Credit Card Fraud Detection using Hybrid Machine Learning Algorithm

Dr Ramani Jaydeep R [1*] ,
Assistant Professor,
Atmiya University,
Rajkot, 360005,
E-Mail jaydeep.ramani@atmiyauni.ac.in

Dr Jayesh N Zalavadia [2]
Associate Professor,
Atmiya University,
Gujarat Rajkot, 360005, Gujarat
E-Mail jayesh.zalavadia@atmiyauni.ac.in

Abstract. Credit card security is one of the needful criteria nowadays, which can cause billions of economic frauds worldwide. Some security firewalls provided by the banks are not up to the criteria in the daily life of the common user. The present work focuses on the fraud detection areas in data links and the scope of finding the accuracy to eliminate fraud while using. Data security with machine learning objectives needs more promising algorithms to improve accuracy; the hybridization of algorithms is a new era where two methods can evolve to find solutions for e-commerce applications. This paper combines the random forest and honeybee algorithms from machine learning to detect fraud. Combining the Random Forest Algorithm with the Honey Bee Algorithm, we developed the best model. Different types of hybridization and algorithms in the credit card space should be the subject of future research

Keywords:*Credit card, Fraud detection, Data security, Hybrid, Machine Learning, Random Forest, Honey Bee Algorithm, SMOTE*

Introduction

To perform fraud is to conduct any fraudulent or criminal conduct to gain some financial or personal gain. A system, typically, uses two processes - prevention of fraud and detection of fraud, to safeguard against financial loss that could be brought on. The best line is that of defense to stop frauds before they occur [1]. Credit card fraud occurs when criminals use stolen card information to make authorized transactions. The fraudster learns the user's password or other sensitive information during a credit card purchase. He then uses that information to fraudulently charge a large sum of money to the victim's card without the victim ever realizing what occurred [2].However, the modern technological world relies on credit cards, contributing to an increase in credit card transactions daily and the growth of the e-commerce sector. Each year, the volume of credit card transactions increases. While more people profit from technological advancements, more credit card theft occurs. In terms of global impact, it is undoubtedly a significant challenge today [3].Since the perpetrators of credit card fraud are often able to conceal information about themselves, such as their identity and whereabouts, online, the issue has far-reaching significance for the financial sector.

The fraud detection process is depicted below in Fig 1. The terminal point validates certain conditions, such as sufficient balance and a valid PIN (Personal Identification Number), and filters transactions based on those conditions [4]. The prediction model uses predefined criteria to determine if a given transaction is legitimate. Following up on each suspicious alert, investigators submit feedback to the prediction model to improve the algorithm's accuracy. This is only about the prediction model.

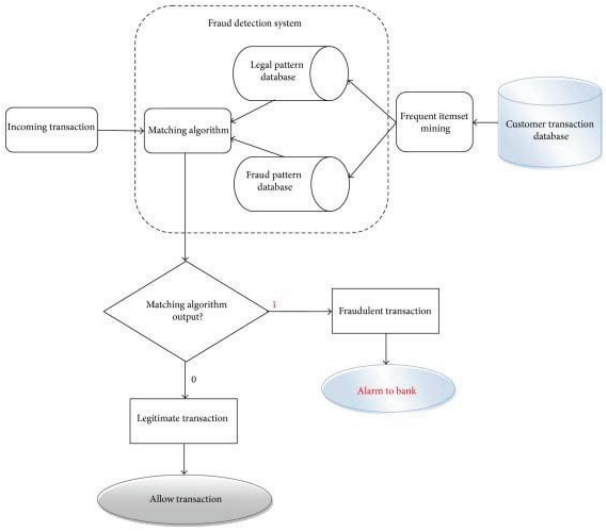


Fig .1. Credit Card Fraud Detection Process [5]

Fraud detection systems are more complicated than they appear to be. In practice, the practitioner must determine which classification technique to apply (decision trees or logistic regression…) as well as how one should cope with the problem of class imbalance (Suspicious cases are exceedingly less in contrast to valid ones). Detection of fraud is not simply a problem because of the disparity between the rich and the poor. Due to a lack of transaction data, many machine learning algorithms fail in the classification job because of the overlap between the real and fraudulent classes. An actual fraud detection scenario involves a model that uses artificial intelligence to identify suspicious transactions and send an alert to the appropriate authorities when one of those transactions is determined to be either authentic or fraudulent. The fraud detection system is improved by investigators who investigate and give their findings back to the system. Furthermore, investigators can only certify a small fraction of transactions using this method in a timely manner. Predictive models typically perform worse when less data points are used to refine the model. Because financial companies infrequently release client data owing to privacy issues, it is challenging to uncover the genuine financial dataset. An important problem in fraud detection systems is overcoming this obstacle [5].

