International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

# BIG DATA-DRIVEN FINANCIAL FRAUD DETECTION AND CLASSIFICATION USING EFFICIENT MACHINE LEARNING TECHNIQUES IN PYTHON SIMULATION

*Akshay Rajshekar Shiraguppi*
*Irvine, California*
*akshayrs1993@gmail.com*

*Abstract*

***Detecting financial fraud as the quantity of digital payments continues to rise, credit card transactions are getting harder. Fraudulent behavior not only causes financial losses but also damages the public's trust in financial institutions. The research conducts its analysis on actual credit card fraud data consisting of 284,807 transactions along with severe class distribution imbalance to explore machine learning solutions for better fraud detection. In the data preparation phase, SMOTE is used as a class balancing tool, while PCA and feature selection are used as dimension reduction methods. Although the authors conducted tests for Naïve Bayes (NB) and Logistic Regression (LR), they concluded that Decision Tree (DT) produced better analytical findings. The models need to be assessed by computing the F1-score, recall, accuracy, and precision. Decision Tree is the most effective model for fraud detection, according to data trials, with 97% accuracy, precision, recall, and F1-score. The research presents evidence about how machine learning fraud detection systems minimize incorrect fraud alarms to let organizations prevent fraudulent activity and promote financial safety.***

***Keywords: Financial Fraud Detection, Machine Learning, Big Data, Credit Card Fraud, Fraud Classification.***

## I. INTRODUCTION

In recent years, the attention toward financial fraud has increased substantially because of the combination of corporate money laundering, credit card fraud, and fraud. Fraud as[1] "intended to produce financial or personal gain through unlawful or criminal deception." characterized fraud as causing a business organization's structure to be abused without always having direct legal repercussions. Defining A crime against the law takes the form of deliberate actions that intentionally harm financial rules with a purpose of obtaining unauthorized monetary benefits.

Nowadays, fraud detection is a crucial corporate activity that minimizes the negative effects of illegal transactions on a company's customer service by utilizing state-of-the-art fraud technology frameworks, financial performance, and reputation[2]. Applications of knowledge discovery techniques include both Neural network training to stop past fraud incidences and the identification of fraudulent transactions in static data. Therefore, transactional instances that

correspond with known patterns of fraudulent behavior may be reported to fraud staff for additional human examination and the start of any necessary preventive measures.

In 2014, Globally, Credit cards became the most popular payment method, surpassing even e-wallets and bank transfers. In order to illegally exploit credit card services, cybercriminals usually target major transactional services[3]. Credit card fraud includes transactions on an inactive card, unusual transaction activity, and unauthorized card use. Three categories of credit card theft are commonly recognized: conventional frauds (such as stolen, fake, and counterfeit), internet scams (such as fraudulent or phoney merchant websites), and frauds connected to merchants.

The advancement of methods for machine learning. One effective technique for detecting fraud has been shown to be machine learning. A lot of data is transmitted during online transaction processes, and the results might be either legitimate or fraudulent. The example fake datasets are used to develop features. These are data points, such as the age and value of the client account and the credit card's place of origin. The probability of fraud is increased by each of the hundreds of characteristics in varying proportions[4]. Note that the machine's artificial intelligence, which is fueled by the training set, determines the degree to which each characteristic adds to the fraud score; a fraud analyst does not make this determination. Big Data is a low-cost supercomputer made up of generic servers. This system uses distributed parallel computing to process and store massive amounts of data[5]. Data that is too big or too costly to process with conventional computer equipment is referred to as large-scale data. The financial industry and other businesses are greatly impacted by the awful issue of financial fraud[6], government and business sectors, as well as private customers. These transaction scams are getting more complex as technology advances. In today's data-driven environment, fraudulent transactions can be found by using Big Data platforms and data mining techniques to analyze large transaction data sets.

