BLEU scores recorded in Table 1. These ratings, starting from 0 to 1, indicate the model's potential to supply captions intently aligned with human-written references. The effects underline the version's talent in know-how photograph content material, as it generated linguistically coherent and contextually applicable captions. Despite the quantitative achievement indicated by means of BLEU rankings, it's essential to know the limitations of this metric in taking pictures of the whole spectrum of caption fine. As highlighted in our discussion, the qualitative richness and creativity inherent in human-generated captions are elements that warrant further exploration and refinement.

Additionally, this takes a look at shedding light on the ethical considerations surrounding AI-generated content. As those technologies emerge as extra established, it's far vital to cognizance of mitigating biases, ensuring fairness, and promoting responsible AI use. Addressing those moral worries is vital for reinforcing the inclusivity and respectfulness of AI-generated captions. Furthermore, the study diagnosed demanding situations associated with the tools used, which include TensorFlow and Keras, emphasizing the need for non-stop improvements in supporting technology for AI research.

In end, whilst our research showcased the version's quantitative prowess thru BLEU scores, destiny efforts should focus on refining the qualitative dimensions of generated captions. By addressing ethical concerns, improving creativity, and improving linguistic richness, AI-pushed technologies can be harnessed efficiently, leading to a destiny in which human-AI collaboration is not most effectively progressive however additionally socially and ethically accountable.

References

[1] "Overview of Image Caption Generators and Its Applications | SpringerLink." Accessed: Oct. 05, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-19-0863-7_8

[2] MD. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Comput. Surv.*, vol. 51, no. 6, p. 118:1-118:36, Tháng Hai 2019, doi: 10.1145/3295748.

[3] J.-H. Huang, T.-W. Wu, and M. Worring, "Contextualized Keyword Representations for Multi-modal Retinal Image Captioning," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, in ICMR '21. New York, NY, USA: Association for Computing Machinery, Tháng Chín 2021, pp. 645–652. doi: 10.1145/3460426.3463667.

[4] S. Mishra, C. S. Minh, H. Thi Chuc, T. V. Long, and T. T. Nguyen, "Automated Robot (Car) using Artificial Intelligence," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jan. 2022, pp. 319–324. doi: 10.1109/ISMODE53584.2022.9743130.

[5] "SATMeas - Object Detection and Measurement: Canny Edge Detection Algorithm | SpringerLink." Accessed: Apr. 19, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-23504-7_7

[6] "Integrating State-of-the-Art Face Recognition and Anti-Spoofing Techniques into Enterprise Information Systems | SpringerLink." Accessed: Oct. 05, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-45140-9_7

[7] "Image Captioning for Information Generation | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10128347

[8] D. Beddiar, M. Oussalah, and S. Tapio, "Explainability for Medical Image Captioning," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Apr. 2022, pp. 1–6. doi: 10.1109/IPTA54936.2022.9784146.

[9] N. Wang *et al.*, "Efficient Image Captioning for Edge Devices." arXiv, Dec. 17, 2022. doi: 10.48550/arXiv.2212.08985.

[10] V. Atliha and D. Šešok, "Image-Captioning Model Compression," *Appl. Sci.*, vol. 12, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/app12031638.

[11] "Image Captioning Using Deep Learning | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9740788

[12] S. Chakraborty, "Captioning Image Using Deep Learning: A Novel Approach," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 6, pp. 3468–3472, Jun. 2023, doi: 10.22214/ijraset.2023.54297.

[13] A. Sen, "Captioning Image Using Deep Learning Approach," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 7425–7428, May 2023, doi: 10.22214/ijraset.2023.53389.

[14] Channasandra, Bangalore, India., N. R. U. S, M. R, and Professor, Department of Computer Science and Engineering RNS Institute of Technology, "IMAGE CAPTIONING: NOW EASILY DONE BY USING DEEP LEARNING MODELS," *Int. J. Comput. Algorithm*, vol. 12, no. 1, Jun. 2023, doi: 10.20894/IJCOA.101.012.001.001.

[15] N. Goel, A. Arora, P. Kashyap, and S. Varshney, "An Analysis of Image Captioning Models using Deep Learning," in *2023 International Conference on Disruptive Technologies (ICDT)*, May 2023, pp. 131–136. doi: 10.1109/ICDT57929.2023.10151421.

[16] "Deep Image Captioning: An Overview | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/8756821

[17] "[1906.05963] Image Captioning: Transforming Objects into Words." Accessed: Oct. 03, 2023. [Online]. Available: https://arxiv.org/abs/1906.05963

[18] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," in *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 26–36. doi: 10.18653/v1/W19-1803.

[19] "[1905.13302] A Survey on Biomedical Image Captioning." Accessed: Oct. 03, 2023. [Online]. Available: https://arxiv.org/abs/1905.13302

[20] L. Panigrahi, R. R. Panigrahi, and S. K. Chandra, "Hybrid Image Captioning Model," in *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)*, Feb. 2023, pp. 1–6. doi: 10.1109/OTCON56053.2023.10113957.

Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 111–114 DOI: 10.15439/2023R12 ISSN 2300-5963 ACSIS, Vol. 38

Unemployment Rate Future Forecasting Using Supervised Machine Learning Models

N. Vinaya
M.Tech Scholar,
Department of CSE,
CMR Institute of Technology,
Hyderabad, Telangana, India
Email:
vinaya.nareddy@gmail.com

Vijender Kumar Solanki,
L Arokia Jesu Prabhu
Department of CSE,
CMR Institute of Technology,
Hyderabad, Telangana, India
Email: spesinfo@yahoo.com,
arokiajeruprabhu@gmail.com

Sivadi Balakrishna
Department of Advanced CSE,
Vignan's Foundation for Science,
Technology, & Research
(Deemed to be University),
Telangana, India. Email:
drsivadibalakrishna@gmail.com

Abstract—**This study sees how well various models can anticipate the jobless rate. The objective of the review is to find the best model for anticipating jobless rates. There is likewise the utilization of a spiral premise neural network and learning vector quantization. While learning vector quantization and an outspread premise capability brain network are utilized together, the outcomes show that none of the other foreseeing models fill in too. It likewise involves techniques like straightforward normal and backing vector relapse as a component of a gathering to obtain significantly more exact outcomes. In our task to sort out state jobless numbers, we presently utilize the SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning methods. This product takes every one of the information from the picked state and uses the ML strategies referenced above to construct a preparation model. This model can then be utilized to anticipate joblessness for the following month or series.**

Index Terms—**Unemployment rate, SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning.**

I. Introduction

With the rapid development of the market economy, unemployment has become an increasingly major issue in today's society, and it is one of the important goals for the government to control the unemployment rate within the range. The government can ensure social stability and economic growth by using the unemployment rate prediction to implement a relevant control plan. Therefore, developing the unemployment prediction model is quite important from a practical standpoint [1].