To recognize illegal financial transactions specific machine learning algorithms are used. This work proposes a Hybrid Model based on the feature selection approach, Honey-Bee, and Random Forest Ensemble Classifier (especially HBRF) for fraud detection, which focuses on the problem of imbalanced datasets in the banking sector.

Problem of Statement

Credit card fraud is a crucial obstacle to expanding the financial services industry. Because of these scams, a large number of companies lost money. However, privacy concerns mean that only some of these studies analyse data collected from real-world transactions to identify patterns of fraudulent behaviour. Specific Machine Learning (ML) Algorithms are utilized in this setup to identify potentially fraudulent financial dealings. Because of the imbalanced dataset problem in the financial sector, this paper proposes a Hybrid Model for fraud detection based on the feature selection approach, Honey-bee Algorithm, and Random Forest Ensemble Classifier (specifically HBRF).

3 Related Work:

Methods such as RF, ANN, SVM, k-nearest neighbours, and others with a Hybrid and privacy-preserving approach to data privacy have been recognized as useful for detecting credit card fraud.

AltyebAltaher et al. [7], a novel cost-conscious decision tree method for detecting fraud. To evaluate its efficacy, it partitions attributes at non-terminal nodes by minimizing the cost of misclassification and comparing the resulting model to a standard one using a dataset of actual credit card transactions. The cost of misclassification is demonstrated in several scenarios. A cost-sensitive algorithm was tested, and the findings demonstrate significant increases in performance relative to established approaches in terms of accuracy and positive rate metrics, while also defining a cost-sensitive metric specific to credit fraud detection. By taking this strategy, economic losses due to fraud can be prevented.
Demerit: Inaccuracy in Fraud Detection

Kurshan, E.; Shen, H et al. [8] Submitted a manuscript focusing on the challenges of fraud detection in credit card transactions and provides a survey of solutions based on natural and ML. The advantages and disadvantages of various ML approaches are analysed. Misuse (Supervised) and anomaly detection are two of the mentioned methods (Unsupervised). The ability to deal with numerical and categorical datasets provides a second classification.
Demerit: Accuracy is too Low

Khaled Gubran Al-Hashedi et al. [9] The Hidden Markov Model (HMM) proposed to model the process flow of credit card transactions and its subsequent use to detect fraudulent activity. Users' typical patterns of conduct are tracked. If a trained HMM rejects a credit card transaction as likely fraudulent, the transaction is cancelled. However, it is crucial to protect normal credit card transactions so that they are not accidentally rejected.
Demerit: Accuracy is only around 80% for a wide range of input data.

Hossain, M.A et al. [10] Their study focuses on employing AI to detect fraud in real-time Self-Organization Map is used to interpret, filter, and analyse customer behaviours patterns to identify potential fraud. This premium approach is utilized to identify fraud red flags in personal finances.
Demerit: More effective models allow for greater scope for advancement in accuracy.

Varun Kumar K S et al. [11] Credit card fraud has been detected using categorization and clustering methods. Indicating that the likelihood of the fraudulent transaction is low but not zero. Therefore, their study aims to evaluate categorization methods and classifiers. With its attention on preventing fraud, this system will not falsely flag legitimate purchases as malware.
Demerit: The model's accuracy is poor compared to other approaches.

Ramyashree. K et al. [12] The performance of seven Hybrid Machine-Learning Models was evaluated on a real-world dataset to detect fraudulent actions. The generated Hybrid Models had two stages: first, State-of-the-Art, Machine Learning techniques were employed to detect credit card fraud; then, the hybrid approach was implemented.

Harsh Harwani et al. 13] Design and maintain complex machine learning models for prediction and understanding of the data set, with a focus on forecasting fraud and fraud-free transactions with regard to time and quantity using classification machine learning algorithms, statistics, calculus, and linear algebra.

