

A BLOCKCHAIN IMMUTABILITY AND MACHINE LEARNING CERTIFICATE: FRAUD DETECTION FOR MITIGATING CREDENTIAL FRAUD

¹Aminu Baffa Ibrahim, ¹Ibrahim Bashir Faruk, ²Eli Adama Jiya

¹Department of Computer Science, Federal University Dutsin-Ma, Katsina, Nigeria

²Department of Computer Science, Federal University of Applied Sciences, Kachia, Kaduna State, Nigeria

*Corresponding Author Email Address: ibrahimbashirf@gmail.com

ABSTRACT

Academic credential fraud is a critical systemic challenge in Nigerian higher education because conventional verification depends on the continued operation, record integrity, and responsiveness of issuing institutions. This study addresses the resulting research gap by developing a verification infrastructure that remains functional independently of institutional continuity. A blockchain-based academic certificate verification system was designed and implemented using two Solidity 0.8.20 smart contracts, InstitutionRegistry and CertificateRegistry, deployed on the Sepolia Ethereum testnet. Certificate documents were stored on the InterPlanetary File System (IPFS) through Pinata, while SHA-256 hashes and IPFS Content Identifiers were permanently anchored on-chain during issuance. A React 18 application provided role-specific interfaces for institutions, graduates, and third-party verifiers. Security evaluation demonstrated resistance across five attack categories, while 200-trial benchmarking recorded an average verification time of 87 ms at zero verifier cost. Polygon Layer 2 deployment reduced certificate issuance costs to approximately \$0.018. A complementary machine-learning fraud-detection component used 1,000 structured certificate records, each containing 16 domain-specific features. Logistic Regression, Random Forest, and XGBoost were optimized using GridSearchCV with five-fold cross-validation and evaluated on held-out data. XGBoost achieved the highest accuracy (93.5%), followed by Random Forest (91.5%) and Logistic Regression (75.5%). Verification complexity, previous rejections, and days to verify were the most predictive features. The complete system and analytical pipeline are packaged in a reproducible Google Colab notebook.

Keywords: Blockchain verification; Smart contracts; IPFS hashing; Machine learning; Credential fraud; Security benchmarking

INTRODUCTION

The problem of academic credential fraud has been a long-standing issue in Nigeria, and the effects are anything but trivial. Fake degree certificates have been discovered in the civil service, hospitals, engineering companies, and even university lecture theaters. Research has documented the extent of the problem, with the West African Examinations Council reporting tens of thousands of malpractice cases in 2019 alone (WAEC, 2019). Apart from the moral implications, the problem of credential fraud has created a distorted labor market that has affected the very people it should be helping the most: the graduates.

What makes this problem difficult to solve is the nature of the systems meant to solve it. In most Nigerian universities, verifying a

certificate involves contacting the university where the certificate was earned, checking a physical log or reviewing a paper document with an embossed seal and a hologram sticker. These are all too easy to forge with modern printing technology (Alammary *et al.*, 2024). It can take weeks to verify a certificate. Moreover, if the university itself does not have its records, which is common in Nigerian universities because of fires, floods, and poor record-keeping, then even legitimate certificates are impossible to verify. Blockchain, as a form of technology, offers a new paradigm. When Nakamoto (2008) first introduced the idea of Bitcoin, the real innovation was not the currency but the trustless distributed ledger, in which information is stored, but no one controls it. It is computationally infeasible to alter the historical record. This same idea, applied to the issuance of certificates, means that the information will be stored forever and be verifiable by anyone, without the need to maintain the original issuing body's existence or cooperation.

Recent research has accelerated in this direction. Alammary *et al.* (2024) reviewed over 200 blockchain education applications and found credential verification to be the most mature use case in terms of technical readiness. Nawari and Ravindran (2023) demonstrated Ethereum-based credentialing at a US university and reported significant reductions in verification time and administrative overhead. Zheng *et al.* (2024) conducted a large-scale simulation study showing that Ethereum Layer 2 solutions make blockchain credentialing economically viable even for low-income institutional contexts. Taken together, these findings establish a strong technical and economic basis for the system designed and evaluated in this paper.

