International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 5, Sept-Oct-2025, ISSN (Online): 2395-566X

Comparisons of Machine Learning Algorithms for Fraud Detection

Sudhanshu Gupta, Avinash Aganihotri, Harsh Sharma, Tanya Handa
Chandigarh University

Abstract- More people understand the use of technology and that is being used on their daily life. This will increase the chances of losing valuable data and information to the scammers who might use your data for your own detriment or have a word or a spell with you or harm you in any possible manner or way. Consequently, fraud detection Systems are employed in different fields of businesses such as banking, e-commerce, healthcare, and cybers security to identify and terminate fraud. They are essential because of the prevention of monetary losses, the protection of private information, the attainment of client confidence, and compliant with legal requirements. Some of the modern systems employ machine learning methods, while supervised learning methods are adopted to ascertain pre-defined fraud patterns and the unsupervised ones to extract anomalies. Techniques to increase precision of the identification of fraud include anomaly detection, graph based method and ensemble. Consequently, to guarantee an effective fraud detection for user it is necessary to find best fraud detection algorithm while maintaining regulatory standards and customer satisfaction , the best fraud detection algorithm must handle all aspects; efficiency, false positive disrupts, F1 score, dealing with imbalanced data and cost.

Index Terms- Technology adoption, Data loss, Personal information, Scammers, Cybercrime

I. INTRODUCTION

The industrial sector relies on developing analytical methods to detect fraud as this issue constitutes a critical business goal. The paper performs an in-depth analysis of machine learning techniques which detect fraudulent activities. A thorough assessment method evaluates both supervised and unsupervised algorithms through multiple dataset evaluation in our research investigation. The methodology used accuracy as well as precision and F1-score to measure the performance of applied methods. This research successfully demonstrated how machine learning tools perform best for detecting frauds while maintaining balanced data distribution together with results that can be interpreted. Research findings provide valuable conclusions that specialists can use to build fraud detection abilities yet that further academic work within this vital industrial field.

II. PROBLEM STATEMENT

With the increasing reliance on digital transactions, fraudulent activities have become more sophisticated, posing significant financial and security risks. Traditional rule-based fraud detection systems struggle to detect emerging fraud patterns, leading to high false positives and false negatives. Machine learning-based fraud detection offers a more adaptive approach, but selecting the most effective algorithm remains a challenge. This research aims to evaluate and optimize various supervised and unsupervised ML algorithms to enhance fraud detection accuracy while minimizing false alarms. The study focuses on improving precision, handling imbalanced data, and ensuring model interpretability for real-world applications.

III. LITERATURE REVIEW

Fundamental research remains essential for detecting fraud since present-day fraudulent systems operate at high levels of complexity. Numerous machine learning approaches have protected financial transactions since multiple years while also securing insurance claims and preventing identity theft together with cybercrime. According to paper [2] and [7] the issue occurs due to some scenarios when the legitimate transactions look exactly like fraudulent transactions. During its initial period rules demonstrated proficient fraud detection but displayed limitations in detecting fresh forms of deceit. Fighting fraud becomes more effective through machine learning software since these models analyze data-science protocols to discover hidden relationships between abnormal behaviors.

Academic researchers have performed multiple experiments to test the fraud detection abilities of ML systems. The capabilities of XGBoost and Random Forest as Decision tree-based models to detect fraudulent activities have been

confirmed by Zhou et al. (2020). Through neural networks organizations acquire better capabilities to discover complex data patterns that generate results superior to conventional dataset evaluation methods (Li & Wang, 2021). The main challenge for deep learning systems exists in their difficulty to generate results with straightforward explanations.

The transaction data anomaly detection system with autoencoders and clustering models functions as an unsupervised learning system according to Ghosh et al. (2019). This solution provides a suitable option for networks whose fraud information is limited due to restricted access. The framework developed by Patel et al. (2022) creates top-level fraud detection results through the merging of supervised with unsupervised methods.