Emmanuel I leberi et al. [14] In their study to the ML method, the best Data Mining Algorithm available at the time, was developed specifically for the task of identifying instances of credit card fraud; hence, it was one of the first models used in this context. The bank's data set is based on the real world; thus, it is being taken and analysed. Support Vector Machine (SVM), K- nearest Neighbour (KNN), Fuzzy Logic, and Decision Trees are just few of the methods that have been used

for fraud detection over the years. All these methods have shown some success, but a hybrid learning approach is required to further enhance accuracy when uncovering fraud.

Zahra Faraji et al. [15] Credit card fraud is just one type of financial crime that has risen in frequency due to the rise of online shopping and payment systems. Because of this, it is essential to set up systems that can identify instances of credit card theft. This paper provides a Genetic Algorithm (GA)-based Machine Learning (ML) based Credit Card Fraud Detection (CCFD) Engine.

4. Material and Methods

This approach is based on a number of basic ML techniques to identify possible Fraudulent Credit Card transactions. To create the Hybrid Algorithm, combine the best features of other algorithms, such as the RF and the HB. As a result, the system's efficacy is owed to the Hybridization Algorithm. To use ML algorithms for timing and monetary transaction detecting fraud on credit cards.

4.1 Decision Tree Algorithm

The Decision Tree method (depicted in Fig .2. below) can be applied to both Classification and Regression issues as it shown in Fig .2 below. The process is the same, however the corresponding equations may differ. Decision trees for a classification issue are constructed using Entropy and Information gain. Entropy and information gain both measures the measure to which data is random and how much we can learn from it. The Gini index and Gini coefficient are used to create a Regression Decision tree.

The root node in a classification problem is chosen based on the Information gain principle, where the node with the most information is preferred over those with the most entropy. The feature with the smallest Gini value is used as the root node in regression situations. By optimizing hyper parameters, we can calculate the tree's depth with the grid search cv algorithm.

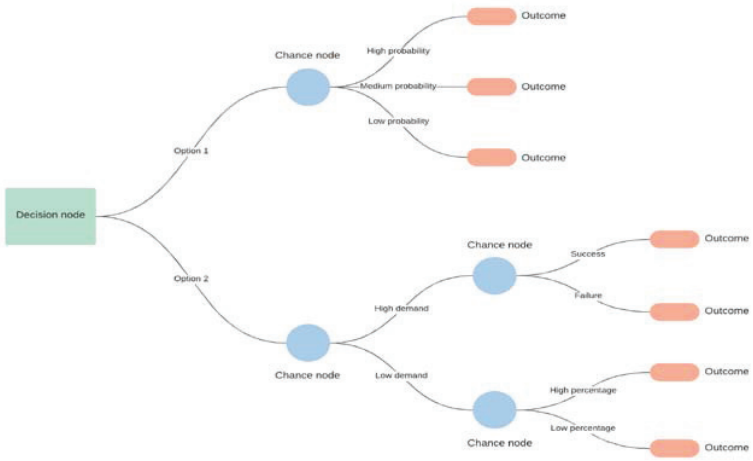


Fig .2. Decision Tree Schematic Layout **[21]**

4.2**Random Forest Algorithm**

Random forests sample rows at random, choose features at random (which are the independent variables), and the number of DT can be optimized for utilizing system parameters. When presented as a classification issue, a random forest's output is the maximum of the DT models' responses. Many models have successfully implemented this well-known ML algorithm. This algorithm accomplishes the goals outlined in most Kaggle computing challenges.

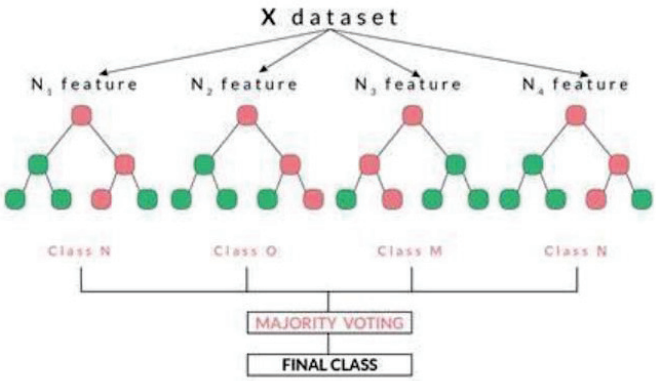


Fig .3. Random Forest Algorithm Data Set Layout **[22]**

4.3**Honeybee Algorithm**

Honey bees are known for their hunting behaviour, which inspired the Bees Algorithm. The bee's algorithm begins with an initiation phase and continues with a main search cycle that repeats for a fixed number of times, T, or until a solution with sufficient fitness is identified. This algorithm is used for optimization and by using this algorithm, the results were optimal in nature**.**

