From Black Box to Transparent Gate: Explainable AI for Certificate Integrity

1st Ibrahim Khormi
Faculty of Engineering and Information Technology
University of Technology Sydny
Sydny, Australia
Jazan University, Saudi Arabia
Jazan, Saudi Arabia
ikhormi@jazanu.edu.sa
ibrahimmohsen.khormi@student.uts.edu.au

2nd Priyadarsi Nanda
Faculty of Engineering and Information Technology
University of Technology Sydny
Sydny, Australia
priyadarsi.nanda@uts.edu.au

3rd Manoranjan Mohanty
Faculty in Information Systems
Carnegie Mellon University
Doha, Qatar
mmohanty@cmu.edu

***Abstract*— The persistent threat of certificate fraud to the education sector underscores the urgent need for a robust and transparent fraud detection system. AI and machine learning technique have proven effective in detecting fraudulent certificates but lack interpretability and transparency. High performing AI models often operate as 'black-boxes' making it difficult for users to trust, or refine their decisions. This paper proposes a XAI-based fraud detection framework with human-understandable interpretation. The proposed framework was empirically validated on certificate behavioral log dataset. Multiple machine learning models including KNN, logistic regression, decision tree and XGBoost were explored to detect fraud cases. Given the imbalanced nature of fraud datasets, SMOTE was used to ensure fairness. The XAI technique, Shapley Additive Explanations (SHAP) was employed to facilitate explainability, enhancing transparency. Experimental results demonstrate the efficacy of the proposed fraud detection framework and reduction in the number of false positives. SHAP technique quantify the contribution of behavioral features to model predictions. This will provide actionable insights for human review and institutional decision support. The performance of the proposed model shows that a robust ML-based mechanism for certificate fraud detection is practicable.**

Keywords*— *Explainable AI, Fraud Detection, SHAP, XGBoost, Interpretability

I. INTRODUCTION

The rapid advancement in digital certificate is becoming popular in the education ecosystem. However, this also represent a significant challenge, not only to secure data but also prevent certificate forgery. Cases of certificate fraud and the number of fake certificates in circulation can only be imagined. A recent study [1] shows that billions of people are affected by certificate fraud world-wide. From fabricated degree certificates to falsified diplomas, certificate fraud has grown into a multi-million-dollar business [2]. More worrisome, is the rising cases of institutional certificate fraud and insider threat [3]. In 2024, certificate forgery syndicates were uncovered in Malaysia [4]. They sold fake degree certificates for RM1,500 to RM4,000 using reputable university names. This scandal affected the reputation of those legitimate universities/colleges. The Justice Department in USA uncovered a shocking 7,600 certificate forgery scandal in South Florida nursing schools. This shows the consequences of certificate fraud to human life and public safety. In Hong Kong University Business, 30 postgraduate students were found to have fabricated foreign degree certificate. In Nigeria, a market for fraudulent certificates was uncovered by a journalist who posed as a buyer. He was able to get a degree certificate within six weeks. This led to the government of Nigeria placing a ban on degrees from 18 foreign universities. Certificate fraud in many ways undermines the credibility of the education system. It damages the reputation of individuals and institutions of learning that are involved. This erodes trust in the system, with substantial financial losses. As a consequence, institutions of learning are facing the challenge of identifying fraudulent activities in the certification process. However, to accurately model such behavioral profiles is not an easy task [5] . Fraudsters usually mimic the system to ensure that their profiles match the real process. To this end, stakeholders institute certificate verification processes to ensure fake certificates are detected.

Traditional verification methods have become inefficient in combating certificate fraud in today's digital landscape. However, the emergence of modern techniques such as AI and machine learning allows for the development of robust schemes. Such schemes are based on historical data. To this end, certificate fraud detection and prevention can be analyzed as a behavioral detection problem. This approach is expected to model the underlying relationship within the data and distinguish anomalies from legitimate activities.

Complex machine learning techniques such XGBoost have emerged as powerful tools for fraud detection. According literature, these methods produce high accuracy for detecting fraud [6]. Despite their high detection accuracy, these models face challenges and potential limitations that warrant further investigation. These challenges are:

(1) The 'black-box' nature of ML models limits their acceptability and deployment. They lack transparency and explainability of how and why decisions are made. They also do not show the importance of each feature to the model's outcome.

(2) Existing AI-based certificate fraud detection systems fail to effectively prevent certificate fraud. They serve as post-issuance verification mechanisms that only detect fake certificates that are already in circulation.

These limitations in the existing AI-based models are considered as a motivation for developing a new approach for enhanced certificate fraud detection. The proposed behavioral-based AI certificate fraud detection system will incorporate explainability. Explainability and transparency of

the model is needed to ensure trust and acceptability by stakeholders. For instance, if a certificate is flagged as fraudulent, the system is expected to provide a rationale for such a decision. In addition, transparency is a legal requirement of data regulatory agencies. For instance, European union citizens have the right to demand explanation of decisions by automated processes under the GDPR [7].