**A. Motivation and Contribution of Study**

In the digital era, Financial transactions have resulted in a rise in fraudulent conduct, endangering both businesses and consumers. Rule-based systems and other traditional fraud detection methods usually fall behind in responding to evolving fraud patterns, which results in inefficiencies and high false positive rates. By finding hidden patterns and abnormalities in massive datasets, A more flexible and scalable approach to identifying fraudulent behavior is offered by machine learning and big data analytics. This project uses cutting-edge ML algorithms to identify financial fraud more accurately and efficiently, and A real-time system provides instant updates, which increases the security level while increasing the dependability of financial systems. The following are this study's primary contributions:

- The study improves high accuracy in identifying illegal transactions using advanced data preparation techniques and ML algorithms in financial fraud detection.
- To lower dimensionality, PCA is used, enhancing computational efficiency while preserving essential fraud detection patterns.

- A detailed comparative analysis highlights the Decision Tree as the proposed model, while the comparison of Naïve Bayes and Logistic Regression offers insights regarding the best algorithm for practical fraud detection situations.
- The results encourage the creation of stronger Financial institutions benefit from fraud detection systems to decrease losses and strengthen transaction security while receiving assistance from these systems.

### B. Structure of Paper

This paper explores Big Data-driven ML methods for detecting and categorizing financial fraud. It reviews existing research on fraud detection models Section II, details dataset preprocessing, feature selection, and model evaluation methods Section III, analyzes experimental results and model performance Section IV, and concludes with key findings while suggesting future improvements using advanced AI techniques Section V.

## II. LITERATURE REVIEW

This study examines ML-based financial fraud detection, emphasizing Big Data, ensemble approaches, and supervised learning for increased accuracy and fewer false positives. Some of the related works are:

Pillai et al. (2018) to use the study created a sophisticated model that uses DL approaches to identify credit card fraud. The model demonstrated that credit card fraud may be identified using both the logistic and hyperbolic tangent activation functions. According to the sensitivity assessment, the logistic activation function operates at its best with ten or one hundred hidden nodes, with 82% and 83% accuracy rates, respectively. The tests show that 1000 nodes achieve the best results with the hyperbolic tangent activation function because it demonstrates an 82% sensitivity across all hidden layer numbers[7].

Zamini and Montazer (2018) have released an autoencoder-based clustering technique for unsupervised fraud detection. In order to evaluate a k-means clustering autoencoder with three hidden layers, 284807 European bank transactions were examined. The results showed that this method worked better than others, with a TPR of 81% and an accuracy of 98.9%. Banks are quite concerned about detecting fraud due to the rise of e-commerce and online payments[8].

Mubalaike and Adali (2018) explore the potential benefits of using DL models to accurately identify transactions that are fraudulent. Once the data has been preprocessed, the best ML approaches, such as EDT and DL techniques, including SAE and RBM classifiers, are used. Accuracy, sensitivity, specificity, precision, ROC scores, and the assessment of performance rely on the confusion matrix data of the created classifier models. The corresponding ideal accuracy results are 90.49%, 80.52%, and 91.53%[9].

Rizki, Surjandari and Wayasti (2017) Data mining methods, including SVM and ANN, were applied. The study results tell auditors that efficiency and profitability are key indicators of

financial wrongdoing. The accuracy of the SVM approach has increased to 88.37% with the use of feature selection. Without feature selection, ANN achieves an accuracy of 90.97%. The three categories of professional fraud identified by the field of financial statement fraud, asset theft, and corruption are the primary emphasis of the Association of Certified Fraud Examiners[10].

Yaram (2017) includes relevant industrial use examples and emphasizes the application of a group of classification methods (DT, RF, and NB) in conjunction with the document clustering approach. When it comes to handling massive volumes of both structured and unstructured data, machine learning is crucial. In order to make educated business decisions, it is possible to utilize a set of algorithms to extract useful information from the data. A collection of algorithms may be applied to obtain valuable information from the data that aids in the process of making wise business decisions[11].

Chen (2016) A reliable and precise model development for false financial statement detection constitutes the main objective of this research. ANN and SVM, together with Bayesian belief networks CART and multiple additional technologies, form the models for detecting false financial statements during the second developmental phase. At 92.69%, Based on the results, the CHAID-CART model outperforms all others in terms of detection accuracy, coming in at an impressive 87.97%[12].