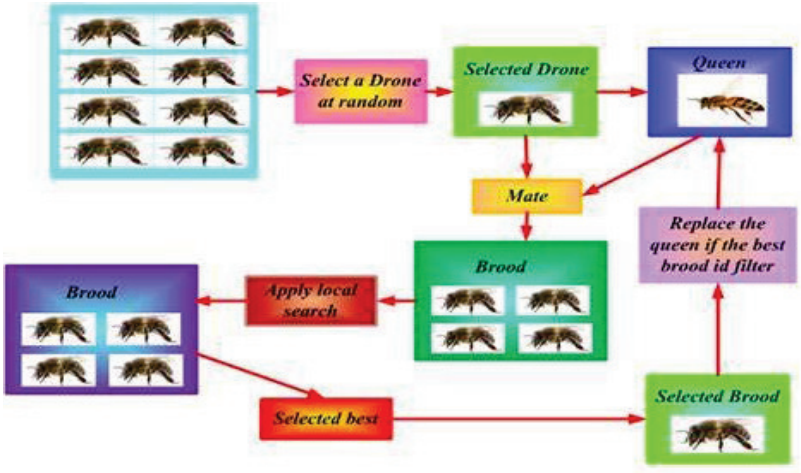


Fig .4. Honey Bee Algorithm Systemic Layout**[23]**

4.4 SMOTE (Synthetic Minority Oversampling Technique)

An effective statistical method for resolving imbalanced data is the SMOTE. In order to achieve a more equitable distribution of data, it is necessary to artificially generate new minority cases while randomly increasing the number of existing minority instances. Also, it assists in minimizing the overfitting issue that comes with using a large sample size.

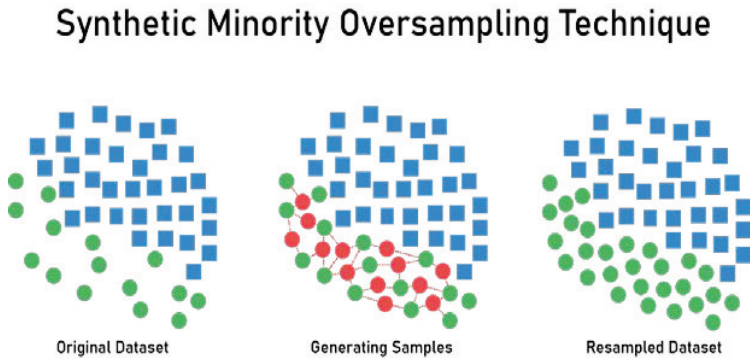


Fig .5.SMOTE Balancing Data **[24]**

4.5**Confusion Matrix and Related Metrics**

The model's accuracy may be very high yet misleading if it incorrectly categorized all the incidents as normal transactions based on extremely skewed fraud data. Due to this, accuracy is not a reliable indicator of performance when dealing with fraud data. This analysis measures accuracy in terms of precision and recall, which are derived from the confusion matrix. Confusion matrices reveal how various inputs were distributed across classes**.**

Table 1. Confusion Matrix

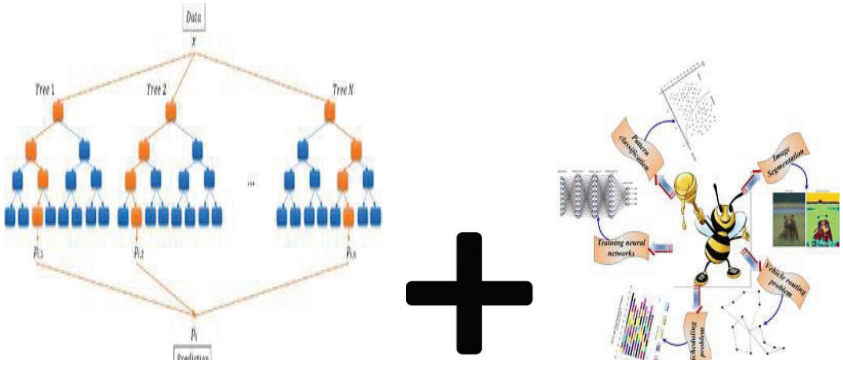
Actual/ Predicated	**Fraud**	**Not Fraud**
Fraud	True Positive	False Negative
Not Fraud	False Positive	True Negative

