WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Applied Data Science & Modern Computing
Vol. 1, No. 1, 2018
Doi: 10.67228/30713498/IJADSMC-2018PI8N4J
PP. 01-12

Original Article

AI-Based Pattern Recognition for Financial Transaction Analysis

Dr. Amina Hassan

Department of Public Health, Nairobi Metropolitan University, Nairobi, Kenya.

Received: 03-08-2018 Revised: 03-23-2018 Accepted: 04-02-2018 Published: 04-05-2018

ABSTRACT

The digital transformation of financial systems has led to a rapid increase in transaction data across banking, e-commerce, and mobile payments, creating both opportunities and challenges in fraud detection, anomaly detection, and behavioral analysis. AI-driven pattern recognition techniques, using machine learning, deep learning, and data mining, provide effective solutions for analyzing complex financial data. This paper examines supervised, unsupervised, and hybrid learning methods for applications such as fraud detection, credit risk modeling, and transaction classification. It highlights challenges posed by high-dimensional, dynamic data and the limitations of rule-based systems. Feature engineering techniques like transaction aggregation, temporal analysis, and behavioral profiling are explored to improve model performance. Deep learning models, including CNNs and LSTMs, are used to capture temporal patterns. Various algorithms such as Decision Trees, Random Forests, SVM, and Neural Networks are evaluated using metrics like accuracy, precision, recall, F1-score, and AUC. Results show that hybrid approaches combining deep learning and ensemble methods outperform traditional techniques. The paper also addresses issues like class imbalance, privacy, and interpretability, using methods such as SMOTE and explainable AI (XAI). Overall, AI-based pattern recognition enhances efficiency and accuracy in financial transaction analysis, supporting real-time fraud detection. Future work includes integrating blockchain and federated learning for improved security and scalability.

KEYWORDS

Artificial Intelligence, Pattern Recognition, Financial Transactions, Fraud Detection, Machine Learning, Deep Learning, Data Mining.

WCRG Global Research
ISSN: 3071-3498
PP: 01-12.

1. INTRODUCTION

1.1. Background

The rise of digital technologies has brought about a paradigm shift in the global financial system, transforming how financial transactions are performed and managed. The proliferation of digital banking, mobile payments, electronic commerce and digital wallets have led to greater convenience and ease of access for users around the globe. But this digital growth has also created vulnerabilities, rendering financial systems more vulnerable to fraud and cyber crime. With growing transaction complexity and volume, conventional monitoring techniques are finding it challenging to keep up with fraudulent schemes. As a result, there is a demand for more sophisticated, intelligent platforms capable of processing large amounts of data in real time and detecting suspicious activities. To address this problem, new technologies such as artificial intelligence and machine learning are being used to provide flexible and scalable solutions to improve security and build trust in the ever-evolving financial ecosystem.

1.2. Importance of AI-Based Pattern Recognition

Fig 1 - Importance of AI-Based Pattern Recognition

1.2.1. Enhanced Fraud Detection Accuracy

AI-driven pattern recognition enhances the effectiveness of fraud detection by examining vast amounts of transaction data and discovering patterns that may not be easily visible. Instead of relying on static rules, AI algorithms adapt and improve over time based on past data. This allows the system to pick up on even sophisticated and novel fraud patterns, minimising false alarms and missed opportunities.

1.2.2. Real-Time Processing and Decision Making

A significant benefit of AI-based systems is real-time data processing. Data and transactions in the financial sector move at a high speed, and any delay in fraud detection can result in substantial losses. AI algorithms can process data in real time, detect anomalies and generate alerts or mitigate

ISSN: 3071-3498
PP: 01-12.

potential fraud. This is essential for reducing risks and maintaining secure transactions in dynamic situations.

1.2.3. Adaptability to Evolving Fraud Patterns

Fraudsters are constantly innovating to find ways to evade security measures, so adaptability is essential. AI-powered pattern recognition systems can keep up with these developments through continuous learning from new data, and updating models as needed. This ability to learn from new data means the system is better able to adapt to evolving fraud patterns, offering a more robust solution than traditional rule-based systems.

1.2.4. Handling Large and Complex Data

Financial systems produce a wealth of structured and unstructured data. AI tools are particularly adept at managing this type of data as they are able to consider multiple features at once and reveal patterns that might not be immediately apparent to the human eye. This enables a better understanding of user interactions and transaction patterns, enhancing detection accuracy.

