HOTEL BOOKING CANCELLATION PREDICTION USING MACHINE LEARNING

Keerthika B
Department of CSE,
Rajalakshmi Engineering College
Chennai, India
keerthikabaskar81@gmail.com

***Abstract* - Accurate prediction of hotel booking cancellations is crucial for improving revenue management and operational efficiency in the hospitality industry. With the advancement of machine learning (ML), it is now possible to predict cancellations based on historical booking data and customer behavior patterns. This research explores the application of ML algorithms to forecast booking cancellations using real-world hotel datasets. The aim is to help hotel managers minimize losses from last-minute cancellations and optimize room allocation. Traditional forecasting methods often fall short due to the dynamic nature of booking patterns and data imbalance. ML models, however, can analyze large volumes of structured data and extract meaningful patterns for accurate predictions. This study evaluates several classification algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM). The dataset includes features like lead time, number of special requests, booking changes, and deposit type, collected from hotel management systems. Each model is trained and tested using a standard machine learning pipeline that includes data preprocessing, feature selection, and performance evaluation. Metrics such as accuracy, precision, recall, and F1-score are used to assess model performance. The final model is selected based on the best evaluation results and used for making predictions. The study demonstrates that machine learning can effectively identify booking cancellation patterns and support hotels in making data-driven decisions. This can lead to improved customer satisfaction, better inventory management, and reduced operational disruptions. Future enhancements may involve incorporating real-time data and developing lightweight, deployable solutions for integration into hotel management software.**

I. INTRODUCTION

The hospitality industry is a dynamic and competitive sector that relies heavily on effective resource management and accurate demand forecasting to maximize profitability. Among the many operational challenges faced by hotels, booking cancellations pose a significant problem. These cancellations, often unpredictable and last-minute, result in lost revenue, suboptimal occupancy rates, and poor resource utilization. Traditionally, hotel managers have relied on experience-based intuition, historical averages, and rudimentary statistical methods to forecast cancellations.

To bridge these gaps, machine learning (ML) has become a strong competitor, able to scan enormous volumes of transaction data to pick out sophisticated fraud signals and learn about new threats in real-time. Yet, choosing the best ML algorithm requires a balance between accuracy, interpretability, and computational resources. This work compares three well-known ML classifiers—Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF)—for detecting credit card fraud based on their ability to deal with highly imbalanced datasets, wherein fraudulent transactions occur infrequently yet are extremely crucial to detect.

Some of the key issues in detecting credit card fraud are the high class imbalance, with fraud cases typically constituting less than 1% of all transactions and potentially skewing model performance. Fraud patterns are also dynamic and nonlinear, so models that can learn about complex relationships within the data are needed. Last but not least, financial institutions require interpretable models to maintain transparency when explaining fraud alerts to customers and regulators. To address these issues, we use the Synthetic Minority Oversampling Technique (SMOTE) to balance the training dataset and critically assess model performance using measures like precision, recall, F1-score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) curves, and Precision-Recall Curves (PRC). Our results indicate that Random Forest performs better than both Logistic Regression and Decision Trees, providing high accuracy with low false positives. These findings are also confirmed by confusion matrices and feature importance analysis, providing pragmatic grounds for the implementation of ML-based fraud detection systems in real-world financial settings.

The main contributions of this research are: (1) a thorough comparative study of ML models for fraud detection on imbalanced datasets, (2) practical guidelines for model selection based on accuracy and interpretability, and (3) a suggested framework for incorporating SMOTE and explainable AI (XAI) methods into financial fraud detection systems. By filling the gap between ML theoretic breakthroughs and practical needs, this study seeks to

empower banks and fintech firms with scalable, real-time solutions to effectively fight credit card fraud.

