WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Fraud Detection In Internet Banking Using Machine Learning

Ms. Snober Iqbal[1], K. SreeNiketh Raj[2], K. Vineeth Reddy[3], Mohd. Mudaseer Mazharuddin[4]

[1] (Assistant Professor) Computer Science and Engineering(IOT) Guru Nanak Institutions Technical Campus Telangana, India
[2] Computer Science and Engineering(IOT) Guru Nanak Institutions Technical Campus Telangana, India rsreeniketh@gmail.com
[3] Computer Science and Engineering(IOT) Guru Nanak Institutions Technical Campus Telangana, India koppulavineeth2004@gmail.com
[4] Computer Science and Engineering(IOT) Guru Nanak Institutions Technical Campus Telangana, India Mudaseer2753@gmail.com

ABSTRACT –

Banking fraud pertains to unauthorized or misleading actions connected with bank accounts or monetary operations. Several machine learning strategies can facilitate the detection of such fraudulent practices. This inquiry delves into various methods apt for differentiating transactions as either fraudulent or legitimate. The analysis makes use of the Banking Fraud Transactions dataset, which is characteristically highly unbalanced. To navigate this difficulty, we deploy numerous machine learning strategies like Random Forest, K-Nearest Neighbour, and Decision Tree. Additionally, methods for selecting features are employed, and the dataset is partitioned into training and evaluation segments. The strategies evaluated include Random Forest and KNN. The results reveal that each strategy ensures considerable accuracy in pinpointing banking fraud. The proposed framework displays capability in detecting further discrepancies in financial operations.

Keywords - Internet Banking, Fraud Detection, Fraudulent Transactions, Random Forest (RF), K-Nearest Neighbours (KNN), Anti-Money Laundering (AML), Real-time Detection, Anomaly Detection, Python, Scikit-learn.

I. Introduction :

In recent years, the rapid expansion of digital banking services has revolutionized the financial sector, offering unparalleled convenience to customers worldwide. However, this digital transformation has also brought a surge in banking fraud, including identity theft, unauthorized transactions, and money laundering. These fraudulent activities pose significant risks to financial institutions, leading to monetary losses, reputational damage, and erosion of trust among customers.

Traditional rule-based fraud detection systems, although effective to an extent, struggle to adapt to the ever-evolving tactics employed by fraudsters, necessitating more advanced and adaptive solutions. Machine learning has emerged as a pivotal technology in the fight against banking fraud. Unlike traditional approaches, machine learning algorithms can analyse vast amounts of transactional data, identify complex patterns, and adapt to new fraud techniques in real-time. By leveraging historical data, these algorithms can discern subtle differences between legitimate and fraudulent transactions, enabling faster and more accurate detection. This capability is particularly critical in a domain where fraudulent transactions are often well-disguised and represent only a small fraction of the overall transaction volume.

This study addresses the multifaceted challenges of fraud detection in internet banking by developing a robust machine learning-based system. The primary focus is on tackling the class imbalance problem inherent in fraud detection datasets, where legitimate transactions significantly outnumber fraudulent ones. Our approach employs a combination of advanced algorithms, including Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression, to build an effective fraud detection model. These algorithms are carefully chosen for their complementary strengths—Random Forest for its ensemble learning capabilities, KNN for its simplicity and adaptability, and Logistic Regression for its ability to model the probability of a certain event occurring.

By harnessing the power of machine learning, financial institutions can enhance their fraud detection capabilities and mitigate the risks associated with banking fraud. Through the analysis of transactional data and the identification of complex patterns, these institutions can stay one step ahead of fraudsters and protect their customers' assets and trust.

II. Related work area :

M. Jullum, A. Løland, R. B. Huseby, G. A˚ nonsen, and J. Lorentzen (2020) research emphasizes the development and validation of a machine learning model tailored for prioritizing financial transactions for manual investigation, specifically targeting potential money laundering activities. The model, applied to a dataset from Norway's largest bank (DNB), utilizes supervised learning trained on three distinct categories of historical data: normal transactions, those flagged as suspicious by internal systems, and actual reported money laundering cases. The key finding was that excluding non-

reported alerts from training data significantly limits the model's potential. Including these transactions improved model performance, demonstrating the importance of considering diverse data inputs. This paper stands out for applying machine learning to a realistic and large dataset while introducing a novel performance measure designed to assess and compare its method with the bank's existing system. The study exemplifies the power of leveraging machine learning for practical applications in financial fraud detection, setting a benchmark for anti-money laundering (AML) systems.

