Grenze International Journal of Engineering and Technology, June Issue

Transact Safe: A Machine Learning Shield against Online Fraud

Ajit Huke[1], Archana Jadhav[2], Md Suhel Shaikh[3], Sameer Pathan[4], G.S. Mate[5] and Chandan Prasad[6]
[1-6]Department of Information Technology, Jspm's Rajarshi Shahu College of Engineering, Pune, Maharashtra, India
Email: ajithuke63@gmail.com, ajjadhav_it@jspmrscoe.edu.in, mdsuhelshaikh22@gmail.com, athansameer007@gmail.com, gsmate_it@jspmrscoe.edu.in, chandanprasad_it@jspmrscoe.edu.in

***Abstract*— Online transactions have become a fundamental part of daily life in the current digital age, acting as a crucial conduit for important financial activity. The prevalence of fraudulent activities in online transactions, however, presents a serious risk to the reliability and security of digital financial systems. This survey article provides a thorough examination of modern approaches used to identify online transaction fraud. Through an exploration of the most recent developments in statistical and machine learning methods, we outline a terrain of creative strategies intended to reduce fraudulent activity. In addition, we carefully analyze the built-in constraints and new difficulties that face existing fraud detection techniques, opening the door for further research. We discuss various fraud detection algorithms and compare them, explaining how they perform in terms of important parameters like accuracy, sensitivity, and specificity. In the end, this survey report proves to be a priceless tool that helps practitioners and scholars alike successfully and precisely traverse the complex world of online transaction fraud detection.**

***Index Terms*— Fraud detection, accuracy, sensitivity, specificity.**

I. INTRODUCTION

Online transaction fraud is a significant problem that affects millions of people worldwide. Fraudsters use various techniques to steal personal information, money, and other assets, causing considerable financial losses to individuals, businesses, and governments. The increasing scale and complexity of online transactions[Fig. 1] have made fraud detection a critical challenge for cybersecurity experts. Following this research study, we'll examine the latest studies and advancements within the field of online transaction fraud detection. Statistics and multidimensional analysis are used in conventional fraud detection methods to identify fraudulent activities[12]. However, these methods have limitations in detecting complex patterns and hidden relationships in the transaction information. With the growth of massive data sets and machine learning technologies, researchers have developed sophisticated fraud detection models that can analyze large volumes of transaction data and identify fraudulent activities with high accuracy. Several studies have shown that machine learning models outperform traditional methods in detecting online transaction fraud.

In the realm of online transaction fraud detection, neural networks stand out as a popular machine learning technique. Researchers have introduced a range of neural network models, including fully connected neural networks, recurrent neural networks[17], and convolutional neural networks[3][18][19], to uncover and flag fraudulent activities.

Grenze ID: 01.GIJET.10.2.99

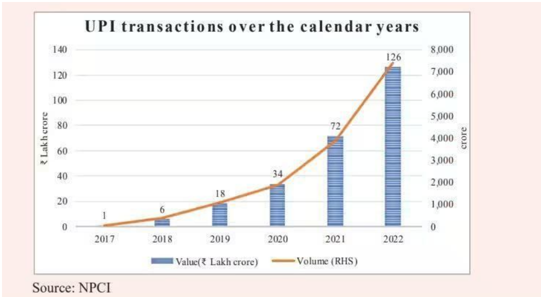


Fig. 1: Statistics for Transaction[23]

These models employ diverse architectures and learning algorithms to uncover irregular patterns indicative of fraudulent behavior within transaction data. For instance, in 2018, Zhang[3] introduced a prediction model that utilizes feature convolutional neural networks for transaction verification. This model demonstrated superior stability and classification accuracy when compared to other CNN models, showcasing its effectiveness in identifying fraudulent transactions.

II. LITERATURE SURVEY

Vaishnavi Nath et al. [1] propose a novel method for machine learning techniques for detecting credit card fraud. It focuses on streaming transaction data and involves clustering cardholders employing range partitioning, according to their transaction amount, classified as high, medium, and low. A sliding window strategy is used to extract behavioral patterns[15]. Different classifiers are trained on each group, and the classifier with the best rating score is chosen for fraud prediction. Importantly, the paper incorporates a feedback mechanism to tackle the concept drift challenge, where evolving transaction behaviors can impact the model's effectiveness over time. The Matthews Correlation Coefficient (MCC) is used to evaluate the classifiers, Random forest, decision tree, and logistic regression algorithms are found to perform well in detecting fraud[19]. The study also covers the utilization of SMOTE[21] to balance the dataset and improve classifier performance. Overall, the proposed method shows promise in the detection of credit card fraud.