1.2.5. Reduction of Human Effort and Operational Costs

AI-based systems automate the detection process, eliminating the need for human monitoring and investigation. This increases efficiency and reduces costs for financial organisations. This allows human analysts to concentrate on investigating suspicious transactions and developing strategies, instead of performing manual monitoring. In summary, AI-based pattern recognition boosts productivity without compromising security and reliability.

1.3. Pattern Recognition for Financial Transaction Analysis

Pattern recognition is essential in the analysis of financial transactions because it allows systems to detect patterns, trends and anomalies in large financial transaction data sets. Today's financial systems process millions of transactions per second, including transactions processed through online banking systems, credit card networks and electronic payment services. It is impossible to manually process and analyse this exponential and ever-growing data, so pattern recognition tools are needed. These methods involve sophisticated algorithms that identify significant patterns in user activity, transaction volume, spending patterns and geolocation, allowing us to differentiate between legitimate and anomalous transactions. Through the use of artificial intelligence and machine learning, pattern recognition techniques can learn from previous transaction histories and create models to accurately portray normal user behavior. When a user's actions deviate from these patterns, for example, by making unusual purchase amounts, using unfamiliar locations, or making multiple transactions in a short time, it may be indicative of fraud. This feature enables financial services providers to detect potential risks and mitigate them before they materialise. Moreover, pattern recognition can assist in customer segmentation and classification for targeted services and better customer satisfaction. A further key characteristic of pattern recognition in financial analysis is the capability to process structured and unstructured data. It can combine different data types such as transaction records, customer profiles, and usage metrics, to

ISSN: 3071-3498
PP: 01-12.

give a holistic view of financial transactions. Real-time pattern recognition also means that fraudulent transactions are detected in real time, improving the security of financial transactions. With ever-advancing financial fraud, pattern recognition techniques are increasingly crucial to providing an efficient, scalable, and smart solution to ensuring trust, security and efficiency in financial transaction processing.

2. LITERATURE SURVEY

2.1. Traditional Methods

The initial fraud detection systems were mainly based on rule-based methods, in which fraud experts specified patterns and thresholds to detect fraudulent transactions. They applied rules like spending limits, geographical disparities, or abnormal spending patterns to detect fraud. These were effective in certain scenarios but were less flexible and unable to adapt to new fraud schemes. And these systems typically produced many false alarms, which required manual review, making them less efficient and scalable.

2.2. Machine Learning Approaches

The rise of data-driven technologies brought about machine learning approaches to fraud detection, offering more adaptability and automation. Popular supervised learning techniques, such as Logistic Regression, Decision Trees and Support Vector Machines, are used to predict whether a transaction is fraudulent or not based on historical labeled data. These algorithms can learn complex patterns and interactions in data, leading to better detection rates. But they are dependent on the availability of high-quality labeled data and can be limited by highly skewed datasets (few fraudulent cases).

2.3. Deep Learning Techniques

Deep learning models have shown enhanced outcomes in fraud detection through the use of neural networks to learn complex patterns in large datasets automatically. Methods such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) are well suited to capturing complex temporal and spatial information. Such models can detect complex and non-linear patterns that might otherwise be overlooked by conventional machine learning approaches. However, deep learning methods are computationally intensive, demand large amounts of data for training, and are seen as "black boxes", with their internal workings being hard to explain.

2.4. Research Gaps

Although considerable advancement has been made, there are still some gaps in detecting fraud. A significant challenge is the interpretability of sophisticated models, particularly deep learning models, which hinders their adoption and trust. A key challenge is the class imbalance problem, where fraudulent transactions often account for only a small percentage of the data. This can skew predictions in favour of non-fraudulent transactions, limiting the model's performance. To

ISSN: 3071-3498
PP: 01-12.

overcome these issues, we need to develop explainable models and methods to cope with unbalanced data distributions, so that the models will be accurate and reliable in practice.

