VOLUME - 14, ISSUE - 04, APRIL - 2025 • PRINT ISSN No. 2277 - 8160 • DOI : 10.36106/gjra

Original Research Paper | Computer Science

COMPARATIVE ANALYSIS OF TREE-BASED ALGORITHMS FOR ENHANCED FRAUD DETECTION IN TRANSACTIONAL DATA

PVV Satya Eswara Rao	Computer Science & Engineering Sasi Institute Of Technology & Engineering Tadepalligudem, India
S . Asha Pallavi	Computer Science & Engineering Sasi Institute Of Technology & Engineering Tadepalligudem,india
D. Manoj Kumar	Computer Science & Engineering Sasi Institute Of Technology & Engineering Tadepalligudem, India
B.N.S.S.R. Krishnaveni	Computer Science & Engineering Sasi Institute Of Technology & Engineering Tadepalligudem, India
K. Aditya Brahmendra	Computer Science & Engineering Sasi Institute Of Technology & Engineering Tadepalligudem, India

ABSTRACT With the rapid adoption of Unified Payments Interface (UPI) for digital transactions, fraudulent activities have also increased significantly. To address this challenge, we propose a machine learning-based approach for UPI fraud detection. Our system analyzes transaction details such as the bank book name, transaction ID, and transaction amount to classify transactions as either fraudulent or legitimate. The model integrates multiple algorithms, including Decision Tree, Random Forest Gradient Boosting Machines (GBMs), and XGBClassifier, to enhance detection accuracy. The Random Forest classifier provides robustness against overfitting, while GBMs and XGBClassifier improve precision and recall by identifying complex fraud patterns. By leveraging these techniques, the proposed system enhances the security of UPI transactions and reduces unauthorized activities, ensuring a safer digital payment ecosystem.

KEYWORDS : UPI Digital Payments, Machine Learning, Random Forest, Decision Tree, Gradient Boosting Machines (GBMs), XGBClassifier, Fraud Detection, Financial Security.

INTRODUCTION

Unified Payments Interface (UPI) has revolutionized digital transactions by providing a fast, seamless, and efficient payment mechanism. However, the increasing adoption of UPI has also led to a surge in fraudulent activities, posing significant security threats to users and financial institutions. Fraudsters exploit vulnerabilities in transaction processes, leading to unauthorized transactions and financial losses. Traditional fraud detection methods often fail to adapt to the evolving nature of the fraud, necessitating advanced approaches for real-time and accurate detection.

To adress these challenges, we propose a machine learning-based fraud detection system that analyzes transaction details such as the bank book name, transaction ID, and transaction amount. The system utilizes multiple machine learning algorithms, including Random Forest, Decision Tree, Gradient Boosting Machines (GBMs), and XGB Classifier, to classify transactions as either fraudulent or legitimate. These models are choosen for their ability to capture complex fraud patterns while maintaining high accuracy and efficiency.

Random Forest is a robust classifier Known for its resilience against overfitting, making it well-suited for fraud detection. Decision Trees provide interpretability, allowing clear classification boundaries, while GBMs enhance predictive accuracy by iteratively refining fraud detection capabilities. XGBClassifier, an optimized version of GBMs, improves speed and scalability, making it highly effective in processing vast transaction datasets.

By integrating these techniques, the proposed system enhances fraud detection accuracy while minimizing false positives. The multi-algorithm approach ensures real-time transaction analysis, allowing legitimate transactions to be processed smoothly while blocking fraudulent activities. This model aims to strengthen the security of UPI transactions, providing users with a reliable and efficient fraud prevention mechanism.

II. Literature Survey

I) Cardwatch: a neural network based database mining system for credit card fraud detection

https://ieeexplore.ieee.org/document/618940

Here we introduce CARDWATCH, a database mining method that can identify credit card fraud. An intuitive graphical user interface, connectivity to several commercial databases, and a neural network learning module from the basis of the system. Very high success rates in detecting fraud were found in tests using autoassociative neural network models and synthetically created credit card data.