Dheepa V [2] et al., published in the International Journal of Recent Trends in Engineering in 2009, analyzed various strategies for detecting credit card fraud. The paper discusses the problem of identifying credit card fraud and presents a trio of approaches for doing so. The research looks into the value of using various methods to solve the issue. One of the approaches discussed is the clustering model, which uses data clusterization to classify legal and fraudulent transactions. The model breaks down the data into related components to identify patterns and order. It uses 24 parameters of transactions for classification. Another approach is the probability density estimation method, which models the likelihood density of a credit card user's previous actions to spot irregularities in current behavior. The third approach is the model based on Bayesian networks[10], which describes the statistical data for a certain user and the statistical data for several fraud situations. The paper suggests that combining these three methods could improve the fraud detection system.

Zhang [3] et al. proposed a model for online transaction fraud detection using CNN. The model includes a feature sequencing layer that optimizes the sequence of transaction features, allowing the CNN to learn derivative features that are beneficial for classification[18][19]. The model aims to identify fraudulent transactions by analyzing user transaction behavior[15]. The experiments are conducted on a dataset of 5 million transaction records from a bank. The outcome demonstrates that without using derivative features the model reaches a good fraud detection performance, with precision found is around 91% and recall found is around 94%. The model is more suitable for low-dimensional transaction data and outperforms current CNN models detection of fraud. The paper also discusses the potential future work of discovering the sequence characteristics of transactions and applying the Long Short Term Memory (LSTM) algorithm for more accurate fraud detection[17][21].

Jain [4] conducted a prospective analysis of telecommunication fraud detection using neural network classification[13]. The paper discusses the application of data stream analytics and data mining in this context and compares their accuracy. The results indicate that data stream analytics outperforms data mining for telecommunication fraud detection[22]. The paper highlights the benefits of using data stream analytics, such as real-time processing of high-velocity data. On the other hand, data mining is more suitable for processing stored data that doesn't require urgent processing. The paper suggests integrating data stream analytics with sensors in various industries. Furthermore, it compares the output and other parameters between data stream analytics and data mining, suggesting that data mining can be enhanced by applying it to real-time dynamic data, eliminating the need to store data before processing.

Bocheng Liu et al. [5] suggested a machine learning-based approach for identifying fraud in online transactions. The paper presents 2 authentication methods: one utilizing a fully connected neural network and the other based on XGBoost[13][19]. The algorithms achieve high AUC values of 0.912 and 0.969, respectively. The paper also discusses the optimization of the XGBoost classifier using Hyperopt, which improves its performance in fraud detection. The optimized classifier achieves an AUC of 0.969, significantly enhancing the accuracy of fraud detection in network transactions. Continuous features are transformed using logarithmic transformation and Z-score standardization, while categorical features are encoded using one-hot encoding. The paper presents the design of an XGBoost-based solution for detecting fraud in transactions over the internet that allows users to upload transaction data and receive fraud detection results.

Shayan Wangde et al. [6] present a system that employs machine learning to detect fraud transactions within the e-commerce domain. The research focuses on the creation of a web-based platform designed to identify and block potentially fraudulent transactions in e-commerce settings. The paper commences by offering an overview of credit card fraud and the preventative measures in place, such as Verification of Address Service and Card Value, and Device Verification Analysis. The proposed system takes into account multiple actors, including user behavior[15], geographical location, and transaction specifics, in order to detect unusual patterns that could potentially signify fraudulent behavior. Detailed information is provided concerning the software and hardware prerequisites necessary for implementing this system, with a specific mention of the utilization of the Banksim dataset, which contains financial data to aid in the detection of fraudulent transactions.

Lahari Madabhattula et al. [7] proposed a fraud detection system for online transactions. The authors introduce the Behavior and Location Analysis (BLA) method to analyze the spending habits of credit card owners and identify any irregularities. The system aims to protect legitimate users from financial loss and enhance the security of electronic payments. The paper includes various diagrams such as E-R diagram, use case diagram, sequence diagram, activity diagram, class diagram, and data flow diagram to illustrate the design of the proposed system. Data Flow Diagrams (DFDs) and their implementation using Python and the Django framework are discussed. It explains the rules, types, and features of DFDs. The features and benefits of using Django as a web development framework are discussed, including WAMP server, PHP Admin, and SQL Server. The different levels of testing in software development are explained, including unit testing, validation testing, system testing, integration testing and output testing.

