MACHINE INTELLIGENCE RESEARCH
ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

REVIEW ON MACHINE LEARNING APPROACHES ON CREDIT CARD FRAUD DETECTION

Dr.Revathi.R[1]*, Dr.Subha.R[2], Geetha.K[3], Rajendiran.P[4]

[1]*Assistant Professor Department of Computer Science-Information Technology, PSGR Krishnammal College for Women, Coimbatore

[2]Assistant Professor Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore

[3]Assistant Professor & Head Department of Computer Science, G.T.N. Arts College Dindigul

[4] Assistant Professor Department of Computer Science, G.T.N. Arts College Dindigul

[1]*Corresponding author: revathilakshay@gmail.com

[2]Corresponding author: subharj2013@gmail.com

[3]Corresponding author: geethchouthri@gmail.com

[4]Corresponding author: eieraja.09@gmail.com

Abstract

Credit card fraud is a serious threat to both individuals and financial institutions, advanced machine learning algorithms must be used to detect and stop fraudulent transactions. This thorough analysis explores the body of research on credit card fraud recognition, with a particular emphasis on big data and machine learning technology. The analysis showcases advancements in the field, underscoring the importance of hybrid models that amalgamate diverse algorithms to enhance detection accuracy. Additionally, the integration of big data technologies facilitates the management of vast datasets, enabling the extraction of valuable insights crucial for effective fraud detection. This review underscores the ongoing imperative for research and development, urging exploration of innovative approaches and algorithm combinations.

Keywords

Apache Spark, Fraud Detection, Machine Learning

1 INTRODUCTION

The convenience and security of commerce have unquestionably been improved by the rise in online transactions and the widespread use of credit cards. However, this Convenience brings with it the potential risk of credit card fraud, which can result in significant losses for financial institutions as well as individuals. Real-time fraud detection is essential to reducing these losses and protecting users. Machine learning techniques have shown themselves to be useful instruments for fraud detection in recent times. Facilitating the analysis of extensive datasets and recognizing discernible patterns indicative of deceptive activities is made more accessible through the application of decision trees and clustering algorithms. Notably, methodologies relying on these approaches have exhibited significant potential in the realm of identifying and addressing fake activity.

A number of techniques for detecting credit card fraud have been put out recently (CCFD). Machine learning algorithms identify attributes that indicate normal/fraudulent behaviour by using historical transaction data, which includes both normal and fraudulent transactions. These characteristics are then employed to evaluate the likelihood of fraud in a transaction.

This review article aims to provide an in-depth exploration of diverse machine learning techniques employed in the detection of credit card fraud. The focus centers on leveraging Apache

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Spark as a robust platform for large-scale data processing in this critical domain. After a thorough analysis and synthesis of the results from twenty pertinent research papers, this review intends to elucidate the strengths, limitations, and advancements in fraud detection techniques. The analyzed papers encompass a broad spectrum of methodologies, including single classifier-based techniques, clustering methods, and hybrid approaches that combine multiple models. These techniques take advantage of machine learning algorithms' natural capacity to learn from past data and spot patterns suggestive of fraudulent activities.

In addition, the incorporation of big data technologies, such Apache Spark, provides efficient and scalable processing capabilities that allow large amounts of transactional data to be analysed in real time.

This study aims to highlight significant developments and obstacles in the field of credit card fraud detection using machine learning and Apache Spark through a comprehensive assessment of the evaluated publications. Compared to Hadoop's MapReduce, Apache Spark, which is well-known for its ability to handle massive volumes of data efficiently, runs much faster. One noteworthy feature is its sophisticated DAG execution engine, which supports acyclic data flow and does in-memory processing[1]. Popular programming languages like Scala, Python, R, and Java are easily used by developers to create applications, and it allows for integration with a wide range of other sophisticated languages.