II) Understanding Telephony Fraud As An Essential Step To Better Fight It

https://www.semanticscholar.org/paper/Understanding-telephony-fraud-as-an-essential-step-Sahin/4f88de8f9ffb34aa147b2c10d4bf08b350ae917b

The first large-scale network, which reached around 7 billion people over a century ago, was the telephone network. The complex nature of telecommunications and the possibility of monetising several services make it an attractive target for fraudsters. Academic studies on these networks are few because of their complexity and closed nature. A comprehensive examination of fraud in telecommunication networks is the first part of this thesis. We present a taxonomy that differentiates between basic reasons, weaknesses, ways of exploitation, types of fraud, and benefits to fraudsters. We break it down for you and show how our taxonomy sheds light on CAller NAMe (CNAM) revenue sharing fraud. We look at two types of wholesale billing fraud that operators face head-on in the second part of the article. New forms of interconnect telecom fraud include Over-The-Top (OTT) bypass fraud. Directed via IP to a voice chat app on a smartphone, rather than terminating it over the telco infrastructure, is how OTT bypass works. We analyse the effects of this fraud on a small European country and how to identify it using more than 15,000 test calls and a thorough user survey. A big wholesale scam, the International Revenue Share fraud (IRSF), will be discussed later. Calls from IRSF numbers are being diverted

to "international premium rate services" by con artists. In order to gain a deeper understanding of the IRSF ecosystem, we examine data from several third-party premium rate service providers' worldwide premium rate tests. Using this data, we suggest characteristics for IRSF-detection for both the source and destination numbers of calls. The latter section of the thesis delves into consumer-side telephonic fraud, namely voice spam. A recent approach to stop unsolicited phone calls includes linking the spammer with a phone bot ("robocallee") that impersonates a genuine person. Lenny, a bot, plays pre-recorded audio messages to communicate with the user.

III) Fraud Detection System: A Survey

https://www.sciencedirect.com/science/article/abs/pii/S1084804516300571

Credit card, telecommunication, and healthcare insurance systems are just a few examples of the many forms of electronic commerce that have emerged as a result of the proliferation of both personal computers and large corporations. There are legitimate and dishonest individuals on these networks, unfortunately. There were a number of methods in which fraudsters gained access to e-commerce platforms. Fraud prevention systems (FPSs) do not adequately safeguard e-commerce networks. Electronic commerce systems may be protected, nevertheless, through FDS-FPS cooperation. Several challenges, including as concept drift, real-time detection, skewed distribution, and huge data, impede the performance of FDS. In a systematic fashion, this survey study examines these worries and roadblocks that impede FDS operation. Our five e-commerce platforms include online auctions, credit card processing, telecommunications, healthcare insurance, and auto insurance. We will go over the two main categories of online shopping fraud. Additionally, selected E-commerce sites' modern FDSs approaches are showcased. Here is a concise synopsis of the patterns and conclusions that will shape future study.

IV) Network Analysis (From Criminal Intelligence Analysis, P 67-84, 1990, Paul P Andrews, Jr and Marilyn B Peterson, ed. -- See NCJ-125011)

https://www.ojp.gov/ncjrs/virtual-library/abstracts/network-analysis-criminal-intelligence-analysis-p-67-84-1990-paul-p

Relational matrices and maps of relationships are described together since they both provide the same information. You may find the frequency, intensity, and strength of the relational linkages between the subjects of investigation using the diagrams and matrices. The presentation of mathematical models serves to demonstrate how conclusions may be derived from network models. We take a look at a sophisticated use of network analysis by the police. twelve figures and fourteen references.

V) Modelling Different Types Of Automobile Insurance Fraud Behaviour In The Spanish Market

https://www.sciencedirect.com/science/article/abs/pii/S0167668798000389

Controlling insurance fraud necessitates in-depth understanding of insureds' conduct, according to microeconomic theory. We quantify the impact of insured and claim variables on the risk of fraud using discrete-choice models that we propose in this research. The data is based on a sample from Spain. Oversampling of fraud claims necessitates correction for choice-based sampling in the estimation. Also covered is the organisation of the Spanish vehicle insurance sector. Our findings vary depending on the specific form of fraudulent activity being examined.

VI) Application Of Credit Card Fraud Detection: Based On Bagging Ensemble Classifier

https://www.sciencedirect.com/science/article/pii/S1877050915007103

More and more people are committing credit card fraud as a result of today's technology and global communication superhighways. Credit card fraud is a billion dollar industry because con artists are always thinking of new ways to steal money from unsuspecting customers and banks. Banks and other financial organisations require fraud detection systems in order to decrease losses. Researchers have limited access to credit caed transaction datasets, which has led to a dearth of literature on credit cadr fraud detection. Naïve Bayes(NB), Support Vector Machine(SVM), and K-Nearest Neighbour algorithms are some of the most used methods for detecting fraud. You may build classifiers using these approaches alone, or in conjunction with ensemble and meta-learning. Not because to its ease of implementation, but rather to its outstanding anticipated performance in real-world scenarios, ensemble learning has become the most preferred technique. In this study, we used design criteria to train and assess data mining techniques for detecting credit card fraud. We compared and tested several options before deciding that the bagging classifier from choice three would work best with the fraud detection model. Real credit card transactions are used to test the bagging ensemble approach and demonstrate its benefits.