Unemployment rate forecasting is a critical aspect of economic analysis that involves predicting the future trends and levels of unemployment within a given region or country. This process relies on a combination of quantitative models, statistical techniques, and the interpretation of various economic indicators [2]. The goal is to provide insights into the potential labor market conditions and inform decision-makers in government, business, and other sectors

The prediction results from the conventional early warning approaches, such as the time series method, regression model, support vector regression analysis model, and so forth, aren't always as expected. We are utilizing the methods of SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning to forecast the highest accurate results in this model. With the help of the previously described machine learning techniques, this application gathers all the data from the designated state and creates a training model that can be used to predict unemployment for the upcoming month or series of months [3].

II. Literature Review

A. Forecasting the Unemployment Rate by Neural Networks Using Search Engine Query Data

A neural network-based data mining technique that uses search engine query data to forecast the unemployment rate. The suggested approach mines the features of the unemployment rate time series and search engine query data using several feature selection techniques, such as the grid search algorithm, genetic algorithm, and correlation coefficient. The right neural network predictor with the right feature subset and training function is chosen after multiple neural networks with various training functions are trained and tested [4]. The empirical results showed that the GA- BPNN-Oss model outperforms the other neural network models in terms of assessment criteria when it comes to predicting the unemployment rate.

B. Predictive analysis and data mining among the employment of fresh graduate students in HEI

Better tutoring The load-up struggles with ensuring that all graduates can take care of the issues of the industry, and the industry struggles with finding capable alumni who can tackle their concerns. This is mostly because there is certainly not an effective method for testing decisive reasoning abilities, and there are defects in how decisive reasoning abilities are tried. The objective of this survey is to recommend a decent grouping model that can be utilized to give assumptions and assess the highlights of the student's dataset to satisfy the decision guidelines of work expressed by the scholarly field's graduates [5-7]. In this survey, ML estimations like K-Nearest Neighbour, Nave Bayes, Decision Tree, Neural Network, Logistic Regression, and Support Vector Machine were utilized, as well as ones that were constrained by a PC. The proposed strategy will assist school organizations with concocting better long-haul objectives for turning out graduates who are talented, and proficient, and address industry issues.

C. A Machine Learning Approach for Detecting Unemployment Using the Smart Metering Infrastructure

Changes in how power is conveyed and utilized are changing how clients and service organizations cooperate. Data assembled by savvy meters as a component of a bigger undeniable level global positioning framework could be valuable for various gatherings, like government organizations, and could likewise enable help to mind the condition of their own business all alone. Since the information is effectively open, the information is nitty gritty, and the splendid meter is continuously running, the judicious examination can be utilized to profile clients unpleasantly and exactly [8-10]. For instance, the number of individuals residing in a house, the kind of machines being utilized, or the length of stay are instances of how this should be possible. This study takes a gander at how ML models can be utilized to foresee joblessness among single- home tenants by utilizing information from brilliant meter energy gauges. The consequences of various nonlinear classifiers are checked out and contrasted with a straight model overall. We utilize normal cross-endorsements to take a look at the strength of the calculations[11]. The outcomes showed that a multi-layer perceptron cerebrum network with dropout can foresee employability status with Area Under Curve (AUC) = 74%, Responsiveness (SE) = 54%, and Particularity (SP) = 83%, firmly followed by the outcomes from a good way weighted isolation with polynomial piece model. This shows how states could utilize data assembled from a refined and broadly disseminated Internet of Things (IoT) sensor organization to offer new free administrations like inactive observing.

D. Covid-19 Pandemic And Unemployment Rate Prediction For Deploying Countries Of Asia: A Hybrid Approach.

Using unemployment data from seven developing Asian countries—Iran, Sri Lanka, Bangladesh, Pakistan, Indonesia, China, and India—this study uses an advanced hybrid modeling approach to investigate the impact of the COVID-19 pandemic on the unemployment rate in a subset of Asian nations [12]. The results are then compared with conventional modeling approaches. The results indicate that, for growing economies in Asia, the hybrid ARIMA-ARNN model performed better than its rivals. Furthermore, the unemployment rate five years ahead of time was predicted using the best-fitting model.

E. Neural Networks: A Review from a Statistical Perspective

Based on BP (Back Propagation) neural networks, a prediction model for Nanyang's unemployment rate in Henan province has been developed in this study. The MATLAB program has simulated the prediction model, and the training samples come from the data in the Nanyang statistics yearbook. The findings indicate that using a BP neural network to estimate the unemployment rate is entirely doable, and this offers some insight into future employment[13-15].

III. Methodology

The technique joins the pieces of a facial scene individually by checking Long Short Term Memory (LSTM) and profound drawings of face characteristics. In the plan, both request and straight return are utilized. The recommended technique was superior to the example model as far as endorsement and test set. Takashi and Melanie Swan investigated how individual hereditary informatics and machine learning (ML) can be utilized to all the more likely comprehend satisfaction and abundance studies.

Previously, it was difficult to figure the joblessness number. It is made in an extreme manner. The projected joblessness rate could go up by up to 30% more assuming another monetary model is utilized. It is significant for legislators, monetary specialists, and the business local area to have a smart thought of the joblessness rate since it shows where the economy is solid and the way that the financial cycle is going.

Disadvantages:

- It requires a great deal of investment and is hard to do.
- The journalists observed that joy is better perceived as a "major information issue."

It utilizes the portrayal strategy, and the model that emerges from it can take a gander at specific pieces of the enlightening list that can be utilized to sort out whether or not an alumnus is working, getting more instruction, working on their abilities, hanging tight for a task, or is jobless. In our venture to sort out state jobless numbers, we currently utilize the SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning techniques. This product takes each of the information from the picked state and uses the ML strategies referenced above to fabricate a preparation model. This model can then be utilized to anticipate joblessness for the following month or series.

Advantages:

- Utilizing ML, we can without much of a stretch speculate the jobless rate.

A. Application flow related work:-

1) SVM:

SVM is a type of machine learning named "directed machine learning" that maybe secondhand for both arrangement and relapse. No matter what we call ruling class, they are better organized. The SVM procedure is used to find a hyperplane in an N-hide scope that puts the facts centers in a clear order. SVMs are used to recognize writing, label interruptions, recognize faces, sort emails, group characteristic, and constitute pages. In ML, SVMs are secondhand by way of this. Characterization and return maybe finished on both undeviating and nonlinear dossier.

$$\begin{gathered}\min w,\xi,b\ \{\ 1\ 2\ \|w\|\ 2 + C \sum \xi i\ n\ i{=}1\ \},\qquad \text{s.t.}\\ \forall i{=}1\ \ : yi(w{\rightarrow}{\rightarrow}\ Tx{\rightarrow}i + b) \geq 1 - \xi i\ ;\ \forall i{=}1\ n : \xi i > 0\end{gathered} \quad (1)$$

2) Random forest:

The Random Forest Method is a type of directed machine learning namely frequently secondhand in machine learning to resolve questions accompanying arrangement and return. We see that a jungle contains many trees, the more trees skilled are, the more active the thicket is. Information experts use random forests in many various fields, in the way that investment, stock business, dispassionate study, and netting-located trade. It's used to resolve belongings like consumer behavior, patient education, and strength, which helps these trades run flatly.

$$Y\hat{}(x)=N1\sum i=1\ N\ hi(x) \qquad (2)$$

Where, N is the number of trees in the Random Forest, hi(x) is the forecast of the i-th decision tree for input x, and Y^(x) is the expected output for the input x.