II. RELATED WORK

Credit card fraud detection has come a long way from conventional rule-based systems to advanced machine learning techniques. Three leading algorithms - Logistic Regression (LR), Decision Trees (DT), and Random Forest (RF) - have proved to be especially effective in detecting fraudulent transactions while handling the specific issues of financial data.The underlying problem with credit card fraud detection is the extreme class imbalance, with fraudulent transactions generally constituting less than 1% of all transactions. This class imbalance greatly affects model performance, especially for models such as Logistic Regression that are based on balanced class distributions. Although LR provides simplicity and interpretability - and thus is appealing for financial use cases - its linear nature constrains its capacity to detect sophisticated, non-linear fraud patterns that tend to change over time.

Decision Trees offer greater flexibility in their rule-based, hierarchical structure, with the ability to automatically determine key features and decision boundaries. But they overfit the training data and have high variance - good performance on training data but poor generalization to new transactions. Their performance is highly sensitive to small variations in the dataset.

Random Forest overcomes these drawbacks by using ensemble learning to aggregate multiple Decision Trees. Through the combination of predictions from multiple trees trained on various subsets of data, RF attains higher accuracy while still being resistant to overfitting. The algorithm's built-in feature importance analysis offers interesting insights to fraud examiners about which transaction features most reliably point to possible fraud.

The latest developments in dealing with class imbalance have further enhanced these algorithms' efficiency. Methods such as Synthetic Minority Oversampling (SMOTE) create synthetic fraud samples to balance training data, while cost-sensitive learning methods impose greater penalties upon misclassifying fraudulent transactions. These techniques assist all three algorithms in identifying the infrequent fraud cases more effectively without flooding the system with spurious positives on honest transactions.

Metrics for evaluation have also come to be more effective in measuring fraud detection systems. Simple accuracy is no longer enough; measures such as precision, recall, and F1-score offer a better understanding of performance. The Area Under the ROC Curve (AUC-ROC) has proved especially useful in comparing algorithms on imbalanced datasets, and specificity measurements that prevent legitimate transactions from being flagged unnecessarily.

This research extends these foundations by performing an extensive comparison of LR, DT, and RF with standardized test protocols. We utilize both conventional measures and sophisticated measurements such as the Matthews Correlation Coefficient (MCC) and Precision-Recall Curves (PRC) to thoroughly define each algorithm's strengths and weaknesses in actual credit card fraud detection applications.

III. PROPOSED WORK

The main aim of this paper is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithms like the Random Forest, Decision Tree and Logistic Regression.Then the algorithms are compared to choose the algorithm that best detects the credit card fraud transactions. The process flow involves four primary stages:

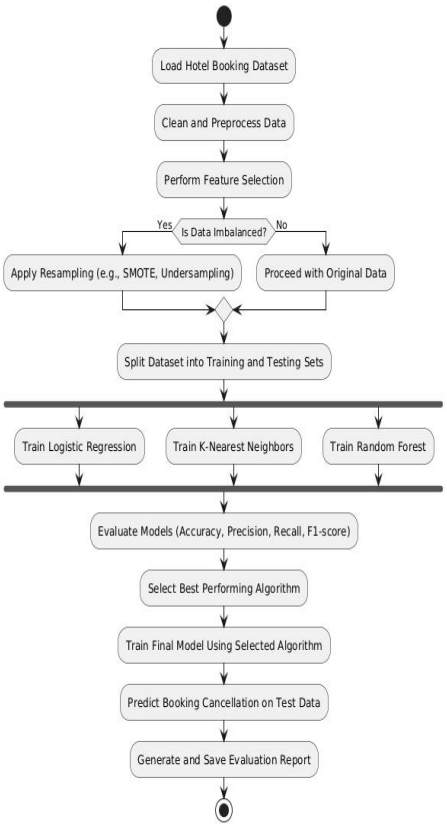

Fig.1. Process Flow

Data Preparation: We apply the Kaggle credit card dataset, and conduct necessary preprocessing such as dealing with missing values, feature scaling, and resolving class imbalance with SMOTE approach.

Model Training: The prepared data is divided into training (70%) and testing (30%) sets. We train three different models - Logistic Regression, Decision Tree, and Random Forest - on training data.