The proposed solution will serve as a pre-issuance mechanism to analyze behavioral profiles. This will essentially screen certificates before issuance. User behavior such as IP address, location, timestamps can form patterns. It is expected that any deviation from the normal user behavioral pattern can be indicative of fraud. The influence of each user behavior to the final model's decision is very important in understanding the model. Therefore, the research question considered in the paper is - *Can behavioral indicators contribute most to fraud detection, and how can explainable AI tools such as SHAP be used to quantify their influence on the model's decisions?*

To answer this question, we seek to leverage explainable AI and behavioral analytics for interpretable AI-based behavioral fraud detection system. To this end, the contribution of this work is in three-fold.

(1) To propose an enhanced machine learning based certificate fraud detection mechanism.

(2) To incorporate explainable AI tools, including SHAP to facilitate transparency in decisions making in traditional black-box ML models.

(3) To explore user behavioral profile for a certificate pre-issuance mechanism to prevent circulation of fake certificates.

The remainder of this paper is organized as follows. Section 2 reviews the related work on certificate fraud detection with respect to methods, and the use of explainable AI. Section 3 outlines the proposed methodology for certificate fraud detection. Section 4 provides the experimental setup and the metrics for evaluating the proposed technique. Section 5 presents the results of the evaluation and discusses the significance and implication of the proposed methodology. Section 5 gives the conclusion, providing directions for further research.

II. Related works

A. AI in Fraud Detection

AI have been applied in fraud detection across various domains including finance, healthcare, e-commerce and education. This section will provide a review of related works on AI fraud detection.

Authors in [8] conducted an analysis of different techniques for fraud detection with emphasis on mobile payment platform. They concluded that machine learning based fraud detection outperformed traditional techniques. The leverage device fingerprinting, user behavioral analytics and anomaly detection to develop a multilayer methodology. The result indicates ML mitigates false positives and significantly improves detection accuracy. In [9], they authors proposed a solution for insider threat prevention using machine learning to detect subtle behavioral patterns and deviations to identify fraudulent employee. The work reported in [10] presents a novel idea in fraud prevention by exploring user behavioral patterns to detect anomalies in online banking. Autoencoder machine was employed to analyze behavioral data. Their results demonstrate the efficacy of behavioral models in detecting fraudulent actions.

B. Certificate Verification

Digital certificate verification and fraud detection represent important areas where blockchain and AI have shown tremendous successes. Legacy techniques have been overshadowed by blockchain and AI techniques due to their immutability and dynamic learning ability respectively. This section explores existing studies on the application of blockchain and AI techniques in the context of certificate forgery, highlighting their strengths and limitations.

A blockchain-based platform for verification of digital certificate has been proposed by [11]. First, all certificates are uploaded onto the blockchain with unique identifiers. The authors leverage smart contracts to automate the verification process. To ensure authenticity of the uploaded certificates, smart contracts execute predefined rules and conditions to check for signs of fraud. Third parties can easily verify certificates on the blockchain using the unique identifier. However, this system does not consider the legitimacy of certificates.

In [12], the authors proposed a robust solution to the overarching issue of academic certificate fraud. They integrate AI in blockchain system to validate the digital certificates that are stored on the blockchain.

Blockchain technology has been used in [13] to automate the certificate verification in the hiring process. E-certificates of graduating students are uploaded to blockchain and a unique digital transcript is generated for each student. Potential employers can validate a candidate's certificate by comparing the submitted transcript with the one stored in the blockchain. By this, the authenticity of the certificate can be easily verified. However, this method does not justify the legitimacy of the certificate.

The authors in [14] combined blockchain and quick response code (QR) to facilitate certificate verification. Each candidate has a QR code that is linked to their records on the blockchain. This enables quick and accurate certificate verification. However, the authors noted concerns over security of QR code. They suggested encryption technique to prevent manipulation and redirection.

In [15], the authors proposed a blockchain-based cross-institutional certificate verification architecture using Hyperledger Fabric. They emphasize decentralization of certificate storage. This prevent unauthorize access and data tampering. However, they noted that implementation can be costly and requires extensive stakeholder engagement and coordination.

An integrated platform for certificate verification and fraud detection using blockchain and AI has been proposed by [16]. While blockchain enables verification of certificate authenticity, AI analyses the verification patterns to detect anomalies. This improves certificate fraud prevention. This shows that the addition of AI layer in blockchain can enhance fraud detection and prevention. However, the authors noted additional computational cost and data privacy concerns. Table I gives an overview including limitations of key studies that have considered technology for certificate verification.