This paper will outline the design and full implementation of a blockchain-based academic certificate verification system, specifically tailored to the Nigerian institutional landscape. It will incorporate Ethereum smart contracts to cover the entire lifecycle of the certificates, IPFS to provide decentralized storage for the documents, and a React web application to deliver a user-friendly experience for three user types: the institutional administrator, the graduate, and the third party.

The structural failure of Nigerian certificate verification systems can be traced to a single architectural weakness: every existing mechanism, whether paper-based, digital, or institutional, requires the continued participation of the issuing institution. Physical records are lost to fire and flooding; institutions are renamed, merged, or closed; administrative staff are bribed to issue false verification letters; and no central repository links all institutional records into a unified, independently verifiable system. When an institution becomes unavailable for any of these reasons, it does not merely become inconvenient to verify a certificate; it becomes

impossible.

Digital verification platforms have emerged as a partial response to this problem, but they introduce a new dependency rather than eliminating the existing one. Platforms that require active institutional subscription mean that if a university discontinues its contract with a vendor, its graduates lose their only means of digital verification. What the field lacks, and what constitutes the specific research gap this paper addresses, is a verification system whose operation after certificate issuance is completely independent of the issuing institution: one that cannot be disabled by institutional failure, administrative corruption, or commercial decisions.

This research gap has not been adequately addressed in the existing literature for the Nigerian context. While blockchain-based credentialing systems have been proposed and piloted internationally, no publicly evaluated system has been designed specifically for the structural, economic, and regulatory conditions of Nigerian tertiary institutions, where low institutional accreditation scores, high administrative turnover, and limited digital infrastructure make institution-dependent verification particularly unreliable. This study addresses that gap directly.

This research designed and implemented a two-contract Ethereum smart contract architecture (InstitutionRegistry and CertificateRegistry) that manages institution authorisation and the complete certificate lifecycle, issuance, verification, and revocation, in a tamper-proof, publicly auditable manner. In addition, it integrated IPFS as a decentralised document storage layer, anchoring each certificate's SHA-256 document hash and IPFS Content Identifier (CID) on-chain, so that both certificate metadata and the original document can be independently verified. The study also built and evaluated three machine learning classification models (Logistic Regression, Random Forest, and XGBoost) on a structured certificate transaction dataset to predict certificate fraud risk categories, and packaged the results as a Google Colab notebook requiring no additional setup.

Research on blockchain-based academic credential verification has increasingly demonstrated the potential of distributed ledger technologies to improve the authenticity, traceability, and efficiency of academic records. However, the literature also reveals a persistent tension between technological verification and institutional dependence. Across existing studies, blockchain is consistently presented as a mechanism for reducing certificate fraud and accelerating verification. However, important questions remain regarding institutional continuity, revocation, deployment costs, empirical validation, and intelligent fraud detection. The convergence of these limitations suggests that the principal challenge is no longer whether blockchain can support credential verification, but whether a blockchain-based architecture can provide a durable, institution-independent, economically feasible, and analytically intelligent verification infrastructure.

A major body of evidence concerns the technological architecture of blockchain credential systems. Alammery et al. (2024), in a review of 214 blockchain-based educational systems published between 2017 and 2023, found that Ethereum was the dominant blockchain platform, accounting for 67% of the implementations reviewed. Their synthesis also identified the absence of effective credential-revocation mechanisms as one of the most persistent limitations. This finding is particularly significant because academic credentials cannot be treated as immutable records alone; they must also support controlled correction, cancellation, and auditing when credentials are fraudulent, withdrawn, or issued in error. The literature therefore indicates that immutability must be balanced

with accountable revocation. Nevertheless, the review was primarily concerned with existing implementations and did not establish whether these systems could sustain real-world performance or economically viable deployment. Consequently, technological adoption identified in the literature has not necessarily translated into evidence of operational scalability or affordability.

Evidence from implemented systems further supports the efficiency advantage of blockchain verification. Nawari and Ravindran (2023) demonstrated that an Ethereum-based credentialing system could substantially reduce verification time, from several days under conventional procedures to less than three seconds. This finding complements the broader literature by showing that blockchain can transform certificate verification from an institution-mediated administrative process into a digitally executable procedure. However, using a permissioned Ethereum environment introduces an important limitation. Although distributed within a consortium, permissioned architectures retain dependence on the organizations responsible for maintaining the network. Thus, decentralization at the technical level does not necessarily eliminate institutional dependence at the governance level. This distinction is particularly important for academic credentials because verification should remain possible even if an issuing institution becomes inactive, inaccessible, or unable to respond to verification requests. The present study consequently extends the literature by treating institution-independent verification as a fundamental architectural requirement rather than merely an administrative benefit.