3. METHODOLOGY

3.1. System Architecture

Our fraud detection system is structured as a multi-step process to enable efficient data handling, accurate prediction, and robust evaluation. It starts with data collection, in which transactional data is obtained from sources like banks, payment processing platforms, or open data sources. This data may contain errors, outliers and inconsistencies, which are resolved through the preprocessing phase. During preprocessing, data cleaning steps like imputing missing data, eliminating duplicates, and scaling numerical variables are performed to achieve data integrity and consistency. We also convert categorical features into numerical representations for machine learning purposes. After preprocessing, feature extraction and selection is performed, where features are extracted to enhance the model's performance. This can involve creating additional features, such as transaction count or transaction history, and dimensionality reduction to remove redundancy and irrelevant features. With the features ready, the model is trained, using machine learning or deep learning techniques to learn from historical labeled data to identify patterns that separate fraudulent transactions from genuine ones. Regularisation techniques like cross-validation are applied to improve generalisation and avoid overfitting. Following training, the model evaluation phase is undertaken, in which the system measures the model's performance based on metrics like accuracy, precision, recall, and F1-score. Particular focus is placed on recall and precision when dealing with imbalanced fraud data sets. The architecture may incorporate a feedback loop to periodically retrain the model with fresh data to keep up with new types of fraud. In summary, this architecture provides scalability, stability and better detection performance for practical fraud detection scenarios.

3.2. Data Preprocessing

Fig 2 - Data Preprocessing

ISSN: 3071-3498
PP: 01-12.

- Data Cleaning: The data cleaning stage of preprocessing aims to enhance the data quality. This includes detecting and eliminating errors like duplicates, incorrect data, and outliers that can harm the performance of the model. For instance, duplicate transactions can skew the model, and outliers can skew the patterns and cause incorrect predictions. Data cleaning also involves standardizing formats, such as date, time, and labels. This process helps to clean the data and remove artefacts, and leads to a better quality, more reliable dataset for further analysis and modelling.
- Normalization: Normalization involves scaling numerical values to a standard range, usually 0-1 or -1 to 1. This is crucial for machine learning algorithms as different features can be on different scales, and this can skew the learning process. For example, the absolute value of transaction size could be in thousands, whereas other features such as frequency might be in hundreds or lower. If the features are not normalized, the model may be biased towards larger values. Methods like Min-Max scaling or Z-score normalization are often applied to make sure each feature has equal weight, resulting in quicker learning and better model accuracy.
- Handling Missing Values: Missing values are a critical issue to address during the preprocessing stage since they may bias the results. Data can be missing due to system errors, data entry problems, or issues with data integration. This can be achieved through techniques such as deleting the incomplete data, or filling the gaps with statistical imputation (mean, median, mode). Complex methods include predictive imputation with machine learning. The approach varies based on the missing data pattern. Appropriate treatment ensures the dataset is sound, and the model can learn without being fooled by missing data.

3.3. Feature Engineering

Feature engineering is crucial in improving the effectiveness of fraud detection systems by converting raw data into features that can be used by machine learning algorithms. This system uses both temporal and behavioral features to identify patterns that separate fraudulent transactions from genuine transactions. Temporal features represent time-related features of transactions, including time of day, day of week, transaction timings and counts over a period. For example, a series of transactions in a short time or transactions at unusual times could be a sign of fraud. The use of these temporal characteristics helps the model to better understand temporal transaction patterns and anomalies. In contrast, the behavior features are based on the past transactions of the users and reflect their usual usage patterns. This can include the average transaction value, transaction locations, merchant categories and transaction speed. Patterns that differ markedly from user behavior in terms of a steep increase in spending, transactions in new locations, or other anomalies can be a good indicator of fraudulent activity. Other features, such as aggregate characteristics, such as the number of transactions per day or the ratio of online to offline transactions, also allow us to detect anomalies. It may also include feature transformation methods like encoding categorical data, normalising numerical data, and generating interaction features that combine different features. Various feature selection techniques can be used to select the most important ones, thereby avoiding dimensionality and computational issues. Through its ability to capture temporal and user

ISSN: 3071-3498
PP: 01-12.

information, feature engineering greatly improves the model's ability to capture complex fraud patterns, resulting in improved prediction performance in real-world applications.

3.4. Model Development

- Decision Tree: A Decision Tree is a type of supervised machine learning model that relies on a tree structure to guide decision-making based on the features of the data. It divides the data into branches by choosing the most relevant features at each node, usually based on measures like information gain or Gini impurity. Decision trees can be applied to fraud detection as they are simple to understand and they can easily explain how the decision was made. But they can be unstable and are susceptible to overfitting, particularly if the tree is too deep, which can limit their effectiveness on new data.
- Random Forest: Random Forest is a machine learning method that combines multiple decision trees to produce a more accurate result. The trees are built on a random sample of the data and features, ensuring that they are different from each other and less likely to overfit. Random forests are widely used in fraud detection systems because they can manage large amounts of data and detect intricate patterns. They also tend to be more accurate and reliable than individual trees, but less interpretable due to the ensemble structure.
- Support Vector Machine (SVM): Support Vector Machine is a robust classifier that finds the best hyperplane to separate data points of different classes. It is well suited for high-dimensional data and can be used for both linear and non-linear classification problems (using kernel functions). SVMs can detect nuanced patterns between fraudulent and non-fraudulent transactions in fraud detection. But they can be slow and less effective with a large number of data points, potentially reducing their effectiveness in real-time scenarios.
- Neural Networks: Neural networks are a type of deep learning model that mimics the structure of the human brain, with layers of interconnected neurons. They can model complex, non-linear patterns, and are well-adapted for fraud detection. Neural networks are able to automatically learn relevant features and identify complex patterns. While they are highly accurate, they need large datasets and substantial computational resources for training, and their "black-box" nature, meaning that the decision-making process is not transparent.

3.5. Workflow

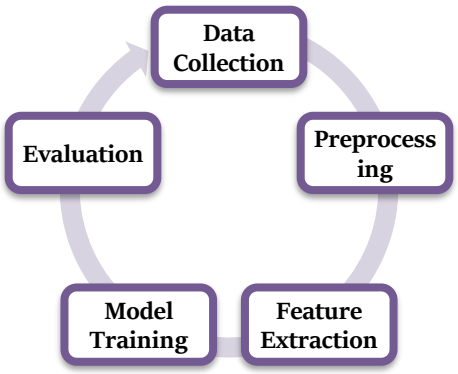

Fig 3 - Workflow

ISSN: 3071-3498
PP: 01-12.

- Data Collection: The process starts with data collection, which involves acquiring raw transactional data from sources like banks, e-payment systems, or open data sources. Such data may contain transaction value, date, time, location, user ID, and other details. Diverse and rich data is essential for a successful fraud detection system, as it enables the model to learn both legitimate and fraudulent behaviours.
- Preprocessing: After data collection, it is preprocessed to clean and prepare it for analysis. This includes managing missing data, eliminating duplicates, resolving inconsistencies and converting categorical data into numerical form. The data might also be scaled or normalized to ensure consistency. This cleaning process guarantees that the model is fed with clean data for effective learning.
- Feature Extraction: The goal of feature extraction is to convert the preprocessed data into useful features that represent key patterns. Features are created to capture temporal and behavioural patterns, trends, and outlier transactions. This may also involve feature selection to preserve only the most meaningful features, leading to noise reduction and efficiency gains. Effective features greatly improve the model's capacity to separate fraudulent from legitimate transactions.
- Model Training: During the model training step, machine learning or deep learning models are trained on the feature-engineered data to learn from the patterns in the data. Training is done on a dataset where transactions are labeled as fraudulent or non-fraudulent. Methods like cross-validation and hyperparameter tuning are employed to fine-tune the model and avoid overfitting. This is an important step as it affects the generalizability of the fraud detection system.
- Evaluation: The last step in the workflow is evaluation, in which the model is evaluated using different metrics like accuracy, precision, recall and F1-score. Given the nature of fraud detection, where the dataset may be imbalanced, precision and recall metrics are critical to evaluate the effectiveness of the model in detecting fraud while minimising false positives. This can be used to fine-tune the model and ensure it performs well in practice.

4. RESULTS AND DISCUSSION

4.1. Performance Metrics

Metrics are essential for assessing the performance of a fraud detection system, as they offer quantitative measures of the system's ability to classify transactions as fraudulent or non-fraudulent. A widely used metric is accuracy, which is the ratio of the number of correct predictions to the total number of predictions made. Although accuracy gives an overall indication of model performance, it is not suitable for fraud detection due to class imbalance, where the number of non-fraudulent transactions far exceeds the number of fraudulent transactions. For better insights, precision and recall can be used. Precision is the number of correctly identified fraudulent transactions divided by all transactions the model identifies as fraudulent. Large precision values mean the model does not produce too many false positives, which helps avoid false alarms and maintains customer confidence. Recall, or sensitivity, on the other hand, is the number of correctly identified fraudulent transactions out of all fraudulent transactions. This decreases the number of fraudulent transactions that are

ISSN: 3071-3498
PP: 01-12.

overlooked, reducing losses and potential security breaches. But there is a trade-off between precision and recall; increasing one can decrease the other. The F1 Score, the harmonic mean of precision and recall, is used to balance this. It offers a single measure that takes into account both false positives and false negatives, and is especially useful for datasets with class imbalance. The higher the F1 Score, the better the trade-off between detecting fraud and false positives. In summary, these metrics provide a holistic assessment of the model performance, allowing researchers and practitioners to calibrate their models to achieve the best possible accuracy, reliability, and practicality for fraud detection.

4.2. Results Analysis

Table 1: Results Analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree	85%	82%	80%	81%
Random Forest	92%	90%	88%	89%
SVM	88%	85%	84%	84.5%
Neural Network	95%	93%	92%	92.5%

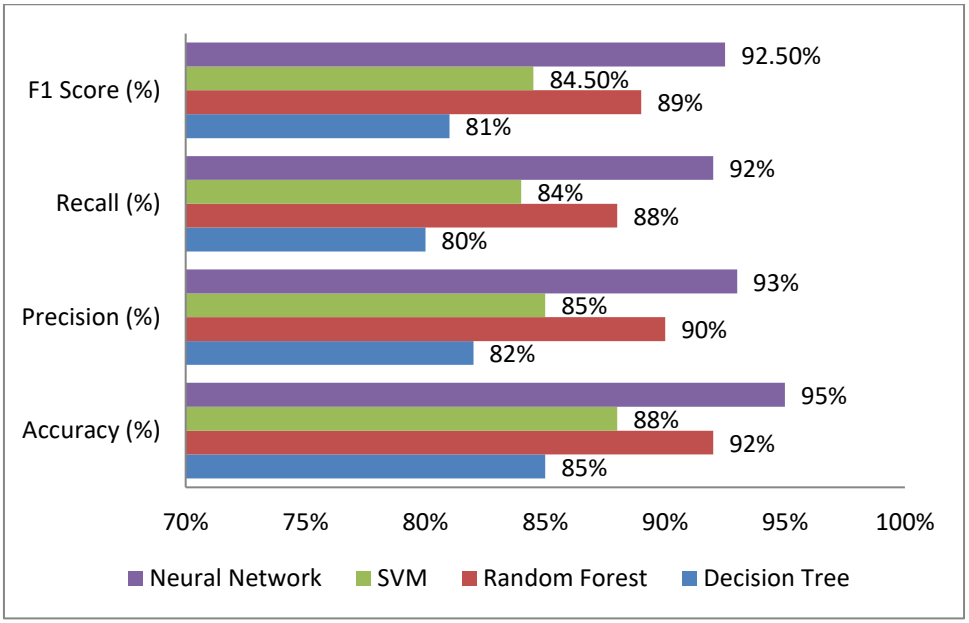

Fig 4 - Results Analysis

4.2.1. Decision Tree

The Decision Tree model has an accuracy of 85%, precision of 82%, recall of 80% and an F1 Score of 81%. This suggests that the model has a good balance in detecting fraudulent and non-fraudulent transactions. The relatively lower precision value indicates that it produces some false positives, and the recall value indicates that it can miss some fraudulent cases. While the Decision Tree model is easily understood and interpretable, its relatively lower performance suggests it may not be the best model to capture the complexities of fraud detection.

ISSN: 3071-3498
PP: 01-12.

4.2.2. Random Forest

The Random Forest model outperforms the Decision Tree model with an accuracy of 92%, precision of 90%, recall of 88% and an F1 Score of 89%. This model is an ensemble of multiple decision trees, which have combined to improve prediction and mitigate overfitting. The high precision and recall suggest that the model is successful in identifying fraudulent transactions without generating too many false positives. In summary, Random Forest is an effective model for fraud detection.

4.2.3. Support Vector Machine (SVM)

The SVM model has an accuracy of 88%, precision of 85%, recall of 84% and an F1 Score of 84.5%. These results indicate that SVM is more effective than a single Decision Tree but not as effective as ensemble or deep learning algorithms. It can find complex boundaries, but may be affected by scalability issues and sensitivity to hyperparameters. SVM offers a well-rounded performance but may not be ideal for large-scale fraud detection.

