iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52214

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Credit Card Fraud Detection Using Machine Learning and Blockchain

Mr. Soham Patil[1], Mr. Amey Godse [2], Mr. Prasad Gawade[3], Mr. Prajwal Halkare[4], Mr. Abhinay Dhamankar[5]
Pune Institute of Computer Technology

Abstract: ***This paper proposes a secure fraud detection model that combines machine learning algorithms and blockchain technology. According to the study it notes that while blockchain is considered a secure method of integration into finance, fraud and anomalies are still common in the network. To address this, the we use two machine learning algorithms, XGboost and KMeans classifier, to classify transactions based on patterns of fraudulent and legitimate transactions .The transaction data is also classified by supervised machine learning algorithms like Random forests and Decision Tree Classifier . The proposed model integrates these algorithms with blockchain technology to detect fraudulent transactions in the Etherium network . The paper includes a security analysis of the proposed smart contract and an attacker model to protect the system from potential attacks and vulnerabilities. The precision and AUC of the models are also calculated to measure the accuracy of the system. Overall, the paper proposes an innovative approach to addressing fraud and anomalies in the network using machine learning and blockchain technology. The integration of these technologies has the potential to improve the security of online transactions and e-banking systems.***

Keywords: anomaly detection; blockchain; fraud detection; machine learning; random forest; XGBoost

I. INTRODUCTION

It is clear that technological advancements have led to the modernization of various industries, including the financial sector, where traditional currencies are being replaced by digital currencies. However, these transactions are vulnerable to digital attacks, and detecting fraud and anomalies in digital transactions is critical for maintaining the integrity of the financial system. Anomaly detection techniques are used to detect illegal and fraudulent activities in financial transactions, but existing methods are designed for centralized systems. Blockchain technology offers a decentralized and immutable ledger that can address security issues in centralized systems. Blockchain-based systems have the potential to provide secure and transparent financial transactions while maintaining privacy. However, malicious actors can still exploit vulnerabilities in the blockchain network, and detecting fraudulent transactions in the blockchain is a challenging task. The application of machine learning techniques, such as XGBoost and random forest, to blockchain data has the potential to improve the detection of fraudulent transactions. In this proposed system, machine learning models are directly linked to the blockchain, and a blockchain-based smart contract is deployed to classify incoming transactions as fraudulent or legitimate. The proposed system also includes two attacker models to protect against blockchain attacks .Overall, the proposed system has the potential to improve the security and integrity of financial transactions in the digital age. However, it is important to note that the effectiveness of the system will depend on the quality and quantity of data used to train the machine learning models, as well as the ability to adapt to evolving attack patterns.

A. Abbreviations and Acronyms

ANN	Artificial Neural Network
DBF	Deep Blockchain Framework
RF	Random Forest
AUC	Area Under Curve
PoW	Proof of Work
BoW	Bag of Words
PCA	Principle Component Analysis
SMOTE	Synthetic Minority Oversampling Technique
XGboost	eXtreme Gradient Boosting

B. Related Work

It is true that blockchain technologies are being deployed in different public and private regions for various objectives, particularly in protecting and monitoring auditing systems. Blockchain allows for secure and private queries from auditors without exposing their identities to unauthorized users. However, using blockchain alone for fraud detection may not be sufficient, as it may not efficiently identify fraudulent transactions. To address this problem, new solutions such as machine learning algorithms are being used.

Supervised machine learning techniques are particularly useful in detecting fraudulent transactions. Different methods have been tested, and a comparative analysis of these methods has been presented in various studies. For instance, in [8], the authors proposed different supervised machine learning solutions for detecting fake businesses and tested them using random forest and XGBoost classifiers on over 300,000 accounts. XGBoost was also used in [9] for accurate results. In [10], the authors addressed the problem of an imbalanced dataset, which is a common issue in fraud detection, by applying specific techniques to balance the dataset.

Overall, combining blockchain technologies with machine learning algorithms can provide more robust solutions for fraud detection and auditing systems. Fraudulent activities in credit card transactions are data mining issues because identifying fraudulent transactions requires analyzing large amounts of data. However, real-time data for fraud detection is often not readily available to researchers due to the confidential nature of customer data and banks' privacy policies.