3) *Gradient boosting:*

Gradient boosting is a type of machine learning namely frequently secondhand in apps for relapse and order. It restores an anticipation model as a group of feeble forecast models, that are mostly decision trees. When a choice sapling is secondhand as the feeble undergraduate, the method is named "gradient-boosted trees," and it frequently beats "uneven forest." A gradient-boosted trees model is innate the alike gradual habit as different habits of helping, but it expands on additional habits by admitting relaxing of some various deficit skill.

$$F(x)=\sum m=1M\ \alpha m\ hm(x) \qquad (3)$$

Where, hm (x) is the prediction of the m-th weak learner for input x, F(x) is the anticipated output for the input x, M is the number of weak learners (trees) in the ensemble, and αm is the learning rate (or shrinkage factor) for the m-th weak learner.

4) *Extreme machine learning:*

An extreme learning machine (ELM) is a method for fitting a single hidden layer feedforward neural network (SLFN) that everything a lot faster than projected arrangements and produces good results. The extreme learning machine (ELM) is frequently secondhand in cluster learning, successive knowledge, and stable knowledge cause it can discover fast and well, is fast, has fields of substance for congregation competency, and is easy to use.

$$Y\hat{}(x) =\sum i=1N\ \beta i\cdot g(wi\cdot x+bi) \qquad (4)$$

Where N is the number of hidden neurons or nodes; βi are the output layer weights; wi are the input layer weights; bi are the biases for each hidden neuron; and g(·) is the activation function, which in the context of ELM is typically a sigmoid or radial basis function (RBF).

B. *The flow of Prediction*

In this part, the client adds a rundown of individuals who are unemployed.

- Extract Selected State Data

This program peruses just the state records from the dataset and utilizes them to make ML models and charts.

- Run SVM Algorithm

In this illustration, we use information to prepare the model and use SVM to anticipate the rate.

- Run Random Forest

In this part, information are utilized to prepare the model and utilize Random Forest to figure the rate.

- Run Gradient Boosting

In this example, information is utilized to prepare the model and use Gradient Boosting to anticipate the rate.

- Run Extreme Machine Learning

In this example, you'll figure out how to utilize Extreme Machine Learning to prepare a model and foresee rates.

- Comparison Graph

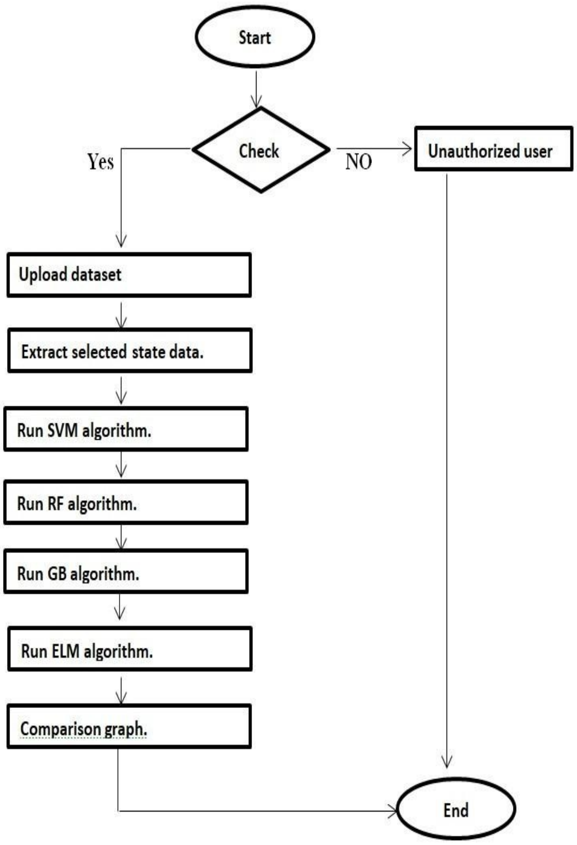


Fig.1: The flow chart

This part shows a diagram that looks at two things.

- Exit

This segment will destroy the application interaction.

IV. Experimental Results and Analysis

This section deals with the results obtained by various machine learning classifiers over the state graph dataset. These algorithms have been implemented on Jupiter notebook, python 3.2, windows 11, 8GB RAM, 500 GB SSD, and i5 processor.

Fig.2 shows the extracted selected state graph to anticipate joblessness for the following month or series.

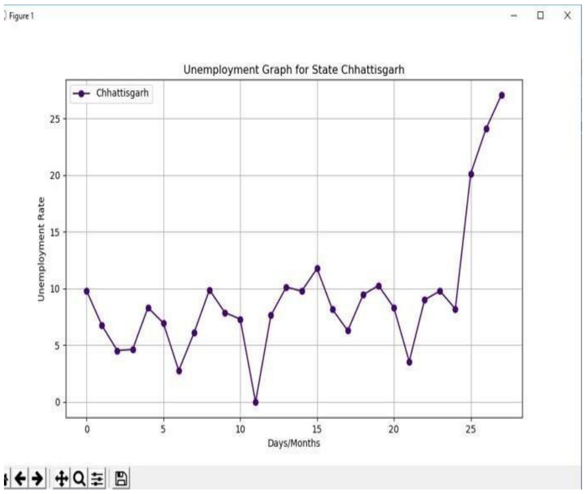


Fig.2: Extract selected state data graph

Fig.3 deliberates the comparative results obtained over the machine learning models.

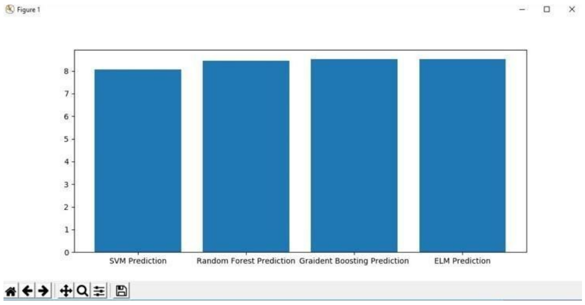


Fig.3: Comparison graph

V. Conclusion

We involved a prepared ML model to anticipate joblessness for the following month or series. In our undertaking to sort out state jobless numbers, we currently utilize the SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning strategies. This product takes every one of the information from the picked state and uses the above ML techniques to fabricate a preparation model that can then be utilized to foresee joblessness.

References

[1] Wei Xu School of Information, Renmin University of China, “Forecasting the Unemployment Rate by Neural Networks Using Search Engine Query Data”, [Online; accessed 2012. https://ieeexplore.ieee.org/document/6149133/

[2] Yiyuan Cheng:Tao Hai:Yangbing Zheng:Baolei Li, “Prediction Model of the Unemployment rate for nanyang in henan province based on BP neural network”, [Online; accessed 2017. [Online]. Available https://ieeexplore.ieee.org/document/8392903/

[3] R. Barnichon and C. J. Nekarda, “The ins and outs of forecasting unemployment: Using labor force flows to forecast the labor market,” Brookings Papers on Economic Activity, Oct 2012. [Online]. Available: http://www.brookings.edu/∼/media/Projects/BPEA/Fall-2012/2012bBarnichon.pdf?lang=en

[4] Wikipedia, “Okun’s law — wikipedia, the free encyclopedia,” 2015, [Online; accessed 3-June- 2015].