TABLE I. Comparative overview of key studies in certificate verification

Study	Description	Technique	Benefit	Limitations
Arvindan T. P., et al., 2024 [11]	Blockchain platform for certificate Verification	Smart contract Predefined rules &conditions Unique identifiers	Automated verification system Immutability of certificate	No use of explainable AI (XAI) Certificate Legitimacy not guaranteed
Rai et al., 2024 [12]	Integration of AI with blockchain for certificate Verification	Machine learning & Smart contract	AI helps detect fraud within the block	Lack of explainability of the result
Shwetha A. N., et al., 2024 [13]	Use blockchain for certificate verification in hiring process	Blockchain technology	Automated verification system Authenticity of certificate	Certificate Legitimacy not guaranteed No integration of XAI
Abdullahi M. U., et al, 2022 [14]	Integrating blockchain with QR code for certificate verification	Blockchain QR code	Ease of use Immutability of certificate Accurate verification	QR code is vulnerable No use of XAI
Teja M. V., et al, 2024 [15]	Blockchain-based cross-institutional certificate verification	Hyperledger Fabric	Decentralized storage Immutability of certificate	No use of XAI High cost of implementation
Kasukurthi N., et al., 2024 [16]	Blockchain and AI integrated platform for certificate verification	AI technique Blockchain Technology	Added AI layer to authenticate the verification process	Lack of explainability of the result Computational cost privacy concerns

C. User Behavioral-Based Fraud Detection

Digital online platforms usually keep history of logins and transactions perform by users for as long as the user is active. This generates a wealth of insightful user behavioral data. These data can be leveraged for significant analysis. User behavior has been employed in various domains for the purposes of fraud detection and prevention. In e-commerce, user behavior may include spending patterns, customer profile such as delivery address, frequently used IP address, and mode of ordering. User behavior data have been explored for fraud detection in e-commerce [17]. In finance industry (banking or credit card), user behavioral data have been used to detect fraudulent activities [18]. These include transaction history (amount, time and location), user actions (mouse or keystroke), device and login information. Behavioral data have been used for fraud detection in mobile and communication networks [19]. The behavioral patterns here include OS type and version, web browser, language and time zones, IP address and user identifier. In summary, Table II shows user behavioral attributes used in different domains.

.

TABLE II. User Behavioral Attributes in Other Domains

S/No	Domain	User behavioral attributes
1	E-commerce	a. frequently used IP address
		b. delivery address
		c. mode of ordering (device type)
		d. transaction amount
2	Finance	a. disparity in withdrawal amount
		b. time of transaction
		c. user actions
		d. mouse and keystroke movement
3	Communication networks	a. OS and web browser type and version
		b. language and time zone
		c. IP address
		d. user identifier

D. Research Gap

Despite the high performance of machine learning models for certificate fraud detection, there exist a notable research gap in the area of interpretability and data. The work reported in this paper aims to fill the following gaps.

1) Interpretability and transparency: To address the opacity and black-box nature of traditional ML models, this work employ SHAP-based XAI technique to explain model's decision. While ML models haven proven effective, their internal mechanics can be opaque, hindering explainability and trust. The proposed framework prioritizes model interpretability, allowing users to understand how and why specific predictions are made.
2) Behavioral Data: Existing studies focus on specific data types such as images of certificates, or signatures. This work proposes a novel framework that integrates different attributes including user behavior. This comprehensive approach allows for a more holistic understanding of fraudulent activities and fraud detection.

E. Explainable AI

Artificial intelligence (AI) and Machine Learning (ML) state-of-the-art models have achieved high performance in different applications. However, their opacity is a concern for critical areas such as fraud detection and prevention. They often operate as black-boxes, lacking transparency and explainability [20]. Even developers do not understand why some decisions are made by the models. This affects the acceptability and deployment of traditional machine learning models. As a result, explainable AI (XAI), a tool for providing interpretability to ML models has become important. This work seeks to leverage the power of XAI to provide transparency and integrity to AI-based certificate fraud detection. Existing studies in the application of XAI in fraud detection will be discussed in the next section.

F. Explainable AI in Fraud Detection

Explainable AI has been applied in the banking sector to provide transparent and accountable models for fraud detection [21]. Realistic datasets were used to evaluate XAI techniques (SHAP, LIME and PDP) in hybrid ensemble models. The experimental results indicate that the XAI-based

framework achieved 99% detection accuracy while providing interpretable predictions.

Authors in [22] developed a credit card fraud detection system that is based on explainable AI techniques. SHAP and LIME XAI techniques were used to provide explanation to the outcome and decision of the system. For comprehensive analysis, the model was evaluated using diverse performance metrics including recall, precision, F1-score and AUC-ROC. The empirical results show that the system improved the transparency and trust of ML model while maintaining the high accuracy.

In [23], the author proposed a novel framework for fraud detection in the financial sector. The framework combines XAI and Federated Learning (FL) to facilitate fraud detection. They proposed to integrate SHAP technique to ensure accurate and interpretable predictions. Providing understanding on the level of contribution and influence of each feature on model's outcome and justification for decisions made. This essentially enhance transparency and trust in FL-based fraud detection systems.

In another work by [24], the problem of financial fraud in mobile money transactions was addressed by integrating ensemble learning and XAI. They authors used SHAP technique with a focus to balance the trade-off between accuracy and interpretability of models.

In related research, [25] proposed an explainable federated learning (XFL) framework for financial fraud detection. With SHAP and LIME, the framework can improve and interpret fraud predictions while maintaining accuracy, data privacy and compliance. Their work presents a robust AI-based solution to the problem of fraud in the banking sector.