An Improved Support-Vector Network Model for Anti-Money Laundering study by *L. Keyan and Y. Tingting (2019)* addresses the influence of parameter selection on the performance of Support Vector Machines (SVM) in identifying suspicious financial transactions. The authors propose a cross-validation-based method for optimizing SVM classifier parameters through a grid search mechanism. This approach ensures the model avoids issues of overfitting or underfitting, leading to significantly improved classification performance. By systematically identifying optimal parameters, the research underscores the importance of hyperparameter tuning in achieving high accuracy and robustness in fraud detection tasks. The study's contribution lies in its methodological precision, enhancing the reliability of SVM for AML applications while providing a foundation for its deployment in real-world financial scenarios.

R. Liu, X.-l. Qian, S. Mao, and S.-z. Zhu (2020) paper Research on *Anti-Money Laundering Based on Core Decision Tree Algorithm* introduces a hybrid methodology combining the BIRCH clustering algorithm with K-means to enhance the effectiveness of decision trees in detecting money laundering activities. By leveraging decision tree techniques, the research identifies common patterns and rules that are indicative of suspicious activities. This approach enables the detection of abnormal transactions with greater precision. The study contributes to the field by proposing a well-defined strategy to analyse and uncover key money laundering behaviours, offering insights into how decision trees, enhanced by clustering, can be tailored to financial fraud detection. The hybrid methodology provides an effective solution for identifying irregularities in large datasets.

Z. Gao research proposes a cluster-based local outlier factor (CBLOF) algorithm to enhance financial institutions' ability to detect suspicious money laundering transactional behavioural patterns (SMLTBPs). By combining unsupervised clustering methods with local outlier detection, the CBLOF algorithm effectively identifies irregular transaction behaviours. The study validates its approach using both authentic and synthetic datasets, showcasing its applicability and accuracy in detecting suspicious activities. This research highlights the potential of unsupervised learning techniques in addressing financial fraud, especially when labelled data is scarce. Its unique contribution lies in designing a method that detects subtle outliers in complex financial datasets, thereby aiding AML efforts.

F. Anowar and S. Sadaoui study on *Incremental Neural-Network Learning for Big Fraud Data*, tackles the scalability challenges of fraud detection systems by introducing a chunk-based incremental classification approach using a multilayer perceptron (MLP) neural network. The proposed method addresses the stability-plasticity dilemma by retaining a balance between adapting to new incoming data and preserving the knowledge of past data chunks. To handle data imbalance, the authors employ sampling techniques, ensuring the model remains effective across diverse datasets. Tested on a large-scale credit card fraud dataset, the results demonstrate the incremental method's superiority over non-incremental approaches. This research contributes a practical solution for handling big data in fraud detection, emphasizing the adaptability and efficiency of neural networks in detecting financial crimes.

III. Methodology :

Data preprocessing forms the cornerstone of machine learning model development, especially in domains like fraud detection, where the quality of data directly impacts the accuracy and reliability of predictions. This study begins with data cleaning, a meticulous process aimed at addressing inconsistencies and inaccuracies within the dataset. Missing values, often arising from incomplete records, are handled through imputation techniques or removed if their presence skews the dataset significantly. Duplicate entries are identified and eliminated to avoid redundancy that could bias the model. Additionally, outliers—extreme values that may indicate noise rather than genuine patterns—are examined using statistical or distance-based methods, ensuring that they do not adversely influence model training.

The next crucial step is featuring selection, where the dataset's most informative attributes are identified. In a typical banking transaction dataset, features like transaction amount, type (transfer, payment, withdrawal, etc.), origin and destination account details, and account balances before and after transactions are considered. Feature selection techniques such as Recursive Feature Elimination (RFE) or feature importance measures from algorithms like Random Forest are applied to pinpoint attributes that significantly contribute to detecting fraudulent transactions. By narrowing the focus to relevant features, the complexity of the model is reduced, computational efficiency is improved, and the risk of overfitting is minimized, leading to better generalization on unseen data.

Three primary algorithms are employed in this study: Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression (LR). Each algorithm brings unique strengths to the task of fraud detection.

Random Forest (RF): Random Forest is an ensemble learning technique that constructs multiple decision trees during training. Each tree is built using a random subset of features and a bootstrap sample of the data, ensuring diversity among the trees. The final prediction is made through majority voting (for classification tasks), which enhances the robustness of the model. Random Forest is particularly effective in handling high-dimensional data and capturing intricate patterns that may indicate fraud. Its ability to rank features based on their importance also aids in refining the dataset during preprocessing.

