Performance Evaluation of Hybrid Machine Learning Models For Credit Card Fraud Detection

Subarnaa. S[1], Mrs. V. Gomathi[2]
[1]Dept of Computer science And Engineering with specialization(AI&ML)
[2]Assistant Professor, Dept of Computer science And Engineering with specialization(AI&ML)
[1, 2,]CARE College of Engineering

Abstract- *Credit card fraud cares with the illegal use of master card information for purchases. Credit card transactions are often accomplished either physically or digitally. In the manual transactions, the credit card is included during the transactions. In digital transactions, this will happen over the phone or the web. Cardholders might be providing their card number, expiry date, and the verification of the card number through telephone or website. Billions of dollars are lost thanks to master card fraud per annum. Machine learning techniques are wont to detect master card fraud. Standard models are first used. Then, hybrid methods which use random forest and xgboost segmentation and popular voting method are applied. Then, a real-world master card data set from a financial organization is analysed. In addition, noise is added to the info samples to further assess the robustness of the algorithms. Here random forest segmentation and xgboost algorithm will give the 94% percent accuracy.*

Keywords: Machine Learning, Credit Card Fraud, detection, Random forest, xgboost algorithm

I. INTRODUCTION

Machine learning as a buzzword for the last few years, the reason would be might be the high level of amount and the data production by the applications, the growth of the computation power in the last few years and the growth of enhanced algorithm. You may already be using a device that using it. For example, a wearable fitness tracker likes Fit bit, or an intelligent home assistant like Google Home. But there are much more examples of ML in use.

Resurging interest in machine learning is thanks to an equivalent implementation that have made the info mining and Bayesian analysis most familiar than ever. Things like growing the density and varieties of the available data, will be computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to rapidly and automatically produce the models that can analyze bigger, more complex data and deliver it faster, will be more accurate output.

Fraud is cheating or wrongful or criminal activity, its main aim is to focus on financial or personal signs. This proposed system uses two mechanisms namely, (i) fraud prevention and (ii) fraud detection, for avoiding loss from fraud, that detecting details from fraud. The first fraud prevention mechanism is a most protective and proactive method, it stops the fraud from the beginning. Then, the second mechanism of fraud detection is guessing the fraudster. This mechanism is needed for a fraudulent transaction, but it is guessing the fraudster, in the time transaction attempted by the fraudster. Edit card fraud is related to the illegal use of credit card information for purchases that is credit card amount is used in product purchases. In the purchasing time, the user uses the credit card, the fraudster traces out the password or user-oriented important details, then it will be applied in our transaction easily use the credit card cash amount but cannot find out that person, that is a fraudster. The credit card transaction is completed physically or digitally. The physical transaction-based credit card is used during transactions, but the digital transaction-based credit card is used only on the telephone or internet. The cardholders are basically provided the important details such as card number, expiry date, and card verification number via telephone or website.

II. LITERATURE SURVEY

In this paper[1]Credit card fraud detection is a critical task in the financial industry, where the objective is to identify fraudulent transactions in real-time, minimizing financial losses and safeguarding customer trust. Traditional rule-based systems often struggle to detect sophisticated fraud patterns, prompting the adoption of machine learning techniques for more accurate and dynamic detection. This paper explores the application of machine learning models in credit card fraud detection, focusing on transaction data analysis to differentiate legitimate activities from fraudulent ones.A wide variety of features are considered for fraud detection, including transaction amount, time, location, user spending behavior, and device details.

In this paper[2] Credit card fraud is a major concern for financial institutions and cardholders alike, leading to

significant financial losses annually. The growing volume of transactions, coupled with the growing sophistication of fraudulent activities, necessitates the development of more effective and scalable fraud detection system. This paper presents a machine learning-based approach to detect credit card fraud in real-time.

In this paper[3] Credit card fraud detection using machine learning is an important application that aims to identify suspicious and potentially fraudulent transactions in real-time, helping to prevent financial losses. Here's an overview of how this system typically works and the key components involved:

In this paper[4] Credit card fraud is a growing concern that leads to significant financial losses for both consumers and financial institutions. Traditional rule-based systems often fail to detect sophisticated fraud patterns, especially with the increasing volume of transactions. This paper proposes an enhanced machine learning-based approach to credit card fraud detection, utilizing advanced techniques such as ensemble learning, deep learning, and anomaly detection. By leveraging historical transaction data, including transaction amount, time, merchant details, and user behavior, the system is designed to identify subtle and emerging fraud patterns.

In this paper[5] Credit card fraud remains a significant challenge for the banking and financial sectors, leading to billions in annual losses. The traditional methods of fraud detection, relying on manual rules and basic models, struggle to keep up with the increasing complexity and volume of transactions. This paper explores the use of machine learning and data science techniques to enhance the detection of fraudulent activities in credit card transactions.

In this paper[6] With the rise of online transactions, credit card fraud has become a major concern for consumers and financial institutions alike. Traditional fraud detection methods are often inadequate for real-time, online transaction monitoring, leading to increased financial risks. This paper presents an online credit card fraud detection system powered by machine learning techniques, designed to identify and prevent fraudulent activities in real-time. By analyzing a variety of transactional data such as purchase amount, merchant type, user behavior, geographical location, and device information, the system is able to detect suspicious patterns that deviate from typical spending behavior. Advanced machine learning algorithms, including Logistic Regression, Random Forests, xgboost and Neural Networks, are employed to classify transactions as either legitimate or fraudulent based on historical transaction data.

In this paper[7] Credit card fraud is a pervasive issue that causes significant business losses worldwide, affecting both customers and financial institutions. With the cumulative volume of credit card transaction, detecting fraudulent activities has become more complex and challenging. This research paper presents an in-depth exploration of various approaches to credit card fraud detection, focusing on the application of machine learning and data analytics techniques.

In this paper[8] Credit card fraud has become a major concern for financial institutions and consumers due to its potential to cause significant financial losses. Traditional fraud detection systems, relying on predefined rules and thresholds, are often ineffective in identifying sophisticated fraudulent activities. By utilizing transaction data, such as the amount, time, location, merchant, and user behavior, the study explores various machine learning.

In this paper[9] The ANN model is trained on a real-world dataset containing various features such as transaction amount, time, and anonymized user behavior patterns. Due to the inherent class imbalance in fraud datasets, specialized techniques like SMOTE (Synthetic Minority Over-sampling Technique) and data normalization are employed to enhance model performance. The neural network architecture is optimized using appropriate activation functions and regularization methods to reduce overfitting and improve generalization.

In this paper[10] Credit card fraud has become a major concern with the increasing use of online transactions and digital payments. Detecting fraudulent transactions in real-time is critical to ensure financial security and protect customer data. This project focuses on fraud detection in credit card transactions using advanced machine learning techniques.

The system is designed to analyze and classify transaction patterns to distinguish between legitimate and fraudulent activities. It makes use of a real-world dataset containing anonymized transaction records. Due to the imbalance between genuine and fraudulent transactions, data preprocessing techniques such as oversampling, normalization, and feature selection are applied to improve model accuracy.

III. METHODOLOGY

Despite significant advancement in fraud detection systems, here are still several challenge and gaps that need to be addressed. One key gap is the evolving nature of fraud techniques, where fraudsters constantly devise new ways to

bypass security measures, making it difficult for static or outdated models to keep up. Existing models often suffer from high false-positive rates, which can lead to legitimate transactions being flagged as fraudulent, causing inconvenience to customers and financial institutions. Real-time fraud detection remains a difficult task due to the need for immediate responses, which many systems struggle to achieve efficiently without compromising accuracy.

3.1 Random Forest Algorithm

Random Forest is a common deep learning algorithm that belong to the supervised learning technique. It can be used for both Classification and Regression problem in DL. It is grounded on the concept of collective learning, which is a procedure of combining multiple classifiers to resolve a complex problem and to improve the performance of the model.

USE RANDOM FOREST

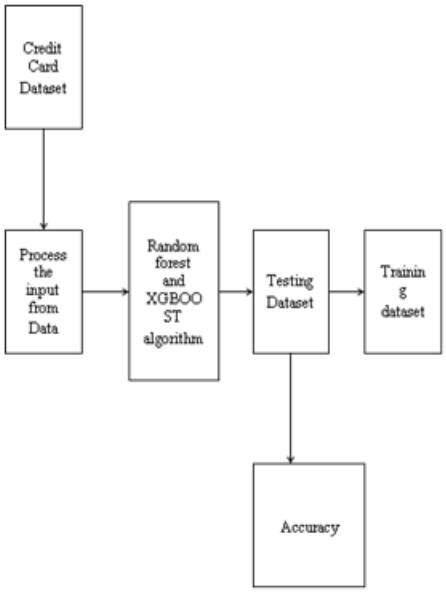
- Below are some point that clarify why we should use the Random Forest algorithm:
- It takings less training time as equaled to other algorithm.
- It predicted productivity with high accuracy, even for the huge dataset it run efficiently.
- It can also keep accuracy when a large proportion of data is lost.

XG BOOST

XGBoost stands for Extreme Gradient Boosting. It uses more accurate approximations to find the best tree model. Boosting: N new training data sets are formed by random sampling with replacement from the original dataset, during which some observations may be repeated in each new training data set.For each node, there is a factor γ with which hm(x) is multiplied. This accounts for the difference in impact of each branch of the split. Gradient boosting helps in predicting the optimal gradient for the additive model, unlike classical gradient descent techniques which reduce error in the output at each iteration.

$$\Omega(f) = \gamma * T + (1/2) * \lambda * \Sigma w_j^2$$

3.2 SYSTEM ARCHITECTURE



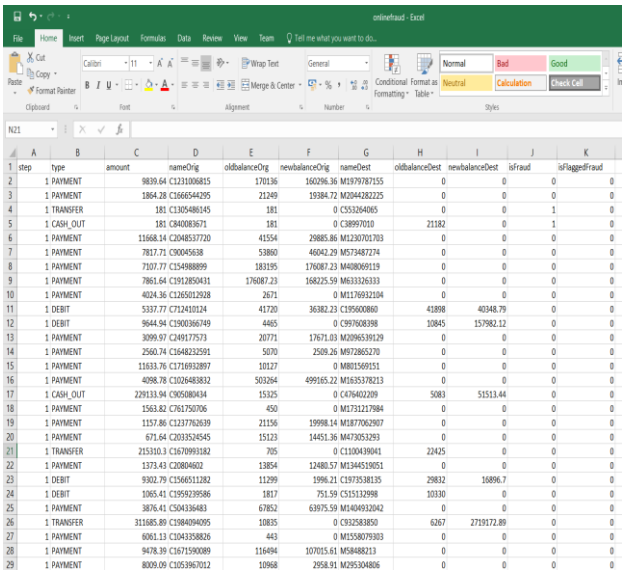
IV. EXPERIMENTAL RESULTS

In this section, a collection of knowledge (dataset) is collected that may be a set of report articles, To classify online payment fraud with machine learning, we want to train a machine learning model for classifying fraudulent and non-fraudulent payment. A dataset containing data about online payment fraud, so that we can know what type of transaction lead to fraud. For this task, I collected a dataset from Kaggle, which contain historical information about fraudulent transaction which can be used to detect fraud in online payments.

DATASET

Dataset : www.kaggle .com

- Dataset Name: Fraud Transaction
- Description: A dataset containing examples of phishing and legitimate emails.
- Source: UCI Machine Learning Repository
- Size:
  - Number of Records: 1,000 emails
  - Features: 30 attributes
- Types of Data: Text, Numeric, Categorical
- Target Variable: Fraud: Yes/No



step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
1	PAYMENT	9839.64	C1231006815	170136	160296.36	M1979787155	0	0	0	0
1	PAYMENT	1864.28	C1666544295	21249	19384.72	M2044282225	0	0	0	0
1	TRANSFER	181	C1305486145	181	0	C553264065	0	0	1	0
1	CASH_OUT	181	C840083671	181	0	C38997010	21182	0	1	0
1	PAYMENT	11668.14	C2048537720	41554	29885.86	M1230701703	0	0	0	0
1	PAYMENT	7817.71	C90045638	53860	46042.29	M573487274	0	0	0	0
1	PAYMENT	7107.77	C154988899	183195	176087.23	M408069119	0	0	0	0
1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0	0	0	0
1	PAYMENT	4024.36	C1265012928	2671	0	M1176932104	0	0	0	0
1	DEBIT	5337.77	C712410124	41720	36382.23	C195600860	41898	40348.79	0	0
1	DEBIT	9644.94	C1900366749	4465	0	C997608398	10845	157982.12	0	0
1	PAYMENT	3099.97	C249177573	20771	17671.03	M2096539129	0	0	0	0
1	PAYMENT	2560.74	C1648232591	5070	2509.26	M972865270	0	0	0	0
1	PAYMENT	11633.76	C1716932897	10127	0	M801569151	0	0	0	0
1	PAYMENT	4098.78	C1026483832	503264	499165.22	M1635378213	0	0	0	0
1	CASH_OUT	229133.94	C905080434	15325	0	C476402209	5083	51513.44	0	0
1	PAYMENT	1563.82	C761750706	450	0	M1731217984	0	0	0	0
1	PAYMENT	1157.86	C1237762639	21156	19998.14	M1877062907	0	0	0	0
1	PAYMENT	671.64	C2033524545	15123	14451.36	M473053293	0	0	0	0
1	TRANSFER	215310.3	C1670993182	705	0	C1100439041	22425	0	0	0
1	PAYMENT	1373.43	C20804602	13854	12480.57	M1344519051	0	0	0	0
1	DEBIT	9302.79	C1566511282	11299	1996.21	C1973538135	29832	16896.7	0	0
1	DEBIT	1065.41	C1959239586	1817	751.59	C515132998	10330	0	0	0
1	PAYMENT	3876.41	C504336483	67852	63975.59	M1404932042	0	0	0	0
1	TRANSFER	311685.89	C1984094095	10835	0	C932583850	6267	2719172.89	0	0
1	PAYMENT	6061.13	C1043358826	443	0	M1558079303	0	0	0	0
1	PAYMENT	9478.39	C1671590089	116494	107015.61	M58488213	0	0	0	0
1	PAYMENT	8009.09	C1053967012	10968	2958.91	M295304806	0	0	0	0

3.3 IMPLEMENTATION

Credit card dataset upload

Load all the libraries like pandas, math plot lib and numpy, then load the credit card data. The Credit Card Fraud Detection dataset was used, which can be taken from Kaggle. This dataset contains businesses, occurred in two days, made in September 2013 by European cardholder. The dataset contains 31 numerical features. Since some of the input variables contains commercial information, the PCA transformation of these input variables were implemented in order to keep these data anonymous.

CSV FILE PROCESSING

Feature selection is a fundamental method, which select the variable that are most relevant in the given dataset. Carefully selecting appropriate features and removing the less important one can reduce over fitting, increase accuracy and reduce training time. Visualization method can be helpful in that procedure. Feature selector tool by Will Koehrsen was used in this testing for that purpose. By using this device it has been determined which features are the most important. Furthermore, features that do not contribute to the cumulative importance of 95% were remove.

DETECT THE CREDIT CARD

Model Evaluation is an essential share of the model development procedure. It help to find the best model that represent our data and how well the selected model will work in the future. Evaluating model presentation with the data used for training is not satisfactory in data science because it can effortlessly generate overoptimistically and over fitted model.

RESULT

RF is the supervised learning. It is a kind of ensemble learning with majority voting techniques where mixture of expert's decision is taken into account. In ensemble learning, where predictions by applying different models. Multiple decision trees are constructed with optimal data points as next node. Especially this algorithm work better in improper scaling and missing data values. This algorithm uses the bagging concept where bootstrap samples are randomly selected and decision tree is constructed by training the data.

IV. RESULT & ANALYSIS

Performance Analysis

The results indicate that [briefly summarize findings, e.g., RF achieved the highest accuracy in detecting fraud detection methods outperformed traditional classification models]. The experimental results are presented in terms of various performance metrics, such as exactness, precision, recall, and F1-score, to evaluate the strengths and faintness of each model. In addition, the analysis provides insights into the most common types of fraud dataset and the trends observed over time.

Fig : 4.1 Table Performance Analysis

MODEL	ACCURACY
SVM	85
KNN	96
XG	95
RF	99.8

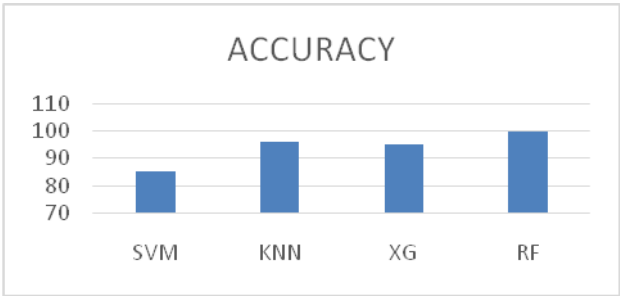


Fig : 4.1 Figure Performance Analysis

V. CONCLUSION

An original credit card data set from a financial institution has also remained used for evaluation. The same single and hybrid model have been employed. A perfect random forest and xgboost segmentation score of 1 has been achieved using ad boost and majority voting method. To

further procedure the hybrid models, noise from 10% to 30% has been added into the data example. The majority of the voting way has yielded the best accuracy score of 0.942 for 30% noise added to the data set. This show that the majority voting method is stable in presentation in the presence of noise. Random forest algorithm is the best algorithm.

REFERENCES

[1] Sahithi, G.L.; Roshmi, V.; Sameera, Y.V.; Pradeepini, G. Credit Card Fraud Detection using Ensemble Methods in Machine Learning. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1237–1241.

[2] Federal Trade Commission. CSN-Data-Book-2022. no. February 2023. Available online: https://www.ftc.gov/system/files/ftc_gov/pdf/CSN-Data-Book-2022.pdf (accessed on 11 March 2023).

[3] UK Finance. Annual Report and Financial Statements 2022. Available online: https://www.ukfinance.org.uk/annual-reports (accessed on 20 November 2023).

[4] Gupta, P.; Varshney, A.; Khan, M.R.; Ahmed, R.; Shuaib, M.; Alam, S. Unbalancd Credit Card Fraud Detection Data: AMachine Learning-Oriented ComparativeStudy of Balancing Techniques. *Procedia Comput. Sci.* 2023, *218*, 2575–2584.

[5] Mondal, I.A.; Haque, M.E.; Hassan, A.-M.; Shatabda, S. Handling imbalanced data for credit card fraud detection. In Proceedings of the 2021 24th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 18–20 December 2021; pp. 1–6. [Google Scholar]

[6] Ahmad, H.; Kasasbeh, B.; Aldabaybah, B.; Rawashdeh, E. Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS). *Int. J. Inf. Technol.* 2023, *15*, 325–333. [Google Scholar] [CrossRef] [PubMed]

[7] Bagga, S.; Goyal, A.; Gupta, N.; Goyal, A. Credit card fraud detection using pipelining and ensemble learning. *ProcediaComput. Sci.* 2020, *173*, 104–112. [Google Scholar] [CrossRef]

[8] Forough, J.; Momtazi, S. Ensemble of deep sequential models for credit card fraud detection. *Appl. Soft Comput.* 2021, *99*, 106883. [Google Scholar] [CrossRef]

[9] Karthik, V.S.S.; Mishra, A.; Reddy, U.S. Credit card fraud detection by modelling behaviour pattern using hybrid ensemble model. *Arab. J. Sci. Eng.* 2022, *47*, 1987–1997. [Google Scholar] [CrossRef]

[10] Sudjianto, A.; Nair, S.; Yuan, M.; Zhang, A.; Kern, D.; Cela-Díaz, F. Statistical methods for fighting financial crimes. *Technometrics* 2010, *52*, 5–19. [Google Scholar] [CrossRef]

[11] Data, S. Descriptive statistics. *Birth* 2012, *30*, 40. [Google Scholar]

[12] Walters, W.H. Survey design, sampling, and significance testing: Key issues. *J. Acad. Librariansh.* 2021, *47*, 102344. [Google Scholar] [CrossRef]

[13] Lee, S.; Kim, H.K. Adsas: Comprehensive real-time anomaly detection system. In Proceedings of the Information Security Applications: 19th International Conference, WISA 2018, Jeju, Republic of Korea, 23–25 August 2018; pp. 29–41. [Google Scholar]

[14] Sengupta, S.; Basak, S.; Saikia, P.; Paul, S.; Tsalavoutis, V.; Atiah, F.; Ravi, V.; Peters, A. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowl. Based Syst.* 2020, *194*, 105596. [Google Scholar] [CrossRef]

[15] Muppalaneni, N.B.; Ma, M.; Gurumoorthy, S.; Vardhani, P.R.; Priyadarshini, Y.I.; Narasimhulu, Y. CNN data mining algorithm for detecting credit card fraud. In *Soft Computing and Medical Bioinformatics*; Springer: Singapore, 2019; pp. 85–93. [Google Scholar]

[16] Roy, A.; Sun, J.; Mahoney, R.; Alonzi, L.; Adams, S.; Beling, P. Deep learning detecting fraud in credit card transactions. In Proceedings of the 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 27 April 2018; pp. 129–134. [Google Scholar] [CrossRef]

[17] Fiore, U.; De Santis, A.; Perla, F.; Zanetti, P.; Palmieri, F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Inf. Sci.* 2019, *479*, 448–455. [Google Scholar] [CrossRef]

[18] Somvanshi, M.; Chavan, P.; Tambade, S.; Shinde, S.V. A review of machine learning techniques using decision tree and support vector machine. In Proceedings of the 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 12–13 August 2016; pp. 1–7. [Google Scholar]

[19] Shah, R. Introduction to k-Nearest Neighbors (kNN) Algorithm. Available online: https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8 (accessed on 20 November 2023).

[20] Jadhav, S.D.; Channe, H.P. Comparative study of K-NN, naive Bayes and decision tree classification techniques. *Int. J. Sci. Res.* 2016, *5*, 1842–1845. [Google Scholar]

[21] A cost-sensitive decision tree approach for fraud detection, y. Sahin, S. Bulkan, and E. Duman expert systems with applications, vol. 40, no. 15, pp. 5916–5923, 2013.

[22] A survey of machine-learning and nature-inspired based credit card fraud detection techniques a. O. Adewumi and A. A. Akinyelu, "," international journal of system assurance engineering and management, vol. 8, pp. 937–953, 2017.

[23] "credit card fraud detection using hidden markov model," A. Srivastava, A. Kundu, S. Sural, A. Majumdar, IEEE transactions on dependable and secure computing, vol. 5, no. 1, pp. 37–48, 2008.