IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A New Framework for Fraud Detection in Bitcoin Transaction through ensemble stacking model in smart cities

M.S. Bennet Praba1, Rahul Jaiswal2, Shobhit Kumae3, and Aman Sachan4

[1] SRMIST, Ramapuram, Chennai
[2] SRMIST, Ramapuram, Chennai
[3] SRMIST, Ramapuram, Chennai
[4] SRMIST, Ramapuram, Chennai
bennetms09@gmail.com

Abstract. Bitcoin has garnered a notorious reputation for its association with unlawful activities, ranging from money laundering to facilitating dark web transactions and serving as a preferred method of payment for ransomware in the context of smart cities. While blockchain technology provides inherent security features that prevent unauthorized modifications to transaction records, it falls short in actively detecting and thwarting illicit transactions. Traditional detection techniques, such as heuristic and signature-based approaches, formed the foundation of early detection methods. However, these methods proved inadequate in tackling the intricate complexity of anomaly detection within the Bitcoin network. Recognizing the limitations of conventional approaches, the adoption of machine learning (ML) has emerged as a promising avenue for enhancing anomaly detection capabilities. ML algorithms can be trained on vast datasets containing known malware samples, enabling them to discern subtle patterns and features indicative of illicit transactions. By leveraging ML's ability to analyze large volumes of transaction data, smart cities can bolster their defenses against fraudulent activities in the Bitcoin ecosystem, thereby fostering a safer and more secure digital landscape.

Keywords: illicit transaction, Anomaly detection, Signature-based approach, Support Vector machine, Machine Learning.

1 Introduction

A blockchain is a decentralized and distributed ledger that securely records transactions in a public manner. Each block within the chain contains a sequence of transactions that have been verified and accepted by the network. The fundamental principle of blockchain technology lies in its consensus mechanism, whereby network participants must collectively agree on the validity of transactions. Once a block has been added to the chain through consensus, it becomes immutable, meaning it cannot be altered or deleted. Bitcoin, as a prominent example of blockchain implementation, relies on this decentralized network rather than centralized institutions to control and validate transactions. This decentralized nature ensures that transactions conducted through Bitcoin are safe, swift, and cost-effective, bypassing the need for intermediaries such as banks or payment processors.

However, despite its advantages, blockchain technology is not impervious to risks and vulnerabilities. One of the primary concerns surrounding Bitcoin is its association with unlawful activities. While the block chain itself provides transparency and traceability, it can also be utilized for illicit purposes due to its pseudonymous nature, which can make it challenging to link specific transactions to real-world identities. Consequently, Bitcoin has gained notoriety for being utilized in activities such as money laundering, facilitating transactions on the dark web, and as a preferred method of payment for ransom ware attacks.

Nevertheless, ongoing efforts are being made to enhance the security and resilience of blockchain networks. Innovations such as improved consensus algorithms, advanced encryption techniques, and regulatory frameworks aim to mitigate these risks and promote the responsible use of blockchain technology. By addressing these challenges, the potential of blockchain to revolutionize various industries while maintaining integrity and security can be fully realized..

2 Literature Survey

The transition to advanced metering infrastructure (AMI) in modern electricity grids has created a pressing need to detect electricity theft. This detection is facilitated by studying energy consumption (EC) data from smart meters (SMs). While machine learning