[5] G. Zhang, B. E. Patuwo, and M. Y. Hu, “Forecasting with artificial neural networks:: The state of the art,” International Journal of Forecasting, vol. 14, no. 1, pp. 35 – 62, 1998.

[6] S.-C. Huang, N.-Y. Wang, T.-Y. Li, Y.-C. Lee, L.-F. Chang, and T.-H. Pan, “Financial forecasting by modified kalman filters and kernel machines,” Journal of Statistics and Management Systems, vol. 16, no. 2-03, pp. 163–176, 2013.

[7] F. zkan, “Comparing the forecasting performance of neural network and purchasing power parity: The case of turkey,” Economic Modelling, vol. 31, 2013.

[8] M. Fernandes, M. C. Medeiros, and M. Scharth, “Modeling and predicting the CBOE market volatility index,” Journal of Banking and Finance, vol. 40, 2014.

[9] G. Szpiro, “A search for hidden relationships: Data mining with genetic algorithms,” Computational Economics, vol. 10, no. 3, 1997.

[10] A. L. Montgomery, V. Zarnowitz, R. S. Tsay, and G. C. Tiao, “Forecasting the u.s. unemployment rate,” Journal of the American Statistical Association, 1998. [Online]. Available: http://www.tandfonline.-com/doi/abs/10.1080/01621459. 1998.10473696

[11] P. Rothman, “Forecasting asymmetric unemployment rates,” Review of Economics and Statistics, 1998.

[12] K. G and P. SM, “Dynamic asymmetries in u.s. unemployment,” Journal of Business and Economic Statistics, 1999.

[13] J. Skalin and T. Tersvirta, “Modeling asymmetries and moving equilibria in unemployment rates,” Macroeconomic Dynamics, vol. 6, 2002.

[14] F. Liang, “Bayesian neural networks for nonlinear time series forecasting,” Statistics and Computing, vol. 15, no. 1, pp. 13–29, 2005.

[15] E. Olmedo, “Forecasting spanish unemployment using near neighbor and neural net techniques,” Computational Economics, vol. 43, no. 2, pp. 183– 197, 2014.

DOI: 10.15439/2023R75
ISSN 2300-5963 ACSIS, Vol. 38

Utilizing Flex Sensors for the Evaluation of Parkinson's Disease

Quan Vu
Department of Biomedical Engineering Le Quy Don Technical University
Ha Noi, Viet Nam
quanvu42@lqdtu.edu.vn

To-Hieu Dao
Faculty of Electrical and Electronic Engineering
Phenikaa University
Ha Noi, Viet Nam
hieu.daoto@phenikaa-uni.edu.vn

Manh-Cuong Nguyen
Department of Biomedical Engineering Le Quy Don Technical University
Ha Noi, Viet Nam
nmcuong@mta.edu.vn

Duc-Tan Tran*
Faculty of Electrical and Electronic Engineering
Phenikaa University
Ha Noi, Viet Nam
tan.tranduc@phenikaa-uni.edu.vn

***Abstract*—Parkinson's disease is a neurodegenerative disorder with symptoms such as tremors, stiffness, and issues with balance and coordination. Detecting and monitoring the signs of the disease is of great significance. By combining flex sensors and Arduino, we have designed a simple and effective system capable of recording, detecting, and evaluating early signs of the disease. The electronic components used are: an Arduino Nano, two flex sensors, and a glove with sensors attached to each finger to capture movements and flexion. The patient wears the glove and whenever tremors are detected, the sensor sends a signal to the Arduino which is converted into an angle of flexion by changing resistance. The tremor signals are initially transmitted as resistance and subsequently transformed into voltage. This voltage is then graphed according to the sensor's bending angle. By analyzing abrupt and rapid tremors, a threshold is established to deduce the severity and progression stage of the illness.**

***Index Terms*—Flex Sensor, Glove, Microcontroller, Parkinson’s Disease.**

I. Introduction

Parkinson's disease is a gradual deterioration of the nervous system resulting in involuntary and unmanageable movements, including tremors, rigidity, and challenges with balance and coordination. It ranks as the second most prevalent neurodegenerative condition, impacting 2-3% of individuals aged 65 and above [2]. Symptoms typically start gradually and worsen over time. The impact of this condition can be minimized through early diagnosis and appropriate symptom monitoring, improving the management and medical treatment, and enhancing the quality of life for patients. However, detecting Parkinson's disease in its early stages often presents many challenges [26].

The tremor manifestations observed in Parkinson's disease may impact a specific region of the body (rest tremor), multiple adjacent areas (segmental tremor), one side of the body (hemitremor), or the entire body (generalized tremor). Typically, two categories of tremors are recognized based on the circumstances that elicit them: resting tremor, occurring when the affected body part is at rest, and action tremor (which includes kinetic, postural, or isometric tremor), arising when the individual initiates voluntary movements or sustains a particular posture against gravitational forces. Tremor characteristics include frequency (typically between 4-8 Hz) and amplitude. In theory, EMG is a gold standard for assessing and monitoring tremors. However, the drawback of EMG is its lack of suitability for continuous monitoring and frequent evaluation of tremor features. With technological advancements, various solutions have been proposed to enable continuous monitoring of disease symptoms and important signs. Published studies are usually focused on wearable devices [19], to provide patients with the ability to perform activities of daily living, and therefore to analyze the extent of their disease without the supervision of a doctor in a laboratory, It can be classified into two main groups: (1) devices that assess tremor features, and (2) devices that monitor tremors and the effectiveness of therapies.

Equipment designed to assess and delineate tremor attributes have demonstrated notable efficacy in clinical settings, particularly in the evaluation, diagnosis, and management of tremors associated with Parkinson's disease (PD). Subsequently, alternative wrist or forearm devices were introduced to evaluate rest and action tremors (postural and movement) in individuals with Parkinson's disease (PD) using inertial measurement units (IMUs), which include three-axis accelerometers and three-axis gyroscopes, or a combination of four three-axis accelerometers. Hssayeni et al. [16] utilized an IMU to assess tremor severity in PD, while Mahadevan et al. [22] and Dai et al. [5] employed it to distinguish between bradykinesia and tremors. Shawen et al. [31] compared the efficacy of smartwatches and skin-mounted IMUs in categorizing tremor and bradykinesia severity in PD, demonstrating that smartwatches can perform comparably to specialized sensors. Additionally, Huo et al. [15] introduced a more intricate array of devices comprising force sensors, three IMUs, and four custom mechanical sensors (MMG). These wearable solutions can accurately estimate tremor frequency and amplitude, laying the groundwork for the advancement of more sophisticated diagnostic and therapeutic monitoring devices for tremor disorders.