III. THE PROPOSED FRAMEWORK

Certificate forgery is a major issue that has impacted individuals, education and the society at large. Legacy and traditional verification techniques fail to combat behavioral fraud. This work proposes to use behavioral attributes to develop an explainable ML framework. Integrating ML with SHAP-based XAI technique provides a robust transparent and accurate certificate fraud detection system. The framework is as shown in Figure 1.

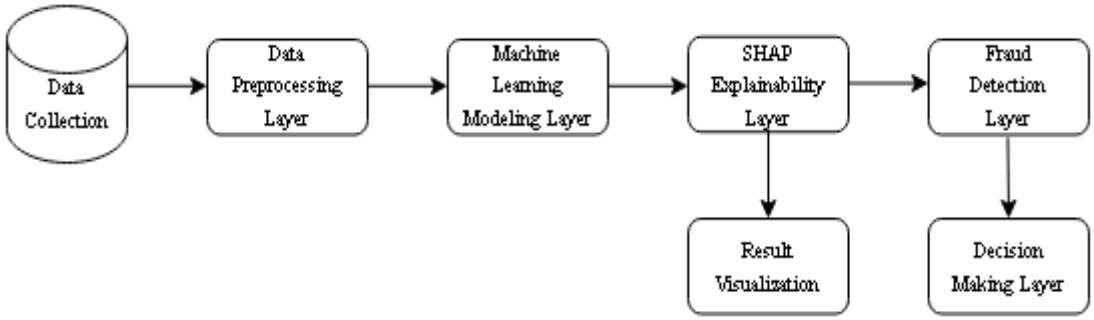

Fig. 1. The Proposed Framework

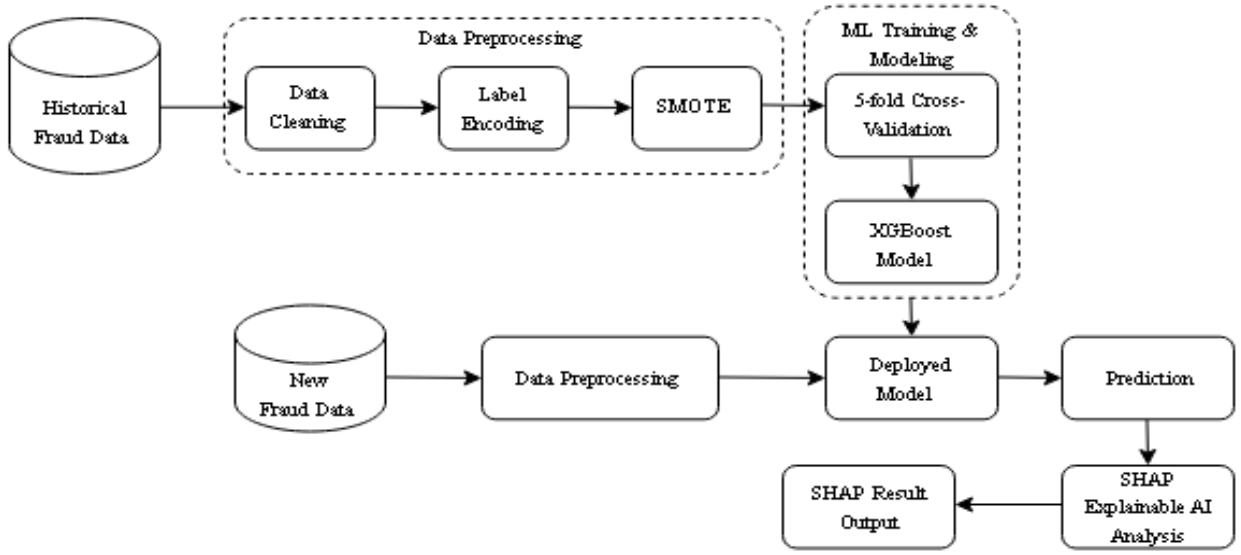

Fig. 2. Workflow for the Proposed Framework

A. Architecture

The architecture consists of a multilayer architecture. This comprises a data layer, ML model layer, fraud detection layer, and explainability layer.

Data Layer: The first step in the development of the proposed framework is the identification and collection of relevant data.

AI Model Layer: This layer employs ML classification algorithms for model development. These algorithms include KNN, logistic regression, decision tree, support vector regression and XGBoost.

Explainability Layer: Here, the SHAP method of explainable AI is adopted to provide explanation and interpretation to the model's decisions. This layer is important as it brings insight and quantifies the contribution of each behavioral feature to detecting certificate fraud.

B. Workflow for the proposed framework

Figure 2 shows the complete workflow of the proposed certificate fraud detection framework. The process starts from data collection, preprocessing to model development and evaluation. The final phase is on SHAP-XAI interpretability of the model's predictions. This ensures transparency and trust.