Based on the Confusion Matrix, the performance metrics can be determined:
True Positive (TP) = Estimated Number of Unauthorized charges
True Negative (TN) = number of commercial transactions considered to be authentic
False Positive (FP) = Prediction of fraud in legal transactions
False Negatives (FN) = Predicts fraud transactions will be considered legal.
To calculate the following, we use these parameters.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

5. Proposed Model

The combination of RF Algorithm and HB Algorithm with SMOTE is the Hybrid Approach. It helps in predicting output with high accuracy and is optimal in nature. This Algorithm Detects Fraud and decreases Fraud Transactions and Controls Misuse of Data.



Algorithm:

Input: No. of Transactions, Fraud transactions based on threshold value rate for input data, combined transactional data.

Output: Prediction of optimized accuracy

Explore the initial text of user transactions data

Arrange in N-dimensional array with different combinations

// Classification of Data presentation based on Random-forest

Based on the pre-processing rate of each transaction, performed on n-dimensional re-production with different notations.

For *i*= 1 to *n*-dimensions

Select randomly appeared transactions,

Calculate the threshold rate for each transaction and destroy the non-matched transactions

Commit the relations of each transaction

End for

//optimization process

Based on threshold matched values, select optimal rates for fraud

For each *i*=1 to *n*-dimensions

Select an optimized solution for each transaction.

Generate and store solutions for each transaction.

Save optimized transactions

End for

Update accuracy, classification parameters

Return best-optimized solution.

6. Results and Discussions

The data used in this study is available for free on Kaggle. There were 284,807 transactions over two days, and only a small percentage were flagged as potentially fraudulent. In total, 28 characteristics were transformed from the dataset. In addition to Time and Amount, the dataset comprises 30 more characteristics (V1, ., V28).There are no non-numerical attributes in this dataset. The final column indicates the category with 1 indicating a fraudulent transaction and 0 indicating a legitimate one. Features V1-V28 cannot be revealed due to the risk of compromising data privacy or security. To address the issue of class imbalance, the SMOTE technique has been implemented in the first stage of the proposed architecture depicted in Fig 1, Data Pre-processing. SMOTE is a technique for generating new members of a marginalized group by picking a random point along a line connecting nearby samples in the classifier.

In this case, the proposed model significantly improved above Random Forest. In contrast, it is easy to see that our dataset has a severe "class imbalance" issue. Over 99% of all transactions are legitimate (i.e., not fraudulent), while only 0.17 % are fraudulent. Suppose we train our model with such a distribution without addressing the imbalance issues. In that case, it will assign more weight to

legitimate transactions (because there are more data about them) and produce more accurate predictions. Class differences can be addressed using a variety of methods.

One such label is "oversampling." Random selection of the minority group is one method for redressing skewed data sets. The fastest process requires just generating minority class examples without changing the model. Instead, we can create new examples by synthesizing them from existing models. The "Synthetic Minority Oversampling Technique" (or "SMOTE" for short) is a method of data augmentation that focuses on underserved populations—implementing an oversampling technique to RF and DT.

The methods we took to arrive at our prediction included

1. reading the problem statement,
2. analyzing the data through statistical analysis and visualization, and
3. Analyzing the distribution of the data.

Due to its uneven distribution, this data set underwent balancing via oversampling, standardized scaling via standardization and normalization, and finally was put through a battery of ML algorithm evaluations. For data science projects, NumPy, numeric python, and pandas are essential, as are matplotlib and seaborn, which improve on matplotlib.

.

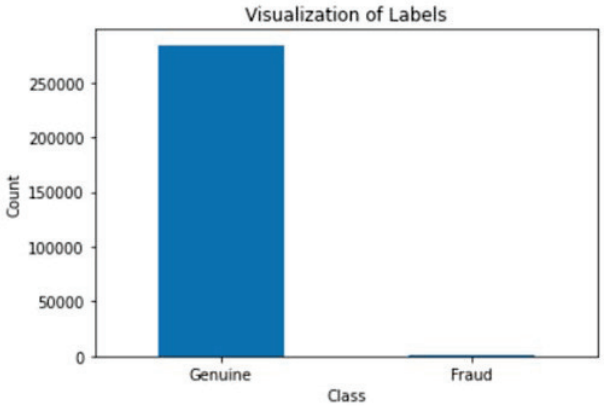


Fig .7. Visualize the "Labels" column in our Dataset

Fig .7 below shows Genuine transactions and Fraud transactions of our dataset. Here, we were comparing Genuine transactions and Fraud transactions, and from this, we have 284,315 Genuine transactions and 492 Fraud transactions in our dataset. In a nutshell, 0.1727 percent of fraud transactions are in our dataset.