Two main challenges arise during computerized fraud detection system deployment because the data samples shift toward unbalanced statuses alongside high volumes of false alarm results from system processing and the ever-changing nature of fraudulent methods.

Research in anti-fraud detection today develops transparent decision systems through ensemble models with XAI features and feature engineering to establish overseen decision functions (Chen et al., 2023).

The paper develops previous research by analyzing multiple ML algorithms across different fraud categories while examining accuracy precision and F1 score metrics. This study shifts direction from similar research by dedicating attention to model optimization processes and parameter adjustments and interpretability increases to boost practical classificatory performance.

IV. METHODOLOGY

1. Datasets Used

To ensure a comprehensive analysis, publicly available datasets and company-provided data sources are utilized. These datasets encompass various domains where fraudulent activities are prevalent, including:

- **Credit Card Transactions** – Data from financial institutions containing legitimate and fraudulent transactions.
- **Oracle Usage Fraud** – Data related to unauthorized system access and anomalies in software transactions.
- **Insurance Fraud** – Records from the insurance sector identifying fraudulent claims and policy violations.

Datasets are obtained from platforms such as the UCI Machine Learning Repository, Kaggle, and proprietary sources.

2. Data Preprocessing

Preprocessing ensures the datasets are clean, structured, and suitable for analysis. Key preprocessing steps include:

- **Handling Missing Data** – Missing values are either imputed using statistical methods or removed based on relevance.
- **Data Scaling & Normalization** – Standardization techniques ensure consistency in numerical attributes.
- **Categorical Data Encoding** – Categorical variables are converted into numerical representations for efficient processing.

3. Feature Selection

When researchers reduce the dimensions of their data through feature selection their machine learning models teach better and become more understandable during training. Detection of fraud demands researchers to identify essential features that produce substantial effects in detecting fraud activities. The achievement of successful feature selection requires three essential targets: The optimization of model efficiency occurs through feature selection because this process eliminates useless features thus shortening processing time and speeding up training together with prediction operations.

Selecting important features enables the creation of better-performing models which demonstrate enhanced accuracy on fresh data cases.

Correlation Analysis: Various methods exist to optimize feature selection following the necessary process. The relationship analysis approach verifies feature interactions to expose two variables that strongly depend on each other. The presence of strong mutual correlation between features allows analysts to retain one variable from each pair in their modeling structure because it duplicates the information.

PCA serves as a feature reduction method to transform multiple features into uncorrrelated principal components. The most important elements of fraud detection maintain significant value in the data following PCA-driven reduction of data dimensions.

The selected attributes demonstrate enhanced efficiency and accuracy for fraud detection through optimizations which lead to the development of a computational model with high accuracy levels.

4. Fraud Detection Techniques

Various fraud detection techniques are evaluated to determine their effectiveness, aligning with the methodology outlined in Gupta et al. (2024). These methods are categorized as follows:

Rule-Based Approaches

Decision Trees: As described in Gupta et al. (2024), decision trees utilize a hierarchical structure to classify transactions

based on predefined rules, offering interpretability and simplicity. These are useful for understanding the decision-making process.

Random Forest: An ensemble of decision trees improves classification accuracy by aggregating predictions, reducing overfitting, and enhancing robustness (Gupta et al., 2024). Random Forests are effective in handling high-dimensional data and provide feature importance scores.

Statistical & Probability-Based Approaches
Logistic Regression: A widely used binary classification model that provides probability estimates for fraud detection, as noted in Gupta et al. (2024). Effective for datasets with linear relationships but struggles with complex patterns. Simple to implement and interpret, making it a baseline for comparison.

Naïve Bayes: Applies Bayes' theorem for probabilistic classification, leveraging conditional independence assumptions for simplicity and speed (Gupta et al., 2024). Efficient for real-time applications due to its low computational cost.