To address these challenges, various approaches have been proposed. In [13], a distributed data mining model was used to address problems of slanted delivery of credit cards and non-uniform expenditures. In [14], a fraud detection algorithm was presented that can identify fraud without relying on any fraudulent historical instances, overcoming the cold-start problem. In [15], the authors suggested and demonstrated the application of uncertain association law mining to extract useful data from credit card transactions.

Other techniques, such as support vector machine models [16] and a combination of Bayesian learning, rule-based learning, and Dempster–Shafer theory [17], have also been used to decrease wrong identifications of fraud. In [18], a transaction aggregation technique was used to interpret customer behavior before any transaction is performed and then used to identify fake transactions. This model can work with unknown datasets and can identify fraudulent transactions while maintaining customer privacy.

It is crucial to ensure the privacy and security of data in cloud-based systems, especially for cyber-physical systems. In [23], the authors proposed an anomaly detection system that uses machine learning algorithms to detect both insider and outsider attacks. The system is based on a distributed architecture that preserves the privacy of the data by encrypting it before sharing it with other nodes in the network. The authors in [24] proposed a privacy-preserving machine learning framework for edge computing. The framework uses differential privacy and federated learning to train machine learning models on data that is distributed across different edge devices. The authors demonstrated that the proposed framework can achieve high accuracy while preserving the privacy of the data.

In [25], the authors addressed the issue of privacy in location-based services (LBS) by proposing a privacy-preserving framework based on blockchain technology. The framework uses a combination of homomorphic encryption and smart contracts to ensure the privacy of user data while still allowing LBS providers to offer personalized services. In [26], the authors proposed a secure and privacy-preserving data sharing framework for healthcare applications. The framework uses blockchain technology to ensure data integrity and privacy, and differential privacy to protect sensitive information. The authors demonstrated the effectiveness of the proposed framework on a real-world dataset.

Finally, in [27], the authors proposed a privacy-preserving data analysis framework for smart grids. The framework uses homomorphic encryption and secret sharing to ensure the privacy of the data while allowing the utility company to perform various data analysis tasks. The authors demonstrated the effectiveness of the proposed framework on a real-world dataset from a smart grid testbed. Adversarial attacks pose a significant threat to the security and robustness of machine learning models, especially in sensitive domains such as finance and cyber security. In recent years, researchers have proposed various techniques to mitigate the impact of such attacks. For example, in [31], the authors proposed a model-agnostic defense approach called Adversarial Training with Ensemble Diversity (ATED), which combines adversarial training and ensemble learning to improve the model's robustness against adversarial attacks. The authors of [32] proposed a defense mechanism that uses a conditional generative adversarial network (cGAN) to generate adversarial examples that are indistinguishable from real examples, thus fooling the attacker's model. In [33], the authors proposed a method based on gradient regularization to improve the robustness of deep neural networks against adversarial attacks. Despite these efforts, adversarial attacks remain a challenging problem, and new defense mechanisms need to be developed to mitigate their impact. In [34], the authors proposed a novel approach that combines multiple defense mechanisms, including adversarial training, feature squeezing, and gradient masking, to improve the model's robustness against various types of attacks. In [35], the authors proposed a method based on gradient obfuscation, which modifies the gradient of the model to make it harder for the attacker to generate effective adversarial examples.

Other researchers have explored the use of game theory and reinforcement learning to develop more effective defense mechanisms against adversarial attacks [36].

Overall, adversarial attacks pose a significant challenge to the security and robustness of machine learning models. Although researchers have proposed various defense mechanisms to mitigate their impact, this remains an active research area with significant room for improvement.

II. PROBLEM STATEMENT AND SYSTEM DESIGN

A. *Problem Statement*

It allows secure transactions without the need for a centralized authority. However, the PoW algorithm requires a significant amount of computational power and energy consumption ,leading to environmental concerns. To address these issues, alternative consensus algorithms have been proposed, such as proof of stake (PoS) and delegated proof of stake (DPoS), which are more energy-efficient. In addition to blockchain technology, other AI techniques such as machine learning and natural language processing (NLP) can also be used to combat financial fraud. These

techniques can analyze vast amounts of data in real-time and identify patterns and anomalies that may indicate fraudulent activity. Furthermore, NLP can be used to analyze text-based data such as emails and chat messages to detect fraudulent behavior .Overall, the financial sector faces significant challenges in combating cybercrime and fraud. However, with the use of advanced technologies such as blockchain, AI, and NLP, it is possible to develop more robust and secure systems to prevent and detect fraudulent activity. Ongoing research and development in these areas will continue to improve the security and resilience of financial systems in the face of evolving cyber threats.