VII) Fraud Detection Using Machine Learning And Deep Learning

https://www.researchgate.net/publication/339411416_Fraud_Detection_using_Machine_Learning_and_Deep_Learning

Frauds are difficult to detect because they are always changing and never follow a pattern. New technology is being used by con artists. Losing millions of dollars, they evade security inspections. Irregularities and dishonest financial dealings can be detected by data mining. transactions. This study compares several deep learning techniques with ML techniques such as autoencoders, CNN, RBM, and DBN, and with ML techniques such as random forest and SVM. Databases from Germany, Australia, and the European Union (EU) will be utilised. Alpha, beta, and cost of failure are the three metrics used for assessment.

VIII) A Synthesis Of Fraud-Related Research

https://www.researchgate.net/publication/277509802_A_Synthesis_of_Fraud-Related_Research

For two reasons: first, to get a feel for the scope and character of fraudulent financial reporting, and second, to spot where the current body of knowledge is lacking. This research expands upon earlier work by Hogan et al. (2008), who synthesised comparable data in 2005 for the AICPA's Auditing Section. To create this synthesis, we combed through the literature on fraud in accounting and auditing since Hogan et al. (2008) and summed up relevant studies from other domains. Review topics include accounting and related disciplines such as sociology, psychology, criminology, ethics, and organisational behaviour. In order to draw conclusions from the research, the auditor employed their fraud approach model. Auditors consider potential fraud schemes and concealment techniques, as well as the client's anti-fraud measures (such as corporate governance mechanisms and internal controls), when assessing the client's fraud risk. They also use the fraud triangle, which consists of management's incentive, attitude, and opportunity to commit fraud. Additionally, the model demonstrates how auditors may utilise this evaluation to uncover fraud by adhering to procedures for fraud detection. We suggest more research after reviewing the most recent literature on each part of the model.

IX) Financial Fraud, Scandals, And Regulation: A Conceptual Framework And Literature Review

https://www.researchgate.net/publication/328109001_Financial_fraud_scandals_and_regulation_A_conceptual_framework_and_literature_review

The purpose of this views essay is to provide a theoretical framework for the analysis of fraudulent and other dubious financial practices by reviewing relevant business history publications. Characteristics of individuals, corporate

management and oversight, economic conditions, and laws determine the breadth and depth of the analyses. Modern accounts, often presented as scandals, shed light on questionable actions and have the potential to inspire changes in regulations. The social and economic impact of dubious actions determines whether regulations are tightened. Reform agendas have a significant impact on regulations.

X) Histopathological Image Classification And Vulnerability Analysis Using Federated Learning

https://www.semanticscholar.org/paper/Histopathological-Image-Classification-and-Analysis-Vyas-Patra/4aa14ab241c56a98df2d8ab838c61d575e17d267

The healthcare industry is a major user of machine learning. In the old days of machine learning, central servers would collect data from all over the place and use it to train models that could predict new data outcomes. Since models could get their hands on personal user information, this is a major breach of privacy. One such solution is federated learning (FL): When clients train global models, the models receive their copies, and the clients then feed them information about their weight changes. Over time, global models become more and more precise. Data privacy is safeguarded by local training on customers' devices. Contamination of data can impact the worldwide model. We show how to attack a skin cancer dataset using data poisoning and build a privacy-preserving FL approach. While ten customers train the model, one intentionally uses flipped labels to attack. The accuracy of the global model declines. The accuracy drops as the number of label flips increases. Through optimisation using stochastic gradient descent, we get the highest possible accuracy for the model. Even though FL safeguards user privacy for medical diagnostics, data poisoning still has to be addressed.