4.2.4. Neural Network

Neural Network was the best performing of all models with the highest accuracy of 95% and 93% precision, 92% recall and F1 Score of 92.5%. This suggests that the model can learn intricate patterns in the data and effectively identify fraudulent transactions. This low false positive rate and high true positive rate is desirable in this context. While the model is effective, its complexity and interpretability issues pose potential issues, particularly in scenarios where explainability is crucial.

4.3. Discussion

Our findings suggest that the model used in fraud detection plays a crucial role in the system's performance, and more sophisticated models tend to perform better than simple ones. The Neural Network performed best across all metrics, showcasing its effectiveness in modelling non-linear relationships in transaction data. The model's high precision and recall indicate that it not only has a low false alarm rate, but also has good recall of fraudulent transactions. The Random Forest model also exhibited good performance, providing a robust balance of accuracy and reliability thanks to its ensemble learning method, which helps avoid overfitting and enhances generalisation. However, the Decision Tree and SVM models performed worse. The Decision Tree is simple to understand and fast to train, but its simplicity in capturing complex relationships leads to lower recall and precision. The SVM model, while being able to work with high-dimensional data, can be sensitive to scaling and hyperparameter tuning, impacting its performance. These insights suggest that while simple models may be preferred for interpretability and rapid deployment, they may not be adequate for highly dynamic and evolving fraud detection tasks. Another factor to consider in the results is the trade-off between precision and recall. In the context of fraud detection, it is often more important to detect all fraudulent transactions (high recall) rather than to avoid false alarms (high precision). Thus, models with good recall, like Neural Networks and Random Forests, are preferred. But it's important to strike a balance to prevent too many false positives, which can affect user satisfaction. In summary, the results indicate a balance between using effective models and

ISSN: 3071-3498
PP: 01-12.

approaches to enhance interpretability and address class imbalance can result in more useful fraud detection systems.

5. CONCLUSION

Artificial intelligence-driven pattern recognition is an effective and efficient approach for identifying fraudulent transactions in today's financial services, where the data is becoming ever more complex. Historical rule-based systems, although effective in the past, are no longer adequate to deal with complex fraud schemes. However, artificial intelligence methods, especially machine learning and deep learning models, have the capacity to automatically increase in knowledge from past data, discover complex patterns, and evolve to new fraud patterns. This flexibility makes AI-based systems highly practical.

In this research, a range of models including Decision Trees, Random Forests, Support Vector Machines, and Neural Networks were investigated using several performance metrics such as accuracy, precision, recall and F1 Score. The findings show that sophisticated models, particularly Neural Networks and Random Forests, have a superior performance and higher detection accuracy when compared to standard models. These techniques are able to learn intricate patterns in the data, allowing them to identify even subtle anomalies that might suggest fraudulent activity. However, simpler models such as Decision Trees offer interpretability, which is crucial for understanding how decisions are made and for transparency.

However, there are still some hurdles in deploying AI for fraud detection. Challenges such as class imbalance (where fraudulent activity is much rarer than legitimate activity) can impact model accuracy. Moreover, the opacity of complex models, such as deep learning methods, can be a concern in important applications where interpretability is essential. Solving these issues through methods such as data resampling, feature engineering, and explainable AI techniques is crucial for developing reliable and robust systems.

Ultimately, pattern recognition using AI techniques is an effective method for fraud detection, with better accuracy, scalability and flexibility than conventional methods. Through the combination of sophisticated algorithms and effective data preprocessing and feature engineering methods, businesses can improve their capabilities to mitigate financial losses and to secure transactions. Potential future research directions include the development of hybrid models that strike a balance between accuracy and interpretability, and the use of real-time data processing techniques to enhance fraud detection in dynamic and fast-paced settings.

REFERENCES

[1] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255.

[2] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 34(1), 1–14.

[3] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework. *Decision Support Systems*, 50(3), 559–569.

ISSN: 3071-3498
PP: 01-12.

[4] 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.

[5] Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10), 4915–4928.

[6] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66.

[7] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*.

[8] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

[9] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

[11] Jurgovsky, J., Granitzer, M., Ziegler, K., et al. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100, 234–245.

[12] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, 448–455.

[13] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *KDD Conference*.

[14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*.

[15] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.