B. *Proposed Model*

T he proposed system model that integrates blockchain and machine learning for fraud and anomaly detection in the financial sector is an innovative approach. The blockchain layer initiates transactions, and then machine learning models are used to classify them as legitimate or malicious, based on their characteristics.

Binary classification is used to determine if a transaction is fraudulent or not. The machine learning models are trained on a dataset of bitcoin transactions, which is a popular cryptocurrency used in the financial sector. The dataset is used to identify unusual and suspicious events that deviate from the normal data patterns.

The random forest and XGboost classifiers are used to classify transactions as legitimate or malicious. These classifiers are also used to predict incoming transactions, which can help prevent fraudulent activities in the financial sector.

The proposed model is trained and tested using the given dataset to identify legitimate and malicious data patterns. The model's performance can be evaluated based on metrics such as precision, recall, and F1-score.

Overall, the proposed system model can be a useful tool for fraud and anomaly detection in the financial sector, particularly for cryptocurrency transactions. The integration of blockchain and machine learning can provide an added layer of security and help prevent fraudulent activities.

1) SMOTE Analysis

Algorithm 1: Data balancing through SMOTE
1: Initialization
2: Inputs: Minority data M(D)= mi 2 X, Where i = 1,2,3
3: Outputs: Synthetic Data S
4: Number of minority samples (D)
5: Percentage of SMOTE (P)
6: Number of (k) nearest neighbors
7: for n = 1 to D do
8: Find the K nearest neighbors of Di
9: Check P = P/100
10: While P 6= 0 do
11: Select a random sample m in minority class
12: Find neighbor of m
13: Pick a random number *a* 2 [0, 1]

14: m = mi + *a*(m□ mi)
15: While Append m to S
16: Check P = P □ 1
17: end while
18: end for
19: End

The imbalance of data in machine learning can be a significant problem, as it can result in biased models with poor performance. The SMOTE (Synthetic Minority Over-sampling Technique) algorithm is a widely used method to solve this problem, by generating synthetic data for the minority class. SMOTE works by randomly selecting data points from the minority class and then generating new synthetic samples based on their nearest neighbors in the feature space. This process helps to balance the data distribution and improve the effectiveness of machine learning algorithms. Algorithm 1 shows the steps involved in using SMOTE to balance data with an imbalanced class distribution. The input, output, and variables are initialized in lines 1-6. Then, in lines 7-16, SMOTE is applied to generate synthetic data points for the minority class. The SMOTE algorithm works based on the k-nearest neighbor approach. In the first step, a data point is randomly selected from the minority class. Then, the k-nearest neighbors for that point are determined. Finally, synthetic data points are generated by selecting random combinations of features from the minority class and its nearest neighbors. The SMOTE algorithm is a powerful tool for addressing the issue of imbalanced data in machine learning. By generating synthetic data, it helps to balance the class distribution and improve the accuracy of models. However, it is important to use SMOTE with caution, as it can also result in overfitting and other issues if not applied properly.

2) Fraudulent Transaction Detection

The rise of online businesses has led to an increase in fraudulent activities, which can be challenging for organizations to combat using traditional fraud detection systems. These systems often rely on static rules created by human experts, which may not be effective in detecting new or evolving forms of fraud. In this study, the focus is on detecting fraudulent transactions involving Bitcoins, a popular cryptocurrency used in the financial sector. Anomaly detection is used to identify unusual patterns in Bitcoin transactions that do not conform to expected behavior. The proposed model is based on a dataset of Bitcoin transactions and is trained using machine learning algorithms. The dataset used in this study is based on Bitcoin transactions, but since the transaction patterns of cryptocurrencies like Ethereum (Ether) are similar to Bitcoin, the model is expected to perform well on Ethereum transactions too. The proposed model is well-suited for the financial sector, where blockchain-based cryptocurrencies are commonly used. It provides an efficient way to detect fraudulent transactions and can help organizations minimize the impact of fraudulent activities on their business. Overall, the proposed model can be a valuable tool for organizations in the financial sector, where fraud prevention is crucial for maintaining the integrity of transactions and building trust with customers.