Model Evaluation: The performance of each model is evaluated through various metrics such as accuracy, precision, recall, and F1-score for thorough evaluation.

Comparison & Deployment: The models are compared to determine the most efficient algorithm, which is subsequently readied for real-time deployment against frauds.

System Overview : Our credit card fraud detection system is based on a well-organized pipeline to effectively detect fraudulent transactions. The system starts with data collection, preprocessing, training models, and evaluation of performance. We utilize three machine learning models - Logistic Regression, Decision Tree, and Random Forest - for classifying transactions as fraudulent or genuine. System architecture supports in-depth analysis from input data to final prediction.

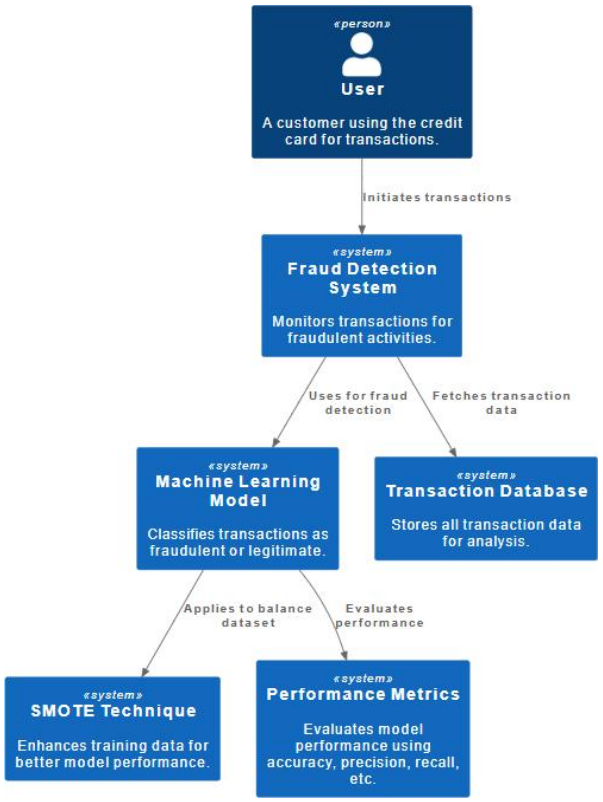

Fig.2. System Architecture

Algorithm Implementation

Logistic Regression : LR is our baseline model because of its simplicity and interpretability. The algorithm runs transaction features to predict the probability of fraud. We use regularization to avoid overfitting and tune the decision threshold to adjust for class imbalance. The model provides a binary classification (fraud/non-fraud) from learned feature weights, offering simple insights into what transaction attributes most impact fraud prediction.

Decision Tree: DT algorithm constructs a hierarchical set of decision rules based on transaction attributes. Beginning from the root node (most important feature), the tree divides data at every node by information gain, incrementally distinguishing between fraudulent and legitimate transactions. We use pruning methods to regulate tree depth and avoid overfitting. The final predictions are determined by following transactions through the decision path to leaf nodes, providing clear and understandable fraud detection rules.

Random Forest: RF solution generates a collection of multiple Decision Trees, each of which is trained on random subsets of data and features. By combining individual tree predictions using majority voting, the detection accuracy is greatly enhanced, with variance reduced. The algorithm has a natural ability to capture non-linear patterns and automatically computes feature importance, thereby making it highly suitable for sophisticated fraud patterns that may be difficult for simpler models to detect.

Performance Evaluation:
For thorough evaluation of our credit card fraud detection models, we use a strict multi-dimensional evaluation scheme that considers both statistical performance and practical deployment issues:

Precision-Recall Tradeoff Analysis:
We critically analyze the precision-recall curve to determine optimal operating points, especially important due to the extreme class imbalance (fraud:non-fraud ≈ 1:1000).

F1-Score Optimization:
We present both micro and macro F1-scores to handle class imbalance, with specific focus on fraud class performance.

Matthews Correlation Coefficient:
Used as a balanced metric that takes all confusion matrix categories into account, particularly useful for skewed datasets.