1) Data Description

The unavailability of public real-world certificate fraud dataset poses a serious challenge in undertaking this study. To overcome this constraint, we simulated certificate fraud data from an "Enterprise Access Log" dataset. The enterprise access log data is a publicly available dataset hosted on GitHub by Manas-stack13 [26] [27]. Though the enterprise dataset was not designed for this purpose, it comes with characteristic user behavioral activities that are typical of certificate regimes. These activities include access times, source IP address, device type, and geolocation. The original enterprise dataset contains 5000 records, with 8 attributes and two (2) target classes. For the purpose of this work, a modified version of the dataset was created.

TABLE III. Key features

S/No	Feature name	Description
1	IP address	Access origin of the activity
2	Location	Geographic origin from where activity was performed
3	Device	Type of device used (desktop, mobile, tablet)
4	Off-hour	Unusual activity or event times such as weekends

2) Modification/Simulation

The procedure for modifying the original enterprise data is as follows.

a) The columns of user behavioral attributes that are not relevant to certificate fraud problem were dropped.

b) Drop all records that were tagged malicious in the original dataset.

c) Create certificate fraud scenarios and inject them into the modified data

d) The entire data was then labeled. The normal records were labeled as 0 while the abnormal (fraud) cases were labeled as 1.

This modified dataset will be benchmarked for evaluating machine learning based certificate fraud detection models. The redefined features are listed in Table III. The use of such simulated data has limitations which include lack of generalization. Despite this, it provides data control, quality and consistency.

3) Data Preprocessing

The first step is to transform the raw data into a ML understandable form. This transformation or preprocessing can be achieved by performing some important operations. This includes feature selection and engineering, data encoding, feature reduction, fraud labeling.

4) Handling Class Imbalance

The class distribution of the dataset as shown in Fig. 3 is skewed towards normal or legitimate activity with only less than 1% labeled as suspicious. This class imbalance represents the real-world fraud problem where fraudulent cases are usually rare compared to legitimate ones. However, class imbalance will affect the model's results, creating biases and overfitting. SMOTE oversampling technique [28] is used to create class balance. Here, the data of the minority class is oversampled by injecting synthetic in proportion to the majority class. SMOTE technique provides better generalization and prevents overfitting.

5) Model Training and Validation

For machine learning modeling, the preprocessed data is usually split into training and testing datasets. The training dataset is used to develop the machine learning-based fraud detection model. The training involves an iterative procedure for optimization of model's parameter. These parameters continue to be adjusted until an optimal set of parameters that meet the required accuracy is achieved. The trained model is then stored. The testing dataset is used to validate the trained model, which enables the evaluation of fraud detection performance. This ensures that the model identifies legitimate and suspicious activities correctly, thus minimizing false positives and negatives. In order to develop a robust model, cross validation technique is employed in this work to build ML model.

6) Cross Validation

Cross validation [29] is a technique that is used to train and validate ML models in a rigorous manner. This ensures the generalizability of the model, and mitigate overfitting. Here, the dataset is randomly divided into disjointed k sub-groups, otherwise known as k-fold. The ML model is trained on k-1 set of the data and validated on the fold that is not part of k-1 group. This process continues, with a different set of training data and validation dataset until every fold is used in training and validation. The results of all the folds were then averaged across the k-folds. In this paper, a 5-fold cross validation was used. In this way, a stable ML model is generated.

7) SHAP-Explainable AI

SHAP [30], an explainability technique used in XAI is employed in this work to interpret the output of machine

learning model. SHAP is intuitive and simple to compute. It uses the principles of game theory to allocate scores to features according to their importance in the model prediction. SHAP is model-agnostic and can provide both local and global explanations to ML predictions.

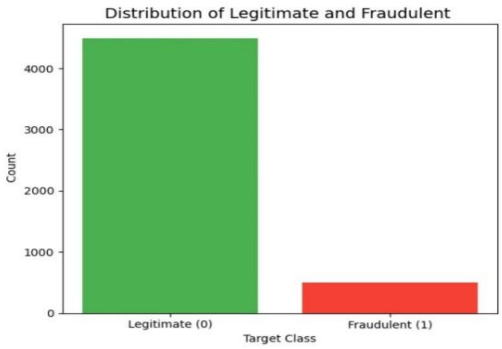

Fig. 3. Data Class Distribution

IV. Evaluation of the Proposed Methodology

In this section, the performance of the proposed approach will be analyzed with regards to detecting fraudulent activities. The detail results of the considered models will be based on various metrics that will be discussed in the next section.

A. Perfomance Evaluation Matrics

To evaluate the effectiveness of ML classifiers, the following metrics are used to measure different aspects of a model's performance.

Accuracy: This measures how well the model correctly classifies both fraudulent cases and non-fraudulent cases.

Precision: This measure how much of identified fraudulent cases are actually fraud. Low precision value means the model has high false positive rate.

Recall: This measures the ability of the model in identifying all true fraud cases. A high value of recall means only a few cases of fraud are misclassified.

F1-Score: This provides a balance in the overall performance of a model by combining precision and recall. A low F1 means there is a trade-off between false negatives and false positives.