Abbassi presents a cutting-edge viewpoint on fraud detection metamethods in his discussion of the benefits and difficulties of big data technology [2]. This review also looks at the performance measures used to evaluate the efficacy of different models, offering important insights into the advantages and disadvantages of various strategies[3]. This review intends to add to the body of knowledge already in existence, spur more developments, and identify possible directions for future research in credit card fraud detection by illuminating these features.

2 METHODOLOGY

This section provides an overview of the dataset used in the study as well as an explanation of the preprocessing methods applied to the dataset. A number of fraud detection models are also shown, with the ultimate objective being the identification of fraudulent actions using machine learning techniques on the Spark framework. Fig. 1 provides a diagrammatic illustration of the suggested mode1.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

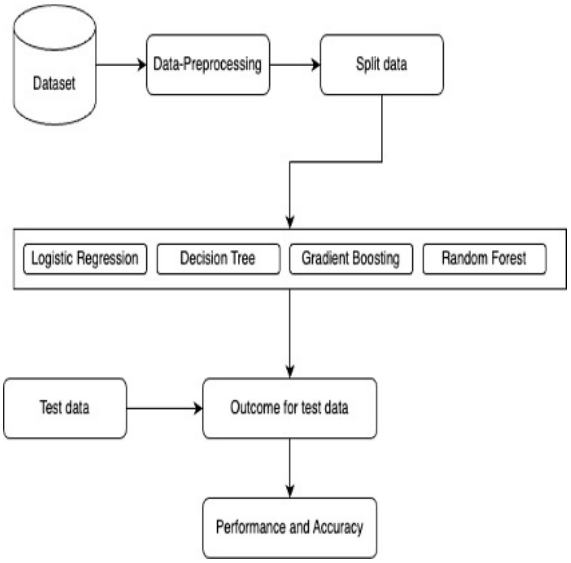


Fig.1 General Flow Chart for Proposed System.

2.1 DATASET DESCRIPTION

Obtaining an appropriate dataset to detect credit card fraud is an arduous and time-consuming procedure. Thankfully, there is a publicly accessible dataset designed just for this. This dataset, which is specifically intended for credit card fraud detection, consists of numerical input variables that are transformed using Principal Component Analysis (PCA). Notably designated as V1 through V28, these variables capture principal component values that were retrieved via PCA rather than directly representing the original data features. Unfortunately, it is not possible to do feature analysis or pre-analysis on this credit card fraud detection dataset due to a lack of information describing the original features.

It is noteworthy that there are no missing data points in the dataset. The qualities 'Time' and 'Amount' were thoroughly examined. The dataset consists of three unaltered attributes: 'Class' denotes the type of transaction ('0' indicates normal transactions, '1' indicates fraudulent ones), 'Time' indicates the amount of time in seconds that elapses between a given transaction and the first transaction in the dataset, and 'Amount' represents the total transaction value.

A summary of the credit card dataset's general statistics is provided in Table 1, and Fig.2 shows a histogram that illustrates the attributes' distribution.

A heatmap is a type of 2D data visualization where discrete values within a dataset are represented by colors, often ranging from 0.00 to 1.00. This heatmap is a useful tool for displaying the correlations between the attributes in the dataset in an intuitive manner[4]. In this instance, it makes it easier to understand the attribute correlations present in the credit card information and offers insights into the relationships and structure of the data

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

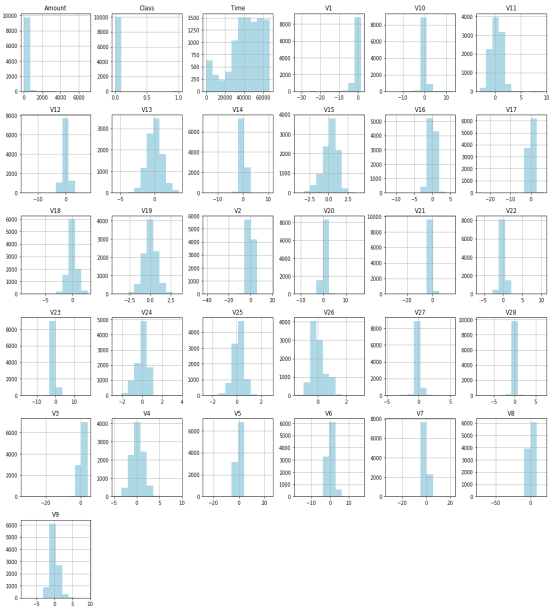


Fig. 2 Histogram f or attributes presented in the Dataset.