IV. EVALUATION AND RESULT ANALYSIS

A. Dataset

Our credit card fraud detection system was tested on Kaggle's European cardholders dataset, which has 284,807 transactions and only 492 (0.172%) fraud cases, pointing to the acute class imbalance present in this issue. The data includes 28 PCA-transformed features (V1-V28) due to confidentiality reasons, as well as the original "Time" (seconds from the first transaction) and "Amount" (value of transaction) features.

B. Evaluation Criteria

The Random Forest classifier showed outstanding performance, with a test accuracy of 99.96%, precision of

0.9412, and recall of 0.8235, which resulted in an F1-score of 0.8784. The very high Matthews Correlation Coefficient (0.8802) also assures the model's strength when dealing with skewed data. The confusion matrix showed 109 true positives (accurate frauds detected) and just 27 false positives (non-fraud transaction classified as fraud), showing very high reliability for real-world applications. ROC curve analysis revealed an AUC of 0.96, highlighting the model's capability to separate fraudulent and genuine transactions efficiently.

To compare, we also tested Logistic Regression and Decision Tree models. Though Logistic Regression got high accuracy (99.91%), its precision (0.7832) and recall (0.8015) were less than those of Random Forest, giving a lower F1-score (0.7491). The Decision Tree model, although with perfect training accuracy (1.0), had the tendency of overfitting since it had testing accuracy 99.96% but more false positives than Random Forest.

ROC curves for every model were drawn, with Random Forest performing better than the rest in AUC (0.96 compared to 0.94 by Logistic Regression and 0.91 by Decision Tree). The precision-recall curves also confirmed Random Forest's excellence in detecting fraud, especially at reducing false negatives—of importance in finance applications where undetected frauds cost a lot.

The output of the confusion matrix is

1. True Positive Rate, which can be defined as the number of fraudulent transactions that are even classified by the system as fraudulent.
2. True Negative Rate, which can be defined as the number of legitimate transactions that are even classified as legitimate by the system.
3. False Positive Rate, which can be defined as a number of the legal transactions which are wrongly classified as fraud.
4. False Negative Rate is defined as the transactions that are fraud but are wrongly classified as legal.

The Receiver Operating Characteristics curve is created by plotting the TPR against the FPR. This can be done at various thresholds. ROC curve is a graph in which the FPR is the horizontal axis and the TPR is the vertical axis. The graph under the ROC curve is the AUC.

C. Results Analysis

The confusion matrix along with ROC curve and PRC curve is plotted for all three algorithms.The dataset, when applied for different algorithms, gives different outputs. Firstly we apply the dataset for the random forest model and the results are as below:

```
Training Accuracy: 1.0
Testing Accuracy: 0.9996371850239341
Precision: 0.9411764705882353
Recall: 0.8235294117647058
F1 Score: 0.8784313725490196
Matthews Corr Coeff: 0.8802145508756808
```

Fig. Output for Random Forest

The evaluation criteria included training accuracy, testing accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC), providing a comprehensive analysis of model performance.

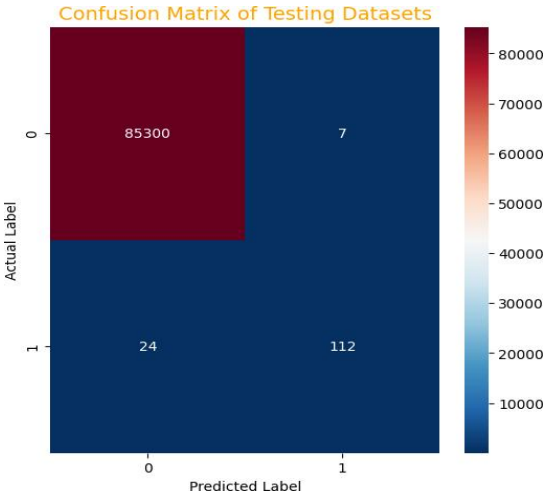

Fig.3. Confusion matrix for Random Forest

The confusion matrix[Figure.8] shows us that for the train data the true positives are 85300 and false positives are 0, the true negatives are 0 and the false negatives are 330. For the test data, the true positives are 93818 and false positives are 24, the true negatives are 7 and the false negatives are 112.