Receiver Operating Characteristic-Area Under Curve (ROC-AUC): This measures the model's ability to distinguish legitimate cases from fraudulent cases.

Confusion Matrix: This provides a pictorial summary of all the predictions made, both correct and incorrect predictions by class (legitimate or fraudulent).

B. Experimental Results and Discussion

The performance of the developed models will be evaluated based on academic certificate behavioral dataset. The process of digital certification is expected to involve activities that if properly harness can be explored for fraud analysis. Such activities could be termed user behavior. The processed data is used as input to ML algorithms for the purposes of fraud detection.

1) Performance of the developed models

The test dataset was used to evaluate each developed ML model. The models include k-nearest neighbor (KNN), decision trees, logistic regression, and XGBoost algorithms. Table IV gives the comparative performance analysis of the models based on the evaluation metrics. The aim is to achieve high accuracy while minimizing false positives.

Figure 4 shows the confusion matrix of the XGBoost model's predictions on the test dataset. This provides a detailed breakdown of the classification results as follows:

There are 867 True negatives, indicating the number of legitimate cases that are correctly identified as non-fraudulent.

The number of false positives is 30, which represent legitimate cases that are wrongly flagged as suspicious.

The model failed to detect only 7 fraudulent cases (False negatives). These were incorrectly classified as non-fraudulent.

Finally, the confusion matrix shows 96 true positives. This means 96 of the fraudulent cases were correctly detected by the XGBoost model.

The extremely low false positive rate (30 out of 897 legitimate cases) and low false negative rate (7 out of 103 fraud cases) illustrate the model's strong performance in both detecting fraud and avoiding false alarms.

TABLE IV. Performance Comparison of ML Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
KNN	88.5	93.6	88.5	90.0	93.5
Logistic Regression	85.7	92.9	85.7	87.8	94.3
Decision Tree	91.3	94.4	91.3	92.2	95.4
XGBoost	96.3	96.8	96.3	96.5	96.9

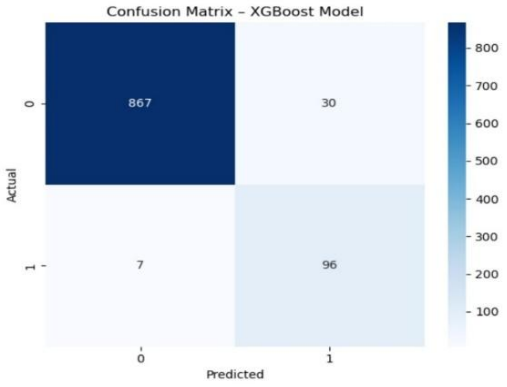

Fig. 4. Confusion matrix for XGBoost

Secondary evaluation metrics such as recall, precision, and F1-score can be derived from the confusion matrix. These metrics further demonstrates the model's ability to accurately identify fraudulent cases with minimal false alarms.

- An accuracy of 96.3% means that 96.3 times out of 100 predictions, XGBoost model classified fraudulent and legitimate cases correctly.
- A precision of 96.8% shows that the XGBoost model was able to identify 96.8 actual fraudulent cases among 100 predicted as fraudulent.
- A recall of 96.3% shows the XGBoost model's ability to successfully detect fraudulent cases.

- A F1-Score of 96.5% shows the reliability of the model.
- An ROC-AUC of 96.9% means that the model is very good at discriminating fraudulent cases from legitimate cases.

Overall, XGBoost model has proven effective in discriminating fraudulent cases from legitimate cases in the face of large class imbalance. The model also demonstrates how robust it is by the consistently high performance across all the various scenarios created by the 5-fold cross-validation.

2) *SHAP-XGBoost Analysis*

The evaluation of the explainability of the best performance model - XGBoost was based on the SHAP technique. SHAP uses the principles of cooperative game theory to compute and assign SHAP values to each feature. SHAP value represents the contribution of a feature to the model's outcomes. It provides global interpretation of feature impact to model's output and local explanation to individual predictions.

SHAP Global Feature Impact

Feature importance is a key element in the explainability of ML models. This shows the impact of each feature to the overall outcome or performance of a model.

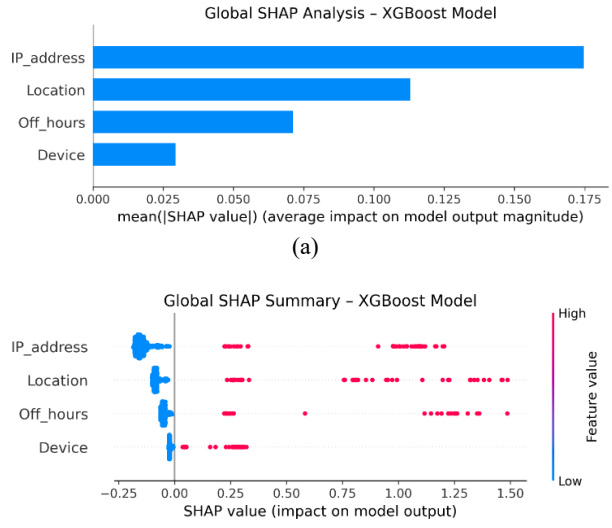

Fig. 5. SHAP global description of feature contribution in XGBoost fraud detection model

Here, the SHAP values of each feature are shuffled to see how much the performance of the model degrade. Fig. 5 shows the global SHAP plot of the feature importance across the whole dataset. The feature with the longer bar produces greater drop in performance when shuffled, which implies its importance. This means that IP address contributes more in detecting fraud across all cases in the dataset. And in the order of length of the bars follows other top influential features. The SHAP summary plot in Fig. 5b illustrates the impact of features on the global model's classifications. The SHAP values on the horizontal axis represent each features contribution to fraud or legitimate case prediction. The vertical axis ranks the features by importance. IP_address and location are the most influential factors in detecting fraudulent cases. The global SHAP analysis enhances model transparency, and helps identify the most important fraud indicators.

SHAP Local Prediction Explanation

Local SHAP plots provide explanations to individual predictions. Figure 6 describes the local SHAP prediction for a single record (4581). This record was flagged as suspicious without an understanding as to why it was so. However, with SHAP, it is known that off_hours with the highest positive value contributes the most to its classification. This kind of insight is vital in terms of review or further investigations or scrutiny.

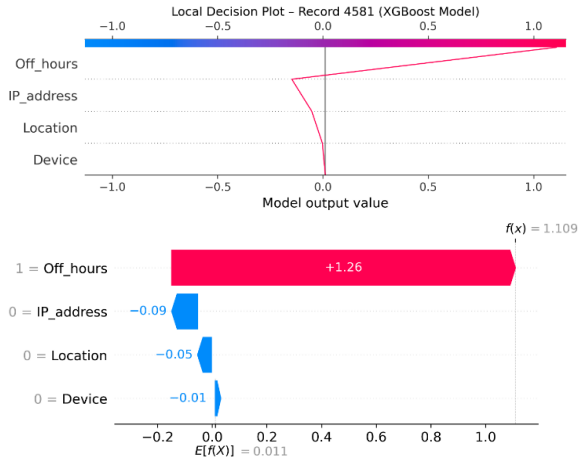

Fig. 6. SHAP local description of feature contribution in XGBoost fraud detection model for single case (Record 4581)

V. CONCLUSIONS

This work addressed the overarching issue of lack of model interpretability in machine learning-based certificate fraud detection systems. A novel framework that integrates explainable AI with ML to enhance fraud detection has been developed. SHAP technique was incorporated into XGBoost to facilitate model interpretability while minimizing false positive rate. The developed SHAP-XGBoost model performed well with high accuracy and recall of 96.3% and 96.8% respectively. Through the use of SHAP analysis, the contribution of individual fraud indicator features to model's outcome is quantified. This is essential for reviews and further scrutiny of cases that have been flagged as suspicious. This enhanced trust in certificate fraud detection decision-making. Educational institutions can integrate this system in their digital certification pipeline just before certificates are issued or registered on blockchain. This ensures certificate screening and avoid circulation of fraudulent certificates.

A. *Limitations*

A major limitation of this work is the lack of access to real-world certificate fraud data. There are no known publicly available certificate fraud data.

B. *Future Works*

Future plans include validation of the developed framework on real-world academic certificate data in collaboration with higher education institutions.

VI. References

[1] Eaton S. E., and Carmichael J., "Fake Degrees and Credential Fraud, Contract Cheating, and paper Mills: Overview and Historical Perspectives," in *Fake Degrees and Fradulent credentials in Higher Education*, 2023, pp. 1-22.

[2] Sardar L., "Fake me id you can: Unforgeable Digi-Physical Academic Certificates with Instant Verifiability," *IEEE Access,* vol. 1, no. 1, pp. 1-20, 2025.

[3] Tariq A., Binte Haq H., & Taha Ali S., "Cerberus: A Blockchain-Based Accreditation and Degree Verification System," *arXiv,* pp. 1-14, 2019.

[4] Arifin R. B. et al., "Tackling Counterfeit Certificate Problems with Blockchain Technology: A Review," *International Journal of Advanced Technology and Engineering Exploration,* vol. 11, no. 19, 2024.

[5] Dornadula V, & Geetha S., "Credit Card Fraud Detection using Machine Learning Algorithms," *Procedia Computer Science,* vol. 165, pp. 631-641, 2019.

[6] Hajek P., Abedin M. Z., & Sivarajah U., "Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework," *Information Systems Frontiers,* pp. 1-19, 2022.

[7] Wall L., "Some Financial Regulatory Implications of Artificial Intelligence," *Journal of Economics and Business,* vol. 100, pp. 55-63, 2018.

[8] Choi D., & Lee K., "Machine Learning Based Approach to Financial Fraud Detection Process in Mobile Payment," *IT CoNvergence PRActice (INPRA),* vol. 5, no. 4, pp. 12-24, 2017.

[9] Koutsouvelis V., Shiaeles S., Ghita B., & Bendiab G., "Detection of Insider Threats using Artificial intelligence and Visualization," in *6th IEEE Conference on Networks Softwarization (NetSoft)*, 2020.

[10] AbouGrad H., & Sankuru L., "Online Banking Fraud Detection Model: Decentralized Machine Learning Frmaework to Enhnace Effectiveness and Compliance with Data Privacy Regulations," *Mathematics,* vol. 13, no. 3, p. 2110, 2025.

[11] Arvindan T. P., Balaji M. et al., "Digital Certificate Fraud Detection Using Blockchain Technology," *International Journal of Engineering Research & Technology,* vol. 12, no. 2, 2024.

[12] Rai B. K., Bilal A., Priyadarshi A., Bej A. K., and Vivek S. H., "VaLiDiFy - Certificate Validation using Blockchain and AI," in *International Conference on ICT in Business Industry & Government*, 2024.

[13] Shwetha A. N., Ashwini B. P. et al., "An Automated Certificate Validation System Using Blockchain Technology for the Hiring Process," *International Journal of Engineering Trends and Technology,* vol. 72, no. 8, pp. 112-127, 2024.

[14] Abdullahi M. U., Aimufua G. I. O. & Muhammad A. A., "Certificate Generation and Verification System Using Blockchain Technology and Quick Response Code," *IOSR Journal of Computer Engineering,* vol. 24, no. 1, pp. 37-47, 2022.

[15] Teja M. V., Raju M. S. et al., "E-certificate Validation Using Blockchain," *International Journal of Advance Research and Innovative Ideas in Education,* vol. 10, no. 2, pp. 2014-2022, 2024.

[16] Kasukurthi N., & Kancharla G. R., "Integrating Machine Learning and Blockchain for Fraud Prevention in Education Records," *Nanotechnology Perceptions,* vol. 20, no. 15, pp. 1876-1888, 2024.

[17] Zhang Z., Yin H., & Rao S. X., et al., "Identifying E-commerce Fraud Through User Behavior Data: Observations and Insights," *Data Science and Engineering,* vol. 10, no. 1, pp. 24-39, 2025.

[18] Baabdullah T., Rawat D. B., Liu C., et al., "Analysis of Cardholder Spending Behavior and Transaction Authentication to Enhance Credit Card Fraud Detection," in *International Conference on Machine Learning and Applications (ICMLA)*, Jacksonville, FL, USA, 2023.

[19] Alejandro G. M., Alberto F., Isaac M. D., et al., "A Survey for User Behavior Analysis Based on Machine Learning Techniques: Current Models and Applications," *Applied Intelligence,* pp. 1-43.

[20] Chaquet-Ulldemolins J., "On the Black-Box Challenge for Fraud Detection Using machine Learning: Linear Models and Informative Feature Selection," *Applied Sciences,* vol. 12, no. 7, p. 3328, 2022.

[21] Amirineni S., "Explainable Artificial Intelligence (XAI) Models for Transparent and Accountable Fraud Detection in Banking Ecosystems," *International Journal of Scientific Research and Management ,* vol. 13, no. 8, pp. 2493-1, 2025.

[22] Suriya S., & Sireesha R. M., "Credit Card Fraud Detection using Explainable AI Methods," *Journal of Information Systems Engineering and Management,* vol. 10, no. 24, pp. 1-14, 2025.

[23] Awosika T., Shukla R. M., & Pranggono B., "Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection," arXiv Preprint, 2023.

[24] Vijayanand D., & Smrithy G. S., "Explainable AI - Enhanced Ensemble Learning for Financial Fraud Detection in Mobile Money Transactions," *Intelligent Decision Technologies,* vol. 19, no. 1, pp. 52-67, 2024.

[25] Aljunaid S. K., Almheiri S. J., Dawood S., & Khan M. A., "Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection," *Journal of Risk Financial Management,* vol. 18, no. 4, p. 179, 2025.

[26] Udayakumar S. K., Ragothaman H., & Khare K. M., "A Novel Dataset and a Hybrid Ensemble Approach for Anomaly Detection in Enterprise-Access-Logs," in *International Conference on Data Science, Agents, and Artificial Intelligence*, Chennai, India, 2025.

[27] Manas-stack13, "Access-Log-Anomaly-Detection-Dataset," 20 May 2025. [Online]. Available: https://github.com/Manas-stack13/Access-Log-Anomaly-Detection-Dataset.

[28] Obimbo C., Mand D., & Singh S., "Oversampling Techniques in Machine Learning Detection of Credit Card Fraud," *Journal of Internet Technology and Secured Transactions (JITST),* vol. 9, no. 1, pp. 741-746, 2021.

[29] Allgaier J., & Pryss R., "Cross-Validation Visualized: A Narrative Guide to Advanced Methods," *Machine Learning Knowledge Ex,* vol. 6, no. 2, pp. 1378-1388, 2024.

[30] Salih A. M., Raisi-Estabragh Z.,Galazzo I. B., et al., "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems,* vol. 7, pp. 1-8, 2025.