Unified Payments Interface Fraud Detection using Machine Learning

P. Sirisha[1,] S. Jaheda[2], G. Afra Tahaseen[3], K. Divya Madhuri[4], P. Mohammad Arshad[5]
1 Assistant Professor Dept. of CSE, Srinivasa Ramanujan Institute of Technologv, Anantapur, India;
2,3,4,5.Students, Dept. of CSE(Data Science), Srinivasa Ramanujan Institute of Technologv, Anantapur, India;
Email: sirishap.cse@srit.ac.in

***Abstract*— Digital transaction fraud has become more common due to the widespread use of the Unified Payments Interface (UPI). Our proposed fraud detection system employs six machine learning algorithms to address this issue: XGBClassifier, Decision Tree, Random Forest, and Gradient Boosting Machines (GBMs). A logical framework for classifying transactions is provided by the Decision Tree algorithm. By enhancing accuracy and resilience, Random Forest successfully detects fraud. By combining weak learners, GBMs are able to identify changing fraud trends over time and capture intricate fraud patterns. training the model to converge efficiently. When it comes to classification jobs, XGBClassifier is a formidable gradient boosting method. Quick, it deals with missing values, and it stops overfitting. Enhancing UPI security, this multi-algorithm technique processes UPI transactions securely and accurately, differentiating between fraudulent and authentic ones. We are on the cusp of seeing this paradigm implemented in actual financial systems.**

Keywords— Machine Learning, XGBClassifier, Decision Tree, Random Forest, UPI Digital Payments, and Fraud Detection.

I. Introduction

The Unified Payments Interface (UPI) has transformed the digital payment landscape, enabling instant and seamless financial transactions. UPI allows users to send and receive money using their mobile devices without requiring traditional banking details such as account numbers and IFSC codes. Its convenience and accessibility have led to rapid adoption across individuals, businesses, and financial institutions. However, as digital transactions increase, so do the risks associated with fraudulent activities. Cybercriminals exploit vulnerabilities in the system using tactics like phishing, fake payment requests, identity theft, and transaction manipulation. These frauds pose a serious threat to both users and financial institutions, leading to monetary losses and decreased trust in digital payments.

Traditional fraud detection methods rely on rule-based approaches, where predefined conditions flag potentially fraudulent transactions. However, these static rules often fail to detect new and sophisticated fraud patterns. Fraudsters continuously evolve their tactics, making it essential to implement intelligent and adaptive fraud detection systems. This is where machine learning (ML) comes into play. ML-based fraud detection systems can analyze vast amounts of transaction data, learn from historical fraud patterns, and automatically identify suspicious activities. By leveraging advanced ML algorithms, we can enhance fraud detection accuracy, reduce false positives, and improve real-time security measures in UPI transactions.

In this study, we propose a machine learning-driven UPI fraud detection system that utilizes six powerful algorithms: Decision Tree, Random Forest, Gradient Boosting Machines (GBMs), and XGBClassifier. These models help in detecting fraudulent transactions with high precision. Decision Tree provides a clear decision-making process for transaction classification. Random Forest enhances detection accuracy by combining multiple decision trees, making it robust against fraudulent anomalies. Gradient Boosting Machines (GBMs) use weak learners to identify evolving fraud patterns, while XGBClassifier, a fast and efficient gradient boosting algorithm, prevents overfitting and handles missing values effectively.

By implementing this multi-algorithm approach, our system ensures secure, accurate, and real-time fraud detection in UPI transactions. Financial institutions can integrate this model into their existing security infrastructure to prevent fraudulent activities and protect users. The proposed system not only improves fraud detection efficiency but also contributes to a safer digital payment ecosystem, enhancing user confidence in UPI transactions.

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 form 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.

