Secure UPI: Machine Learning-Driven Fraud Detection System for UPI Transactions

Rupa Rani
Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, Ghaziabad, U.P., India.
Department of Computer Science & Engineering, Banasthali Vidyapith, Rajasthan
Email: rupachoudhary2010@gmail.com
ORCID ID- 0000-0002-4783-3857

Adnan Alam
Department of Computer Science & Engineering- Data Science, ABES Institute of Technology, Ghaziabad, U.P., India.
Email: alamadnan256@gmail.com

Abdul Javed
Department of Computer Science & Engineering- Data Science, ABES Institute of Technology, Ghaziabad, U.P., India.
Email: abdul2020csds033@abesit.edu.in

2024 2nd International Conference on Disruptive Technologies (ICDT) | 979-8-3503-7105-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDT61202.2024.10489682

***Abstract*— SecureUPI specializes in developing an advanced fraud detection gadget the usage of the effective XGBoost device getting to know set of rules to create an advanced fraud identity device. XGBoost is a properly-proper alternative for enhancing the precision of fraud detection fashions because of its reputation for dealing with tricky datasets with efficiency and its music record of success throughout multiple industries. In order to extract pertinent records, like transaction quantity, frequency, and place, our approach preprocesses UPI transaction facts. This article makes use of a labelled dataset to train the XGBoost version in order that we may also take gain of its sturdy prediction talents and capacity to handle imbalanced datasets. To help create a system that is less difficult to apprehend and use, feature importance evaluation is used to discover essential symptoms of feasible fraud. After training, the model is covered right into a real-time UPI transaction tracking device, in which it maintains an eye fixed out for any suspicious traits in incoming transactions. In order to lessen the results of fraudulent activity, the system is constructed with 98.2 % accuracy to send out instant notifications and take preventive steps. This challenge allows in improving UPI transaction security and advancing economic era are accomplished through demonstrating the performance of machine learning in fraud detection.**

Keywords*— *UPI, Fraud Detection, SMOTE, PCA, XGBoost, Imbalanced Datasets, Dimensionality Reduction, Machine Learning.

I. Introduction

In a global in which virtual transactions rule, the Unified Payments Interface (UPI) has become a current platform that streamlines and expedites monetary transactions [1]. But the quick development of digital bills also gives the undertaking of fighting state-of-the-art fraudulent sports activities. Through the advent of a sophisticated UPI Fraud Detection System and the utility of current pre-processing, feature extraction, and algorithm choice techniques, this venture tackles this essential assignment [2] [3].

Improving virtual transaction security and dependability via the UPI framework is the principle intention of the UPI Fraud Detection System. The tool pursuits to hit upon and decrease fraudulent sports activities via using modern techniques, making certain the safety and reliability of the UPI surroundings [4]. During the UPI Fraud Detection System's design and deployment, protection and privateness of sensitive economic records are of utmost importance. Strict methods that comply with felony requirements and industry first-rate practices are used to shield and anonymize the statistics that has been gathered. The system is made to paintings in the constraints of privateness and statistics safety laws, selling person self assurance and criminal compliance [5].

Fig. 1. Interface of Secure UPI

Figure 1 shows the interface of Secure UPI website which is supposed to be advanced constantly it isn't static [6] This structure of the device consists of non-stop mastering, which lets in it to alter to converting fraud techniques and customer conduct. By mechanically retraining the version, the device makes use of the most modern records to beautify its predicting capabilities. By incorporating non-save you studying, the UPI Fraud Detection System is assured to remain a sincere protector within the path of latest threats and to hold its efficacy over the years [7].

For the cause of resolving the inherent elegance imbalance in UPI transaction datasets, the Synthetic Minority Over-sampling Technique (SMOTE) choice made for the duration of pre-processing is essential [8] [9]. Since fraud times are commonly unusual, SMOTE creates fictitious examples to even out the commands just so the following set of rules can teach on a greater representative dataset. By doing this, the model's capacity to identify fraudulent patterns is improved and bias against the majority class is eliminated. [10] Because Principal Component Analysis reduces dimensionality while effectively collecting the most important information in the dataset, it is used. principle component analysis (PCA) helps identify patterns that significantly contribute to the variation

in the data by converting the original features into uncorrelated principle components. As a result, features are represented more effectively, which is crucial for developing reliable fraud detection models [11].