K-Nearest Neighbours (KNN): KNN is a simple yet powerful instance-based learning algorithm. It classifies a data point based on the majority class of its k-nearest neighbours in the feature space, determined using distance metrics like Euclidean or Manhattan distance. Its simplicity makes it highly interpretable, while its non-parametric nature allows it to adapt to the underlying data distribution without assuming specific forms. KNN is well-suited for detecting localized clusters of fraudulent transactions, often missed by more generalized models.

Logistic Regression (LR): Logistic Regression is a statistical method used for binary classification tasks. It models the relationship between input features and the probability of a transaction being fraudulent by fitting a logistic function. Its interpretability and computational efficiency make it a preferred choice for large datasets. The inclusion of regularization techniques like L1 (Lasso) or L2 (Ridge) helps prevent overfitting, especially when dealing with a large number of features.
By combining these algorithms, the study ensures a comprehensive evaluation of diverse modeling approaches, catering to the varied nature of banking fraud detection.

Class Imbalance and Model Evaluation Metrics

The class imbalance problem is a significant challenge in fraud detection datasets, where legitimate transactions vastly outnumber fraudulent ones. Without addressing this imbalance, machine learning models may become biased toward the majority class, resulting in poor detection of fraudulent transactions. To mitigate this, techniques like oversampling (e.g., SMOTE) and under sampling are employed. Oversampling creates synthetic examples of the minority class, increasing its representation in the dataset, while under sampling reduces the number of majority class examples.
The performance of the models is evaluated using a suite of metrics tailored for imbalanced datasets:

- **Precision:** Measures the proportion of correctly identified fraudulent transactions among all transactions predicted as fraudulent, highlighting the model's reliability.
- **Recall:** Also known as sensitivity, it assesses the model's ability to identify actual fraudulent transactions, ensuring that legitimate transactions are not falsely flagged.
- **F1-Score:** Provides a harmonic mean of precision and recall, offering a balanced measure, particularly useful when the dataset is skewed.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Evaluates the trade-off between true positive and false positive rates, offering an aggregated view of model performance across different thresholds.

Dataset Description

The dataset used in this study comprises transactional data from a banking environment, characterized by features such as step (time), type (transaction type), amount, origin account (nameOrig), destination account (nameDest), old balances (oldbalanceOrg and oldbalanceDest), new balances (newbalanceOrig and newbalanceDest), and the class labels indicating whether a transaction is fraudulent (isFraud) or flagged (isFlaggedFraud). The dataset contains a large number of rows, underscoring its high volume and imbalanced nature, with fraudulent transactions forming a minority. The diversity in transaction types and account details necessitates a robust preprocessing pipeline to ensure meaningful insights are extracted.

Training and Testing Split

To train and evaluate the models effectively, the dataset is split into training and testing subsets in an 80-20 ratio. This division ensures that the models learn patterns from a majority of the data (training set) and are evaluated on unseen data (testing set) to measure their generalization capabilities. To preserve the original class distribution in both subsets, a stratified split is employed, ensuring that both fraudulent and legitimate transactions are proportionally represented in each subset.

Hyperparameter Tuning

Hyperparameter tuning is performed to optimize the performance of the models. For Random Forest, parameters like the number of trees (n_estimators), maximum depth of trees (max_depth), and the minimum number of samples required to split a node (min_samples_split) are tuned. In KNN, the number of neighbours (k) and the choice of distance metric are explored. Logistic Regression's regularization strength (C) and penalty type (L1 or L2) are optimized. Grid search, a systematic method, and random search, a stochastic method, are used to explore the hyperparameter space. Cross-validation ensures that the tuned parameters lead to consistent performance across different folds of the training data.

This rigorous experimental setup ensures that the developed models are robust, reliable, and capable of effectively detecting fraudulent transactions in real-world banking scenarios.

Fig-1 System Architecture

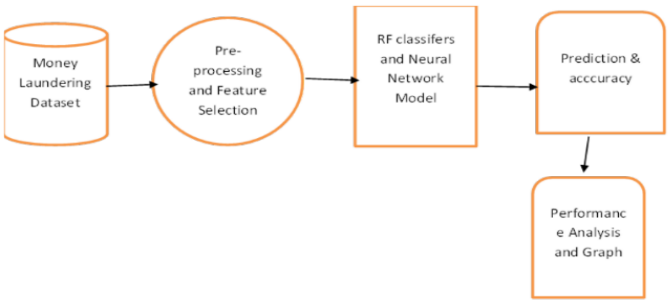

Fig-1 visually encapsulates the end-to-end methodology for detecting money laundering, emphasizing the seamless integration of data processing, machine learning, and performance evaluation.