Fig .8. below represents the confusion matrix of the Decision Tree, Random Forest and the Hybrid approach. False Negative value is lesser for decision trees which means Fraud Transactions are predicted as Legal. The confusion matrix for Random Forest shows a False Positive value lesser which means Legal Transactions are predicted as Fraud. The confusion matrix for the hybrid approach after oversampling shows a False Negative value lesser, which means Fraud Transactions are predicted as Legal.

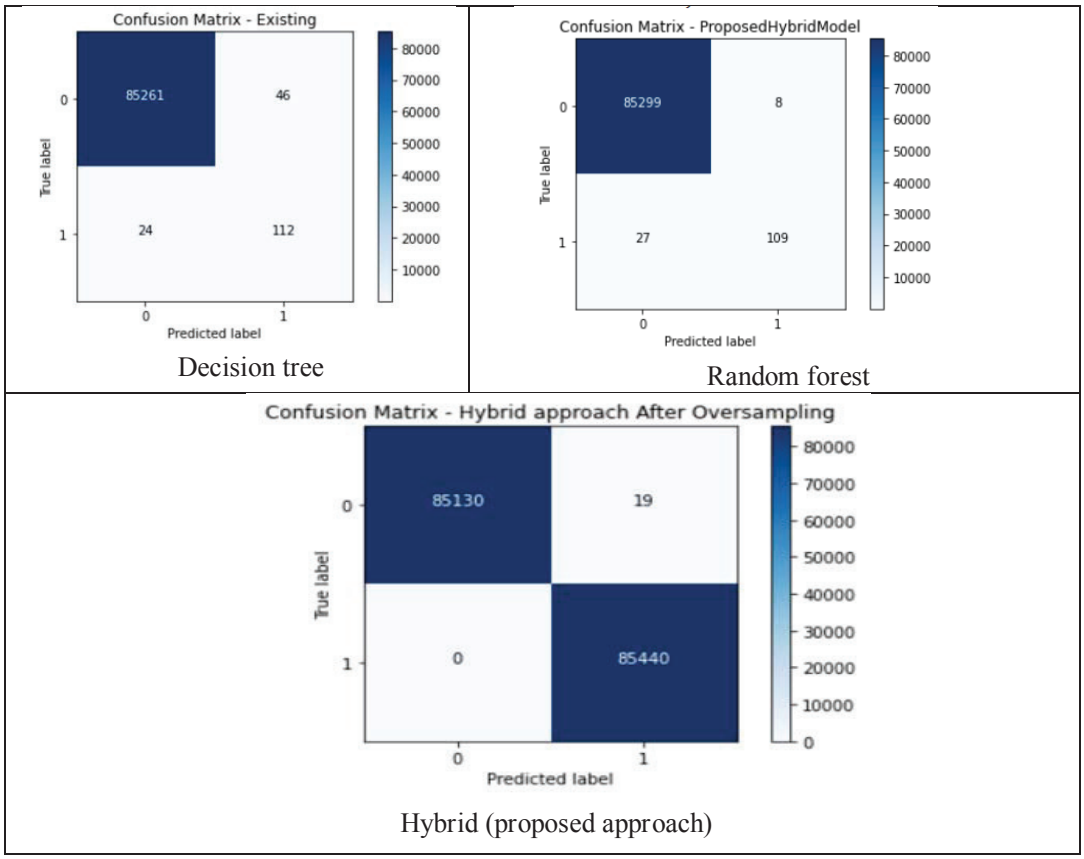


Fig .8. Confusion Matrix of Methods

The table below compares the effectiveness of the three approaches using the metrics Accuracy, Precision, Recall and F1-score. The proposed Hybrid Model outperforms the other two approaches.