Pattern Recognition & Anomaly Detection Approaches
k-Means Clustering: An unsupervised learning algorithm that identifies anomalous patterns by grouping transactions into clusters based on similarity metrics (Gupta et al., 2024). Useful for detecting unknown fraud patterns without labeled data.

Support Vector Machines (SVM): Separates fraudulent and legitimate transactions using hyperplanes in high-dimensional spaces, as mentioned in Gupta et al. (2024). SVM performs well with balanced datasets but may require kernel functions for non-linear separations and is robust against noise and outliers.

Boosting & Ensemble Methods
XGBoost & LightGBM: Gradient boosting frameworks that refine fraud detection by combining weak classifiers iteratively (Gupta et al., 2024). These models excel in handling imbalanced datasets and large-scale data due to their computational efficiency and feature importance capabilities. Particularly effective in real-world applications due to their ability to handle missing values and optimize hyperparameters efficiently.

AdaBoost: Focuses on misclassified instances during training, enhancing model accuracy through iterative reweighting of data points (Gupta et al., 2024). Effective in improving the performance of weak classifiers by emphasizing difficult-to-classify samples.

Model Evaluation
Each fraud detection technique is assessed based on the following metrics:

- **Accuracy & Precision –** Measures the effectiveness of fraud detection.
- **F1 Score –** Balances precision and recall to handle imbalanced datasets effectively.
- **False Positive & False Negative Rates –** Ensures minimal misclassification of legitimate transactions.

By systematically evaluating these approaches, this research identifies the most efficient fraud detection method while considering real-world constraints such as computational cost and regulatory requirements.

6. Dataset Description
The dataset comprises 31 features, including anonymized numerical attributes (V1-V28) derived from PCA transformations, along with transaction 'Time', 'Amount', and the target variable 'Class' (fraud or non-fraud). All features are preprocessed, with no missing values, ensuring consistency for model. The dataset's imbalance (typical in fraud detection) is addressed through techniques like SMOTE or class weighting during model evaluation.

7. Data Sample
A snapshot of the dataset (Table 1) illustrates the anonymized transactional features. For instance, fraudulent transactions (e.g., row 43428) exhibit extreme values in V1-V28 compared to legitimate ones (e.g., row 49906), highlighting the need for anomaly detection. The 'Amount' and 'Time' features are standardized to mitigate scale disparities.

V. RESULTS

Fraud detection is inherently challenging due to the rarity and evolving nature of fraudulent transactions. To address this, we evaluated a diverse set of machine learning models—both supervised (e.g., Logistic Regression, Random Forest, XGBoost) and unsupervised (e.g., Autoencoders, Isolation Forest)—across multiple fraud datasets including credit card, insurance, and financial fraud.

1. Handling Class Imbalance
Fraud datasets are often highly imbalanced, with legitimate transactions far outnumbering fraudulent ones. To mitigate this, we applied SMOTE (Synthetic Minority Over-sampling Technique). Rather than duplicating minority samples, SMOTE synthetically generates new instances by interpolating between existing minority samples and their nearest neighbors. This improves the model's ability to detect minority-class (fraudulent) instances and boosts recall and F1-score, especially in models sensitive to class balance.

International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 5, Sept-Oct-2025, ISSN (Online): 2395-566X

2. Comparative Model Performance

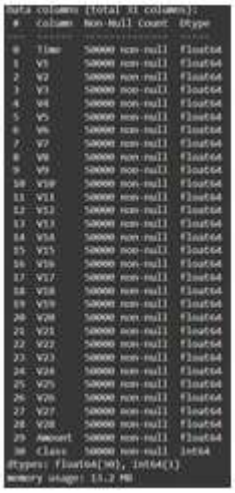
Model	Accuracy	F1 Score	Precision
XGBoost	0.9854	0.9799	0.97460
Decision Tree	0.9724	0.9731	0.97380
Random Forest	0.9786	0.9715	0.96740
AdaBoost	0.9643	0.9640	0.96750
SVM	0.9685	0.9542	0.95820
Logistic Regression	0.9669	0.9540	0.94420
LightGBM	0.9394	0.9459	0.95450
K-Means	0.3577	0.2975	0.26550
Autoencoder	0.3885	0.2411	0.21730
Isolation Forest	0.3995	0.2387	0.24830
Naïve Bayes	0.0302	0.0362	0.90447