Continuous monitoring of tremors and evaluating the effectiveness of such monitoring are essential requirements for home healthcare solutions and intelligent services aiming to

* Corresponding author

alleviate the strain on the National Healthcare System. Battista and Romaniello [3] introduced and validated a device akin to a smartwatch, utilizing a three-axis accelerometer, capable of identifying tremor events by computing statistical indices indicative of motion patterns. San-Segundo and Luis Sigcha [32],[20] employed a wrist-worn accelerometer paired with a smartphone annotation app to gather labeled data in controlled laboratory settings and weakly labeled data during everyday activities. Various AI models have been applied to discern tremor presence and severity from diverse feature sets [10]. However, a key challenge lies in accurately distinguishing tremors from other motions or objects due to the substantial signal variability inherent in normal daily activities. These monitoring devices exhibit improved performance when complemented with self-annotation. Nonetheless, achieving high accuracy in identifying tremors and evaluating their severity during continuous, fully automated monitoring remains an ongoing challenge.

The majority of the proposed systems utilize inertial measurement devices (IMU), including accelerometers, gyroscopes, and surface electromyography, either independently or in conjunction in certain wearable configurations [9]. Additionally, speech analysis [1, 28] serves as another significant diagnostic tool. A novel postural estimation method, integrating Transformers with HRNe, employing machine learning techniques, has been suggested [4]. Furthermore, in [8], the authors introduced a device for tracking motion trajectory and tremor occurrences. This study proposes the utilization of a magnetic tracking system to capture data on translational movements and vibrations within a spatial cubic domain.

With the two device groups mentioned above, we found that the design and use of sensor components are complex. Therefore, in this study, we used a smart glove designed as follows: a glove with two fixed bend sensors on two fingers, used to detect abnormal movements related to motor disorders and record detections [27]. The vibrations in the fingers [29] cause changes in the shape and curvature of the sensors, resulting in resistance changes. These changes are recorded and converted into voltage [12]. Then, the voltage is sent to the Arduino input and plotted against angles. The values obtained from the glove can be used for both detection purposes and for patients recovering function [26], [34].

When designing gloves, it is important that they are both aesthetically pleasing and easy to use, allowing for easy and simple manipulation [14], [19]. During the preliminary phase, the suggested system will undergo simulation using the Proteus Simulation Tool. The system comprises two primary modules: hardware and software. The hardware component encompasses an Arduino Nano, flex sensor, amplifier module, accelerometer, and Wi-Fi. Integration of both software and hardware is achieved through the utilization of embedded C language.

II. Method

A. Glove

The study used gloves made from synthetic fibers. The gloves were cut to remove three fingers, leaving only the index and middle fingers. flex sensors were placed on the outer surface of the glove, above the two fingers, and sewn into it to ensure accurate reading results [6], [24]. The amplification circuit and Arduino are fixed on the back of the hand. Below is the amplification circuit, and above is the Arduino. The output signal of the Arduino is stored in a micro SD card and transmitted to a laptop port via a cable. This is designed to minimize any potential hardware interference with the dexterity and natural movement of the user's hand. Therefore, whenever a finger vibrates or moves, it leads to a change in the resistance of the sensor.

The structure of the glove is shown above. The components will be connected as follows. One end of the sensor is grounded, while the other end is connected to a 3.3V source through a 10 kΩ pull-up resistor as shown in Fig. 1. The low voltage on the sensor is the input to the amplification circuit. The output of the amplification circuit is sent to the Arduino. After being processed on Arduino, the signal will be saved to the micro SD card and displayed on the laptop through the IDE interface.

This design enables the recording of any changes in resistance in response to alterations in the curvature or shape of the sensor [12]. The variation in resistance results in a shift in the voltage drop across the sensor. This voltage alteration is then routed through an amplification circuit, which serves as a receiver circuit, and the output voltage values are delivered to Arduino pins 5 and 7 [7]. Consequently, any movement within the glove's finger leads to a voltage change that is transmitted to the Arduino.

1) *Flex sensor*

The Flex sensor is based on carbon resistive elements and is a type of sensor used to measure bending or deflection. When the sensor is deformed or bent, the resistance of the carbon material changes, and this value can be read [13], [34], [24], [17]. The resistance in this device changes linearly with the bending angle. Therefore, the flex sensor can be used as an input signal generator for devices. The required voltage to detect normal values falls within the range of a DC current of 3.3-5V. The observed resistance change is greater when the bending angle is larger [11].

In reality, Flex sensors often produce different results, so they need to be calibrated before use. To determine the relationship between the curvature of the sensor and the change in resistance value, a test was set up. Successive sensors had flat resistance values of 12.5 kΩ and 11.8 kΩ. The resistance values were 63.51 kΩ and 56.8 kΩ when bent at 180°. When the Flex sensor is bent, the resistance increases linearly, resulting in a corresponding change in output voltage. By describing this relationship, we can use the flexible sensor to determine the movement of an individual's finger in different hand motions. Two flex sensors were used in this study, with each sensor having a linear resistance value but different flat resistance values.

2) *Signal amplifier circuit*

The voltage divider circuit is utilized to convert angle measurements into voltage. The bending of the flex sensor causes a change in resistance, resulting in a corresponding change in voltage output. Arduino board provides a 3.3V power supply voltage through a pull-up resistor (R_M), which is then connected to the nominal resistance (R_{flex}) of the flex sensor. One end of the flex sensor is grounded, and the output voltage (Vout) is measured across it. We can determine the relationship between the voltage on the flex sensor and the output voltage by using the voltage division formula [7]. The output voltage measured on the flex sensor is determined by the following formula (1).

$$V_{in} = \frac{R_{flex}}{R_M + R_{flex}} \times Vcc \qquad (1)$$

In Fig. 1, when the flex sensor is flat, the output voltage Vin is 1.75V when Vcc = 3.3V, $R_M = R_{flex} = R$. To determine the output voltage level, the nominal value of the sensor can be calculated using formula (1). From this, it can be observed that the greater the output voltage, the more the flex sensor bends.

To ensure a stable input signal to the Arduino, we employ an additional module that functions as an amplifier and buffer before transmitting the signal to an Arduino pin. The output voltage will be converted to a digital signal by sending a value to the analog pin of the Arduino. In Fig. 1, the relationship between input and output will be calculated using the following formula (2):

$$Vout = (1 + \frac{R_1}{R_2 + R_3}) \times Vin \qquad (2)$$

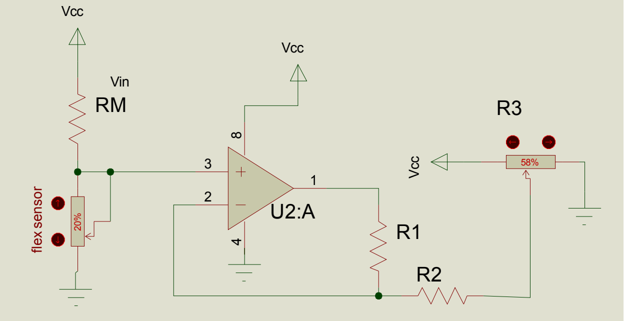


Fig. 1 Signal amplifier circuit.

In addition, we have also designed a module to store data on a micro SD card. This data will be convenient for analysis and evaluation. The gloves are designed with the described components and will function according to the diagram shown in Fig. 2.