Magilan Saravanan et al. [8], talked about using machine learning methods for determining online scams. The authors' method for predicting fraudulent online transactions makes use of five separate algorithmic learning processes. Random Forest, Artificial Neural Network, Decision Tree, Naive Bayes, and SVM are the algorithms employed[16]-[20]. The experiment's dataset was gathered from Kaggle and is made up of credit card transaction data. With accuracy rates of 99.78% and 99.65%, respectively, Random Forest and Support Vector Machine outperforms the other methods, according to the results. The importance of machine learning for fraud detection and its potential to increase the process's accuracy and effectiveness are highlighted in the study.

Aditya Oza [9] utilized the Paysim dataset, a Kaggle-provided collection of simulated mobile-based payment transactions. Their research concentrated on the TRANSFER and CASH OUT transaction categories, known to include instances of fraud, within this highly imbalanced dataset, where only 0.13% of transactions were labeled as fraudulent. The authors explored the effectiveness of three approaches: logistic regression, support vector machine (SVM)[10], and a class weights-based method-for detecting fraud in payment transactions. They outlined their dataset splitting strategy, which involved categorizing the dataset by transaction types (TRANSFER and CASH OUT) that were prone to fraud. For model training and tuning, a class weights-based approach was adopted, assigning different weights to fraud and non-fraud samples for each of the three techniques (logistic regression, linear SVM, and SVM with RBF kernel). The dataset was partitioned into three subsets: 70% for training, 15% for cross-validation, and 15% for testing, with stratified sampling ensuring consistent class proportions across the splits.

In the study conducted by M. Priyadharsini et al. [10], the Support Vector Machine (SVM) algorithm was proposed for the purpose of classifying online fraud transactions. The research underscores the effectiveness of SVM[8][11] in accurately differentiating fraudulent transactions, establishing its significance as a valuable tool in the realm of online fraud detection and prevention. The paper delves into the experimental methodology employed, encompassing several stages: user registration, login authentication, customization of security questions, OTP (One-Time Password) verification, and the use of SVM for classifying online fraudulent transactions. A number of models were assessed and contrasted in the last part, including Bayesian networks, support vector machines (SVM), hidden markov models (HMM), and artificial neural networks (ANN)[17].

Gyanangshu Misra et al. [11] discuss the problem of credit card fraud in online transactions and the need for more efficient fraud detection techniques. It explores various existing systems and techniques used for fraud detection, such as hidden Markov models and Naive Bayes systems[20]. The proposed system aims to detect fraudulent transactions by analyzing user behavior and implementing a multi-layered security approach. It also utilizes IP geolocation to verify transaction details and detect suspicious activity. The paper concludes that the proposed system shows promise in detecting online credit card fraud, although fraud detection is a complex process.
Ruttala Sailusha et al.[14] have harnessed the Random Forest and Adaboost algorithms for the purpose of fraud detection[19]. The assessment criteria encompassed key metrics such as accuracy, precision, recall, and F1-score. The authors generated ROC curves by utilizing information derived from the confusion matrix. Authors found consistent accuracy in both Random Forest and Adaboost algorithms in their analysis. However, a closer examination of precision, recall, and F1-score reveals that the Random Forest algorithm consistently outperforms the Adaboost algorithm, showcasing higher values.

III. METHODOLOGIES

A. Logistic Regression

Drawing from the realm of statistics, machine learning incorporates Logistic Regression(LR) as a powerful classification technique.[20] It provides a straightforward yet effective approach for modeling binary outcomes, where the target variable can belong to one of two categories. Logistic regression's core aim is to estimate probabilities through a logistic function derived from the cumulative logistic distribution, facilitating the quantification of the relationship between predictor variables and the target variable.In logistic regression, we work with a vector $\beta = (\beta 1, \beta 2, ..., \beta n)$ denoting the coefficients and $X = (X1, X2, ..., Xn)$ representing the multiple predictors in the dataset. Additionally, ϵ signifies the error term within the model. Equation (eq. 1) [8] expresses the primary logistic function, wherein Y represents the target variable and is modeled as a linear combination of predictors X1, X2, ..., Xn, featuring respective coefficients β1, β2, ..., βn, alongside an error term ϵ.