3) XGBoost

XGBoost is a powerful boosting algorithm that generates a sequence of decision trees. The goal of each subsequent tree is to reduce the error of the previous tree and update the residual error. This is achieved by building trees sequentially, with each new tree learning from the errors of the previous trees. In the proposed model, XGBoost is used as a classifier to differentiate between legitimate and malicious transactions. The algorithm is trained on a dataset of Bitcoin transactions and can accurately classify new transactions based on their features. Furthermore, the XGBoost algorithm can be connected to a blockchain smart contract to predict new incoming transactions. This can be useful in real-time fraud detection systems, where quick identification of suspicious transactions is critical. Overall, the use of XGBoost in the proposed model is an effective approach for fraud detection in blockchain-based financial transactions. It provides a powerful tool for detecting fraudulent activities, which can help organizations minimize the impact of such activities on their business.

Algorithm 2: Fraud detection using XGboost
1: Inputs: Balanced Dataset S
2: Outputs: Transactions in Blockchain B
3: Initialization of Dataset
4: Spliting of S into training and testing
5: Xtrain input variables from dataset
6: Ytrain target variables to dataset
7: Xtest input variables from test dataset

8: Ytest target variables from test dataset
9: Model = XGBClassifier(nestimators = 100)
10: Model = Model.fit(Xtrain, Xtrain)
11: Ypred = Model.predict(Xtest)
12: Predictions = [round(value) for value in Ypred]
13: if Predictions == 0 then
14: transaction = legitimate
15: B.add (transaction)
16: else if Predictions == 1 then
17: transaction = malicious
18: end if
19: return B
20: End

Algorithm 2 outlines the implementation of XGboost for fraud detection in blockchain-based transactions. The algorithm takes in the input dataset, which is split into training and testing sets. The XGboost model is then trained on the training data and deployed for testing on the test set. The integration of the blockchain technology is also shown in this algorithm, where the algorithm checks the integrity of new incoming transactions by passing them to the trained XGboost model. If the model predicts the transaction as legitimate, it sends the transaction status back to the blockchain with the "0" prediction value, and if the transaction is classified as malicious, it sends the transaction status back to the blockchain with the "1" prediction value.

4) KMeans Clustering

K-means is a clustering algorithm that groups similar data points together. In the context of fraud detection, K-means can be used to group together transactions that have similar patterns. By doing so, it can help to identify groups of transactions that are anomalous or suspicious. We have a dataset of credit card transactions. We can apply K-means clustering to group together transactions that have similar attributes such as transaction amount, time of day, merchant category code, etc. Once we have these clusters, we can analyze them to identify any patterns that are unusual or suspicious. For instance, if we find a cluster of transactions with unusually large transaction amounts or with a high frequency of transactions at merchants with a high risk of fraud, we may flag those transactions for further investigation. K-means can be a useful tool in fraud detection, especially when used in conjunction with other techniques such as anomaly detection and supervised learning algorithms. It is important to note, however, that K-means clustering is not perfect and can have limitations such as sensitivity to initial conditions and the need for the number of clusters to be specified beforehand. Therefore, it is important to use K-means as part of a comprehensive fraud detection system that includes multiple techniques and approaches.

5) ML and Blockchain linkage

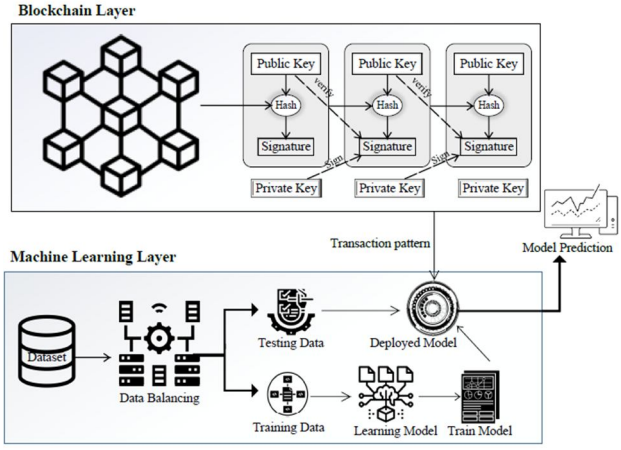


Figure 1. The proposed system mode of blockchain and ML.