B. *Experimental Procedure*

The movement of the hand is a complex motion that encompasses the wrist, hand, and fingers. Various quantities describe this motion, as illustrated in Fig. 3. In Fig. 3-a, α represents the rotation angle of the hand, 3a ,β is the angle of motion of the wrist, and 3c θ_j^i is the angle of motion of the finger relative to the hand, where i has symbols from 1 to 5 corresponding to the fingers, j is the angle of the knuckles from 1 to 3, respectively.

The first step in study, our focus is solely on the finger's motion (θ_3^i angle) relative to the hand. This motion is considered a change in angle, as depicted in Fig. 3c.

The data collection process is as follows: The test subject wears gloves and performs movements with their hand and individual fingers. The resulting output is interpreted as changes in voltage and recorded. Each sensor has a different initial value based on the stiffness of the material and the angle of deviation, so not all initial voltage values are the same.

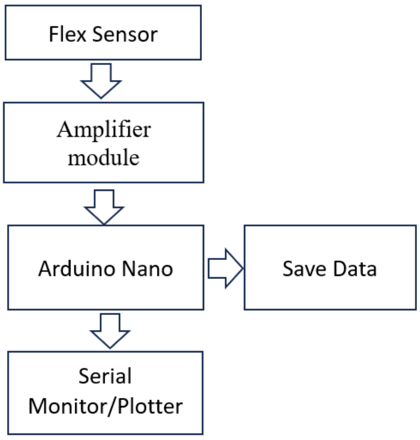


Fig. 2 Flowchart depicting workflow.

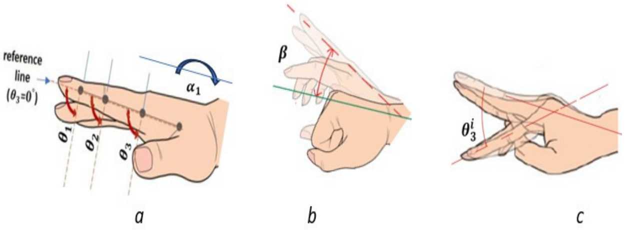


Fig.3 a- The angle of movement of the hand; b- the angle of motion of the hand relative to the forearm; c- the angle of movement of the finger relative to the hand, with i from Eq. 1-2.

The initial position is defined as follows: the gloved hand is placed on a flat surface, with the fingers extended straight and parallel to the hand at an angle (Fig. 4). To observe the glove's response, perform the following actions: gradually bend and tighten the fingers to 180°, then straighten the fingers back to their initial position. The voltage output should fall within the range of 2.2-3.3V. The hand should be held steady with the fingers straightened, gradually bent, and tightened, as shown in Fig. 3b. These movements will determine the output limits of the corresponding signal value. The output signal is then adjusted to fall within the desired range. Multiple repeated measurements are conducted to analyze the sensitivity of the flex sensor [11]. Collect data with the following motions.

- Collect data with a small angle $\theta_3^i < 45^o$: keep your hand steady, move your fingers with different amplitudes and frequencies.
- Collect data with a large angle ($45^o < \theta_3^i < 90^o$) keep your hand steady, vibrate your fingers with a small amplitude and frequency.

The voltage graph recorded from the index finger and middle finger is approximately 2.25V in the initial established state $\theta_3^i = 41^o$. In this experiment, the angles (θ_3^i) were primarily measured from 0° to 90°. The chart illustrates the voltage change when the sensor is bent, as shown in Fig. 6. To test the sensitivity of the device, record the voltage change when the finger vibrates at different speeds, ranging from low to high.

The second step, register the signal of hand vibration with α≠ 0. In this position the hand will place parallel to the tabletop, as depicted in Figure 3c.

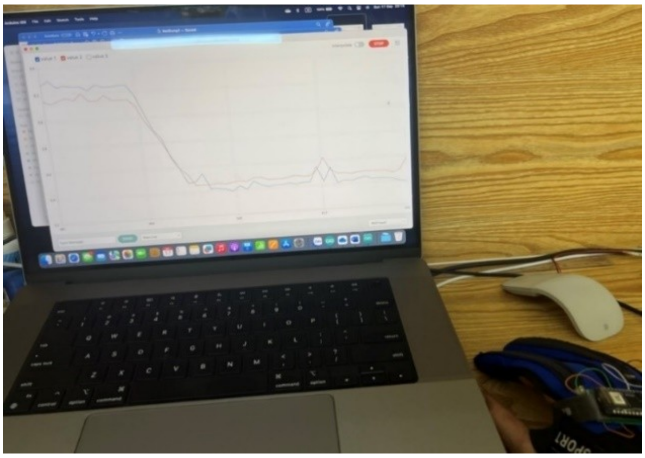


Fig.4 Shows the initial state of the test.

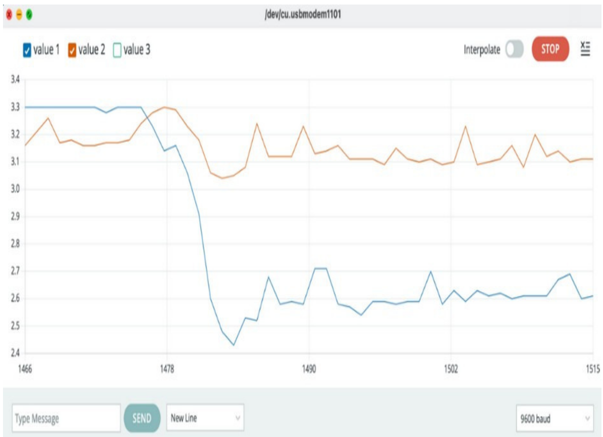


Fig. 5 Demonstrates the voltage changes as the finger moves at different angles.

III. Results and Discussion

The change in voltage over time is depicted in the graph below (Fig. 6). When the finger is held in a fixed position, the voltage change is insignificant. For example, θ_3^1 : 20^o, θ_3^2 : 90^o as shown in Fig. 5, the voltage remains relatively constant. However, when gripping, there is a peak on the chart, with the voltage reaching a maximum value of 3.3V. In Fig. 6, Series 1 and Series 2 represent data collected from the index finger and middle finger, respectively. Small, insignificant changes in the graph indicate minor alterations in finger movement. The voltage is measured based on angles ranging from 0° (flat surface) to approximately 120° (excluding the first fully tightened position). The angles increase gradually from 0° to 90°. This is done to test sensitivity, accuracy, and perform calibration to assess the device's stability. The device's stability enables us to detect subtle vibrations on the finger. The graphs below illustrate the relationship between voltage and flex angle. In Figs. 7 and 8, the graphs display voltage variations when the flex angle is less than 45° and greater than 45°, respectively.