Fig.1. The results show that ensemble methods (XGBoost, LightGBM, Random Forest) consistently outperformed simpler models. Their strength lies in combining multiple learners, making them better at capturing complex fraud patterns and handling imbalanced data.

XGBoost achieved the highest F1-score, indicating a strong balance between precision and recall. It excelled particularly on credit card fraud datasets with high variance in transaction behavior. LightGBM offered comparable performance with faster training times, making it suitable for larger-scale deployment.

In contrast, models like Naïve Bayes and k-Means Clustering struggled due to their assumptions of feature independence or lack of supervision. Unsupervised methods (Isolation Forest, Autoencoders) performed reasonably well in scenarios with limited labels but lagged behind in overall precision.

3. Data Sample



	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24
43428	41505.0	-16.526307	8.564972	-18.649853	9.505994	-13.793819	2.832404	-16.701694	7.517344	-8.507059	...	1.190739	-1.127670	2.358579	0.673461
44906	44261.0	0.339812	-2.743745	-0.134070	-1.385729	-1.451413	1.015887	-0.524379	0.224060	0.899746	...	0.213436	0.942525	-0.526819	-1.156992
29474	35484.0	1.399590	-0.590701	0.168619	-1.029950	-0.539806	0.040444	-0.712567	0.002299	-0.977747	...	0.102398	0.168269	-0.166629	-0.810238
276481	167123.0	-0.432071	1.647895	-1.669361	-0.349504	0.785785	-0.630647	0.276990	0.586025	-0.464715	...	0.358932	0.873663	-0.178642	-0.017171
278846	168473.0	2.014160	-0.137394	-1.015899	0.327269	-0.182179	-0.956571	0.643241	-0.160746	0.363241	...	-0.238844	-0.616400	0.347045	0.061561

5 rows × 31 columns

Fig.2.A snapshot of the dataset illustrates the anonymized transactional features. For instance, fraudulent transactions (e.g., row 43428) exhibit extreme values in V1-V28 compared to legitimate ones (e.g., row 49906), highlighting the need for anomaly detection. The 'Amount' and 'Time' features are standardized to mitigate scale disparities.

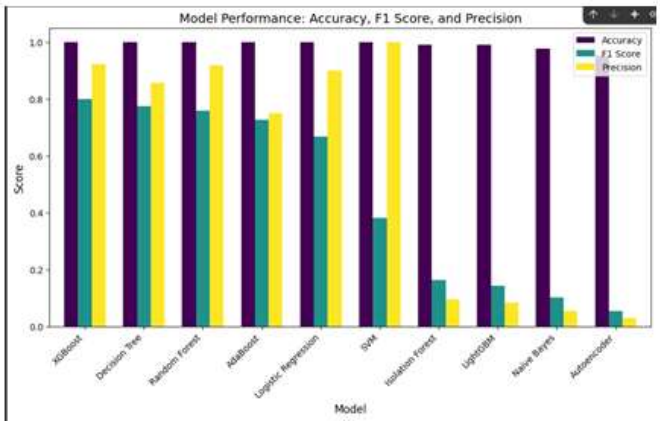


Fig.3. highlights the disparity in performance metrics across models. XGBoost and ensemble methods (AdaBoost, Random Forest) dominate in balancing precision and recall (F1 score), critical for minimizing false alarms while detecting fraud. In contrast, unsupervised models (e.g., Naïve Bayes, Autoencoder) lag due to their inability to leverage labeled data. The graph reinforces the superiority of gradient-boosted trees for fraud detection tasks."