The comparison between fraud detection approaches in financial transactions appears in Table I which shows methods as well as datasets and performance measurements along with projected developments.

TABLE I. SUMMARY OF LITERATURE REVIEW BASED ON BIG DATA FINANCIAL FRAUD DETECTION

| Study | Methodology | Dataset | Techniques Used | Performance Metrics | Key Findings | Future Work |
|---|---|---|---|---|---|---|
| Pillai et al. (2018) | Using deep learning to identify credit card theft | Financial Fraud Dataset | Logistic & Hyperbolic Tangent Activation Functions | Sensitivity: 82% (10 nodes), 83% (100 nodes) for logistic; 82% (1000 nodes) for hyperbolic tangent | Logistic activation performs better with 10 and 100 nodes, while hyperbolic tangent works best with 1000 nodes in all hidden layers. | Extend to different activation functions and architectures like CNNs or LSTMs for better feature learning. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Zamini and Montazer (2018) | Unsupervised fraud detection via clustering based on autoencoders | Financial Fraud Dataset (284,807 transactions from European banks) | Three hidden layer autoencoder with k-means clustering | Accuracy: 98.9%, TPR: 81% | Autoencoder with k-means clustering outperforms other approaches. | For improved fraud pattern identification, investigate other clustering techniques like DBSCAN or hierarchical clustering. |
| Mubalaike and Adali (2018) | Fraud detection with deep learning | Financial Fraud Dataset | Stacking Auto-Encoders (SAE), Decision Tree Ensembles (EDT), and Restricted Boltzmann Machines (RBM) | Accuracy: 90.49% (EDT), 80.52% (SAE), 91.53% (RBM) | RBM achieves the highest accuracy among tested models. | Investigate hybrid models combining deep learning with explainable AI techniques. |
| Rizki, Surjandari and Wayasti (2017) | Financial fraud detection via data mining | Financial Dataset | Artificial neural networks (ANNs) and support vector machines (SVMs) | Accuracy of SVM (with feature selection): 88.37%; accuracy of ANN (without feature selection): 90.97% | Financial fraud is best detected by ANN, and SVM accuracy is increased by feature selection. | Implement real-time fraud detection models using online learning techniques. |
| Yaram (2017) | Implementation of document clustering and classification algorithms | Financial Fraud Data | Options include decision trees, Naïve Bayes, and Random Forest. | Accuracy, Precision, Recall, F1-score | Machine learning aids in processing structured and unstructured data for business decision-making | Extending classification techniques and exploring additional clustering methods |
| Chen (2016) | Fraudulent financial statement detection | Financial Dataset | CART, CHAID, Bayesian Belief Network, SVM, ANN | Accuracy: CHAID–CART (87.97%), FFS Detection Accuracy: 92.69% | The best model for identifying false financial statements is CHAID-CART. | Develop a reinforcement learning-based adaptive fraud detection system. |

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

## III. METHODOLOGY

The project's goal is to use big data-driven ML techniques to enhance financial fraud detection. The first line of defense against fraudulent charges utilizes 284,807 transactions consisting of genuine and fraudulent payment records. In addition to PCA-based dimensional reduction, feature selection and cleaning procedures improve model efficiency. After the training data makes up 80% of the set and the testing data makes up 20%, the suggested model DT is used. The NB and LR models are used for comparison. These models have been altered to improve the fraud categorization accuracy. The optimal fraud detection model for real-world financial applications is determined by performance assessment. A combination of accuracy, precision, and recall with F1-score should be used together with comparative analysis to evaluate results. The Figure 1 flowchart provides an illustration of the study design.