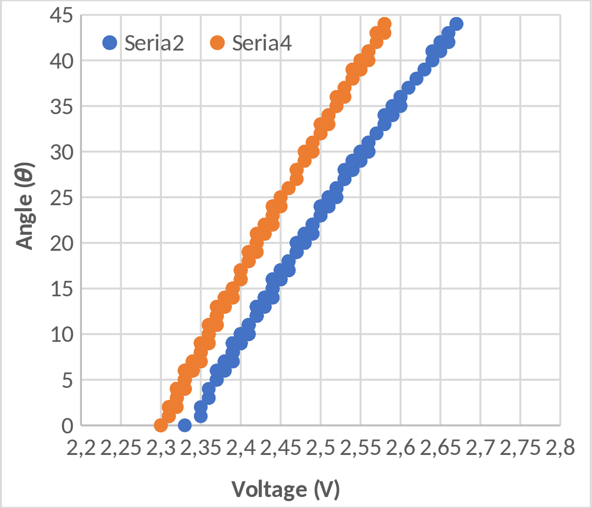


Fig. 6 Shows the relationship between voltage and flex angle

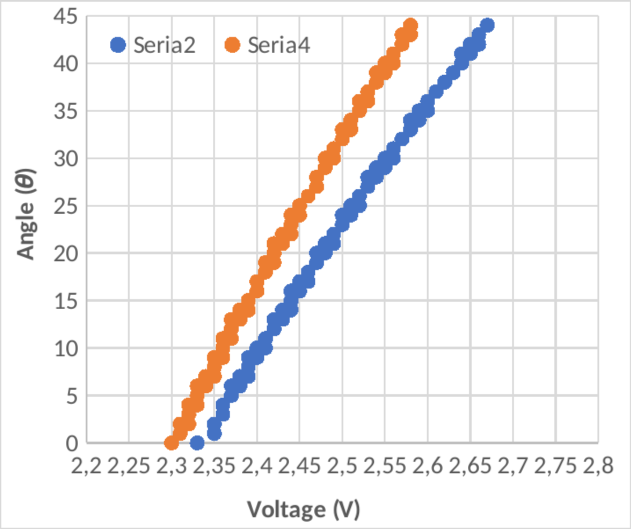


Fig. 7 Voltage and angle of lex when θ_3^i <45°.

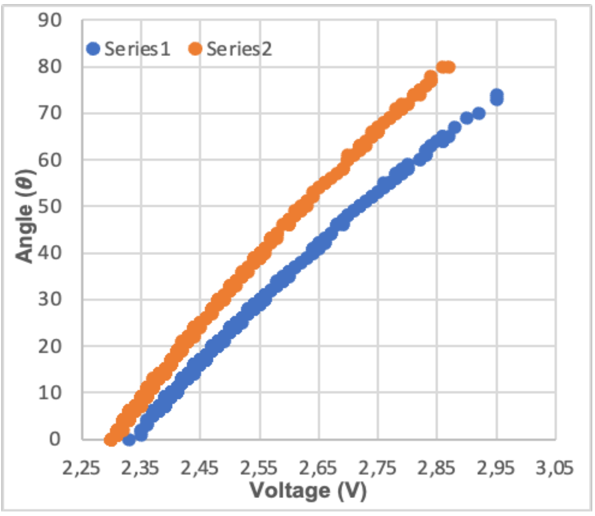


Fig. 8 Voltage and angle of lex when θ_3^i <90°

Using the formula, we will calculate the resistance value R3 for calibration with voltage input into Arduino. It can be observed from the graph above that the voltage changes linearly with the angle. This demonstrates that when the fist is clenched, the voltage undergoes significant changes in small increments. This device model can be utilized for individuals with Parkinson's disease. Additionally, the device is capable of detecting and evaluating the early stages of Parkinson's disease.

IV. Conclusion

Parkinson's disease is a condition of the nervous system that leads to involuntary or uncontrollable movements, including tremors, stiffness, and issues with balance and coordination. Typically, symptoms start off mild and worsen over time. As the condition progresses, people may have difficulty speaking and moving. This study primarily focuses on Parkinson's patients who are currently affected by the disease or experiencing early symptoms. The goal is to develop a reliable testing and diagnostic method that allows us to classify the signs and symptoms of the disease. The study also aims to make automatic monitoring and severity assessment easier. The flex sensors and Arduino Nano controller have been successfully integrated into the electronic glove system. Each component works well and provides accurate data from the flex sensors. This system has the advantages of affordability, minimal parts with small size, quick response, reliability, and easy management. There was a slight variation in angles during the repeatability testing process, but the project's objective has been achieved. This is demonstrated by the linear transformation of voltage and resistance when detecting vibrations using 2 flex sensors. This model can be used to detect early signs as well as monitor and assess the condition of Parkinson's disease. This study has focused on developing an electronic glove system that effectively detects and monitors the symptoms of Parkinson's disease, with a particular emphasis on early detection. The successful integration of the flex sensor into the wearable device has delivered a cost-effective, compact, and reliable solution for this purpose. In the future, we tend to develop embedded algorithms on wearable devices [6], [18], [21] [23], [25], [35], [36] integrated with constrained-performance microcontrollers to analyze the complex variations of Parkinson's patients in various postures and intricate actions in daily life.

References

[1] Almeida, J. S.; Filho, P. P. R.; Carneiro, T.; Wei, W.; Damaševičius, R.; Maskeliunas, R.; de Albuquerque, V.H.C. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognit. Lett.* 2019, *125*, 55–62.

[2] B I. Y. Abdi, S. S. Ghanem, and O. M. El-Agnaf, "Immune-related biomarkers for Parkinson's disease," Neurobiol Dis, (2022) vol. 170, p. 105771.

[3] Battista L,Romaniello A. A wearable tool for selective and continuous monitoring of tremor and dyskinesia in Parkinsonian patients. *Parkinsonism Relat Disord.* (2020) 77:43–7. doi: 10.1016/j.parkreldis.2020.06.020

[4] Chenbin Ma, Lishuang Guo , Longsheng Pan , Xuemei Li , Chunyu Yi n , Rui Zong , Zhengbo Zhang " Tremor detection Transformer: An automatic symptom assessment framework based on refined whole-body pose estimation" Elsevier: Amsterdam, The Netherlands, 2020.

[5] Dai H, Cai G, Lin Z, Wang Z, Ye Q. Validation of inertial sensing-based wearable device for tremor and bradykinesia quantification. *IEEE J Biomed Health Inform*. (2020) 25:997–1005. doi: 10.1109/JBHI.2020.3009319

[6] Duc Hung Pham, Viet-Ngu Nguyen, Thi Minh -Le," Fuzzy Brain Emotional Controller for Heart Disease Diagnosis" Proceedings of the 2022 Seventh International Conference on Research in Intelligent and Computing in Engineering Annals of Computer Science and Information Systems, Volume 33

[7] F. Salman, Y. Cui, Z. Imran, F. Liu, L. Wang, and W. Wu, "A Wireless-controlled 3D printed Robotic Hand Motion System with Flex Force Sensors," Sens Actuators A Phys, (2020) vol. 309, p. 112004

[8] Filippo Milano, Gianni Cerro, Francesco Santoni,Alessio De Angelis,Gianfranco Miele, And Paolo Carbone, "Parkinson's Disease Patient Monitoring: A Real-Time Tracking and Tremor Detection System Based on Magnetic Measurements"

[9] Huo W, Angeles P, Tai YF, Pavese N, Wilson S, Hu MT, et al. A heterogeneous sensing suite for multisymptom quantification of Parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng*. (2020) 28:1397–406. doi: 10.1109/TNSRE.2020.2978197

[10] Hssayeni MD, Jimenez-Shahed J, Burack MA, Ghoraani B. Wearable sensors for estimation of Parkinsonian tremor severity during free body movements. *Sensors (Basel).* (2019) 19:4215. doi: 10.3390/s19194215

[11] G. Saggio, "A novel array of flex sensors for a goniometric glove," Sens Actuators A Phys, (2014) vol. 205, pp. 119–125.