The proposed system combines the use of blockchain and machine learning for fraud detection. The system receives a new transaction from the Ethereum network, and the transaction pattern is analyzed and compared to the pattern of bitcoin transactions stored in the database. The machine learning model is trained on the bitcoin transaction-based dataset and predicts if the new transaction is legitimate or malicious. If the prediction result is legitimate, the transaction is added to the blockchain. Otherwise, it is rejected, and the transaction is not added to the blockchain. The system provides a robust mechanism for detecting fraudulent transactions and ensures the security and privacy of the blockchain network.

III. RESULTS AND DISCUSSION

Let's check the accuracy of our decision tree model.

```
[ ] print('Accuracy score of the Decision Tree model is {}'.format(accuracy_score(y_test, dt_yhat)))

    Accuracy score of the Decision Tree model is 0.9991583957281328
```

Checking F1-Score for the decision tree model.

```
[ ] print('F1 score of the Decision Tree model is {}'.format(f1_score(y_test, dt_yhat)))

    F1 score of the Decision Tree model is 0.7521367521367521
```

```
[ ] confusion_matrix(y_test, dt_yhat, labels = [0, 1])

    array([[68770,    18],
           [   40,    88]], dtype=int64)
```

Here, the first row represents positive and the second row represents negative. So, we have 68782 as true positive and 18 are false positive. That says, out of 68782+18=68800, we have 68782 that are successfully classified as a normal transaction and 18 were falsely classified as normal – but they were fraudulent.

Let's check the accuracy of our XGBoost model.

```
[ ] print('Accuracy score of the XGBoost model is {}'.format(accuracy_score(y_test, xgb_yhat)))

    Accuracy score of the XGBoost model is 0.999506645771664
```

Checking F1-Score for the XGBoost model.

```
[ ] print('F1 score of the XGBoost model is {}'.format(f1_score(y_test, xgb_yhat)))

    F1 score of the XGBoost model is 0.8495575221238937
```

```
print('Accuracy score of the Random Forest model is {}'.format(accuracy_score(y_test, rf_yhat)))

Accuracy score of the Random Forest model is 0.999143853096524
```

Checking F1-Score for the Random Forest model.

```
[ ] print('F1 score of the Random Forest model is {}'.format(f1_score(y_test, rf_yhat)))

    F1 score of the Random Forest model is 0.728110599078341
```

Let's check the accuracy of our Logistic Regression model.

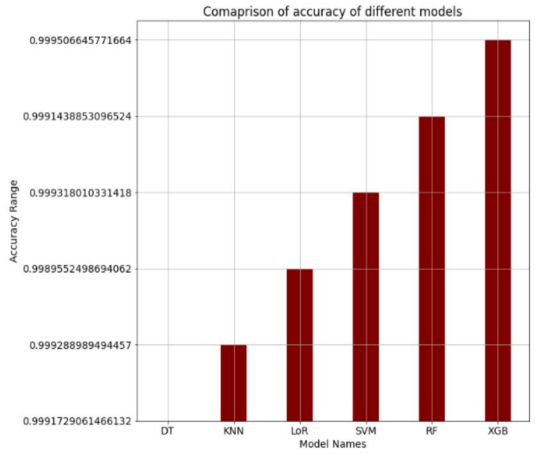
```
[ ] print('Accuracy score of the Logistic Regression model is {}'.format(accuracy_score(y_test, lr_yhat)))
```

```
Accuracy score of the Logistic Regression model is 0.9989552498694062
```

Checking F1-Score for the Logistic Regression model.

```
[ ] print('F1 score of the Logistic Regression model is {}'.format(f1_score(y_test, lr_yhat)))
```

```
F1 score of the Logistic Regression model is 0.6666666666666666
```