III. Methodology

A. *Proposed Undertaking:*

The proposed system enhances UPI fraud detection by integrating multiple machine learning algorithms, including Random Forest, Decision Tree, Gradient Boosting Machines (GBMs), and XGBClassifier. This multi-algorithm approach leverages the strengths of each model to improve accuracy, robustness, and real-time fraud detection. Random Forest, an ensemble method, builds multiple decision trees and aggregates their predictions, making it effective in handling noisy and imbalanced data while capturing intricate feature relationships. Decision Tree, a simple yet interpretable model, provides clear decision paths but is prone to overfitting on complex datasets. However, it serves as a foundational component for advanced ensemble methods. GBMs enhance performance by refining predictions iteratively, making them adept at detecting non-linear fraud patterns within large datasets. XGBClassifier, an optimized version of GBMs, offers high speed, scalability, and regularization techniques to prevent overfitting while handling missing values efficiently. This system ensures enhanced accuracy, robustness against overfitting, real-time detection capabilities, and adaptability to evolving fraud patterns. By leveraging these machine learning techniques, the proposed model significantly strengthens UPI security, ensuring faster, more reliable fraud detection and protecting users from financial risks.

B. *Design of the System:*

The architecture of the proposed UPI fraud detection system is designed to efficiently analyze transaction data, identify fraudulent patterns, and provide real-time security measures. The system follows a structured approach, starting with data collection and preprocessing, where transaction data, including amount, sender and receiver details, timestamps, and transaction types, are gathered and cleaned to remove inconsistencies. Feature engineering is then applied to extract relevant attributes that contribute to fraud detection.

Once the data is prepared, it is fed into multiple ML models, including RF, DT, GBMs, and XGBClassifier. Each model processes the input data to identify anomalies and patterns indicative of fraudulent activities. Random Forest, with its ensemble of decision trees, enhances detection accuracy by reducing overfitting, while Decision Tree provides an interpretable decision-making framework. GBMs and XGBClassifier further refine predictions, capturing complex relationships and improving classification precision.

The system integrates a real-time detection module, which continuously monitors transactions and flags suspicious activities. When a potentially fraudulent transaction is detected, an alert mechanism is triggered, notifying users and financial institutions for further action. Additionally, the model undergoes continuous training and updating to adapt to emerging fraud techniques, ensuring long-term effectiveness. This architecture ensures a robust, scalable, and adaptive fraud detection system, providing enhanced security for UPI transactions.

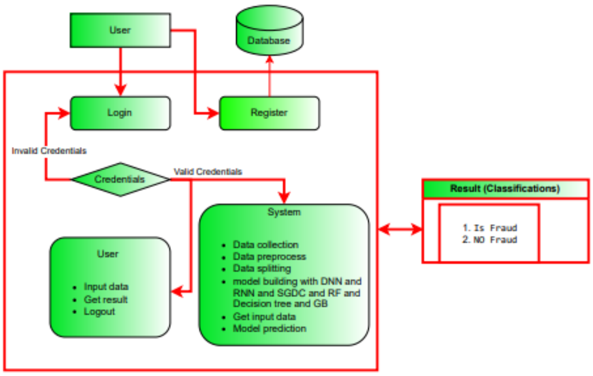

Fig.1. Proposed architecture

IV. Implementation

1. *MODULES:*

The UPI Fraud Detection System is able to identify and stop fraudulent transactions because to its many key components. System primary components:

a) *Data Collection and Preprocessing:* This module gathers transaction data, including sender and receiver details, transaction amount, timestamps, and payment methods. The data is then cleaned, removing duplicates and handling missing values to ensure accurate model training.

b) *Feature Engineering:* In this module, relevant features are extracted and transformed to improve fraud detection accuracy. Key features include transaction frequency, average transaction amount, and user behavior patterns, which help in distinguishing fraudulent activities from legitimate ones.

c) *Machine Learning Model Training:* The system trains multiple ML models—RF, DT, GBMs, and XGBClassifier—on historical transaction data. These models learn fraud patterns and classify transactions based on their likelihood of being fraudulent.

d) *Real-Time Fraud Detection:* This module applies trained models to live transactions, analyzing them in real-time. If a transaction is flagged as suspicious, an alert is triggered for further verification.

e) *Alert and Notification System:* When fraudulent activity is detected, the system sends alerts to the user and financial institutions, enabling immediate action to prevent losses.

f) *Model Update and Continuous Learning:* Fraud patterns evolve over time, so this module updates the machine learning models with new transaction data, ensuring they remain effective against emerging fraud techniques.