$$Y = \beta[\![1X[\![1 + \beta 2X2 + ...+ \beta nXn + \epsilon]\!]\_\ ]\!]\_ \qquad \ldots\ldots \text{(Eq 1)}$$

B. Decision Tree

A decision tree[20] is a widely used hierarchical data structure in machine learning that enables making predictions or drawing conclusions based on a set of conditions. It is commonly utilized for classification tasks, where the leaf nodes represent different classes, and the intermediate nodes are split into sub-nodes based on specific attributes or conditions. The decision tree is built recursively by partitioning the data using attribute values, and decisions are made at each node based on the evaluated attribute. This construction process forms a path from the root to the leaf nodes, which is used to derive classification rules. These rules are based on the attribute values encountered along the path. When new data is presented, it is classified by traversing the decision tree based on the attribute values of each node until reaching a leaf node that determines the final classification. Decision trees offer interpretability and are fundamental components in more complex classification and regression models.

C. Random Forest

Random forests are ensemble techniques that combine multiple decision trees for classification and regression tasks. They reduce overfitting by training individual trees on different subsets of data and aggregating their predictions. The final prediction is obtained through methods like majority voting or averaging. Random forests improve generalization and robustness compared to single decision trees[20].

D. SVM

The Support Vector Machine is among the most popular supervised learning algorithms, serving purposes in both classification and regression tasks. SVM aims to establish the optimal line or decision boundary to facilitate future classification of new data points. This hyperplane is generated by SVM through the selection of extreme points. In the SVM algorithm, it determines which line from each class is closest to the other; these points are referred to as support vectors. To handle data that is not linearly separable, SVM employs the kernel method[18]. This involves using kernel transforming input data into high-dimensional features, making it amenable to linear separation. Common examples of linear functions include polynomial, linear, and radial basis functions.

E. XGBoost

Popular machine learning algorithm XGBoost is renowned for its strong performance and adaptability. Decision trees are combined with the XGBoost algorithm to create a powerful predictive model. Decision trees are used by the XGBoost as base learners. These decision trees are built greedily, choosing the optimum split points based on how much the loss function is reduced. Both classification and regression tasks are supported by XGBoost[21]. Boosting iteratively improves the model performance by focusing on the samples that were previously misclassified. The XGBoost uses a regularized learning objective function, consisting of two parts: the loss function, which helps control model complexity and prevents overfitting.

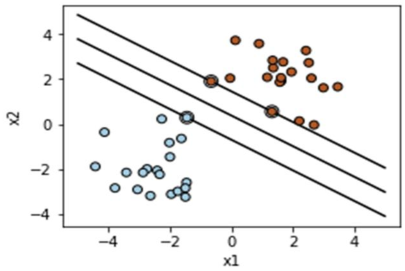


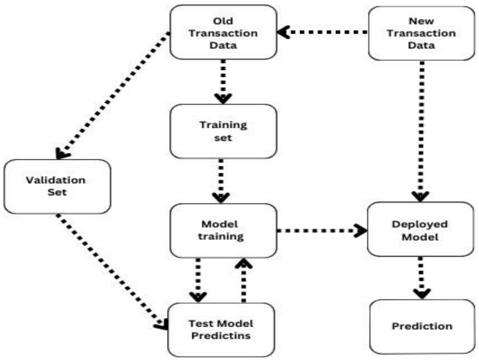
Fig 2. Graph for SVM classification

TABLE I: LITERATURE ON DIFFERENT ALGORITHMS, PERFORMANCE AND FUTURE SCOPE / CHALLENGES

Ref. No.	Methodologies Used	Year	Performance Metrics	Challenges/Future Scope
[1]	LR,DT,RF.	2019	Accuracy : 0.9994 Precision : 0.9310 MCC : 0.8356	Concept drift, i.e., the changes in the underlying data distribution over time, is one of the biggest challenges.Future research can focus on developing methods to adapt to changing patterns.
[2]	Clustering, Probability density estimation.	2009	Max Deviation: Region 1 - 42.5 Region 2 - 37.5	Combining the three presented methods (Clustering, Gaussian mixture model, and Bayesian networks) could be beneficial for improving the fraud detection system.
[3]	Convolutional Neural Networks (CNN)	2018	Precision : 91 % Recall : 94 %	Limited dimensions of transaction data and the need for feature engineering. It also suggests potential future research directions, such as analyzing user behavior from a system perspective.
[4]	Data Stream Analytics (DSA), Data Mining (DM)	2017	Error Percentage : DSA : +16.19 DM : -17.08	The integration of this system with IOT can be beneficial for various industries such as traffic, health, and telecommunications, expanding the application of the framework.
[5]	Fully connected NN & XGBoost algorithm	2021	AUC: 0.969	Does not discuss the robustness of the models against adversarial attacks.
[6]	KNN and random forest.	2022	Recall : 0.99 F1-score : 0.98 Precision : 0.97	Imbalanced dataset, Evolving fraud techniques, False positives.
[7]	Behavior and Location Analysis method	2021	Not Mentioned	Scalability, Performance optimization, and Compatibility could be potential areas of concern.
[8]	LR, Naive Bayes, KNN, DT, SVM, RF and XGBoost	2022	LR : Accuracy - 98.8, Sensitivity - 91.6, Precision - 98.8, F1 Score - 95.5.	Not all fraudulent transactions are caught and reported, leading to misclassification of the models. Developing adaptive machine learning methods that can automatically adapt to dynamic and evolving fraud patterns in real-world data.
[9]	LR, Linear SVM, and SVM with RBF kernel	2018	Precision : 0.4516 AUPRC : 0.8949 f1-measure: 0.6213 Recall : 0.9951	Key challenges include dealing with imbalanced datasets and balancing fraud detection with false positives. Future directions involve exploring decision trees, time series models like CNN, and user-specific models to improve accuracy and efficiency.

[10]	SVM, ANN, Hidden Markov Model(HMM), Bayesian Network	2021	SVM : 94.65 ANN : 97.32 HMM : 94.7 BN : 96.52	In advanced engineering technologies, challenges involve enhancing computing power, integrating diverse technologies, and ensuring security. The future scope includes applying advanced technologies like blockchain, and addressing ethical implications.
[11]	LR, SVM, KNN, Naive Bayes	2022	LR : 74 % NB : 83 % KNN : 72 % SVM : 91 %	One challenge in fraud detection is minimizing false positives, where legitimate transactions are incorrectly flagged as fraudulent. Advancing AI and ML for the Detection of Complex Patterns

VII. Proposed System



V. Challenges

A. Imbalanced Data

Fraudulent transactions are often relatively rare compared to legitimate transactions, leading to imbalance datasets. This can pose a challenge during model training.

B. Evolving Fraud Patterns

Fraudsters continually adapt their techniques to evade detection. It can be challenging to keep up with emerging fraud patterns and update the detection system accordingly.

C. Real-time Processing

Online transactions occur in real-time, requiring the fraud detection system to make quick decisions within a limited timeframe. The system should process transactions promptly without causing any delays or interruptions to the user expertise.

D. Privacy Concerns

Fraud detection systems often process sensitive user information, raising privacy concerns. It’s important to handle data securely, comply with privacy regulations, and implement appropriate measures to protect user data.

VI. Conclusion

The survey paper provides a comprehensive review of the latest research in online transaction fraud detection techniques. The authors have discussed the limitations and challenges of traditional fraud detection methods and proposed potential future research directions. The survey paper presents a comparative analysis of various fraud detection models based on metrics such as accuracy, sensitivity, and specificity. The authors highlight the significance of big data and machine learning technologies in detecting complex patterns and hidden relationships in transaction data.

References

[1] Dornadula, Vaishnavi Nath, and Sa Geetha. "Credit card fraud detection using machine learning algorithms." Procedia computer science 165 (2019): 631-641.

[2] Dheepa, V., and R. Dhanapal. "Analysis of credit card fraud detection methods." International journal of recent trends in engineering 2, no. 3 (2009): 126.

[3] Zhang, Zhaohui, Xinxin Zhou, Xiaobo Zhang, Lizhi Wang, and Pengwei Wang. "A model based on convolutional neural network for online transaction fraud detection." Security and Communication Networks 2018 (2018).

[4] Jain, Vanita. "Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification based data mining." International Journal of Information Technology 9, no. 3 (2017): 303-310.

[5] Liu, Bocheng, Xiang Chen, and Kaizhi Yu. "Online Transaction Fraud Detection System Based on Machine Learning." In Journal of Physics: Conference Series, vol. 2023, no. 1, p. 012054. IOP Publishing, 2021.

[6] Wangde, Shayan, Raj Kheratkar, Zoheb Waghu, and P. S. Lawand. "Online transaction fraud detection system using machine learning & e-commerce." International Research Journal of Engineering and Technology 9, no. 4 (2022).

[7] Lahari Madabhattula, Maridu Manikanta, Pradeep Kumar:"Online Transaction Fraud Detection," 2021.

[8] TP, Rani, Magilan Saravanan, Ashish Kumar Sahu, K. Martin Sagayam, and Ahmed A. Elngar. "Predicting Online Fraudulent Transactions Using Machine Learning." (2022).

[9] Oza, Aditya. "Fraud detection using machine learning." TRANSFER 528812, no. 4097 (2018): 532909.

[10] Mary, I. Mettildha, and M. Priyadharsini. "Online transaction fraud detection system." In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 14-16. IEEE, 2021.

[11] Swapnil malatesh dharnappa goudar and Gyanangshu misra. "Online transaction fraud detection using backlogging on e-commerce websites"(2022).

[12] Jaiswal, Shubham, R. Brindha, and Shubham Lakhotia. "Credit card fraud detection using isolation forest and local outlier factor." Annals of the Romanian Society for Cell Biology (2021): 4391-4396.

[13] Aleskerov, Emin, Bernd Freisleben, and Bharat Rao. "Cardwatch: A neural network based database mining system for credit card fraud detection." In Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr), pp. 220-226. IEEE, 1997.

[14] Sailusha, Ruttala, V. Gnaneswar, R. Ramesh, and G. Ramakoteswara Rao. "Credit card fraud detection using machine learning." In 2020 4th international conference on intelligent computing and control systems (ICICCS), pp. 1264-1270. IEEE, 2020.

[15] Meshram, P. L., and P. Bhanarkar. "Credit and ATM card fraud detection using genetic approach." International Journal of Engineering Research & Technology (IJERT) 1, no. 10 (2012): 1-5.

[16] Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. "A machine learning based credit card fraud detection using the GA algorithm for feature selection." Journal of Big Data 9, no. 1 (2022): 1-17.

[17] Roy, Abhimanyu, Jingyi Sun, Robert Mahoney, Loreto Alonzi, Stephen Adams, and Peter Beling. "Deep learning detecting fraud in credit card transactions." In 2018 systems and information engineering design symposium (SIEDS), pp. 129-134. IEEE, 2018.

[18] Carrasco, Rafael San Miguel, and Miguel-Ángel Sicilia-Urbán. "Evaluation of deep neural networks for reduction of credit card fraud alerts." IEEE Access 8 (2020): 186421-186432.

[19] Alarfaj, Fawaz Khaled, Iqra Malik, Hikmat Ullah Khan, Naif Almusallam, Muhammad Ramzan, and Muzamil Ahmed. "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms." IEEE Access 10 (2022): 39700-39715.

[20] Tanouz, D., R. Raja Subramanian, D. Eswar, GV Parameswara Reddy, A. Ranjith Kumar, and CH VNM Praneeth. "Credit card fraud detection using machine learning." In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 967-972. IEEE, 2021.

[21] Mienye, Ibomoiye Domor, and Yanxia Sun. "A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection." IEEE Access 11 (2023): 30628-30638.

[22] Khine, Aye Aye, and Hint Wint Khin. "Credit card fraud detection using online boosting with extremely fast decision tree." In 2020 IEEE Conference on Computer Applications (ICCA), pp. 1-4. IEEE, 2020.

[23] Moneycontrol News. (2023, January 31). Economic Survey 2023: UPI accounted for 52% of India's total digital transactions in FY22. Moneycontrol. https://www.moneycontrol.com/news/business/economic-survey-2023-upi-accounted-for-52-of-indias-total-digital-transactions-in-fy 22-9970741.html

[24] Jadhav, Archana & Kadam, Vaishnavi & Prasad, Snehal & Waghmare, Nikita & Dhule, Sangameshwar. (2023). An Emotion Recognition from Speech using LSTM. 834-842. 10.1109/ICSCSS57650.2023.10169351.

[25] Mate, Gitanjali S., Abdul K. Kureshi, and Bhupesh Kumar Singh. "An efficient CNN for hand X-ray classification of rheumatoid arthritis." Journal of Healthcare Engineering 2021 (20