PARAMETER NAME	TOTAL COUNT
Number of Transaction	284807
Number of Columns	31
Number of Labels	1
Number of Normal Transaction	284315
Number of Fraud Transaction	492
% of Normal Transaction	99.82
% of Fraud Transaction	0.172

Table 1 A Heat map f or the Features in the Dataset

2.2 PREPROCESSING

Preprocessing involves various data preparation and enrichment steps that have a significant impact on the accuracy and effectiveness of fraud detection models and is critical to identifying credit card fraud. There is a class imbalance in detecting credit card fraud because illegal transactions tend to occur much less frequently than legal transactions. Methods such as under sampling, oversampling, or creating synthetic data (such as SMOTE) are used to balance the class distribution**.**

Data from credit card transactions frequently shows characteristics with different ranges and scales. To bring these variables into a uniform range and avoid biases towards features with bigger magnitudes, normalizing or scaling them is essential.

2.2.1 Robust Scaling

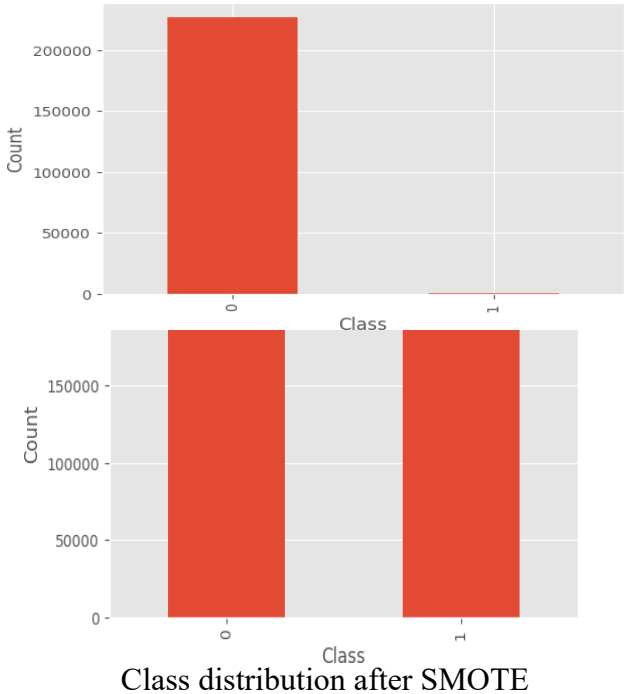
Let's concentrate on the 'Amount' feature in the credit card dataset, which shows the total transaction amount for every transaction. Owing to the nature of financial transactions, values for

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

'Amount' could vary greatly, possibly containing outliers that could have a major influence on conventional normalisation methods. The data is centred around the median and scaled using the Interquartile Range (IQR), a measure that is robust to outliers, in a robust manner. By ensuring that extreme values have less of an impact on scaling, this procedure strengthens the technique's resistance to outlier effects.

To apply Robust Scaling to the 'Amount' feature in credit card fraud detection, take each data point and subtract the median of the 'Amount' values, then divide the result by the IQR. A new column named "scaled_amount" holds the scaled values that are produced. By reducing the impact of anomalies and outliers in the 'Amount' feature, this transformation offers a more accurate and balanced depiction of transaction amounts. The credit card fraud detection model can now anticipate and classify data more accurately thanks to the use of robust scaling, since the'scaled_amount' feature efficiently captures underlying trends while reducing distortion from extreme values.