(ML) and deep learning (DL) strategies have been proposed for identifying power theft in smart grids (SGs), they often struggle with imbalanced data. To address this, a novel approach, the ML boosting classifiers-based stacking ensemble model (MLBCSM) with adaptive synthetic sampling (ADASYN), is proposed. This model, by combining multiple boosting algorithms (AdaBoost, XGBoost, HistBoost, CatBoost, LGBoost), shows superior outcomes in electricity theft detection (ETD) compared to individual models.
[2] Ensemble learning, a method that combines the predictions of multiple classifiers, has been shown to outperform individual classifiers in various applications. In the context of crime prediction, a novel method called assemble-stacking based crime prediction method (SBCPM) is proposed. This method uses support vector machine (SVM) algorithms to improve the accuracy of crime predictions. By integrating learning-based methods and utilizing MATLAB, the proposed model achieves high classification accuracy on testing data, surpassing previous research efforts.
[3] The increasing prevalence of cardiovascular diseases (CVDs) necessitates accurate forecasting of daily hospital admissions (HAs) for CVDs. A stacking ensemble model is proposed to predict the daily number of CVDs admissions using data from HAs, air pollution, and meteorological conditions. By utilizing machine learning models such as linear regression (LR), support vector regression (SVR), extreme gradient boosting (XGBoost), random forest (RF), and gradient boosting decision tree (GBDT), the stacking model outperforms individual models in terms of mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R2). [4] A novel prediction method for metro traction energy consumption using WGAN-GP and stacking ensemble learning is proposed, showing improved prediction accuracy. The model demonstrates the effectiveness of ensemble learning in improving prediction performance.
[5] Decentralized cryptocurrencies have become increasingly popular, leading to the need for accurate trading predictions. Various artificial intelligence (AI) techniques have been applied to address trading challenges in cryptocurrencies. This survey explores recent research in this area, highlighting the application of AI techniques in addressing trading challenges such as price and trend prediction, volatility prediction, portfolio construction, and fraud detection in cryptocurrencies.[6]
Protective structures are essential for spacecraft to resist the impact of micrometeoroids or space debris. A novel method for predicting hypervelocity impact damage of stuffed protective structures is proposed using a stacking ensemble learning framework. By combining different machine learning models (XGBoost, RF, SVM, KNN) and leveraging a meta-model (LSTM), the proposed model achieves high prediction accuracy for impact damage prediction. [7] Early ICU mortality prediction is crucial for identifying high-risk patients and providing timely interventions. A stacking ensemble approach is proposed for ICU mortality prediction, outperforming existing methods in terms of accuracy, F1 score, precision, recall, and area under the receiver operator characteristic (ROC) curve. The model demonstrates the potential of ensemble learning in improving ICU mortality prediction.
[8] Early ICU mortality prediction is crucial for patient care. A stacking ensemble approach is proposed, outperforming existing methods in terms of prediction accuracy. The model demonstrates the potential of ensemble learning in improving ICU mortality prediction.[9] Laser-Induced Breakdown Spectroscopy (LIBS) is a popular technique for elemental quantitative analysis, but existing methods struggle with accurate sample analysis. A heterogeneous stacking ensemble learning model (Hackem-LIBS) is proposed to improve LIBS quantitative analysis accuracy. By combining different heterogeneous component learners and leveraging Genetic Algorithm (GA) and Sequential Forward Selection (SFS) for feature selection, the proposed model achieves better accuracy in determining the concentrations of elements in complex chemical samples via the LIBS technique.

3 Proposed Work

Our proposed research aims to develop a pioneering effort in leveraging advanced technological solutions to address the growing concerns surrounding fraudulent activities within the realm of digital currencies, particularly Bitcoin. As smart cities continue to embrace innovative technologies to enhance efficiency and connectivity, ensuring the integrity and security of financial transactions becomes paramount. The framework revolves around the utilization of ensemble stacking models, a sophisticated machine learning approach, to detect and prevent fraudulent transactions in the Bitcoin network. Unlike traditional heuristic or signature-based methods, which have proven inadequate in navigating the complexities of anomaly detection within blockchain systems, ensemble stacking models offer a more nuanced and robust solution. By leveraging the collective wisdom of multiple machine learning algorithms, each trained on diverse datasets, the framework enhances the accuracy and reliability of fraud detection while minimizing false positives.