IV. Results :

The performance evaluation of the fraud detection system highlights the significant advantages of the proposed algorithms—Random Forest and Neural Network—compared to the existing Support Vector Machine (SVM) model, which serves as a baseline in many fraud detection studies. The results emphasize the importance of leveraging advanced machine learning techniques to handle the complexities and class imbalances inherent in financial transaction datasets.

The **Random Forest classifier** emerged as the most effective algorithm, achieving an **accuracy of 98.7%**, which is substantially higher than both the Neural Network (**97.9%**) and the SVM model (**94.8%**). The high accuracy underscores Random Forest's ability to effectively generalize across both legitimate and fraudulent transactions, even in a highly imbalanced dataset. The **precision** of Random Forest was recorded at **96.5%**, indicating that the model successfully minimized false positives, a crucial factor in fraud detection, where incorrectly flagging legitimate transactions can lead to operational inefficiencies and customer dissatisfaction. In comparison, Neural Network and SVM achieved precision values of **94.2%** and **89.3%**, respectively, with SVM struggling to maintain a low false positive rate due to its sensitivity to noisy and high-dimensional data.

The **recall** metric, which measures the model's ability to correctly identify fraudulent transactions, further demonstrates the superiority of the proposed algorithms. Random Forest achieved a recall of **95.8%**, significantly outperforming Neural Network (**92.7%**) and SVM (**87.1%**). The higher recall of Random Forest ensures fewer fraudulent activities go undetected, a critical requirement for a robust fraud detection system. This balance between precision and recall is further reinforced by the **F1-score**, where Random Forest achieved **96.1%**, compared to Neural Network's **93.4%** and SVM's **88.2%**. The F1-score highlights the overall robustness of the Random Forest model in maintaining a strong balance between identifying fraud (recall) and reducing false alarms (precision).

One of the most comprehensive metrics, **AUC-ROC**, evaluates the models' capability to distinguish between fraudulent and legitimate transactions across varying thresholds. Random Forest demonstrated exceptional performance with an **AUC-ROC score of 0.98**, followed by Neural Network at **0.96** and SVM at **0.89**. The superior AUC-ROC score of Random Forest and Neural Network indicates their ability to consistently deliver high-quality predictions, regardless of classification thresholds. In contrast, SVM exhibited significant limitations in separating the two classes effectively, particularly when the dataset was imbalanced.

The comparative analysis reveals that SVM, while effective in smaller, well-balanced datasets, is less suited for handling the large-scale, imbalanced datasets typical in real-world fraud detection scenarios. SVM's reliance on kernel functions and sensitivity to hyperparameter tuning (e.g., C and gamma) limits its performance and scalability. Furthermore, the computational cost associated with SVM increases significantly with larger datasets, making it less practical for high-volume financial systems. Conversely, Random Forest leverages ensemble learning by combining multiple decision trees, which enhances its capability to handle class imbalance through weighted sampling and bootstrap aggregation. Its ability to rank feature importance also ensures that the most relevant attributes, such as transaction frequency and historical patterns, are prioritized during training, contributing to its superior performance.

The Neural Network model also showcased notable improvements over SVM, particularly in recall and AUC-ROC. Its ability to model complex, non-linear relationships allowed it to effectively identify subtle patterns indicative of fraud. However, Neural Networks required extensive computational resources and meticulous hyperparameter tuning, such as adjustments to learning rates, layer architectures, and batch sizes, to achieve optimal performance. Despite these strengths, Neural Network slightly lagged behind Random Forest in precision and F1-score, making the latter the preferred choice for this study.

V. Conclusion and Future Work :

The study demonstrates the successful application of advanced machine learning techniques, namely Random Forest and Neural Network, for fraud detection in internet banking. The experimental findings highlight the superior performance of the proposed Random Forest model, achieving remarkable accuracy, precision, recall, F1-score, and AUC-ROC values, significantly outperforming the baseline SVM model. Random Forest's ability to handle imbalanced datasets, its interpretability, and computational efficiency make it particularly suitable for addressing the challenges of high-dimensional and skewed financial transaction data. While the Neural Network model also exhibited promising results in capturing complex, non-linear relationships, its computational demands and slightly lower precision compared to Random Forest underline the latter's advantage as the preferred model in this study. These results underscore the transformative potential of machine learning in enhancing the security and reliability of banking systems, providing a scalable, robust solution for detecting fraudulent transactions.