Class distribution before SMOTE



Class distribution after SMOTE

Fig 3 (Representation of Class Distribution)

2.2.2 Handling Class Imbalance

Creating robust machine learning models demands addressing class imbalance, particularly in scenarios where the occurrence of fraudulent transactions is significantly lower than that of valid transactions, as seen in credit card fraud detection[5]. A common solution to this challenge involves the utilization of the Synthetic Minority Oversampling Technique (SMOTE).

SMOTE generates synthetic samples for the minority class (fraudulent transactions),

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

effectively rebalancing the class distribution by interpolating between existing cases. This approach enhances fraud detection by mitigating the risk of the model exhibiting bias in favor of the majority class.

The class distribution may be significantly skewed prior to the deployment of SMOTE, with a disproportionately large number of valid transactions in comparison to SMOTE on the training set entails adding synthetic samples to the minority class, enhancing its representation and attaining a more balanced class distribution.

The fraudulent ones are represented by Fig 3. As a result, the model is better able to identify the fundamental patterns of fraudulent transactions, leading to more precise and trustworthy forecasts.

3.3 CLASSIFICATION MODELS

3.3.1 Logistic Regression

Despite its name, Logistic Regression is a popular and easily understood machine learning method for identifying credit card fraud. Specifically designed for binary classification tasks, it models the likelihood of fraudulent activity based on input features and efficiently discerns between authentic and fraudulent credit card transactions. By using a logistic function on a linear combination of input features, logistic regression calculates the likelihood that a transaction will belong to a specific class and gives a clear picture of the decision boundary.

3.3.2 Decision Tree

The interpretable and flexible Decision Tree technique is used for both regression and classification applications. By dividing the input data into subsets according to feature values, it forms a structure resembling a tree, with interior nodes standing in for decisions and leaf nodes for anticipated results[6]. Decision Trees, which visualise the decision-making process, are notable for their capacity to capture intricate decision boundaries and interactions among components.

3.3.3 By Chance

During training, Random Forest, an ensemble learning technique, builds numerous Decision Trees to improve forecast accuracy and robustness. Voting is used to aggregate predictions, which reduces overfitting and makes it suitable for a variety of data kinds, including complicated datasets.

3.3.4 Gradient Boosting

Gradient Boosting is a useful method for detecting credit card fraud since it trains weak learners iteratively and emphasises the rectification of cases that are misclassified. It is especially useful for datasets that are unbalanced as it iteratively improves predictions by concentrating on difficult cases. This method lowers false positive rates and increases detection rates by utilising the power of gradient boosting.

3.3.5 Classifier for Voting

To improve the accuracy of fraud detection, the Voting Classifier combines the predictive capabilities of Logistic Regression, Gradient Boosting, and Random Forest[7]. It outperforms individual models by combining predictions, offering fraud detection precision and flexibility

.3.3.6 Training and Testing the Models

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

While testing assesses the model's performance on unlabeled data, training entails extracting patterns from a labelled dataset. Metrics like accuracy and AUC score can be used to evaluate a model's performance in real-world scenarios.

4 OUTCOMES AND CONVERSATIONS

The work that is being presented aims to use a variety of models, such as Decision Tree, Random Forest, Gradient Boosting, and Logistic Regression, in order to detect fraud. Finally, a Voting Classifier is used, with Gradient Boosting, Random Forest, and Logistic Regression serving as the classifier's basic learners. This study makes use of a credit card dataset that is openly accessible.

4.1 IMPLEMENTATION DETAILS

The Databricks Community Edition of the DATABRICKS platform is specifically used for this project. A comprehensive cloud-based platform called Databricks offers a unified and cooperative workspace for tasks pertaining to machine learning, data science, and data engineering. It makes advantage of the high-performance distributed data processing platform Apache Spark and combines it with an intuitive user interface and a set of tools to simplify the data lifecycle.

Large and complicated datasets may be processed, analysed, and visualised with ease with Databricks' scalable and effective solution. Advanced machine learning techniques are used in the Databricks platform to improve credit card fraud detection with the use of PySpark. PySpark's integration with the Databricks collaborative environment makes it possible to handle and analyse massive amounts of transaction data quickly and effectively. This in turn makes it easier to use the MLlib and SparkML libraries to create reliable fraud detection models.

This combo uses interpretable algorithms and ensemble approaches to create accurate models by utilising parallel computing. The Databricks collaborative workspace facilitates teamwork and knowledge exchange, which enhances the accuracy and efficacy of fraud detection.

4.2 RESULTS

An extensive study of the performance of multiple classification models is provided by assessment, which is trained on resampled data and tested independently. The ability to distinguish between classes is graphically represented by the AUC-ROC curve, whereas accuracy provides a comprehensive measure of prediction correctness. Examining these outcomes for a variety of models—Decision Trees, Random Forest, Gradient Boosting, and the Voting Classifier, among others—reveals a thorough comprehension of the effectiveness of each model in identifying credit card fraud. The outcomes of every machine learning model that was used with PySpark on the Databricks platform using the credit card dataset are displayed in Table 2.

We give the ROC curve, as seen in Fig 4, to evaluate the prediction models' accuracy. The Voting Classifier's performance is assessed using the ROC curve. The results that are displayed show that the classification performance in the area under the ROC curve (AUC) is 98%. Fig 5 and 6 show the Receiver Operating Characteristic (ROC) curves for the other models.

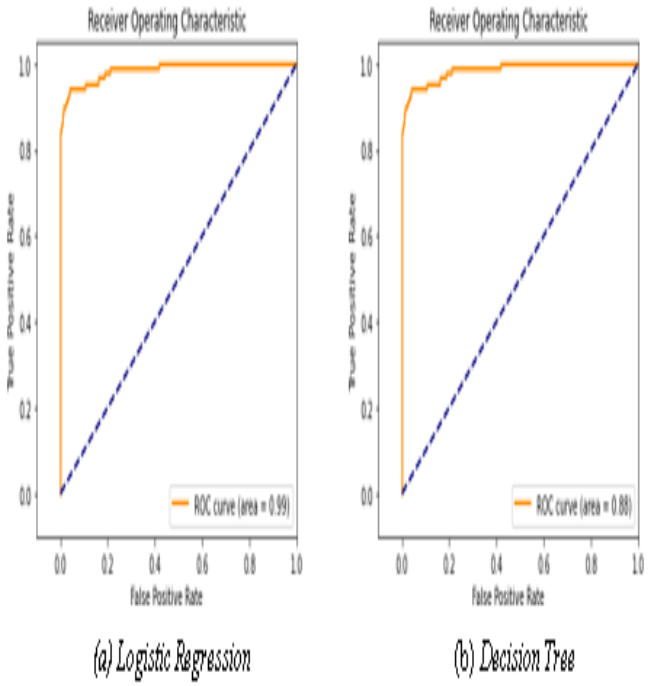
ML MODEL	AUC	ACCURACY
Logistic	**87.2**	**97.8**

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Regression		
Decision Tree	**87.2**	**99.7**
Gradient Boosting	**96.5**	**99.3**
Random Forest	**96.5**	**99.8**
Voting Classifier	**98.6**	**99.6**

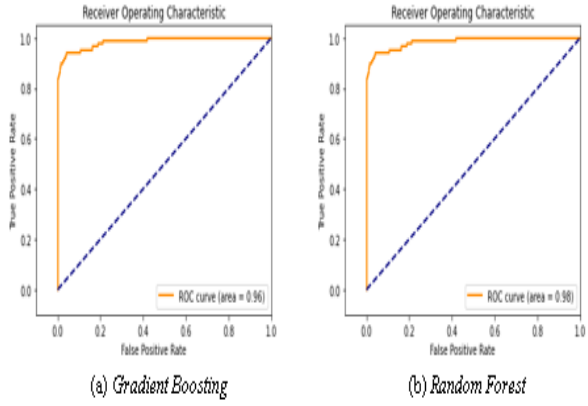
Table 2 Assessing Metrics for Different Models

.