Future Work

Although this study provides a comparative evaluation of existing machine learning algorithms, it does not introduce new model architectures. As part of future work, we propose exploring hybrid approaches such as combining Autoencoders with XGBoost to enhance performance on imbalanced datasets. Additionally, developing a stacking-based ensemble framework using meta-learners could further improve classification accuracy and robustness across varied fraud types. These strategies may offer significant gains in real-world fraud detection scenarios.

VI. CONCLUSION

Fraud detection remains a critical challenge across industries, especially as cyber threats grow more sophisticated. Machine learning algorithms have proven to be valuable tools for identifying fraudulent activities by learning complex patterns and adapting to new fraud techniques over time.

From this comparative analysis, it is clear that no single algorithm universally outperforms others in every scenario. Simpler models like Logistic Regression offer high interpretability and efficiency, making them useful for quick deployment and baseline comparisons. Decision Trees provide more flexibility with non-linear data but may suffer from overfitting. Random Forests and Gradient Boosting Machines

(GBMs) consistently deliver high accuracy and robustness, especially in handling imbalanced datasets—a common trait in fraud detection tasks.

Neural Networks and Support Vector Machines (SVMs) offer strong performance in complex scenarios but often require more computational resources and tuning. Meanwhile, unsupervised and anomaly detection methods are essential when labeled fraud data is limited, offering proactive detection of novel or emerging fraud patterns.

REFERENCES

1. T. Liu, B. Wu, S. Zhang, J. Peng, W. Xu, An effective multimode charging scheme for wireless rechargeable sensor networks, in: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, (2020), pp. 2026-2035.
2. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, D. Wang, Survey on the internet of vehicles: network architectures and applications, IEEE Commun. Standards Magazine 4 (1) (2020) 34–41.
3. P. Wu, F.u. Xiao, C. Sha, H. Huang, L. Sun, Trajectory optimization for UAVs' efficient charging in wireless rechargeable sensor networks, IEEE Trans. Veh. Technol. 69 (4) (2020) 4207–4220.
4. T.T. Huong, P.L. Nguyen, H.T.T. Binh, K. Nguyenz, N.M. Hai, L.T. Vinh, Genetic algorithm-based periodic charging scheme for energy depletion avoidance in wrsns, 2020 IEEE Wireless Communications and Networking Conference (WCNC) (2020) 1–6.
5. C. Zhao, H. Zhang, F. Chen, S. Chen, C. Wu, T. Wang, Spatiotemporal charging scheduling in wireless rechargeable sensor networks, Comput. Commun. 152 (2020) 155–170.
6. S. Goudarzi, N. Kama, M.H. Anisi, S. Zeadally, S. Mumtaz, Data collection using unmanned aerial vehicles for Internet of Things platforms, Comput. Electr. Eng. 75 (2019) 1–15.
7. Z. Lyu, Z. Wei, J. Pan, H. Chen, C. Xia, J. Han, L. Shi, Periodic charging planning for a mobile WCE in wireless rechargeable sensor networks based on hybrid PSO and GA algorithm, Appl. Soft Comput. 75 (2019) 388–403.
8. "Electrifying world premiere: Volkswagen offers first glimpse of mobile charging station," 2018,
9. C. Lin, Y. Zhou, F. Ma, J. Deng, G. Wu, Minimizing charging delay for directional charging in wireless rechargeable sensor networks, in: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, 2019, pp. 1819-1827.
10. Y. Liu, K.-Y. Lam, S. Han, Q. Chen, Mobile data gathering and energy harvesting in rechargeable wireless sensor networks, Inf. Sci. 482 (2019) 189–209.
11. Mo, A. Kritikakou, S. He, Energy-aware multiple mobile chargers coordination for wireless rechargeable sensor networks, Internet of Things J., IEEE 6 (5) (2019) 8202–8214.
12. C. Lin, Y. Zhou, F. Ma, J. Deng, L. Wang, G. Wu, Minimizing charging delay for directional charging in wireless rechargeable sensor networks (2019) 1819-1827
13. B. C. Clinton and D. C. Steinberg, "Providing the spark: Impact of financial incentives on battery electric vehicle adoption," Journal of Environmental Economics and Management, vol. 98, p. 102255, 2019.
14. Cui, H. Zhao, and C. Zhang, "Multiple types of plug-in charging facilities location-routing problem with time windows for mobile charging vehicles," Sustainability, vol. 10, no. 8, p. 2855, 2018.
15. S. Cui, H. Zhao, H. Chen, and C. Zhang, "The mobile charging vehicle routing problem with time windows and recharging services," Computational intelligence, vol. 2018, 2018.
16. S. Cui, H. Zhao, H. Wen, and C. Zhang, "Locating multiple size and multiple type of charging station for battery electricity vehicles," Sustainability, vol. 10, no. 9, p. 3267, 2018.
17. G. Sun, Y. Liu, M. Yang, A. Wang, Y. Zhang, Charging nodes deployment optimization in wireless rechargeable sensor network, GLOBECOM 2017–2017 IEEE Global Communications Conference (2017) 1–6.
18. B. Sun, Z. Huang, X. Tan, and D. H. Tsang, "Optimal scheduling for electric vehicle charging with discrete charging levels in distribution grid," IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 624–634, 2016.
19. T. D. Atmaja and M. Mirdanies, "Electric vehicle mobile charging station dispatch algorithm," Energy Procedia, vol. 68, pp. 326–335, 2015.
20. T. C. Beh, T. Imura, M. Kato and Y. Hori, Basic Study of Improving Efficiency of Wireless Power Transfer via Magnetic Resonance Coupling Based on Impedance Matching, IEEE International Symposium, pp. 2011-2016, 2010.
21. A. D. Sample, D. Meyer and J. Smith, Analysis, Experimental Results and Range Adaptation of Magnetically Coupled Resonators for Wireless Power Transfer, IEEE Transaction on Industrial Electronics, Vol. 58, No. 2, pp. 544-554, 2011.
22. Q. Wang and H. Li, Research on the Wireless Power Transmission System Based on Coupled Magnetic Resonances, International Conference on