The UPI Fraud Detection System operates in real-time, continuously monitoring incoming transactions as they occur [12]. By utilizing the XGBoost algorithm's speed and efficiency, transactions are quickly compared to the learnt patterns, enabling the immediate detection of potentially fraudulent activity. By ensuring that any suspicious activity is immediately addressed, real-time monitoring helps to reduce the negative effects of fraudulent transactions on both users and financial institutions [13].

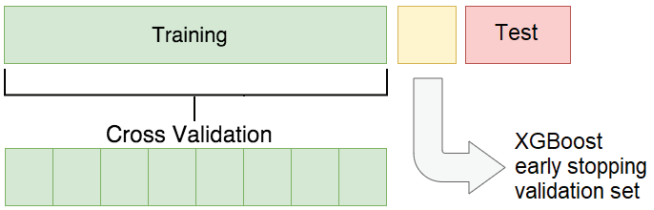

Fig. 2. Data Splitting

Figure 2 shows how we are splitting the data into 2 sections. Then we will search for the best parameter combination for the XGBoost algorithm, using K-fold cross confirmation. Before we get into the details of parameter optimization, a caveat about how I use data for cross confirmation is in order. XGBoost offers a useful point called beforehand stopping, which helps help overfitting due to growing too numerous trees. At every replication, it evaluates the performance of the model so far on a confirmation set. It keeps training until the error on this confirmation set, which I'll call beforehand stopping set, fails to ameliorate for a given number of duplications. This means that we're optimizing the number of trees as we train the model. But that also means that we can not use that same set to estimate the model. else, we'd overfit(the number of trees parameter) and the cross confirmation score will be poisoned overhead. That's why we need a separate confirmation set for early stopping, in addition to the(rotating) confirmation set used in cross confirmation. thus, to help the early stopping point from overfitting, in addition to unyoking the data into a training and a test set, I also set aside a small portion of the training set to be used as the early stopping confirmation set.

The primary classification model is XGBoost, an ensemble learning algorithm selected for its outstanding performance and adaptability. It is especially well-suited for fraud detection due to its capacity to manage intricate linkages, nonlinearities, and interactions within the data base. An extremely precise and flexible classification procedure is guaranteed by the ensemble of decision trees, which have been trained successively to fix the mistakes of the earlier models [14] [15].

The project consists of an in depth hyperparameter tweaking device to get the most predictive power out of XGBoost. Systematic adjustments are made to parameters which incorporates mastering rate, maximum tree intensity, and kind of boosting rounds. Through this improvement, the set of guidelines plays higher and achieves the proper balance among precision and recall— metrics which can be essential to the identification of fraud [16] [17]. In the UPI Fraud Detection System, hyperparameter tuning is a critical step in pleasant-tuning the eXtreme Gradient Boosting (XGBoost) set of policies, which serves because the primary class model. XGBoost comes with severa hyperparameters, together with studying fee, most depth of wooden, extensive form of boosting rounds, and regularization phrases, among others [18].

The system has thorough reporting and assessment capabilities that offer facts about effectiveness, overall performance, and abnormalities determined [19]. Stakeholders can advantage a complete records of the fraud detection panorama, discover the facts of flagged transactions, and realise the features that make contributions to danger scores via the use of interactive dashboards and visualizations. This reporting feature permits with compliance reporting, enables well-knowledgeable preference-making, and gives insightful enter for brought machine development [20]

In cease, a high development round economic safety is the UPI Fraud Detection System. Using behavioural evaluation, anomaly detection, and tool studying, the challenge seeks to gather a strong defence in opposition to fraudulent UPI transactions. This all-encompassing method not exceptional identifies and stops fraudulent interest, however it additionally advances protection capabilities constantly inside the ever-changing region of virtual banking. The undertaking has the functionality to noticeably improve economic safety by means of the use of strengthening client don't forget and the of upi transaction.

II. Literature Survey

M.A Ibrahim [1], This research paper discusses the most common methods of fraud, detection techniques, and recent findings in the field. The researchers used the SMOTE technique to balance the dataset and found that models like Decision Tree, Random Forest, Neural Network, and K-nearest neighbour performed well when fitted and trained with the data. The system allows users to select their preferred model. The Random Forest model achieved an accuracy of 93.58%, but its efficiency decreases when trained with imbalanced transaction datasets.

P. Boulieris [2], The main objective of the work is to present a dataset for online fraud detection that is anonymized and publicly available. We argue that standard evaluation metrics used in existing literature should be complemented with online and offline assessments to evaluate model performance in a real-world business setting. We found that incorporating anomaly detection features improves all metrics except for online detection, highlighting the importance of considering both online and offline evaluation alongside standard metrics. Despite using fewer traditional features, the addition of NLP-based features significantly improved performance compared to a previous study.

B. Baesens [9], This paper discusses the importance of information technology in fraud detection. The authors share information technology for feature design and application design. They argue that intelligent data design is more effective in improving analytical efficiency than developing sophisticated new analytical techniques. The authors demonstrate this using a large European bank payment transaction data set and show that both feature design and request design significantly improve the performance of analytical models. They also point out that simple analytical methods such as logistic regression and classification trees can produce good results if the data are well designed. Although the focus is on fraud in payment transactions, the techniques

discussed can be extended to other frauds in health care, insurance or electronic commerce.

S. Alam [12], The XGBoost Classifier initially produced the best results among the four classification algorithms used. To improve classifier performance, three different data balancing approaches were used, and ROS produced the best results. RUS and the SMOTE method did not perform as well. With Random Over Sampling and the XGBoost Classifier, accuracy, precision, recall, and F1 scores increased significantly. In the future, more effective approaches and strategies should be explored to overcome the limitations of this study.

S. Mohanavalli [13], Data cleaning is a process that deals with various data quality problems such as noise, outliers, inconsistent data, duplicates and missing values. The focus is on detecting and eliminating errors and inconsistencies during the data collection phase. Traditional techniques mostly belong to this phase, but hybrid forms of error correction have yet to be developed to provide noise-free samples for data mining and analysis. Different denoising techniques can lead to different databases and the median of the collected data can be used for homogeneous data sets, the range of values can be expanded for heterogeneous data sets. Future work includes the investigation of hybrid methods for cleaning homogeneous data and reducing noise in various heterogeneous datasets using a voting system that combines multiple techniques. In addition, the impact of denoising techniques on classification performance and mining performance should be investigated for large datasets.

M. Greenacre [21], A popular multivariate analytic method for comprehending and exploring data is PCA (Principal Component analysis). Its ability to extract important information from complicated datasets has been demonstrated by its application in a variety of areas. Large datasets of all kinds may now be analyzed using PCA thanks to recent developments, and this statistical technique should continue to be improved by future breakthroughs. PCA is regarded as a fundamental tool in data science, along with its expansions and variants.

D. Dablain [22], To solve imbalanced data, a novel model called DeepSMOTE combines deep learning and the well-known SMOTE technique. In order to balance the training set and enable bias-free training of deep classifiers, it generates fictitious instances. DeepSMOTE can create effective low-dimensional embeddings, interact with raw photos, and create artificial images of excellent quality. According to experimental research, DeepSMOTE works better than alternative oversampling algorithms and is resilient to different imbalance ratios. Subsequent efforts will focus on augmenting DeepSMOTE with challenges at the class and instance levels, refining the loss function, and expanding its capabilities to accommodate dynamic class ratios and additional data modalities like as text and graphs.

TABLE 1. COMPARATIVE STUDY OF PREVIOUS RESEARCH

Author & Year	Purposed Solution	Technology	Pros	Cons
[1]	Fraud detection model for illegitimate transactions	Questionnaire-Responded transaction model	Works well for old user	Didn't work for new user
[2]	Fraud detection with natural language processing	K-Nearest Neighbours Algorithm	Easy to implement on small datasets	Didn't work on large datasets
[9]	Data Engineering for Fraud Detection Fraud Detection	ROS and RUS	Easy to implement on imbalanced dataset	Loss of valuable data
[12]	Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques	XGBoost	Easily handles imbalanced with maximum accuracy	Risk of overfitting if not appropriately tuned
[13]	Survey of Pre-processing Techniques for Mining Big Data	Logistic Regression	Easily handle large datasets as well as practical for real-time fraud detection	Low accuracy
[21]	Principal Component Analysis	PCA	Easily reduces the noise and redundancy	May lack clear interpretability
[22]	DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data	SMOTE	Preserves information by generating synthetic cases in the dataset	Risk of overfitting due to synthetic cases

The above table 1 shows the comparison of the previous related research.

III. PROPOSED METHODOLOGY

The below figure 3 shows the overall technique that we are implementing on our dataset. After dividing the dataset into training and test sets in which training data set contains 80% of the original data whereas the test set contains the remaining 20% of the original data. Then we will perform all these techniques on our training data set and in last we will use test data set to predict the transaction authenticity.

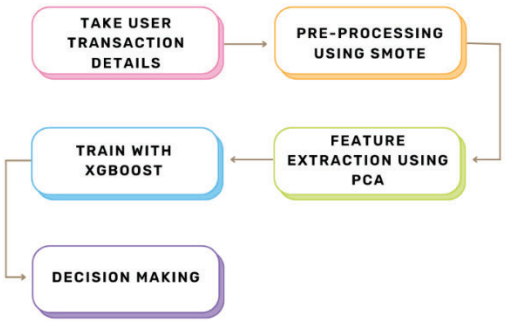

Fig. 3. Procedure for detecting UPI fraud

1. Data Collection (Taking User Transaction Details):

Description:

Relevant information about user transactions is gathered in the first stage. This includes any information deemed necessary for the fraud detection task, including timestamps, transaction amounts, user information, and other facts.

Purpose:

To train a strong fraud detection model, obtaining an extensive dataset is essential. The dataset forms the basis for the process's further stages.

2. Pre-processing with SMOTE Algorithm:

Description:

The dataset is pre-processed inside the 2nd degree using the Synthetic Minority Over-sampling Technique (SMOTE) method. SMOTE is in particular designed to cope with information imbalances, particularly whilst the proportion of legitimate transactions to fraudulent transactions is huge.

Purpose:

Biased models that carry out badly on minority training (fraudulent transactions) might possibly forestall end result from imbalanced datasets. To balance the distribution of commands and enhance the version's fraud detection capabilities, SMOTE creates artificial samples.

3. Feature Extraction with PCA:

Description:

Principal Component Analysis (PCA) is used in the 1/three diploma to extract features. PCA is used to lessen the dataset's dimensionality on the identical time as retaining important data. This aids in overcoming the computational difficulties added on by huge-scale statistics.

Purpose:

By using PCA to lessen dimensionality, the dataset becomes extra practicable for system learning algorithms. The motive of extracted abilties is to extract the most critical information, which improves model performance.

4. Model Training with XGBoost:

Description:

The fourth segment includes schooling a system learning model, XGBoost, the use of the pre-processed and function-extracted dataset. XGBoost is a famous ensemble gaining knowledge of set of policies that excels at dealing with complex, non-linear relationships in statistics because of its tremendous traditional performance and performance.

Purpose:

XGBoost is selected because of its versatility in managing unbalanced datasets, noise resilience, and capability to perceive complex styles inside the records. Based on the skills which have been processed and altered, the version is skilled to distinguish amongst real and fraudulent transactions.

5. Decision (Real or Fake Transaction):

Description:

Making predictions on new, unseen transactions the use of the informed XGBoost version is the final stage. The version uses the styles it has learnt in some unspecified time in the future of training to categorize every transaction as both real or fraudulent.

Purpose:

This stage establishes how the version have to be used in real situations. Accurately figuring out and flagging in all likelihood fraudulent transactions is the reason, giving financial institutions or rate processors a device for preference-making.

This flowchart describes a methodical procedure for detecting UPI fraud that begins with data collection, moves on to correct imbalances, reduces dimensionality, trains a reliable model, and concludes with decision-making based on the model's predictions. Every stage advances the model's precision and efficacy in differentiating between authentic and fraudulent transactions.

IV. RESULT

When we fill all the mentioned transactions details we need to click on the “Click to Detect” button to check whether the transaction is Valid or Fraud transaction.

If the transaction is valid figure 4 will be shown as an output and if the transaction is fraud, then figure 5 will be shown.

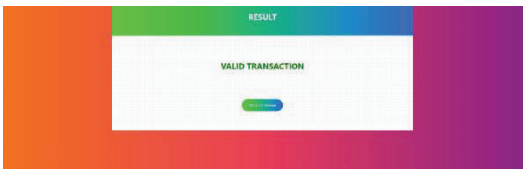

Fig. 4. Figure 1: Valid Transaction

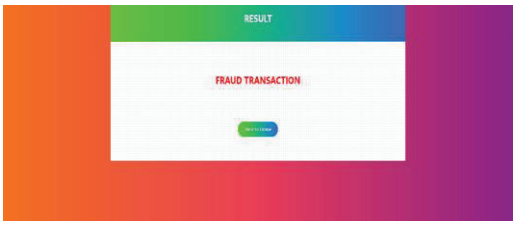

Fig. 5. Fraud Transaction

In figure 6 we can see the comparison of multiple algorithms in which XGBoost outperforms all the algorithms.

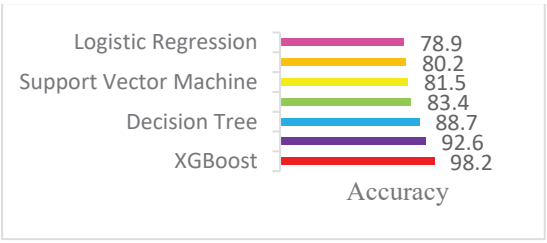

Fig. 6. Accuracy Comparison

The below table 2 shows the comparison of all algorithms.

TABLE 2. COMPARISON OF ACCURACY

Algorithm Name	Accuracy Score (%)
XGBoost	98.2
Random Forest	92.6
Decision Tree	88.7
K-Nearest Neighbors	83.4

Support Vector Machine	81.5
Naive Bayes	80.2
Logistic Regression	78.9

V. CONCLUSION

This UPI fraud detection project has applied innovative methodologies to enhance the robustness and efficiency of our fraud detection model. By addressing the imbalanced nature of the dataset with the Synthetic Minority Over-sampling Technique (SMOTE), we have successfully balanced the representation of fraudulent transactions, boosting the model's ability to recognize subtle patterns indicative of fraudulent operations.

The application of Principal Component Analysis (PCA) for feature extraction has contributed to the reduction of dimensionality, capturing the most salient aspects of the data and mitigating issues related to multicollinearity. This simplified the computing procedure and made it easier to understand the underlying fraud tendencies in the representation.

Our choice of XGBoost as the training model has been effective, exploiting its capabilities in managing skewed data, high predicted accuracy, and robustness. XGBoost's capacity to capture intricate relationships within the data has considerably boosted the model's overall effectiveness in recognizing probable cases of UPI fraud.

For future scope we can reduce the number of transactions details [23] [24] that is needed to detect the transaction authenticity whereas as well as we can track trends in user interactions with the UPI platform by integrating analytics on user behavior. Behavior anomalies may be used as early warning signs of possible fraud.

REFERENCES

[1] Adekunle, I. M., & Ozoh, P. (2023). Fraud detection model for illegitimate transactions. Kabale University Interdisciplinary Research Journal, 2(2), 21-37 https://doi.org/10.1016/j.future.2015.01.001

[2] Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2023). Fraud detection with natural language processing. Machine Learning, 1 22. https://doi.org/10.24321/2394.6539.202012

[3] Mytnyk, B., Tkachyk, O., Shakhovska, N., Fedushko, S., & Syerov, Y. (2023). Application of Artificial Intelligence for Fraudulent Banking Operations Recognition. Big Data and Cognitive-Computing,7(2),93. https://doi.org/10.1016/j.dss.2010.08.008

[4] Ridwan, R., Abdullah, S., & Yusmita, F. (2022). IMPLEMENTATION OF CASHLESS POLICY STRATEGIES TO MINIMIZE FRAUD IN THE GOVERNMENTSECTOR: SYSTEMIC REVIEW. Jurnal Akuntansi, 12(3), 181-201. https://doi.org/10.1007/s10994-023-06354-5

[5] Chang, V., Di Stefano, A., Sun, Z., & Fortino, G. (2022). Digital payment fraud detection methods in digital ages and Industry 4.0. Computers and Electrical Engineering,100, 107734. https://doi.org/10.1145/3394486.3403361

[6] Bandyopadhyay, S. K., & Dutta, S. (2020). Detection of fraud transactions using recurrent neural network during COVID-19: fraud transaction during COVID-19. Journal of Advanced Research in Medical Science & Technology (ISSN: 2394-6539),7(3),16-21. https://doi.org/10.1016/j.compeleceng.2022.107734

[7] Manocha, S., Kejriwal, R., & Upadhyaya, D. A. (2019, September). The impact of demonetization on digital payment transactions: a statistical study. In Proceedings ofInternational Conference on Advancements in Computing & Management (ICACM). https://doi.org/10.1109/TNNLS.2021.3136503

[8] Diadiushkin, A., Sandkuhl, K., & Maiatin, A. (2019). Fraud detection in payments transactions: Overview of existing approaches and usage for instant payments. Complex Systems Informatics and Modeling Quarterly, (20), 72-88. https://doi.org/10.7250/csimq.2019-20.04

[9] Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. Decision Support Systems, 150, 113492.https://doi.org/10.1109/ICICV50876.2021.9388431

[10] Carminati, M., Baggio, A., Maggi, F., Spagnolini, U., & Zanero, S. (2018). FraudBuster: temporal analysis and detection of advanced financial frauds. In Detection of Intrusions and Malware, and Vulnerability Assessment: 15th International Conference, DIMVA 2018, Saclay, France, June 28–29, 2018, Proceedings 15 (pp. 211-233). Springer International,Publishing. https://doi.org/10.1016/j.procs.2023.01.231

[11] Rastogi, S., Sharma, A., Panse, C., & Bhimavarapu, V. M. (2021). Unified Payment Interface (UPI): A digital innovation and its impact on financial inclusion and economic development. Universal Journal of Accounting and Finance, 9(3),518-530. https://doi.org/10.1109/ICCCSP.2017.7944072

[12] Gupta, P., Varshney, A., Khan, M. R., Ahmed, R., Shuaib, M., & Alam, S. (2023). Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. Procedia Computer-Science,218,2575-2584. https://doi.org/10.3390/bdcc7020093

[13] Hariharakrishnan, J., Mohanavalli, S., & Kumar, K. S. (2017, January). Survey of pre-processing techniques for mining big data. In 2017 international conference on computer, communication and signal processing (ICCCSP) (pp. 1-5). IEEE. https://doi.org/10.33050/atm.v5i2.1593

[14] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613. https://doi.org/10.1016/j.softx.2019.100341

[15] Branco, B., Abreu, P., Gomes, A. S., Almeida, M. S., Ascensão, J. T., & Bizarro, P. (2020, August). Interleaved sequence rnns for fraud detection. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3101 3109). https://doi.org/10.33369/jakuntansi.12.3.181-201

[16] Zhu, B., Gao, Z., Zhao, J., & vanden Broucke, S. K. (2019). IRIC: An R library for binary imbalanced classification.-SoftwareX,10,100341. https://doi.org/10.13189/ujaf.2021.090326

[17] Dileep, M. R., Navaneeth, A. V., & Abhishek, M. (2021, February). A novel approach for credit card fraud detection using decision tree and random forest algorithms. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 1025-1028). IEEE. https://doi.org/10.33050/atm.v5i2.1593

[18] Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems, 55, 278-288. https://doi.org/10.1109/ICCCSP.2017.7944072

[19] Lavadkar, M. A., Thorat, P. K., Kasliwal, A. R., Gadekar, J. S., & Deshmukh, D. P. Fingerprint Biometric Based Online Cashless Payment System. IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN, 2278-0661. https://doi.org/10.1016/j.softx.2019.100341

[20] Purnama, S., Bangun, C. S., & Faaroek, S. A. (2021). The Effect of Transaction Experience Using Digital Wallets on User Satisfaction in Millennial Generation. Aptisi Transactions on Management (ATM), 5(2), 161-168. https://doi.org/10.24321/2394.6539.202012

[21] Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. Nature Reviews Methods Primers, 2(1), 100. https://doi.org/10.13189/ujaf.2021.090326

[22] Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning-Systems. https://doi.org/10.1109/TNNLS.2021.3136503

[23] Rani,R., Yogi, K.K., Yadav, S.P.,. Detection of Cloned Attacks in Connecting Media using Bernoulli RBM_RF Classifier (BRRC). Multimed Tools Appl (2024).

[24] Kanaujia, V.K., Kumar, A., Yadav, S.P., Advancements in Automatic Kidney Segmentation using Deep Learning Frameworks and Volumetric Segmentation Techniques for CT Imaging: A Review, Arch Computat Methods Eng 29, 1753–1770 (2024)