(a) Logistic Regression (b) Decision Tree

Fig. 4 ROC Receiver operating characteristic curve



(a) Gradient Boosting (b) Random Forest

Fig-5 Receiver operating characteristic curve for Logistic Regression and Decision Tree.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

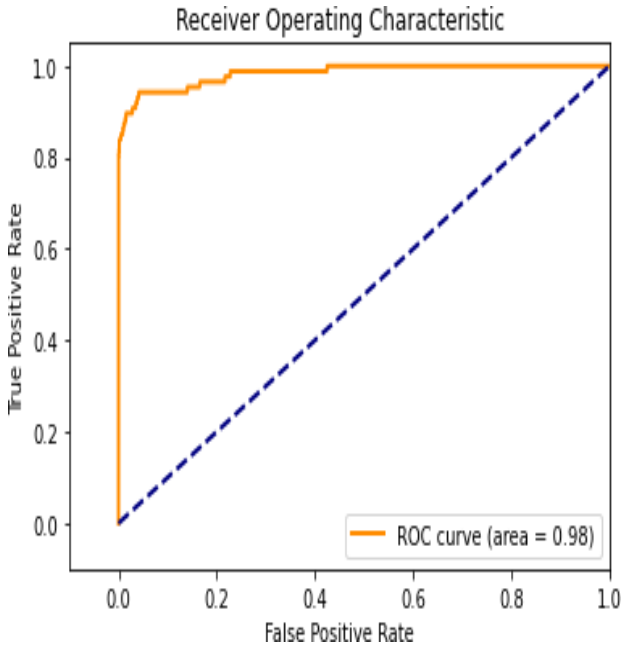


Fig-6 Receiver functional distinctive curve for Gradient Boosting and Random Forest

5 CONCLUSION AND FUTURE ENHANCEMENT

This research presents a major breakthrough in credit card fraud detection by combining big data analysis with machine learning methods. The suggested approach, which makes use of PySpark, performs better in categorization than more modern systems. PySpark and MLflow integration with the Databricks platform has simplified development and deployment while demonstrating an astounding 98% average classification performance.

The system has more potential in the future than it has now. To verify resilience, further study will entail assessing its performance on various datasets. Furthermore, the goal of combining sophisticated machine learning methods and deep learning models is to improve classification accuracy. By investigating Spark's distributed computing capabilities, one may more effectively handle larger datasets and enhance real-time or almost real-time fraud detection. As credit card fraud advances, this study paves the way for increasingly complex and potent detection techniques.

REFERENCES

[1] R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, “Credit card fraud detection using machine learning,” in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 1264–1270.
[2] H. Abbassi, I. El Alaoui, and Y. Gahi, “Fraud detection techniques in the big data era,” 2022.
[3] 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), 2017, pp. 1–9.
[4] P. Save, P. Tiwarekar, K. N. Jain, and N. Mahyavanshi, “A novel idea for credit card fraud detection using decision tree,” International Journal of Computer Applications, vol. 161, no. 13, 2017.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

[5] R. More, C. Awati, S. Shirgave, R. Deshmukh, and S. Patil, "Credit card fraud detection using supervised learning approach," Int. J. Sci. Technol. Res, vol. 9, pp. 216–219, 2021.
[6] M. R. Dileep, A. V. Navaneeth, and M. Abhishek, "A novel approach for credit card fraud detection using decision tree and random forest algorithms," in 2021 Third International Conference on Intelligent Communication Tech- nologies and Virtual Mobile Networks (ICICV), 2021, pp. 1025–1028.
[7] J. I.-Z. Chen and K.-L. Lai, "Deep convolution neural network model for credit-card fraud detection and alert," Journal of Artificial Intelligence, vol. 3, no. 02, pp. 101–112, 2021.