In the context of smart cities, where the seamless flow of digital transactions is essential for sustaining urban infrastructure and services, the proposed framework holds immense potential. By deploying this advanced model, smart cities can bolster their defenses against various forms of financial crimes, including money laundering, illicit transactions on the dark web, and ransomware payments. Moreover, the framework enables authorities to proactively identify and mitigate emerging threats, thereby safeguarding the financial interests and security of city residents and businesses. Furthermore, the adoption of such a framework underscores the commitment of smart cities to embrace cutting-edge technologies while upholding principles of transparency, accountability, and ethical governance. As blockchain technology continues to evolve and reshape the digital landscape, initiatives like the proposed framework serve as a testament to the proactive measures taken to harness its transformative potential for the collective benefit of society. Through collaborative efforts between academia, industry, and government stakeholders, the proposed framework lays the groundwork for a safer, more resilient, and inclusive digital future in smart cities.

3.1 Random Forest

RF is an ML supervised technique which is built by combining hundreds of DTs. All the DTs work in a parallel manner. RF is a bagging ensemble method. The mathematical representation and algorithm of the RF are given below [53] and [54]. In Equation 11, i represents the number of important features calculated for all trees j. While T represents the total number of trees.

$$RFfi_i = \sum_{J} normfi_{ij}/sumT$$

Pseudo code of linear regression

```
Require: Xtrain: Training set with n instances
F: number of features
A: number of classes in a target class B: number of tress
1: Begin
2: for i = 1 to B do
3: Generate the bootstrap samples Xtrain[i] from the training set Xtrain
4: Using random sample from Xtrain[i] and create a tree
5: For a selected node t
• Randomly select m ≈√F
• Find the best point from the subset
• Pass down the data using the best points
Repeat these step until the termination condition are met
6: Construct the trained classifier
7: End
```

, S represents the training set, F is the result of final prediction and B represents the number of DT

3.2 Support Vector Machine

Support Vector Machine (SVM) plays a crucial role in the proposed framework for fraud detection in Bitcoin transactions within smart cities. SVM is a powerful machine learning algorithm that excels in classification tasks, making it instrumental in identifying fraudulent transactions amidst the vast volume of data processed within the Bitcoin network. In the ensemble stacking model, SVM serves as one of the key components alongside other machine learning algorithms. Its importance lies in its ability to effectively classify transactions as either legitimate or fraudulent based on a multitude of features extracted from transaction data. SVM operates by finding the optimal hyperplane that separates different classes in a high-dimensional feature space, maximizing the margin between them and thus enhancing its generalization capabilities. Within the context of smart cities, where the velocity and complexity of financial transactions necessitate advanced fraud detection mechanisms, SVM provides a reliable and efficient solution. By leveraging SVM's ability to discern intricate patterns and anomalies within transaction data, the framework can accurately identify fraudulent activities, including money laundering, illicit transactions, and ransomware payments.

Moreover, SVM enhances the robustness of the ensemble stacking model by contributing its unique decision boundaries, which complement those of other algorithms, thereby improving the overall accuracy and reliability of fraud detection. Its versatility and scalability make SVM well-suited for handling the diverse and dynamic nature of transaction data in smart city environments.

Furthermore, SVM's interpretability allows stakeholders, including financial regulators and law enforcement agencies, to gain insights into the underlying characteristics of fraudulent transactions, enabling more informed decision-making and targeted interventions to combat financial crimes effectively. Overall, SVM's inclusion in the proposed framework underscores its significance as a foundational component in the fight against fraud in Bitcoin transactions within smart cities, contributing to the creation of a safer and more secure digital ecosystem for all stakeholders involved.

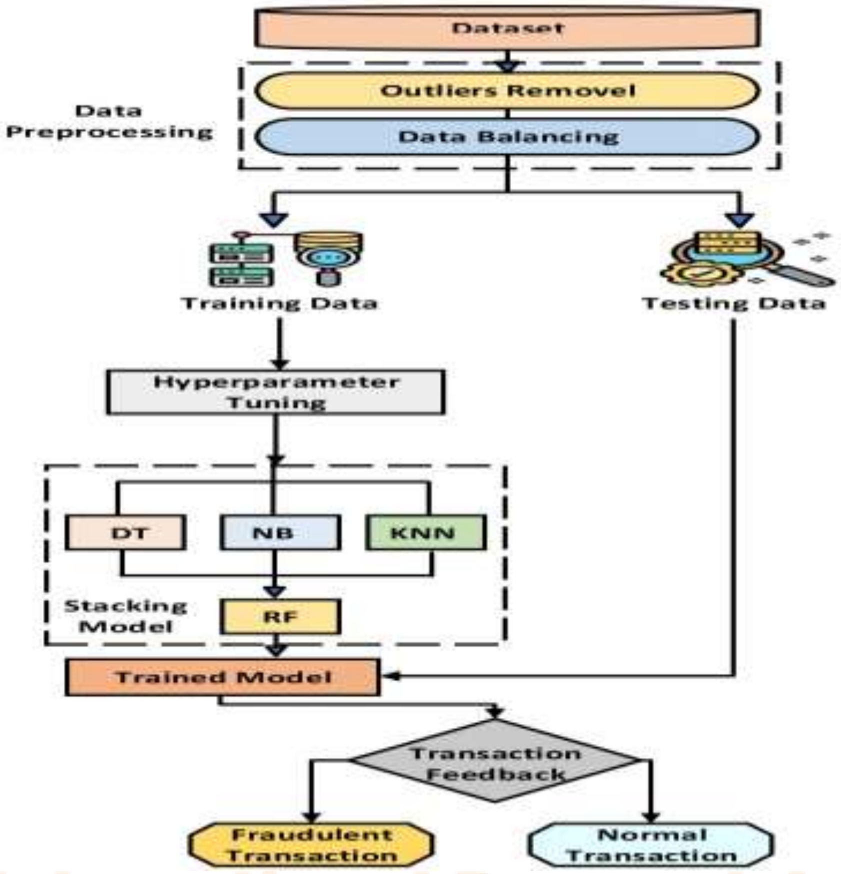
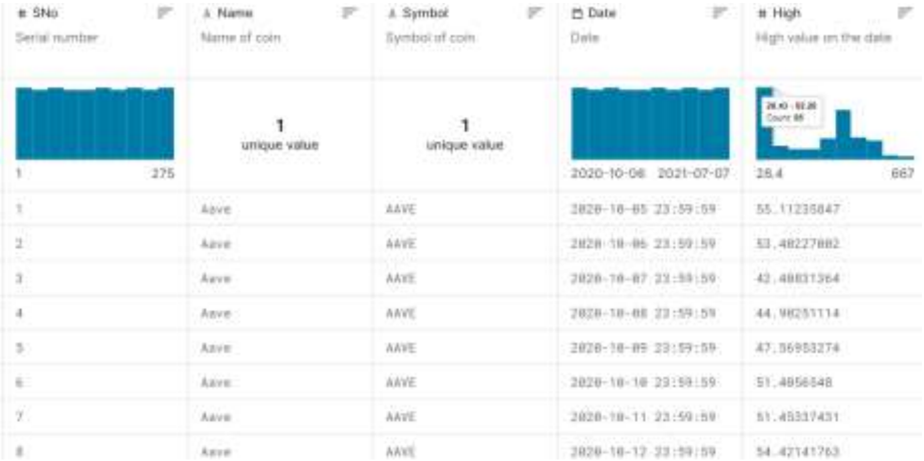


Fig 1: Architecture diagram for proposed model

4 Results and Discussion

There are 23,000 items in the dataset, and 11 different characteristics are included: Date, Close, Open, Market Cap, High, Volume and latitude. and the collection originated on Kaggle. For the purpose of detecting the fraud in Bitcoin, the study uses support vector machines (SVM) and linear regression (LR). SVM manages complex, high-dimensional data, whereas LR provides interpretability and simplicity. In spite of low R-squared values, additional discussion looks at feature importance, adherence to



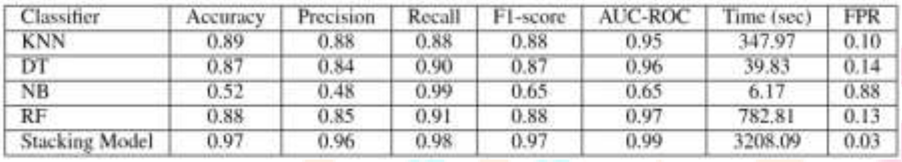
SNo	Name	Symbol	Date	High
Serial number	Name of coin	Symbol of coin	Date	High value on the date
1	Aave	AAVE	2020-10-05 23:59:59	55.11235847
2	Aave	AAVE	2020-10-06 23:59:59	53.40227082
3	Aave	AAVE	2020-10-07 23:59:59	42.40831364
4	Aave	AAVE	2020-10-08 23:59:59	44.90251114
5	Aave	AAVE	2020-10-09 23:59:59	47.56953274
6	Aave	AAVE	2020-10-10 23:59:59	51.4056548
7	Aave	AAVE	2020-10-11 23:59:59	51.45337431
8	Aave	AAVE	2020-10-12 23:59:59	54.42141763

Fig1: Dataset overview

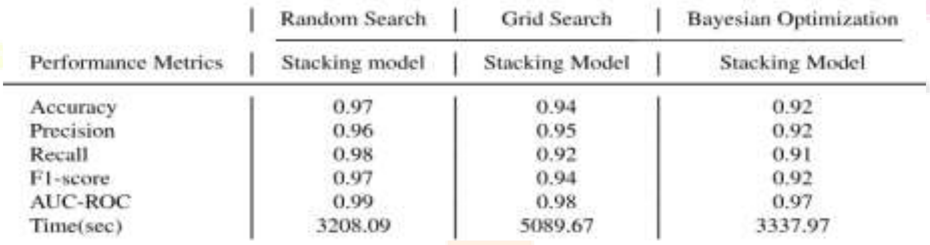
assumptions, parameter adjustment, comparison analysis, and model evaluation to learn more about the model's performance and possible improvements.

The average amounts of error between expected and actual values, as well as the efficacy of regression and classification models, are evaluated using the Root Mean Square Error (RMSE), which is defined in Table 1.

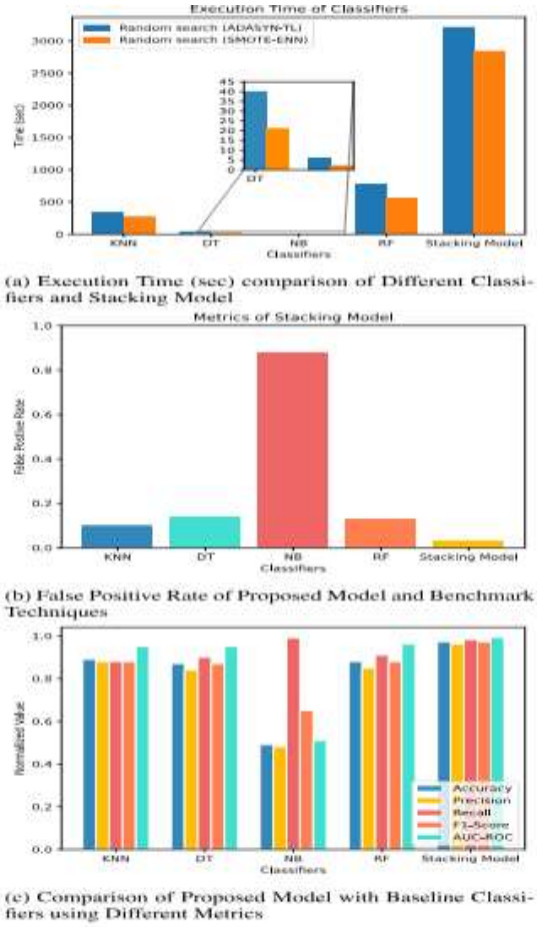
Table 1 Metric Value Table



Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC	Time (sec)	FPR
KNN	0.89	0.88	0.88	0.88	0.95	347.97	0.10
DT	0.87	0.84	0.90	0.87	0.96	39.83	0.14
NB	0.52	0.48	0.99	0.65	0.65	6.17	0.88
RF	0.88	0.85	0.91	0.88	0.97	782.81	0.13
Stacking Model	0.97	0.96	0.98	0.97	0.99	3208.09	0.03



	Random Search	Grid Search	Bayesian Optimization
Performance Metrics	Stacking model	Stacking Model	Stacking Model
Accuracy	0.97	0.94	0.92
Precision	0.96	0.95	0.92
Recall	0.98	0.92	0.91
F1-score	0.97	0.94	0.92
AUC-ROC	0.99	0.98	0.97
Time(sec)	3208.09	5089.67	3337.97



(a) Execution Time (sec) comparison of Different Classifiers and Stacking Model

(b) False Positive Rate of Proposed Model and Benchmark Techniques

(c) Comparison of Proposed Model with Baseline Classifiers using Different Metrics

Fig .4 Overall comparsion of various area type

The results of random search hyperparameter tuning and without hyperparameter tuning for four classifiers (RF, DT, NB, and KNN) and the stacking model using different balancing techniques.

Results show that using random search hyperparameters tuning, the overall performance of the classifiers can be improved. However, the time required to train the classifiers may increase. Comparison of the proposed model with baseline classifiers in terms of different performance metrics. hyperparameter tuning needs more computations. Both balancing strategies perform poorly for NB. For the stacking model, ADASYN-TL defeats SMOTE-ENN and the value of the F1-score of ADASYN-TL is 1 percent greater than SMOTE-ENN.

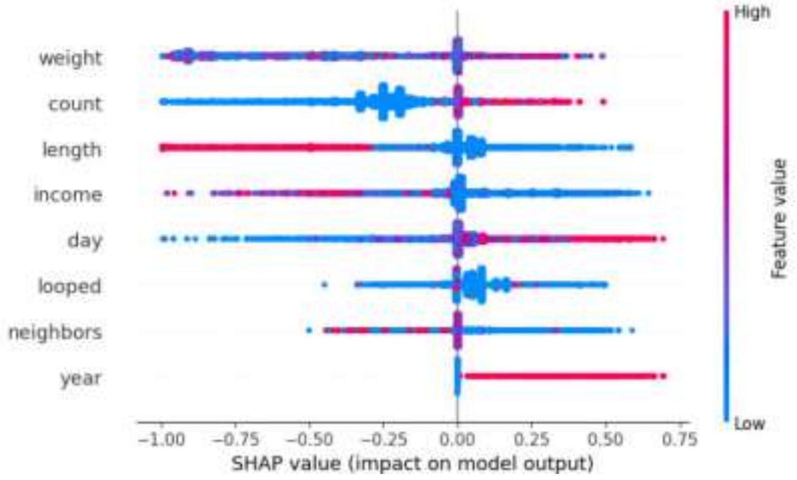


Fig.5 Summary plot of Bitcoin heist dataset.