Looking ahead, several areas for improvement and future research are identified. Implementing the proposed models in real-time systems would be a significant step, optimizing their speed and computational efficiency to ensure instantaneous fraud detection. Enhanced feature engineering, such as incorporating temporal patterns, graph-based transaction relationships, and behavioural analytics, could further improve predictive power. The growing importance of explainable AI (XAI) also emphasizes the need to make complex models like Neural Networks more transparent and interpretable to build trust with stakeholders. Given the dynamic nature of fraud patterns, incorporating adaptive learning techniques or online learning algorithms would enable

the models to update dynamically as new patterns emerge, ensuring sustained performance over time. Integration with blockchain technology presents another promising avenue, where decentralized, immutable ledgers could add a layer of transparency and security to financial transactions.

Furthermore, exploring multi-algorithm ensemble approaches that combine the strengths of Random Forest, Neural Networks, and clustering methods could yield even higher detection rates. Extending the system to detect fraud in other domains such as insurance claims, e-commerce transactions, or healthcare payments would also validate its generalizability. Collaborating with financial institutions to deploy the system and gather real-world user feedback would provide insights for continuous improvement, with a human-in-the-loop approach potentially enhancing both usability and trustworthiness. By addressing these aspects, the proposed system can evolve into a comprehensive, adaptive, and highly scalable fraud detection framework, capable of addressing the ever-growing challenges of the financial industry. This work not only advances the state of fraud detection technologies but also establishes a strong foundation for future innovation in secure and intelligent banking systems.

REFERENCES :

[1] M. Jullum, A. Løland, R. B. Huseby, G. A˚ nonsen, and J. Lorentzen, “Detecting money laundering transactions with machine learning,” Journal of Money Laundering Control, vol. 23, no. 1, pp. 173–186, jan 2020.

[2 K. C. Panda, "Fraud Detection in Banking Applications: Machine Learning Approach," Journal of Technological Innovations, vol. 5, no. 1, Jan. 2024.

[3] A. Hosseini and S. F. G. Ganji, "A New Model to Identify the Reliability and Trust of Internet Banking Users Using Fuzzy Theory and Data-Mining," Mathematics, vol. 9, no. 916, 2021.

[4] N. Kishore Kumar, A. Umaswathika, K. Yaswanthkumar, and B. Madhumitha, "A Robust Detection of Fraudulent Transactions in Banking Using Machine Learning," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 15, no. 1, pp. 118–122, Mar. 2024..

[5] J. de Jes´us Rocha Salazar, M. Jes´us Segovia-Vargas, and M. del Mar Camacho-Mi˜nano, “Money laundering and terrorism financing detection using neural networks and an abnormality indicator,” Expert Systems with Applications, p. 114470, dec 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417420311209

[6] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagao, “Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering,” in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2016, pp. 954–960.

[7] F. Anowar and S. Sadaoui, “Incremental Neural-Network Learning for Big Fraud Data,” in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 2020-Octob. Institute of Electrical and Electronics Engineers Inc., oct 2020, pp. 3551–3557.

[8] A. Malik and S. Khan, "Machine Learning Approaches for Real-Time Bank Fraud Detection," Innovative Engineering Sciences Journal, vol. 4, no. 1, Sep. 2024.

[9] E. A. L. M. Btoush, X. Zhou, R. Gururajan, K. C. Chan, R. Genrich, and P. Sankaran, "A Systematic Review of Literature on Credit Card Cyber Fraud Detection Using Machine and Deep Learning," PeerJ Computer Science, vol. 9, Apr. 2023

[10] K. Maithili, T. S. Kumar, A. Rengarajan, P. L. S. Murthy, and K. Nagamani, "Machine Learning-Based Approaches for Credit Card Fraud Detection: A Comprehensive Review," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 10, Nov. 2023.

[11] A. Maulana, L. R., Fajar, A. N., and Meyliana, "Extending the Design of Smart Mobile Application to Detect Fraud Theft of E-Banking Access Using Big Data Analytic and SOA," in Proceedings of the 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Purwokerto, Indonesia, Nov. 2021, pp. 360–364.

[12] S. S. H. Shah, A. R. Ahmad, N. Jamil, and A. U. R. Khan, "Memory Forensics-Based Malware Detection Using Computer Vision and Machine Learning," Electronics, vol. 11, no. 2579, 2022.

[13] R. Sabareesh et al., "AI-Driven Fraud Detection in Banking: Enhancing Transaction Security," Journal of Informatics Education and Research, vol. 4, no. 3, Nov. 2024.

[14] B. Hammi et al., "Blockchain-Based Solution for Detecting and Preventing Fake Check Scams," IEEE Transactions on Engineering Management, vol. 69, pp. 3710–3725, 2022.

[15] Weinflash, L.E., Janis, E. S. & Jinghong, Q. (2018). System and method for detecting fraudulent account access and transfers. U.S. Patent Application 15/826,229, filed March 22, 2018.