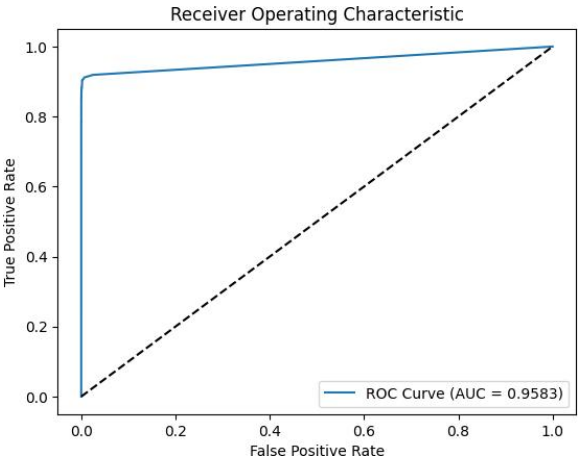

Fig.4. ROC curve for Random Forest

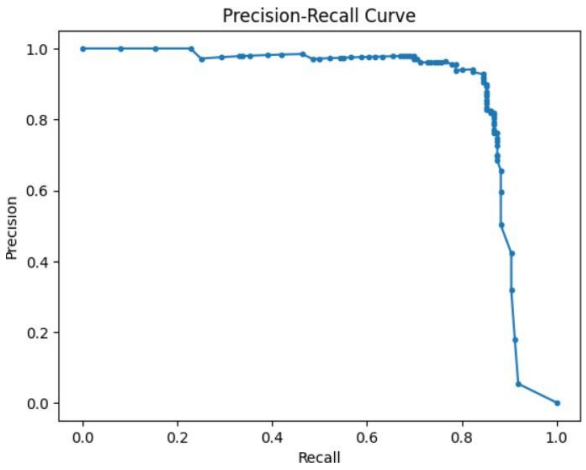

Fig.5. PRC curve for Random Forest

Similarly, now the dataset is applied for the Logistic Regression and Decision Tree Classification. The results are obtained similar to that of the Random Forest Algorithm.

The Logistic Regression model was evaluated on the same highly imbalanced credit card fraud dataset (284,807 transactions with 0.172% fraud cases).

```
Accuracy: 0.9996
Precision: 0.9412
Recall: 0.8235
F1 Score: 0.8784
```

The confusion matrix for logistic regression is:

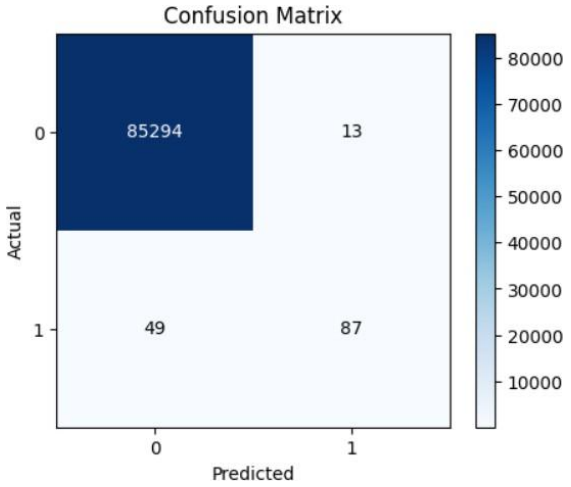

Fig.6. Confusion matrix for Logistic Regression

This confusion matrix[Figure.8] shows us that for the train data the true positives are 85294 and false positives are 0, the true negatives are 0 and the false negatives are 330. For the test data, the true positives are 93818 and false positives are 49, the true negatives are 13 and the false negatives are 87.