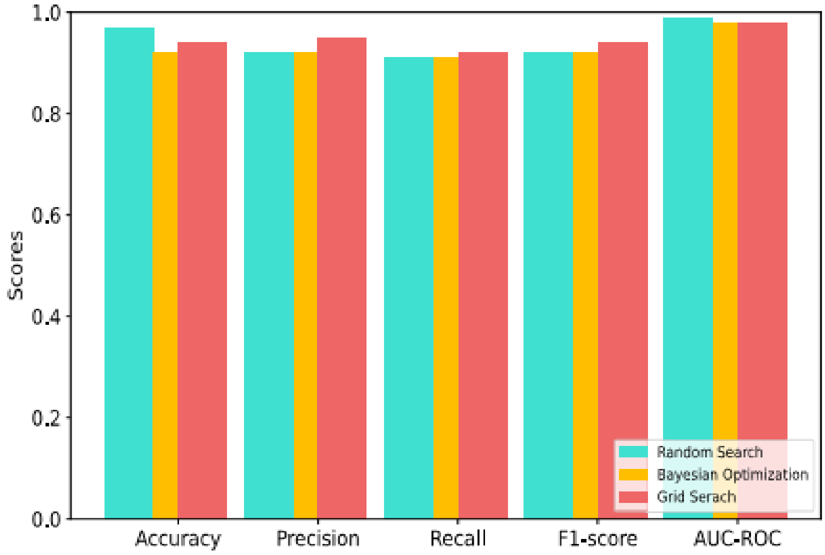


Fig .6 Performance of linear regression and

5 Conclusions & Future Work

In conclusion, the proposed framework presents a pioneering approach to enhancing fraud detection in Bitcoin transactions within smart cities. By leveraging ensemble stacking models and advanced machine learning techniques, it offers a robust solution to address the evolving challenges of financial crime. Future work will focus on refining the model's accuracy, scalability, and real-time processing capabilities to keep pace with the dynamic nature of digital transactions. Additionally, efforts will be directed towards integrating the framework into existing smart city infrastructures and collaborating with stakeholders to ensure widespread adoption and effective implementation, ultimately fostering a safer and more secure digital environment for smart city residents.

Future improvement for the framework entails continuous refinement and optimization of the ensemble stacking model by incorporating additional machine learning algorithms and exploring novel feature engineering techniques. Furthermore, research efforts will focus on enhancing the scalability and real-time processing capabilities of the framework to accommodate the evolving dynamics of smart city environments. Collaboration with industry partners and regulatory bodies will facilitate the integration of the framework into existing financial systems, enabling seamless deployment and adoption. Additionally, efforts will be directed towards enhancing interpretability and transparency in fraud detection outcomes, ensuring stakeholders have access to actionable insights for effective decision-making and policy formulation.

References

•M. U. Hassan, M. H. Rehmani, and J. Chen, “Anomaly Detection in Blockchain Networks: A Comprehensive Survey,” IEEE Commun. Surv. Tutorials, vol. 25, no. 1, pp. 289–318, 2022, doi:10.1109/COMST.2022.3205643.

• K. G. Al-Hashedi and P. Magalingam, “Financial fraud detec-tion applying data mining techniques: A comprehensive review-from 2009 to 2019,” Comput. Sci. Rev., vol. 40, 2021, doi:10.1016/j.cosrev.2021.100402.

•L. Pahuja and A. Kamal, “Enlfade: Ensemble Learning Based FakeAccount Detection on Ethereum Blockchain,” SSRN Electron. J.,2022, doi: 10.2139/ssrn.4180768.

• https://www.cylynx.io/blog/machine-learning-for-fraud-detection/[Accessed on 17-04-2023]

•J. Nicholls, A. Kuppa, and N.-A. Le-Khac, “SoK: The NextPhase of Identifying Illicit Activity in Bitcoin,” 2023 IEEEInt. Conf. Blockchain Cryptocurrency, pp. 1–10, 2023, doi:10.1109/ICBC56567.2023.10174963.

•N. Kumar, A. Hashmi, M. Gupta, and A. Kundu, “AutomaticDiagnosis of Covid-19 Related Pneumonia from CXR and CT-Scan Images,” Eng. Technol. Appl. Sci. Res., vol. 12, no. 1, pp.7993–7997, 2022, doi: 10.48084/etasr.4613.