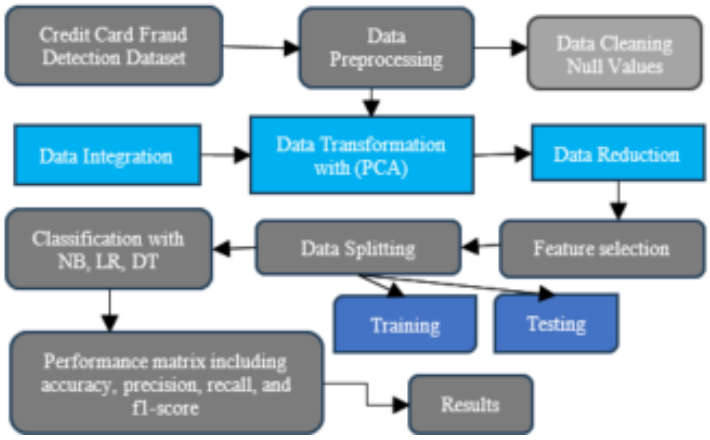


Fig. 1. Flowchart for Financial Fraud Detection in Credit Card Transactions

The overall steps of the flowchart for Financial Fraud in Credit Card Transactions are provided below:

### A. Data Collection

There are 492 illicit transactions among More than 284,807 transactions from two days in September 2013 comprised the highly imbalanced portion of database for the purpose of detecting credit card fraud. Testing uses 20% of the dataset, and training uses the remaining 80%. For security reasons, features V1 through V28 were extracted using Principal Component Analysis (PCA) out of its thirty input characteristics. Quantity and Duration, however, don't alter.

### B. Data Analyses and Visualization

Effective visualization and data analysis are essential for understanding the challenges of identifying fraudulent credit card activity. The dataset visualization is described as follows:

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

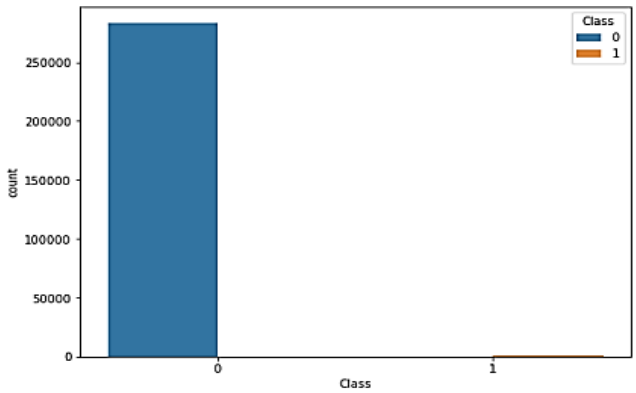


Fig. 2. Unbalanced Data

Figure 2 displays this dataset's class distribution for credit card fraud detection. A value of 0 on the x-axis indicates a legitimate transaction, whereas a value of 1 indicates a fraudulent one. The number of transactions is shown on the y-axis. There is a noticeable disparity in the statistics between the classes depicted in the picture; the majority of transactions are not fraudulent, while fraudulent transactions constitute a small fraction.

### C. Data Preprocessing

Simply transforming data preparation becomes the process that transforms raw data into accessible information. The real-world data sometimes contains various problems like noise and redundancy together with inconsistencies and incompleteness[13]. A number of procedures are involved in data preparation, which helps transform unprocessed data into a logical structure. In this step, the data were processed in the following ways.

- Data Cleaning: Real-world data is frequently noisy, unexpected, and lacking in detail. Data cleaning facilitates the completion of missing values. Eliminates noise and detects extraneous influences and precise unpredictability in the data.
- Null values: A null value indicates missing or unrecorded data in a dataset, affecting analysis and model performance. It is handled through imputation or removal for accuracy.

### D. Data Integration

The process of Combining data from several sources into one cohesive analytical structure is known as data integration. It ensures that credit card fraud detection models are more accurate and perform better by facilitating seamless feature extraction, preprocessing, and transaction record merging.

### E. Data Transformation with Principal Component Analysis (PCA)

Data transformation is the process of extracting, analyzing, and comprehending data before turning it into a form that can be examined. One method for identifying patterns is PCA, This may be applied to assess high-dimensional data that is hard to understand through simple data analysis. For the data analysis, convert the high-dimensional data to low-dimensional data, then

make a plot and examine the results. PCA is used to display the key data in a few straightforward graphs, such as loading and scoring plots.