[12] G. Saggio and G. Orengo, "Flex sensor characterization against shape and curvature changes," Sens Actuators A Phys, (2018) vol. 273, pp. 221–231.

[13] G. Saggio, "Mechanical model of flex sensors used to sense finger movements," Sens Actuators A Phys, (2012) vol. 185, pp. 53–58.

[14] HoudeDai ;Pengyue Zhang and Tim C. Lueth; Quantitative Assessment of Parkinsonian Tremor Based on an Inertial Measurement Unit; Published: 29 September 2015.

[15] Huo W, Angeles P, Tai YF, Pavese N, Wilson S, Hu MT, et al. A heterogeneous sensing suite for multisymptom quantification of Parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng*. (2020) 28:1397–406. doi: 10.1109/TNSRE.2020.2978197

[16] Hssayeni MD, Jimenez-Shahed J, Burack MA, Ghoraani B. Wearable sensors for estimation of Parkinsonian tremor severity during free body movements. *Sensors (Basel).* (2019) 19:4215. doi: 10.3390/s19194215

[17] K. Elgeneidy, N. Lohse, and M. Jackson, "Data-Driven Bending Angle Prediction of Soft Pneumatic Actuators with Embedded Flex Sensors," IFAC-PapersOnLine, (2016) vol. 49, no. 21, pp. 513– 520

[18] K. T. Lee, P. S. Chee, E. H. Lim, and C. C. Lim, "Artificial intelligence (AI)-driven smart glove for object recognition application," Mater Today Proc, (2022) vol. 64, pp. 1563–1568.

[19] Lu R, Xu Y, Li X, Fan Y, Zeng W, Tan Y, Ren K, Chen W, Cao X. Evaluation of wearable sensor devices in Parkinson's disease: a review of current status and future prospects. *Parkinsons Dis.* (2020) 2020:4693019. doi: 10.1155/2020/4693019

[20] Luis Sigcha, Ignacio Pavón,Nélson Costa,Susana Costa Miguel Gago,Pedro Arezes, Juan Manuel López, and Guillermo De Arcas· "Automatic Resting Tremor Assessment in Parkinson's Disease Using Smartwatches and Multitask Convolutional Neural Networks" *Sensors* 2021, *21*(1), 291

[21] M. C. Fennema, R. A. Bloomfield, B. A. Lanting, T. B. Birmingham, and M. G. Teeter, "Repeatability of measuring knee flexion angles with wearable inertial sensors, (2019)" Knee, vol. 26, no. 1, pp. 97–105.

[22] Mahadevan N, Demanuele C, Zhang H, Volfson D, Ho B, Erb MK, et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *NPJ Digit Med.* (2020) 3:5. doi: 10.1038/s41746-019-0217-7

[23] N. Tran Thi Hong, G. L. Nguyen, N. Quang Huy, D. Viet Manh, D.-N. Tran, and D.-T. Tran, "A low-cost real-time IoT human activity recognition system based on wearable sensor and the supervised learning algorithms," Measurement, vol. 218, p. 113231, 2023.

[24] Luu, M. H., Tran, D. T., Nguyen, T. L., Nguyen, D. D., & Nguyen, P. T. Errors determination of the MEMS IMU. Journal of Science, Vietnam National University - Hanoi, (2006) vol. 22, pp. 6-14.

[25] Pham Minh Chuan, Luong Thi Hong Lan, Tran Manh Tuan, Nguyen Hong Tan, Cu Kim Long, Pham Van Hai, Le Hoang Son, " Chronic kidney disease diagnosis using Fuzzy Knowledge Graph Pairs-based inference in the extreme case" Proceedings of the 2022 Seventh International Conference on Research in Intelligent and Computing in En-

gineering Annals of Computer Science and Information Systems, Volume 33

[26] S. Hawi, J. Alhozami, R. AlQahtani, D. AlSafran, M. Alqarni, and L. el Sahmarany, "Automatic Parkinson's disease detection based on the combination of long-term acoustic features and Melfrequency cepstral coefficients (MFCC)," Biomed Signal Process Control, (2022) vol. 78, p. 104013.

[27] Rubén San-Segundo, Ada Zhang, Alexander Cebulla,.. Jessica Hodgins,"Parkinson's Disease Tremor Detection in the Wild Using Wearable Accelerometers" Sensors 2020, 20, 5817; doi:10.3390/s2020581.

[28] Rahman, A.; Rizvi, S.S.; Khan, A.; Abbasi, A.A.; Khan, S.U.; Chung, T.-S. Parkinson's Disease Diagnosis in Cepstral Domain Using MFCC and Dimensionality Reduction with SVM Classifier. *Mob. Inf. Syst.* 2021, *2021*, 1–10.

[29] S. I. Lee, J.-F. Daneault, L. Weydert, and P. Bonato, "A novel flexible wearable sensor for estimating joint-angles," in 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks, (2016) pp. 377–382.

[30] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," J Neuroeng Rehabil, (2012) vol. 9, no. 1, p. 21.

[31] Shawen N, O'Brien MK, Venkatesan S, Lonini L, Simuni T, Hamilton JL, et al. Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. *J Neuroeng Rehabil*. (2020) 17:52. doi: 10.1186/s12984-020-00684-4

[32] San-Segundo R, Zhang A, Cebulla A, Panev S, Tabor G, Stebbins K, et al. Parkinson's disease tremor detection in the wild using wearable accelerometers. *Sensors (Basel)*. (2020) 20:5817. doi: 10.3390/s20205817

[33] S. Huang et al., "Development and evaluation of a novel flex sensor-based glenohumeral subluxation degree assessment for wearable shoulder sling," Sens Actuators A Phys, (2022) vol. 337, p. 113405

[34] Tiboni, M.; Amici, C. Soft Gloves: A Review on Recent Developments in Actuation, Sensing, Control and Applications. Actuators 2022, 11, 232. https://doi.org/ 10.3390/act11080232.

[35] T.-H. Dao, D.-N. Tran, Q.-T. Hoang, H.-D. Vu, D. T. Huy, and D.-T. Tran, "Developing Real-time Automatic Step Detection On A Low-Cost, Performance-Constrained Microcontroller," in *2023 IEEE Statistical Signal Processing Workshop (SSP)*, 2023, pp. 150–154.

[36] T. H. Dao, H. T. H. Yen, V. N. Hoang, D. T. Tran, and D. N. Tran, "Human Activity Recognition System For Moderate Performance Microcontroller Using Accelerometer Data And Random Forest Algorithm," EAI Endorsed Trans. Ind. Networks Intell. Syst., vol. 9, no. 4, pp. 1–18, 2022, doi: 10.4108/eetinis.v9i4.2571.

Author Index