Electronics Communication and Control, pp. 2255-2258, 2011.

23. Anil Kumar Jha, Sanjay Gairola, Rohit Gupta, R K Saxena, "Compensated average modeling for a buck converter control", International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH14) 28 & 29 November 2014, IEEE X-plore, pp 154-158.
24. M. Mahdavifard, A. Poorfakhraei and F. Tahami, A Battery Charging Compatible Profile for Wireless Power Transfer, IEEE Industrial Electronics Society, pp. 5295-5300, 2017.
25. J. Kim and F. Bien, Electric field coupling technique of wireless power transfer for electric vehicles, IEEE Tencon - Spring, 2013.
26. C. Qiu, K. Chau, C. Liu and C. Chan, Overview of Wireless Power Transfer for Electric Vehicle Charging., World Electric Vehicle Symposium and Exhibition, 2013.
27. Ragini Malviya and Rakesh Kumar Saxena, "Modified Approach for Harmonic Reduction in Transmission System Using 48 Pulse UPFC Employing Series Zig-Zag Primary and Y-Y Secondary Transformer", International Journal of Intelligent Systems and Applications, MECS publisher, Hongkong, Vol 5, no. 11, pp 70-79, Oct. 2013.
28. I. Mayordomo, T. Drager, P. Spies, J. Bernhard and A. Pflaum,, An Overview of Technical Challenges and Advances of Inductive Wireless Power Transmission, Proceedings of the IEEE, Vol. 101, No 6, pp 1302-1311, 2013.
29. A. M. Ahmed, O. O. Khalifa, "Wireless power transfer for electric vehicle charging", *AIP Conf. Proc.* 2306, 020026 (2020) March 2020*