**F. Data Reduction**

The procedure for decreasing the size needed to store the data is known as data reduction. Data reduction can save costs And Increase Storage Capacity.

**G. Feature Selection**

After preprocessing, one technique for preparing data is feature selection, which lowers the quantity of features in datasets. Techniques for feature selection look throughout the whole feature space to identify the best feature set, removing unnecessary and redundant features.

**H. Data Splitting**

The process of dataset splitting is seen to be essential and crucial for removing or minimizing bias in training data for ML models. The dataset's training and testing versions were kept separate in the ratio of 80:20 subsets.

**I. Classification with a Decision Tree Model**

A DT is an easy-to-understand tree-based method for creating a model that can classify or forecast dependent variable values using a set of accessible decision rules. The ID3, C4.5, C5.0, CART, and CHAID algorithms are examples of those used in DT analysis methodologies. To conduct this research, the CART algorithm was employed. It is appropriate to compare the outcomes using the CART algorithm because all of the ensemble methods in this study employ it to generate individual DTs with only minor modifications.

As the name implies, CART predicts employing a binary partitioning technique to construct classification and regression trees for continuous dependent variables (regression) and categorical dependent variables (classification)[14]. All independent variables are regularly and recursively divided into subsets using a suitable splitting criterion as part of the CART learning process. The fact that it optimizes both the heterogeneity between and the homogeneity within the subgroups is crucial.

The Gini index, one of the several splitting criteria used by CART, is computed as follows in Equation (1):

$$Gini(p_i,p_2,...,p_n)=\Sigma_{(i\neq j)}\left[p_i\,(1-p_i)\right]=1-\Sigma_j\left[pj\right]^2 \quad (1)$$

where p_i and p_j , respectively, represent the likelihood that landslides will occur in classes i and j. The Gini index has a range of 0 to 0.5.

**J. Performance Matrix**

In many different domains, performance metrics, often known as error measures, are essential parts of the assessment frameworks. The research evaluation utilized the terms accuracy and precision with recall and F-measure and their corresponding definitions. Each matrix originates

from confusion matrices. A summary table of a classification model's performance evaluation that shows how well it predicts results on actual test data is called a confusion matrix.

- True Positive (TP): Cases where fraud is correctly detected as fraud.
- False Positive (FP): Cases where Sometimes legitimate transactions are inadvertently reported as fraudulent.
- True Negative (TN): Cases when it is appropriate to classify transactions as non-fraudulent.
- False Negative (FN): Cases where fraud goes undetected.
- False Positive Rate (FPR): The system determines the fraudulent transactions by analyzing valid transactions that exist in the sample data as FP / (FP + TN).
- True Positive Rate (TPR): It establishes the proportion of actual fraud instances that are discovered; it is sometimes called recall or sensitivity, calculated as TP / (TP + FN).

**1. Accuracy**

The percentage of all samples that the model accurately predicts is known as the accuracy. According to Equation (2).

$$Accuracy=(TN + TP)/(TP + TN + FP + FN) \quad (2)$$

**2. Recall**

The ratio of samples that were expected to be positive to those that were is known as recall. Equation (3) is used to compute it:

$$Recall=TP/(TP+FN) \quad (3)$$

**3. Precision**

The ratio of Precision is the ratio of samples that are expected to be positive to those that are really positive. Equation (4) is used to express it:

$$Precision=TP/(TP+FP) \quad (4)$$

**4. F1 Score**

Throughout F1-Score calculation use the harmonic mean of accuracy and recall values to consider their relative effects on calculations. The following definition applies to the F1 score in Equation (5):

$$F1=(2*(precision*recall))/(precision+recall) \quad (5)$$

**5. ROC**

A two-dimensional graphic known as The TPR and FPR values are presented as ROC curve coordinates using the y-axis and x-axis.

## IV. RESULT ANALYSIS AND DISCUSSION

This study analyzes the findings from ML models applied to financial fraud detection, including NB[15], LR[16], and DT. Experiments were carried out on a 64-bit Windows 10 Pro machine with an Intel i7 CPU (3.60 GHz, four-core) and 16 GB of RAM using Python 3 with Scikit-Learn. The DT fared better than other models in terms of recall, F1-score, accuracy, and precision, according to tests conducted on the Credit Card Fraud Detection dataset. Both feature selection methods and PCA application demonstrate their ability to upgrade fraud detection accuracy while decreasing false alarm rates.

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

TABLE II. PERFORMANCE METRICS OF DECISION TREE MODEL FOR CREDIT CARD FRAUD DETECTION

| Performance Metric | Decision Tree |
|---|---|
| Accuracy | 97 |
| Precision | 97 |
| Recall | 97 |
| F1-score | 97 |

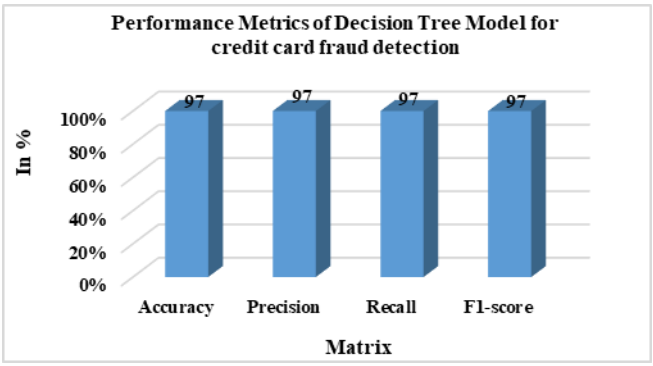


Fig. 3. Performance Metrics of Decision Tree Model for credit card fraud detection

The DT classifier exhibits extraordinarily high recall, accuracy, precision, and F1-score performance features in Figure 3 and Table II, with respective values of 97%, 97%, and 97%. This implies that the model identifies cases accurately while producing a few false negatives and false positives. However, such high scores might indicate overfitting, especially if the dataset is imbalanced. Evaluating with cross-validation and additional metrics like AUC-ROC can help confirm model robustness.

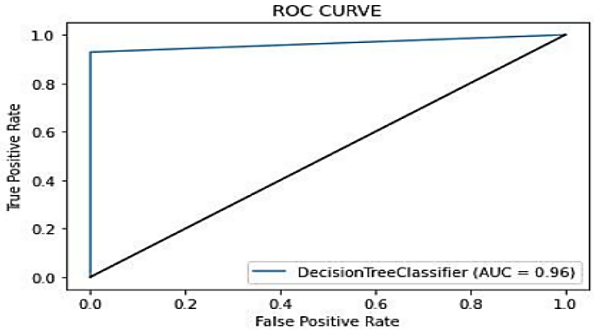


Fig. 4. ROC curve of Decision Tree model

In Figure 4, the DT classifier's ROC curve is displayed. The blue line represents the model's remarkable performance, with an AUC of 0.96. It demonstrates how effectively the model can distinguish between different classes. The black diagonal line represents a random classifier, highlighting the DT superior predictive power.

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

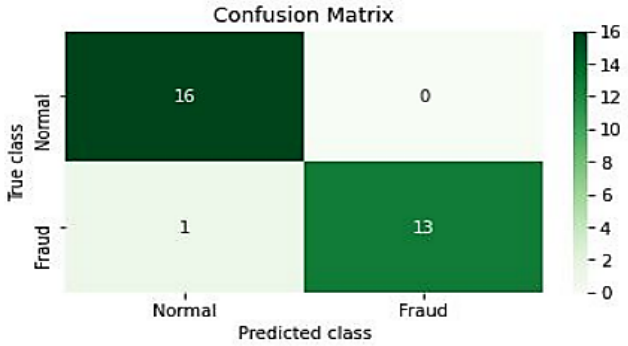


Fig. 5. Confusion Matrix of Decision Tree model

In Figure 5, the confusion matrix used to evaluate a fraud detection model's efficacy is displayed. It shows 16 true negatives correctly classified as "Normal", 13 true positives correctly classified as "Fraud", 1 false negative misclassified "Fraud" as "Normal", and 0 false positives. The color Value intensity is represented by a gradient, where greater numbers are shown by darker hues. Only one misclassification shows that the model is highly accurate.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 1.00 | 0.97 | 16 |
| 1 | 1.00 | 0.93 | 0.96 | 14 |
| accuracy | | | 0.97 | 30 |
| macro avg | 0.97 | 0.96 | 0.97 | 30 |
| weighted avg | 0.97 | 0.97 | 0.97 | 30 |

Fig. 6. Classification Report of Decision Tree Model

Figure 6 presents the classification report. Class 0 exhibits 94% precision, 100% recall, and a 97% F1-score with 16 samples. Class 1 shows 100% precision, 93% recall, and a 96% F1 score with 14 samples. Overall accuracy is 97%. The model appears to be well-balanced and successful for both classes, as seen by the 97% accuracy, recall, and F1-score weighted and macro averages.

### A. Comparative Analysis

A comparison of several ML approaches used to identify financial fraud happens in this portion of the analysis. Table III presents the assessment criteria that evaluate the efficiency of the DT with accuracy measures, precision metrics, and recall factors leading to F1-score calculations, with NB and LR serving as comparison models.

TABLE III. COMPARATIVE ANALYSIS OF CLASSIFICATION ML TECHNIQUES FOR CREDIT CARD FRAUD DETECTION.

| Performance Metric | NB [15] | Logistic Regression [16] | Decision Tree |
|---|---|---|---|
| Accuracy | 90.9 | 54.8 | 97 |
| Precision | 93 | 38.3 | 97 |
| Recall | 93 | 58.3 | 97 |
| F1-score | 93 | - | 97 |

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

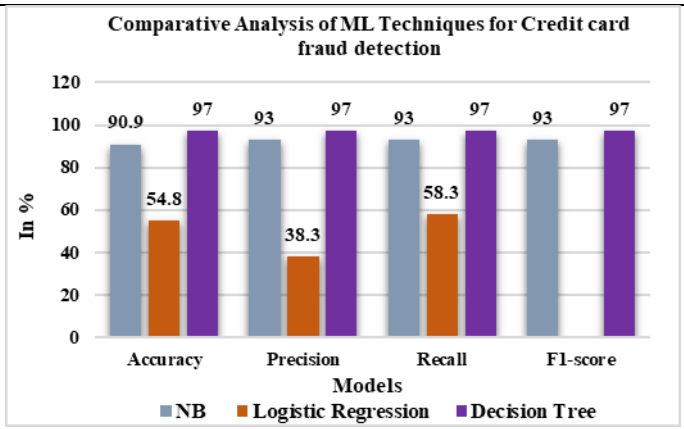


Fig. 7. Comparative Analysis of ML Techniques for Credit Card Fraud Detection.

Table III and Figure 7 shows that when comparing several ML models for financial fraud detection, DT performs better than NB and LR. With an impressive accuracy of 97%, DT significantly outperforms NB at 90.9% and LR at 54.8%; The model shows successful capability in detecting fraudulent payment requests. The DT obtains 97% precision as well as a 97% recall, which corresponds to the highest F1 score, thus developing a balanced fraud detection system. While NB performs well in certain cases, its slightly lower recall, 93% and precision, 93% indicate limitations in detecting complex fraud patterns. The 54.8% accuracy rate for LR poses problems when differentiating between valid and deceptive deals. This comparison highlights DT's superiority as the most dependable and effective model for identifying fraudulent transactions in financial systems.