III. Methodology

A. Proposed Undertaking:

The proposed system enhances UPI fraud detection by leveraging machine learning algorithms, including Random Forest, Decision Tree, Gradient Boosting Machines (GBMs), and XGBClassifier. Random Forest improves accuracy by combining multiple decision trees, reducing overfitting. Decision Tree provides an interpretable model that efficiently splits data based on feature importance. GBMs refine predictions iteratively, increasing precision, while XGBClassifier, an optimized version of GBMs, enhances speed and scalability. These models work together to classify transactions as either "Transaction Failed: Incorrect Details Entered" or "Transaction Successful: Details Verified and Processed."

By integrating these algorithms, the system ensures robust fraud detection while minimizing false positives. It continuously learns from transaction data, adapting to emerging fraud patterns for improved accuracy. The real-time detection capability provides an additional layer of security, ensuring that legitimate transactions are processed seamlessly while fraudulent ones are effectively identified and blocked.

B. Design Of The System:

The architecture of the proposed UPI fraud detection system consists of multiple stages, beginning with data collection and preprocessing. Transaction details such as the bank book name, transaction ID, and transaction amount are gathered from digital payment records. These raw data points undergo preprocessing, including handling missing values, feature selection, and normalization, ensuring that the machine learning models receive clean and structured input for accurate fraud detection.

Next, the processed data is fed into a multi-algorithm classification framework comprising Random Forest, Decision Tree, Gradient Boosting Machines (GBMs), and XGBClassifier. Each algorithm independently analyzes transaction patterns and assigns a classification label. The results from these models are then aggregated to determine whether a transaction is "Transaction Failed: Incorrect Details Entered" or "Transaction Successful: Details Verified and Processed." The final output is then integrated into the UPI system, enabling real-time fraud detection and prevention while ensuring legitimate transactions are processed smoothly.

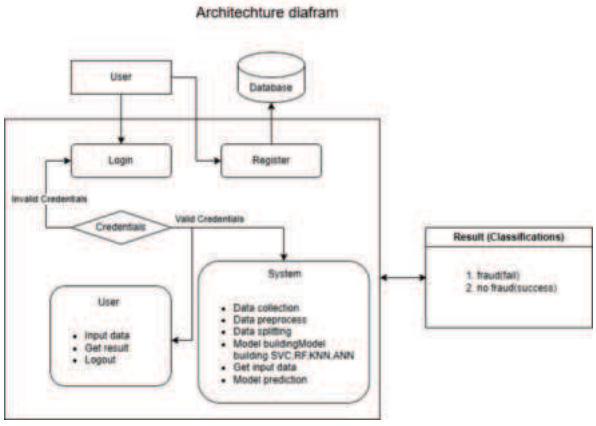

Fig.1. Proposed Architecture

IV. Implementation

1. Modules:

A. User Module

A) Input Model: The user is required to provide specific input values related to UPI transactions, such as the bank book name, transaction ID, and transaction amount. These details are essential for the fraud detection system to analyze and classify the transactions accurately.

B) Upload Dataset: Users can upload a dataset containing multiple transaction records in a structured format (CSV file). This dataset is used by the system to train and evaluate the fraud detection model, allowing it to learn patterns of fraudulent and legitimate transactions.

C) View Results: After processing the data, the system presents the classification results to the user. Each transaction is labeled as either "Transaction Failed: Incorrect Details Entered" or "Transaction Successful: Details Verified and Processed." This helps users in identifying fraudulent activities.

D) View Score: The system calculates the model's performance and displays the fraud detection accuracy as a percentage. This allows users to assess how effectively the machine learning algorithms classify transactions and detect fraud.

B. System Module

E) Dataset Handling: The system first verifies the availability of data and ensures that the uploaded dataset is in the correct format. It loads the transaction details from CSV files and organizes them for further processing. The quality of input data directly affects model performance.

F) Data Preprocessing: This step involves cleaning and transforming the dataset to enhance the accuracy of fraud detection. It includes handling missing values, removing duplicate transactions, normalizing data, and selecting important features. Proper preprocessing ensures that the model learns meaningful transaction patterns and reduces noise in the data.

G) Training The Data: The dataset is divided into training and testing subsets to improve model reliability. The training data is used to teach the machine learning models how to distinguish between fraudulent and legitimate transactions, while the test data is used to evaluate the model's predictive accuracy.

H) Model Building: Various machine learning algorithms, including Random Forest, Decision Tree, Gradient Boosting Machines (GBMs), and XGBClassifier, are applied to build a robust fraud detection model. Each algorithm has unique advantages, such as handling large datasets, capturing non-linear patterns, and reducing overfitting, which together improve detection efficiency.