Table 2. Mode Performance for Unbalanced Data

Methods/Metric	Accuracy	Precision	Recall	F1-score
Decision Tree	0.99925	0.74324	0.80882	0.77465
Random Forest	0.99961	0.94017	0.80882	0.86957
Hybrid Model	0.99989	0.99978	1.00000	0.99989

7. Conclusions and Future Work

The proposed solution uses an approach from machine learning algorithms to prevent credit card fraud. However, none of the current identity verification technologies can reliably identify all frauds in progress. However, they typically detect it after the occurrence. It is because only a small proportion of all transactions are actually fraud. The Random Forest Algorithm improves with additional training data, but the need slows its speed for more time to experiment and put the data into practice. In addition, a hybrid procedure might be put into action with this data. For a hybrid system to be effective, it must combine high-priced training techniques that yield extremely precise results with an improvement strategy that can decrease the overall cost of the system and speed up the machine's learning process. How and where a fraud sensing device is implemented affects which hybrid approaches are used.

For future work, this paper proposes a framework for implementing models of online training. The other training models can be evaluated as well. Cases of potential fraud can be carried along more quickly with online training models. This method of detection can prevent credit card fraud before it begins. Losses are thereby reduced in the financial sector.

References

Disha A. Date, Rugvedi Y. More, Rutuja P. Harne, Mr. Dilip M. Dalgade4 Credit Card Fraud Detection Using Machine Learning International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; Volume 10 Issue XI Nov 2022.

M. Mary, M. Priyadharsini, K. K and M. S. F, "Online Transaction Fraud Detection System," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 14-16, doi: 10.1109/ICACITE51222.2021.9404750

D. Shaohui, G. Qiu, H. Mai and H. Yu, "Customer Transaction Fraud Detection Using Random Forest," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 144-147, doi: 10.1109/ICCECE51280.2021.9342259.

O. Vynokurova, D. Peleshko, O. Bondarenko, V. Ilyasov, V. Serzhantova and M. Peleshko, "Hybrid Machine Learning System for Solving Fraud Detection Tasks," 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), 2020, pp. 1-5, doi: 10.1109/DSMP47368.2020.9204244.

Neha Sethi Anju Gera A Revived Survey of Various Credit Card Fraud Detection Techniques International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 780-791

M. R. Dileep, A. V. Navaneeth and M. Abhishek, "A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1025-1028, doi: 10.1109/ICICV50876.2021.9388431

AltyebAltaher Taha and Sharaf Jameel Malebary "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine", IEEE Access · February 2022.

Kurshan, E.; Shen, H. Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook. Int. J. Semant. Comput. 2020, 14, 565–58

Khaled Gubran Al-Hashedi, PritheegaMagalingam, Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019, Computer Science Review, Volume 40, 2021, 100402, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2021.100402

Hossain, M.A.; Islam, S.M.S.; Quinn, J.M.W.; Huq, F.; Moni, M.A. Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. J. Biomed. Inform. 2019, 100, 103313.

Varun Kumar K S, Vijaya Kumar V G, Vijay Shankar A, Pratibha K "Credit Card Fraud Detection using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 9 Issue 07, July-2020.

Ramyashree. K, Janaki K, Keerthana. S, B.V. Harshitha, Y.V "A Hybrid Method for Credit Card Fraud Detection Using Machine Learning Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S4, April 2019.

Harsh Harwani1, Jenil Jain1, Chinmay Jadhav1, Manasi Hodavdekar "Credit Card Fraud Detection Technique using Hybrid Approach: An Amalgamation of Self Organizing Maps and Neural Networks", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 07 | July 2020.

Emmanuel I leberi, Yanxia Sun, and Zenghui Wang A machine learning based credit card fraud detection using the GA algorithm for feature selection Ileberi et al. Journal of Big Data (2022) 9:24 https://doi.org/10.1186/s40537-022-00573.

Zahra Faraji "A Review of Machine Learning Applications for Credit Card Fraud Detection with A Case study", SEISENSE Journal of Management Vol 5 No 1 (2022): DOI: https://doi.org/10.33215/sjom.v5i1. 770, 49-59.

Rejwan Bin Sulaiman1 · Vitaly Schetinin 1 · Paul Sant "Review of Machine Learning Approach on Credit Card Fraud Detection", Human-Centric Intelligent Systems (2022) 2:55–68 https://doi.org/10.1007/s44230-022-00004-0.

Sanath Kumar Bhat K1, Sreenath N, Divya B S Credit Card Fraud Detection using Machine Learning Methods International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VI June 2020.

Zhang, Y.; Wang, Z. Customer Transaction Fraud Detection Using Xgboost Model. In Proceedings of the 2020 International Conference on Computer Engineering and Application, Guangzhou, China, 18–20 March 2020; pp. 554–558.