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 training 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.

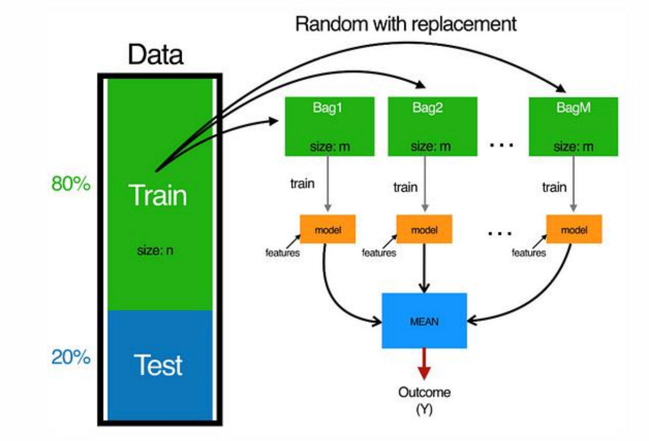

Fig2. Random forest process

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.

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.

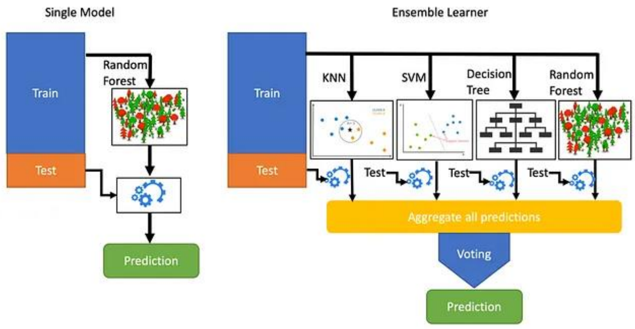

Fig3. XGoost

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:

$$Precision = TP / (TP + FP)$$

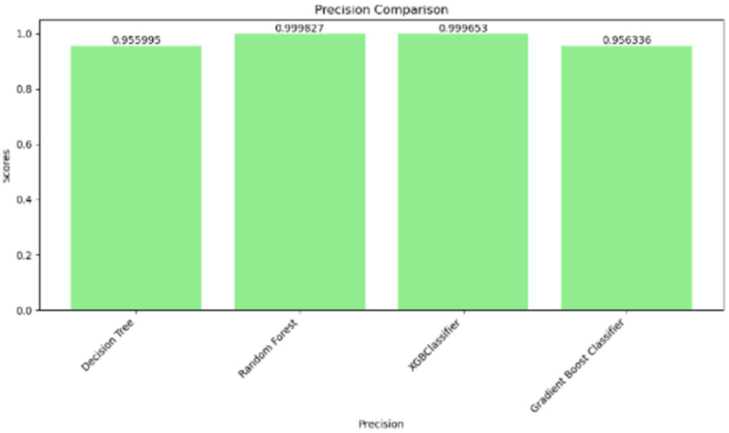

Fig 4 Precision comparison graph

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.