I) Result Generation: Once the model is trained, it classifies incoming transactions based on the learned patterns. The system generates real-time outputs indicating whether a transaction is fraudulent or legitimate. These results help financial institutions and users take immediate action against suspicious activities.

J) Generated Score: The system calculates and displays the final accuracy score of the model as a percentage. This metric helps users determine the model's reliability in detecting fraudulent transactions. A high accuracy score ensures that legitimate transactions are processed smoothly while minimizing false fraud detections.

2. Algorithms:

The UPI Fraud Detection System utilizes multiple machine learning algorithms to enhance fraud detection accuracy and efficiency. The key algorithms used in the system are:

A) Decision Tree: This algorithm classifies transactions based on feature importance by creating a tree-like structure of decisions. It is easy to interpret but may overfit on complex datasets.

An algorithm that uses decision trees is called a random forest. As a decision-support tool, a decision tree resembles a tree. Gaining familiarity with decision trees can facilitate comprehension of random forest techniques.

Nodes at the decision, leaf, and root levels make up decision trees. The transaction dataset is partitioned into additional branches using a decision tree approach. At a leaf node, the series terminates. There is no differentiation at the leaf node.

Nodes in a decision tree represent characteristics that are utilised to make predictions. Decided nodes link leaves. The following diagram shows the three types of nodes in a decision tree.

B) Random Forest: When it comes to regression and classification problems, machine learning employs random forests. In ensemble learning, many classifiers are used to tackle complex problems.

An array of decision trees is employed by a random forest approach. The 'forest' that the random forest approach uses to train is accomplished by bagging or bootstrap aggregation. Machine learning accuracy is improved by using the ensemble meta-algorithm bag.

The outcome is determined by the decision tree projections using the random forest approach. It averages the output of trees to make predictions. When there are more trees, the results are more accurate.

Decision trees are limited, whereas random forests are not. This lessens the likelihood of dataset overfitting and increases accuracy. Like Scikit-learn, it generates predictions without requiring a plethora of package options.

Features Of A Random Forest Algorithm:

- Better at managing missing data and more accurate than decision tree algorithms.
- It is possible to make accurate predictions without adjusting the hyperparameters.
- Addresses problems with decision tree overfitting.
- The splitting point of each node in a random forest tree is used to randomly choose a subset of attributes.

C) Gradient Boosting Machines (GBMs): A boosting technique that improves prediction performance by combining weak learners sequentially. It captures complex, non-linear fraud patterns and enhances classification precision.

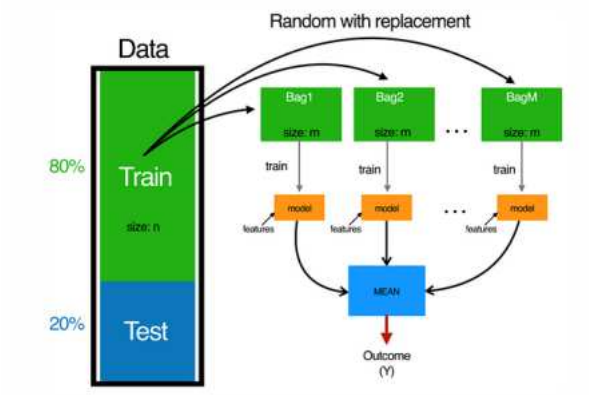

Fig 2. Random forest process

D) XGBClassifier (Extreme Gradient Boosting): Improving GBM performance on massive datasets in terms of speed, scalability, and efficiency. For real-time fraud detection, regularisation is ideal since it manages missing data and reduces overfitting.

XGBoost is a scalable machine learning system that uses distributed gradient-boosted decision trees (GBDTs). With parallel tree boosting, it is the best machine learning software for ranking, classification, and regression.

A solid grasp of supervised ML, decision trees, ensemble ML, and gradient boosting is necessary for fully grasping XGBoost.

Through the use of algorithms, supervised machine learning trains a model to recognise patterns in a dataset that already contains labels and features, and then uses that model to predict labels for new features.