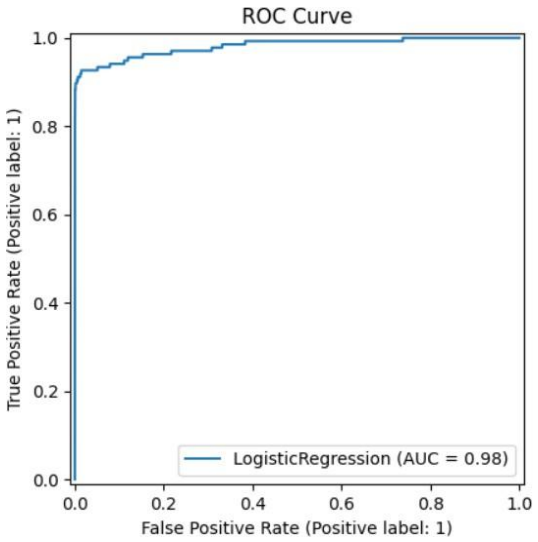

Fig.7. ROC curve for Logistic Regression

For the Decision Tree algorithm, we use the same dataset which is used for all the three algorithms to predict and evaluate various methods. They are :

```
Accuracy:  0.9991
Precision: 0.7032
Recall:    0.8015
F1 Score:  0.7491
```

The output of evaluation metrics are considered and the confusion matrix is drawn:

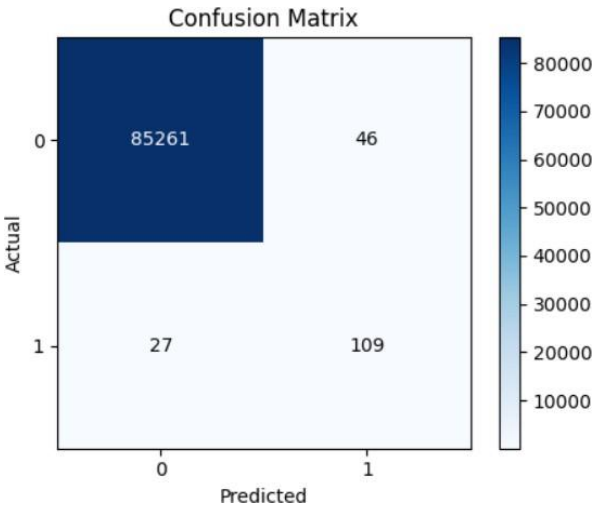

Fig.8. Confusion matrix for Decision Tree

This confusion matrix[Figure.8] shows us that for the train data the true positives are 85261 and false positives are 0, the true negatives are 0 and the false negatives are 330. For the test data, the true positives are 93818 and false positives are 27, the true negatives are 46 and the false negatives are 109.