$$Recall = TP / TP + FN$$

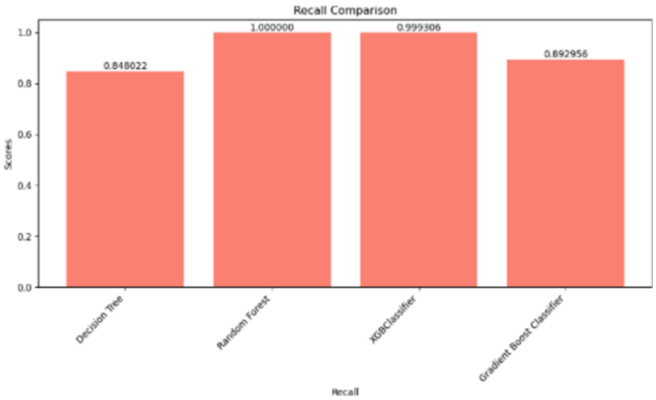

Fig.5. Recall comparison graph

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

$$Accuracy = TP+TN / TP+FP+TN+FN$$

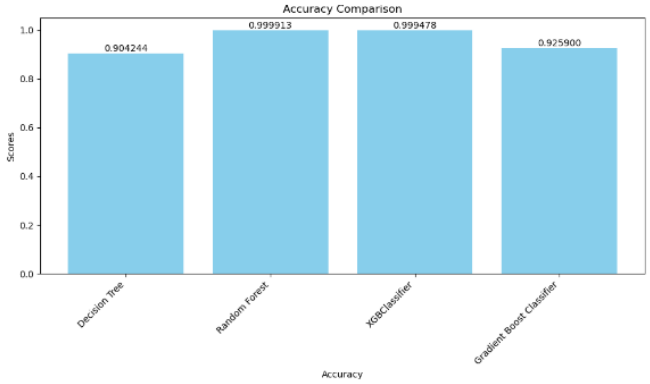

Fig.6. Accuracy graph

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.

*F1 Score=2 * Recall * Precision / Recall + Precision * 100*

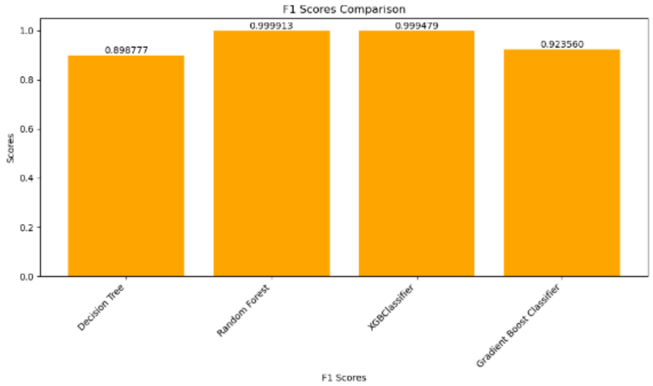

Fig.7. F1 Score

VI. Conclusion

By fixing problems with privacy and openness, Federated Learning (FL) and Explainable AI (XAI) revolutionise the identification of financial crime. Model interpretability and performance are enhanced compared to prior approaches by XGBClassifier, Random Forest, Decision Trees, and Gradient Boosting Machines (GBMs).

Financial organisations can better understand the classification results provided by Random Forest and Decision Trees, while XGBClassifier and GBMs are able to capture complex patterns for improved fraud detection. By eliminating the need to share sensitive financial data, Federated Learning (FL) enables decentralised model training.

This two-pronged approach makes the framework more private and transparent while also enhancing fraud detection. The findings demonstrate that financial fraud detection systems may be enhanced by combining XAI with FL, all while maintaining privacy requirements. A more stable and accountable financial system may be the result of this paradigm shift.

VII. Future Scope

More accurate, interpretable, and private fraud detection systems are available for future usage.

To enhance the accuracy of fraud detection, it is recommended to combine XGBClassifier with other hybrid models such as DT, RF, and GBMs. Transformer models and other deep learning architectures have the potential to detect intricate fraud patterns.

FL frameworks that have their communication strategies fine-tuned can improve the efficiency and accuracy of models used by financial organisations while keeping customer data private.

•By using differentiated privacy techniques, sensitive financial data may be protected in FL, reducing the risk of data leakage and making fraud detection easier.

With the use of streaming data and adaptive learning algorithms, fraud may be detected and prevented in real-time, reducing reaction time and financial losses. It is possible to improve the scalable and fast XGBClassifier and GBMs for use in real-time scenarios.

References

[1] [1] UKFinance. (2022). Annual Fraud Report 2022. [Online]. Available: https://www.ukfinance.org.uk/policy-and-guidance/reports-andpublications/annual-fraud-report-2022

[2] [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] [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] [4] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, ‘‘Data mining for credit card fraud: A comparative study,’’ Decis. Support Syst., vol. 50, no. 3, pp. 602–613, Feb. 2011

[5] [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] [6] S. Vyas, A. N. Patra, and R. M. Shukla, ‘‘Histopathological image classification and vulnerability analysis using ,’’ 2023, arXiv:2306.05980.

[7] [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] [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] [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] [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] [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.