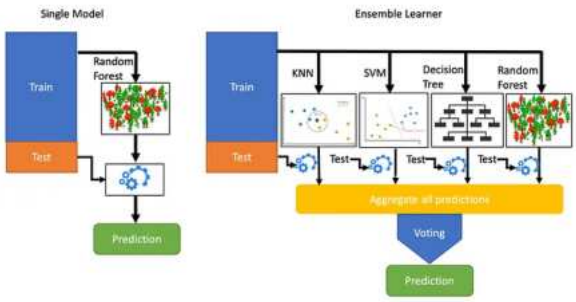

Fig 3. Xgoost

V. Experimental Results

The UPI Fraud Detection System was tested on a dataset containing real and fraudulent transactions to evaluate its effectiveness. The system was trained using Decision Tree, Random Forest, GBMs, and XGBClassifier, and performance metrics such as accuracy, precision, recall, and F1-score were analyzed. Among the models, XGBClassifier demonstrated the highest accuracy due to its efficient handling of missing values and overfitting prevention techniques. Random Forest provided robust fraud detection with reduced false positives, while GBMs effectively captured complex fraud patterns. The system successfully detected fraudulent transactions with high precision and recall, minimizing false alarms while ensuring security. These results confirm that the proposed

multi-algorithm approach significantly enhances fraud detection accuracy, real-time processing, and adaptability to emerging fraud trends in UPI transactions.

I) **Precision:** Accuracy is measured by precision when it comes to positively classifying instances or samples. The correctness formula is:

Precision = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

II) **Recall:** The ability of a model to identify all important instances of a class is measured by its machine learning recall. To determine if a model successfully captures class instances, we compare the total number of positive observations to the number of ones that were accurately predicted.

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

III) **Accuracy:** The accuracy of a model can be measured by looking at the percentage of valid classification predictions.

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

IV) **F1 Score:** Use the F1 Score, a harmonic mean of recall and accuracy, to equalise false positives and negatives if your dataset is not uniform.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

Fig.4. Home Page

Fig.5. Fraud Transaction

Fig.6. Legitimate Transaction

VI. CONCLUSION

The proposed system effectively detects fraudulent transactions using advanced machine learning algorithms such as Random Forest, Decision Tree, Gradient Boosting Machines, and XGBoost. By leveraging pre-processing techniques and model training, the system ensures high accuracy in identifying fraudulent activities. The user-friendly interface allows users to input data, view results, and check model performance with ease. This approach enhances fraud detection efficiency, reducing financial losses and improving security in digital transactions.

VII. Future Scope

Future enhancements can focus on integrating deep learning techniques like LSTMs and CNNs to analyze sequential transaction patterns for better fraud detection. Additionally, real-time monitoring with AI-driven automation can be implemented to detect anomalies instantly. Expanding the dataset with real-world transactions and improving feature engineering techniques will further enhance model accuracy and reliability. Moreover, blockchain technology can be integrated to ensure transaction transparency and security.

REFERENCES

[1] UKFinance. (2022). Annual Fraud Report 2022. [Online]. Available: https://www.ukfinance.org.uk/policy-and-guidance/reports-and publications/annual-fraud-report-2022
[2] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," J. Netw. Comput. Appl., vol. 68, pp. 90–113, Jun. 2016.
[3] A. Pascual, K. Marchini, and S. Miller. (2017). 2017 Identity Fraud: Securing the Connected Life. Javelin. [Online]. Available: http://www. javelinstrategy. com/coverage-area/2017-identity-fraud
[4] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data miming for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, no. 3, pp. 602–613, Feb. 2011
[5] L. T. Rajesh, T. Das, R. M. Shukla, and S. Sengupta, "Give and take: Federated transfer learning for industrial IoT network intrusion detection," 2023, arXiv:2310.07354.
[6] S. Vyas, A. N. Patra, and R. M. Shukla, "Histopathological image classification and vulnerability analysis using ," 2023, arXiv:2306.05980.
[7] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," Stat. Sci., vol. 17, no. 3, pp. 235–255, Aug. 2002.
[8] H. van Driel, "Financial fraud, scandals, and regulation: A conceptual framework and literature review," Bus. Hist., vol. 61, no. 8, pp. 1259–1299, Nov. 2019.
[9] G. M. Trompeter, T. D. Carpenter, N. Desai, K. L. Jones, and R. A. Riley, "A synthesis of fraud-related research," AUDITING, A J. Pract. Theory, vol. 32, no. Supplement 1, pp. 287–321, May 2013.
[10] P. Raghavan and N. E. Gayar, "Fraud detection using machine learning and deep learning," in Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE), Dec. 2019, pp. 334–339.
[11] M. Zareapoor and P. Shamsolmoali, "Application of credit card fraud detection: Based on bagging ensemble classifier," Proc. Comput. Sci., vol. 48, pp. 679–685, 2015.