## V. CONCLUSION AND FUTURE WORK

Financial transactions require the detection of financial fraud to ensure complete safety. Traditional methods encounter several limitations which include high rates of incorrect alarms and modifications in fraud patterns as well as data distribution that is unbalanced. The research demonstrates that ML-based fraud detection technology achieves both high detection accuracy and minimal error rate results. DT proved to be the most effective model from the evaluation with 97% accuracy in fraud classification because it yielded 97% precision and recall and F1-score.. LR and NB served as comparative models. ML algorithms improve detection efficiency by adapting to emerging fraud tendencies, in contrast to traditional rule-based detection systems. Model performance and generalization were further enhanced by preprocessing methods, including feature selection and PCA-based dimensionality reduction for managing unbalanced data. The results demonstrate the value of AI-driven fraud detection and Big Data analytics, offering financial institutions accurate, automated, and scalable fraud protection solutions. The suggested method is less successful in identifying new fraud trends because it depends on past data. Additionally, the Decision Tree model may overfit, reducing generalizability to unseen transactions. Future studies should focus on deep learning-based fraud detection, incorporating NB and Logistic Regression for improved anomaly detection. Real-time fraud detection using streaming data analytics can enhance instant fraud

identification, while XAI will improve interpretability and trust in fraud detection models. Additionally, blockchain technology can ensure secure and tamper-proof transactions and federated learning can protect data privacy while facilitating cooperative fraud detection across organizations.

**REFERENCES**

1. E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decis. Support Syst., 2011, doi: 10.1016/j.dss.2010.08.006.
2. M. E. Edge and P. R. Falcone Sampaio, "A survey of signature based methods for financial fraud detection," Comput. Secur., vol. 28, no. 6, pp. 381–394, 2009, doi: 10.1016/j.cose.2009.02.001.
3. Y. H. Rajarshi Tarafdar, "Finding majority for integer elements," J. Comput. Sci. Coll., vol. 33, no. 5, pp. 187–191, 2018.
4. J. West and M. Bhattacharya, "Intelligent financial fraud detection: A comprehensive review," Computers and Security. 2016. doi: 10.1016/j.cose.2015.09.005.
5. S. Pahune, "Sensor Data Collection and Performance Evaluation using A TK1 Board," University of Memphis, 2019. [Online]. Available: https://digitalcommons.memphis.edu/etd/2372
6. R. Tarafdar, "Algorithms on Majority Problem," Univ. Missouri-Kansas City, 2017.
7. T. R. Pillai, I. A. T. Hashem, S. N. Brohi, S. Kaur, and M. Marjani, "Credit Card Fraud Detection Using Deep Learning Technique," in Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018, 2018. doi: 10.1109/ICACCAF.2018.8776797.
8. M. Zamini and G. Montazer, "Credit Card Fraud Detection using autoencoder based clustering," in 9th International Symposium on Telecommunication: With Emphasis on Information and Communication Technology, IST 2018, 2018. doi: 10.1109/ISTEL.2018.8661129.
9. A. M. Mubalaike and E. Adali, "Deep Learning Approach for Intelligent Financial Fraud Detection System," in UBMK 2018 - 3rd International Conference on Computer Science and Engineering, 2018. doi: 10.1109/UBMK.2018.8566574.
10. A. A. Rizki, I. Surjandari, and R. A. Wayasti, "Data mining application to detect financial fraud in Indonesia's public companies," in 2017 3rd International Conference on Science in Information Technology (ICSITech), 2017, pp. 206–211. doi: 10.1109/ICSITech.2017.8257111.
11. S. Yaram, "Machine learning algorithms for document clustering and fraud detection," in Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016, 2017. doi: 10.1109/ICDSE.2016.7823950.
12. S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," Springerplus, vol. 5, no. 1, pp. 1–16, 2016, doi: 10.1186/s40064-016-1707-6.
13. V. Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis," Int. J. Comput. Appl., 2015, doi: 10.5120/ijca2015907309.
14. S. Park, S. Y. Hamm, and J. Kim, "Performance evaluation of the GIS-based data-mining

techniques decision tree, random forest, and rotation forest for landslide susceptibility modeling," Sustain., vol. 11, no. 20, 2019, doi: 10.3390/su11205659.
15. S. Rajora et al., "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018, no. March 2019, pp. 1958–1963, 2018, doi: 10.1109/SSCI.2018.8628930.
16. J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in 2017 International Conference on Computing Networking and Informatics (ICCNI), IEEE, Oct. 2017, pp. 1–9. doi: 10.1109/ICCNI.2017.8123782.