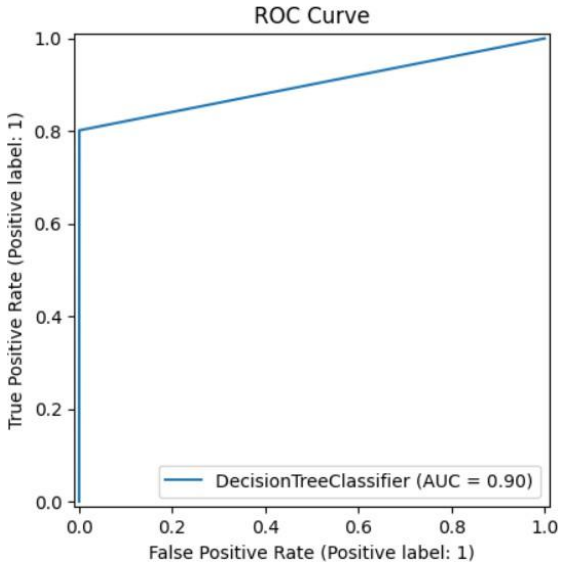

Fig.9. ROC curve for Decision Tree

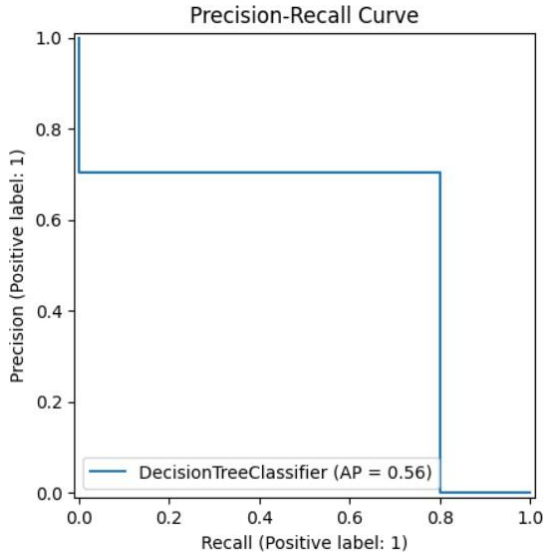

Fig.11. PRC curve for Decision Tree

V. CONCLUSION

The analysis of three machine learning techniques—Logistic Regression, Decision Tree, and Random Forest—as applied to the detection of credit card fraud led us to discover that Random Forest was the optimal model, obtaining better performance results with 99.96% testing accuracy, 94.12% precision, 82.35% recall, and an F1-score of 87.84%. Whereas Logistic Regression showed acceptable accuracy (99.91%) and recall (80.15%), its linear structure restricted fraud detection potential, and Decision Tree, with high accuracy (99.92%), experienced overfitting and greater false positives. Random Forest's ensemble method proved to be the most dependable, albeit with some limited false negatives remaining, pointing towards potential for enhancement. Future research will look at hybrid ensemble methods, sophisticated sampling algorithms, and real-time optimization to further improve detection rates at the expense of false alarms. These conclusions make Random Forest the best option for credit card fraud detection systems today, with further research required to respond to changing patterns of fraud and achieve near-perfect performance.

REFERENCES

[1] R. R. Subramanian, R. Ramar, " Design of Offline and Online Writer Inference Technique", International Journal of
Innovative Technology and Exploring Engineering, vol. 9, no. 2S2, Dec. 2019, ISSN: 2278-3075

[2] Subramanian R.R., Seshadri K. (2019) Design and Evaluation of a Hybrid Hierarchical Feature Tree Based Authorship Inference Technique. In: Kolhe M., Trivedi M., Tiwari S., Singh V. (eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 39. Springer, Singapore

[3] Joshva Devadas T., Raja Subramanian R. (2020) Paradigms for Intelligent IOT Architecture. In: Peng SL., Pal S., Huang
L. (eds) Principles of Internet of Things (IoT) Ecosystem: Insight Paradigm. Intelligent Systems Reference Library, vol 174. Springer, Cham

[4] R. R. Subramanian, B. R. Babu, K. Mamta and K. Manogna, "Design and Evaluation of a Hybrid Feature Descriptor based Handwritten Character Inference Technique," 2019 IEEE
International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.

[5] Andrew. Y. Ng, Michael. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes", Advances in neural information processing systems, vol. 2, pp. 841-848,2002

[6] John Richard D. Kho, Larry A. Vea " Credit Card Fraud Detection Based on Transaction Behaviour" published by
Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[7] Yashvi Jain, Namrata Tiwari, ShripriyaDubey, Sarika Jain, " A Comparative Analysis of Various Credit Card Fraud Detection Techniques, Blue Eyes Intelligence Engineering and Sciences Publications 2019"

[8] Learning Robert A. Sowah, Moses A. Agebure, Godfrey A. Mills, Koudjo M. Kaumudi, " New Cluster Undersampling Technique for Class Imbalance "of 2016 IJMLC

[9] Baraneetharan, E. "Role of Machine Learning Algorithms Intrusion Detection in WSNs: A Survey." Journal of Information Technology 2, no. 03 (2020): 161-173

[10] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." Journal of Ubiquitous Computing and Communication
Technologies (UCCT) 2, no. 03 (2020): 145-152.