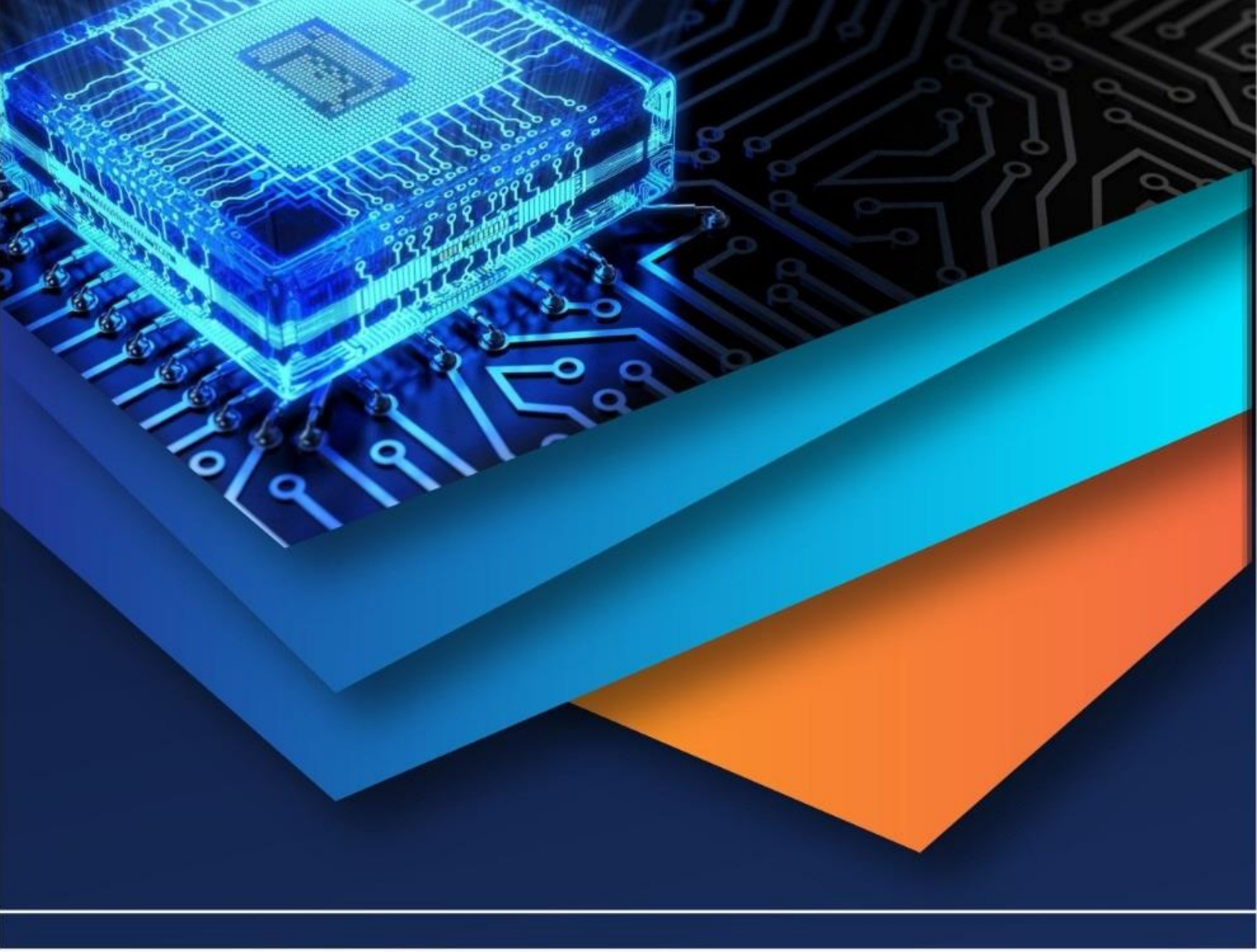
We recently got a 99.95% accuracy rating for detecting credit card fraud. Given that our data was balanced in favour of one class, this figure shouldn't come as a surprise. Our model is not overfitted, which is a positive finding from the uncertainty matrix. Finally, XGBoost is the winner in our situation based on our accuracy score. The data that we have been given for model training is the only problem with this. The PCA-transformed rendition of the data features. We are doing fantastic if the real features follow a similar pattern!

IV. CONCLUSION

It protects financial systems from fraudulent attacks. Therefore, a blockchain-based machine learning algorithm is proposed to secure digital transactions. In this project, various supervised learning approaches to support vector machines, Ada boost and random forest classifier were used. The proposed model predicts whether the incoming transaction in the blockchain is fraudulent or not. The proposed model predicts whether the incoming transaction in the blockchain is fraudulent or not. The supervised learning algorithms allows the model to distinguish between fraudulent and real data. The simulation results show that the proposed algorithm works adequately to find transaction fraud.

V. FUTURE SCOPE

Our model can be made more precise and accurate by using deep learning algorithms in place of supervised machine learning algorithms. Our model can be made more efficient against Sybil Attacks where the malicious attacker uses multiple identities. As we are new into the Blockchain technology and have limited knowledge about it, with proper time given we will be able to explore more about the Blockchain Technology and use it very effectively in our project. We will try to design and build our model for very very big datasets in the future.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY