**Edelweiss Applied Science and Technology**
*ISSN: 2576-8484*
*Vol. 8, No. 5, 864-872*
*2024*
*Publisher: Learning Gate*
*DOI: 10.55214/25768484.v8i5.1781*


# Online transaction fraud detection in the banking sector using machine learning techniques

S. Vasudevan[1], Vediyappan Govindan[2*], Haewon Byeon[3*]
[1,2]Department of Mathematics, Hindustan Institute of Technology and Science, Chennai;
svasudevanidvrs@gmail.com (S.V) govindoviya@gmail.com (V.G)
[3]Department of AI Big data, Inje University, Gimhae, 50834, Republic of Korea; byeon@inje.ac.kr (H.B).

**Abstract:** In this paper Nowadays, online transactions are increasing rapidly and payment scams are also increasing. A hacker is a person who gains access to another person's financial information. Improving feature selection is important to detect transaction risk in large, multidimensional data. Fraud detection and classification performance is divided into two parts: training data and test data. During this process we have focused on data analysis and pre-processing. In a transaction's dataset, the predictive variables influence how well a machine learning algorithm for transaction risk identification classifies and identifies fraudulent transactions. This study examines the application of machine learning techniques in improving fraud detection in the banking industry. Using advanced algorithms and historical transaction data, the model aims to identify anomalous patterns that indicate fraudulent activity. Feature engineering, ensemble methods, and anomaly detection algorithms are used to improve the accuracy and performance of fraud detection systems. reducing false positives and provides an adaptive security mechanism. By harnessing the power of machine learning, this approach aims to strengthen the banking sectors defines against sophisticated fraud schemes. This study not only contributes to the academic understanding of fraud detection, but also provides industry practitioners with a practical framework for implementing proactive measures. In navigating the complex landscape of financial security, this research serves as a guide for building agile and responsive systems that stay ahead of the threats that evolve in the ever-changing landscape of the banking industry.

***Keywords:*** *Fraud Detection, Machine Learning Algorithms, Python.*

## 1. Introduction

The global banking sector, which is essential to maintaining economic stability, is increasingly threatened by fraudulent operations that take advantage of the nuances of online transactions. The growing complexity of transactions resulting from financial institutions' digitization creates a favable environment for hostile actors aiming to obtain illegal access and profit. The conventional approaches to fraud detection, which mostly rely on rule-based algorithms, have demonstrated their inability to adjust to the ever-changing nature of fraudulent schemes. A growing number of people are interested in applying machine learning techniques to improve the security of financial systems against fraudulent activity as a solution to this challenge.

By allowing computers to recognize and adjust based on patterns in data, machine learning techniques provide a paradigm leap in fraud detection. Real-time detection of fraudulent activity and subtle anomalies can be achieved through the use of unsupervised techniques like clustering and supervised learning algorithms like support vector machines and random forests. for deciphering intricate patterns and big information. Machine learning's capacity to recognize and market itself as a potent instrument for raising the precision and effectiveness of fraud detection algorithms.


* Correspondence: govindoviya@gmail.com, byeon@inje.ac.kr

Concerns about security and privacy are crucial when using these ML approaches. Important financial data must be used with strict adherence to ethical standards and data security rules. Verify the integrity and confidentiality of consumer data. To stop fraud, you must keep a delicate balance.

Cooperation between financial institutions will increase the performance of these machine learning techniques. Share Intelligent, Trends and Sample Developments, Helps to create a more comprehensive and integrated protection against the online transaction fraud. This joint approach can lead to the most subtle and accurate models, which also benefit from the entire bank department.

## 2. Literature Review

Pradeepan Raghavan et al. [1] Provide a comprehensive assessment of various machine learning and deep learning models including Restricted Boltzmann Machines (RBM), Autoencoders, Random Forest, Convolutional Neural Networks (CNN), Support Vector Machines (SVM), K-Nearest Neighbors (KNMBL). A method combining KNN, SVM and CNN demonstrates different approaches to tackle complex data analysis challenges. Includes the collection of three different datasets: an Australian dataset, a German dataset and a European dataset, each providing valuable insights for research inquiry. The development of hybrid, dynamic and adaptive systems will create many open areas of future research.

Seyedeh Khasemi et al. [2] proposed weight adjustment as a pre-processing technique for handling unbalanced data. They demonstrated that by employing CatBoost and XGBoost and incorporating a voting mechanism, improvements in LightGBM's performance were achieved. Furthermore, the study acknowledged that due to hardware limitations, utilizing more robust hardware may yield even more favorable results.

D. Kalbande. [3] To address the critical issue of fraudulent transactions in online payment systems, a study titled "Detection of Online Payment Fraud" was conducted. By applying various investigative [8] machine learning techniques, including big data analysis and various machine learning algorithms, the researchers delved into the realm of online payment security. Fraud to protect businesses and their customers from financial losses.

S. Venkatachalam [4] contributed to the field of fraud detection by using domain knowledge to derive intelligent features to improve the performance of ensemble methods such as random forest and gradient boosting. By integrating domain-specific knowledge into the feature engineering process, the researchers aimed to improve the accuracy and robustness of fraud detection models. Furthermore [9], they emphasized the importance of continuous evaluation and refinement through techniques such as cross-validation and A/B testing, highlighting the dynamic nature of fraud detection systems. Their work [10] underscores the need for iterative approaches in updating and improving fraud detection systems to effectively combat emerging fraud activities.

D. Singh [5] highlight the importance of SVMs in classification tasks, emphasizing their ability to improve accuracy while being scalable to manage large-scale transactions. Future research may focus on exploring advanced optimization techniques and hybrid approaches to further improve the performance of SVMs in classification tasks in various domains.

K. G. Krishna [6] where fraud arises in the financial sector through multiple channels such as credit cards, e-commerce, telephone banking, checks and online banking, used a real dataset from a private bank to evaluate our methods. Developing an effective fraud detection system. Each banking session aims to encode deviations from typical customer behavior.

R. Tiwari [7] have some techniques like machine learning logistic regression, random forest, support vector machine algorithm etc. on the dataset and compared their performance to know which one is better among these three. Comparing the results of these three algorithms, the forest algorithm gives the best result.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

*2.1. About the Data*

**Type:** Transactions can be categorized into several types, such as debit, credit, transfer, or other categories, using the "type" column. Analyzing financial activity requires a fundamental understanding of transaction kinds. For example, among particular transaction categories, fraudulent actions may display particular patterns. money are taken out of an account during a debit transaction, whereas money are added during a credit transaction. Transfer transactions entail transferring money between accounts, and since they might involve substantial quantities of money, they are frequently the focus of fraudsters. Financial institutions can recognize patterns and abnormalities in transactional data, which can support fraud detection initiatives.

**Amount:** Each transaction's corresponding monetary value is shown in the "amount" column. It indicates how much money was moved or how much the transaction effected. It is essential to keep an eye on transaction amounts in order to spot odd financial activity. Unauthorized financial transfers or money laundering are two examples of fraudulent activity that might be indicated by large or irregular transaction quantities. Financial institutions can detect unusual trends and reduce the risks connected with fraudulent transactions by examining transaction quantities in conjunction with other characteristics.

**Name Orig:** The sender or originator of the transaction is most usually identified by the "nameOrig" field. It facilitates the differentiation of several account holders or companies who are starting transactions. Comprehending the origin of transactions is crucial for monitoring financial movements and detecting possible deceivers. Financial institutions can evaluate the validity of transactions and identify unlawful activity coming from hacked accounts or fraudulent sources by examining the transaction origins.

**Oldbalance Org:** The account balance of the originator prior to the transaction is shown in the "oldbalanceOrig" column. It shows the amount of money available for transactions as well as the originator's account's starting balance. Keeping an eye on account balances aids in the identification of irregularities and unlawful activity. Unauthorized fund withdrawals or account takeover are examples of fraudulent activity that might be indicated by notable variations in account balances, particularly before and after transactions.

**New balance Org:** Following the completion of the transaction, the originator's revised account balance is shown in the "newbalanceOrig" column. After deducting the transaction money, it shows the original user's account balance. Overdrafts, inadequate money, and illegal account modifications are examples of possible fraud that may be found by analyzing changes in account balances. Inconsistencies between the previous and current balances may indicate fraudulent activity, prompting financial institutions to look into the matter further.

**Name Dest:** The receiver or destination of the transaction is most usually identified by the "nameDest" field. It facilitates the distinction between several account holders or recipients of funding. Tracking money movements and spotting possible fraud requires an understanding of transaction destinations. Financial institutions can determine if a transaction is legitimate and identify any suspect activity involving fake accounts or unapproved beneficiaries by examining the transaction destinations.

**Old balance Dest:** "Old balance Dest" is the column that displays the destination's account balance before the transaction was made. The initial balance of the recipient's account is displayed, along with the amount of money that was accessible before the transaction was received. Monitoring destination account balances facilitates the identification of anomalies and illicit money transfers. Significant variations in the balances of destination accounts may indicate dishonest activities, such as account manipulation or money laundering.

**New balance Dest:** Following the completion of the transaction, the destination's revised account balance is shown in the "new balance Dest" column. After deducting the transaction amount, it displays the recipient's account's current balance. Potential fraud, such as illegal fund injections, account modifications, or unusual activities, might be found by analyzing changes in destination account

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

balances. Differences between the old and new balances may indicate fraud, which will compel financial institutions to look into the matter further and take the necessary measures.

**Is Fraud:** The binary indication in the "is Fraud" column indicates whether or not the transaction is fraudulent (1) or not (0). It helps distinguish between transactions that are suspicious and those that are lawful by acting as the target variable for fraud detection algorithms. Financial institutions must identify fraudulent activity immediately in order to safeguard their clients and minimize damages. Machine learning models may aid in automating fraud detection procedures and improving the general security of financial systems by evaluating transaction data and spotting fraudulent trends.

## 3. Methodology

The methodical process or series of steps used in the creation, testing, and deployment of machine learning models in Figure 1.

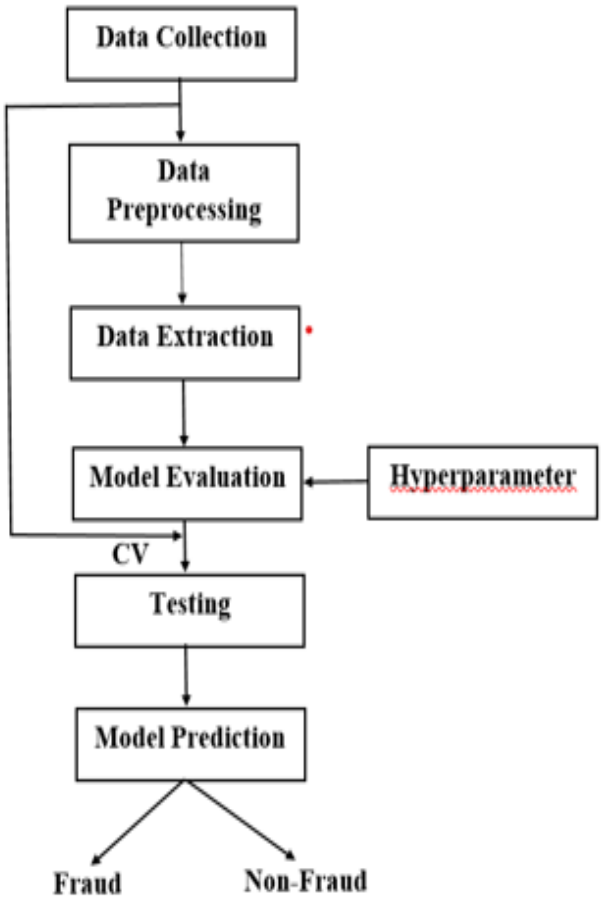


**Figure 1.**
Flow chart of methodology.

### *3.1. Data Preprocessing*

This study could provide light on consumer behavior and transaction patterns. It might also point out areas that need more research, including figuring out what constitutes fraud or figuring out what the preferences of the consumer are.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

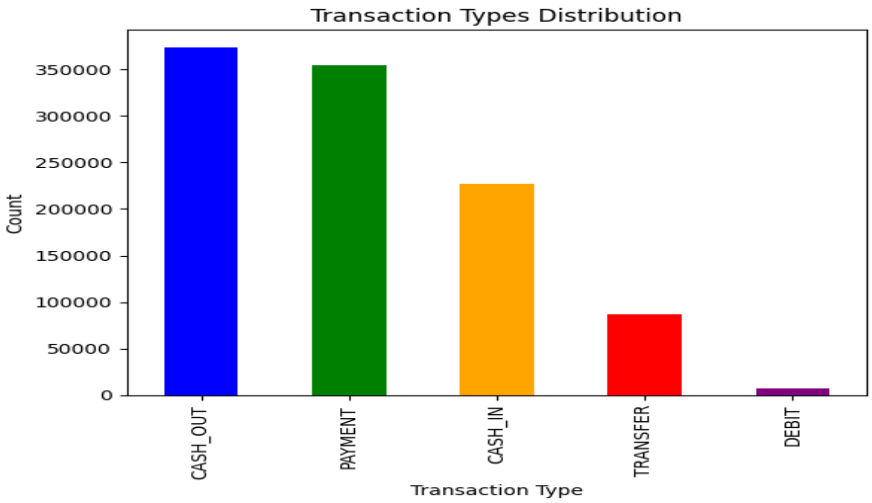


**Figure 2.**
Distributions of transactions.

The most common transaction type, Cash_Out (373,641), shows a high amount of cash withdrawals. Remittance (353,873) The high level of payments also suggests that a significant sum of money is being transferred or spent for a variety of reasons shown in Figure 2. Cash_in (227,130) The frequency of cash deposits is lower than the frequency of cash withdrawals and payments, indicating that more money is leaving accounts than entering them. Transfer (86,753) The fact that transfers make up a lesser percentage of all transactions suggests that, in comparison to other transaction kinds, direct money transfers between accounts happen less frequently. Debt (7,178) The fact that debit transactions are the least common suggests that, in comparison to other transaction kinds, withdrawals from accounts are not common.

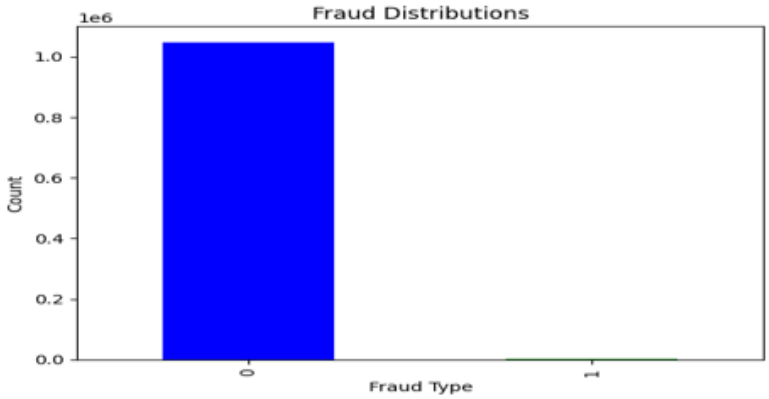


**Figure 3.**
Counts of fraud and non-fraud transactions.

When the model predicts that a transaction is not fraudulent, it is represented by non-fraud (0) shown in Figure 3. There are a lot of transactions that are categorized as not fraudulent, as seen by the total of 1,047,433 that is linked to this category.

Instances where the model expects a transaction to be fraudulent are represented by Fraud (1). 1,142 is the total for this category, meaning that there were a lot less transactions flagged as fraudulent. Examining how average account balances for both fraudulent and non-fraudulent transactions fluctuate over time is part of the process of analysing average balance by fraud type over time shown in Figure 4 & 5.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

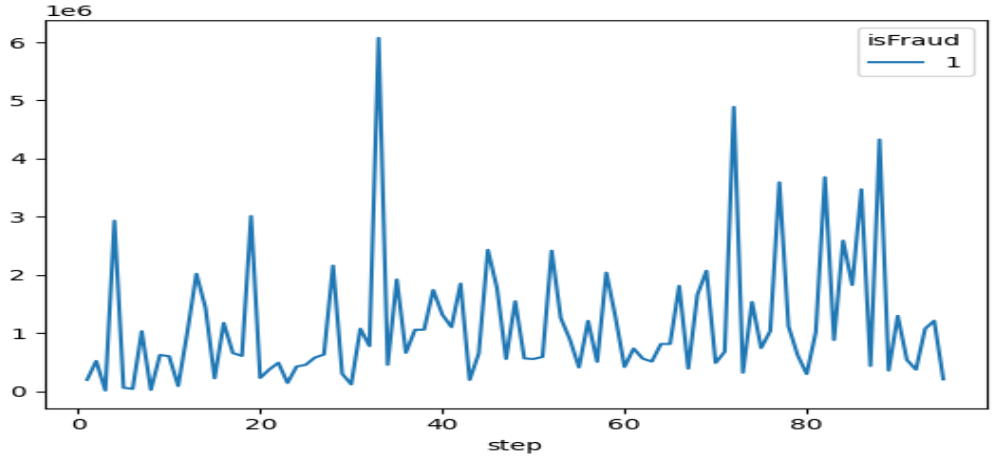


**Figure 4.**
Average account balances for fraudulent.

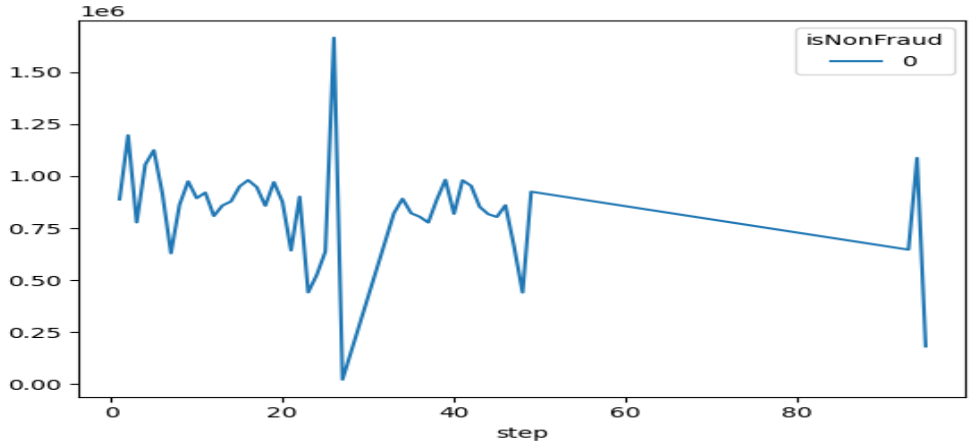


**Figure: 5.**
Average account balances for non-fraudulent.

*3.2. Data Extraction*

The discrepancy between the actual balance determined by utilizing the transaction amount (amount) and the starting balance in the origin account (oldbalanceOrig) and the predicted balance in the origin account (newbalanceOrig) following the transaction is known as errorBalanceOrig. It aids in locating any mistakes or disparities in the balance of the origin account following the transaction. The difference between the predicted balance in the destination account (oldbalanceDest) prior to the transaction and the actual balance (newbalanceDest modified by the transaction amount) after the transaction is computed by errorBalanceDest. It aids in locating any mistakes or disparities in the balance of the destination account following the transaction.

## 4. Model Evaluation

**Logistic Regression:** A binary classification statistical approach where the outcome variable (dependent variable) is categorical and has only two possible values, often represented by the numbers 0 and 1 shown in Figure 6. It is known as "logistic" regression as it is predicated on the logistic function, also sometimes termed the sigmoid function.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

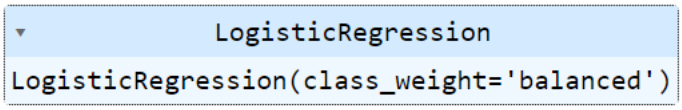


**Figure 6.**
Logistic regression.

**Decision Tree Regressor:** One kind of hierarchical tree structure used in regression and classification issues is the decision tree. Every internal node represents a feature, and every leaf node represents a class label or continuous value. Optimizing homogeneity within each subgroup is the aim of the feature-value-based data splitting technique used in the tree. Recursively, the dataset is split up into smaller subsets until a halting condition is met. Interpretable decision trees are capable of illustrating non-linear relationships in the data. Overfitting is a common problem with complex structures, but it may be prevented by regularizing them using techniques like pruning. Ensemble techniques, like Random Forests in Figure 7, mix several decision trees to improve prediction performance.

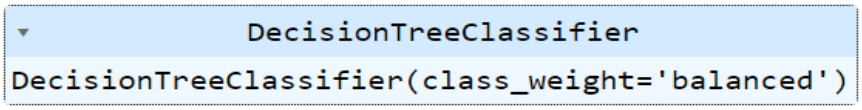


**Figure 7.**
Decision tree classifier model.

**Decision tree-based Random Forest:** is an ensemble learning technique for classification and regression applications. Several decision trees are generated during training, and the predictions of each tree are subsequently integrated using averages (for regression) or votes (for classification). Every tree in the forest is trained using a random selection of features and training data in order to reduce overfitting and improve generalization. The final prediction made by the Random Forest in Figure 8 is based on the sum of the forecasts from each individual tree.

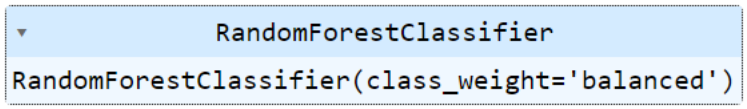


**Figure 8.**
Random forest classifier model

*4.1. Hyper-Parameter*

The process of choosing the ideal collection of hyperparameters for a machine learning algorithm to maximize performance is known as hyperparameter tuning. Hyperparameters are pre-learning parameters, such the size of a Random Forest tree or the intensity of the regularization in a logistic regression, that are determined before the learning process starts.

- To investigate a variety of regularization strength values, employ strategies such as grid search or randomized search. Each set of hyperparameters is usually tested using cross-validation, and the set that performs the best is chosen.
- Use techniques like randomized or grid search to look at different combinations of hyperparameters. Cross-validation is used to evaluate the performance of each combination, and a measurement of the assessment is used to determine which combination is best.

Based on your evaluation metric and the findings of your study, it looks that this set of hyperparameters (C=10, penalty='l2') worked nicely.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

Hyperparameters are set up in the decision tree classifier to maximize its performance. It guides the tree's splitting process by assessing node purity using Gini impurity. The tree's complexity is limited to 15 in order to avoid the tree overfitting to the training set. Furthermore, the minimum number of samples that must exist at a leaf node is fixed at two, guaranteeing that each leaf has enough information to provide trustworthy predictions. Together, these hyperparameters provide a decent compromise between generalization and model complexity, leading to decision tree classification that performs well on the provided dataset.

Enhance the model's performance using untested data. In order to guarantee that the model fits effectively to fresh, untested data and to avoid overfitting, hyperparameter adjustment is used. However, since hyperparameter tweaking may be computationally expensive, particularly for big datasets and sophisticated models, it's crucial to balance time limitations and computing resources shown in Table 1.

**Table 1.**
Findings of accuracy without and with Hyperparameter.

| | Without hyperparameter | | With hyperparameter | | Scores | |
|---|---|---|---|---|---|---|
| **Models** | **Training score** | **Testing score** | **Training score** | **Testing score** | **Precision** | **Recall** |
| Logistic refression | 0.96313 | 0.96 | 0.99929 | 0.99928 | 0.9097 | 0.3954 |
| Decision tree | 0.99993 | 0.99 | 0.999971 | 0.999939 | 0.9714 | 0.9741 |
| Random forest | 0.99997 | 0.98 | 0.999991 | 0.999942 | 0.9832 | 0.9769 |
| SVM | 0.98976 | 0.95 | 0.987812 | 0.979912 | 0.9912 | 0.9836 |

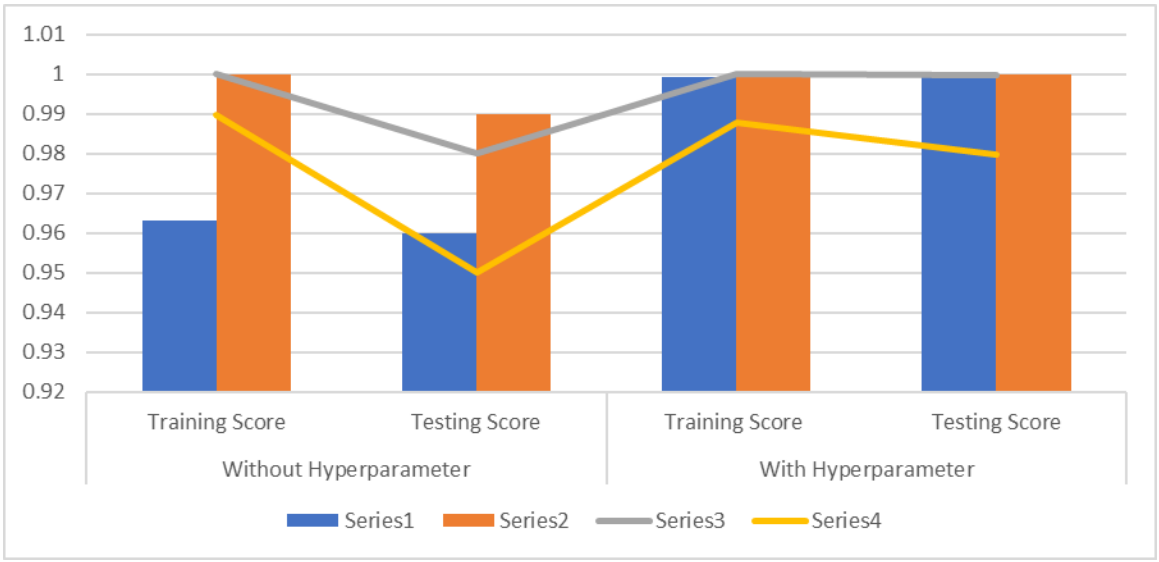


**Figure 9.**
Flow chart of accuracies with and without hyperparameters.

## 5. Conclusions

In summary, this study looked at the performance of four machine learning models: support vector machines (SVMs), logistic regression, random forests, and decision trees. Each model was trained and evaluated using a dataset tailored to our specific research objectives. Following a comprehensive analysis, it was evident that random forest outperformed the other models in terms of accuracy during

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 5: 864-872, 2024*
*DOI: 10.55214/25768484.v8i5.1781*

the training and testing phases. Notably, the predictions made by the random forest model consistently demonstrated excellent levels of reliability and accuracy.

While SVM, logistic regression, and decision tree models performed wonderfully overall, random forest surpassed them in terms of accuracy shown in Figure 9. Despite these drawbacks, the models contributed to a comprehensive understanding of the problem domain and provided valuable insights into the underlying data patterns. Ultimately, the superior performance of the random forest model demonstrates how well it processes complex data and produces accurate predictions. Further investigation into ensemble techniques and optimization tactics may enhance the predictive potential of our models and provide even deeper insights into the problem at hand in the future.

## Funding:
This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF- RS-2023-00237287, NRF-2021S1A5A8062526) and local government-university cooperation-based regional innovation projects (2021RIS-003).



## References
[1]. Raghavan, Pradheepan & Gayar, Neamat. (2019). Fraud Detection using Machine Learning and Deep Learning. 334-339. 10.1109/ICCIKE47802.2019.9004231.

[2]. Seyedeh Khasemi et al., Singh, Jashandeep & Kaur, Prabhjot. (2023). Fraud Detection in Online Transactions Using Machine Learning. 10.13140/RG.2.2.29971.66088.

[3]. D. Kalbande, P. Prabhu, A. Gharat and T. Rajabally, "A Fraud Detection System Using Machine Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021.

[4]. S. Venkatachalam, P. Priyadarsini, K. Jayasree, V. B. Pawar, R. Sasikala and S. Loganathan, "Analysis of Machine Learning Based Credit Card Transaction and its Applications," 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2024.

[5]. D. Singh, "Protecting Contactless Credit Card Payments from Fraud through Ambient Authentication and Machine Learning," 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Kalady, Ernakulam, India, 2023.

[6]. K. G. Krishna, P. Kulkarni and N. A. Natraj, "Use of Big Data Technologies for Credit Card Fraud Prediction," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023.

[7]. S. Ghosh, J. Kumar and T. Pangotra, "Fraud Detection System Analysis," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023.

[8]. R. Tiwari, S. Rautela, S. Sharma, B. Pratap Choudhary, R. Tripathi and P. Singh, "Role of AI for Fraud Detection in Banks: A Bibliometric Analysis," 2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech), Banur, India

[9]. N. Boutaher, A. Elomri, N. Abghour, K. Moussaid and M. Rida, "A Review of Credit Card Fraud Detection Using Machine Learning Techniques," *2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, Marrakesh, Morocco, 2020

[10]. B. Can, A. G. Yavuz, E. M. Karsligil and M. A. Guvensan, "A Closer Look Into the Characteristics of Fraudulent Card Transactions," in IEEE Access, vol. 8, pp. 166095-166109, 2020, doi: 10.1109/ACCESS.2020.3022315.