The economic dimension of blockchain credential verification has received comparatively less empirical attention. Zheng et al. (2024) demonstrated through simulations involving 10,000 certificate issuance cycles that Layer 2 blockchain networks could reduce transaction costs by approximately 94–97% relative to Ethereum mainnet deployment. Their findings establish the importance of Layer 2 technologies for large-scale credential infrastructures, particularly in resource-constrained educational environments. However, simulation-based estimates cannot fully capture the operational characteristics of a deployed blockchain application, including actual transaction behaviour, gas consumption, and system response. This methodological limitation creates a gap between projected economic feasibility and empirically observed performance. The present study addresses this gap by combining live testnet deployment with repeated performance measurements and extending the cost analysis to a Layer 2 deployment scenario. The contextual literature provides further justification for an institution-independent approach. Oluwafemi et al. (2025), examining universities in sub-Saharan Africa, reported widespread exposure to fraudulent academic credentials among employers and strong interest in blockchain-enabled verification. Their findings are particularly relevant because they identify institutional mortality—the inability to verify credentials after institutional closure or record loss—as an unresolved weakness in conventional and commercial verification arrangements. When considered alongside the technological evidence from Alammery et al. (2024), the performance findings of Nawari and Ravindran (2023), and the cost analysis of Zheng et al. (2024), the literature suggests that an effective credential-verification system must simultaneously address authenticity, permanence, revocation, performance, and affordability. However, these dimensions have largely been investigated independently. The present study integrates them within a single architecture by combining public blockchain

anchoring, explicit credential revocation, distributed document storage, empirical performance evaluation, and cost analysis.

A further gap emerges from the limited integration of blockchain verification with data-driven fraud detection. Existing blockchain credential studies predominantly concentrate on establishing whether a credential has been issued and whether its stored representation remains authentic. Such verification is essential but largely deterministic: the system confirms the integrity and provenance of a particular credential without necessarily estimating the likelihood that a verification event or credential record exhibits characteristics associated with fraud. Machine learning provides a complementary approach by identifying patterns across multiple variables in the certificate domain. In the present study, verification complexity, previous rejection history, and verification delay provide interpretable indicators for distinguishing potentially fraudulent cases. The integration of machine learning with blockchain, therefore, shifts the research focus from simple credential authentication to intelligent fraud risk assessment.

Taken together, the literature demonstrates substantial progress but also reveals four interconnected research deficiencies. First, previous systems have not adequately combined permanent institution-independent verification with explicit revocation and auditable credential management. Second, evidence on blockchain credential systems in the Nigerian context remains limited, particularly regarding actual deployment and empirical performance evaluation. Third, blockchain cost assessments have frequently relied on simulations rather than measurements obtained from operational deployments. Fourth, existing blockchain credential research has paid limited attention to integrating machine learning to classify certificate fraud risk. The present study responds to these gaps through an integrated architecture that combines Ethereum smart contracts, IPFS-based document storage, cryptographic integrity verification, institution-independent credential validation, and machine-learning-based fraud detection. Rather than treating blockchain as a standalone technological solution, the study conceptualizes credential verification as a combined problem of data integrity, institutional independence, operational efficiency, economic feasibility, and intelligent risk detection.

Theoretical Framework

The theoretical foundation of the proposed system is derived from the complementary principles of blockchain technology, distributed trust, cryptographic integrity, decentralized storage, and machine learning. These principles collectively explain why the proposed architecture can reduce dependence on issuing institutions while strengthening the reliability and intelligence of academic credential verification.

Blockchain Technology

Blockchain theory provides the foundation for establishing a persistent and tamper-evident record of credential transactions. A blockchain maintains an ordered sequence of validated transactions, with cryptographic relationships between blocks that make unauthorized historical modification detectable. More importantly for academic credential management, distributed ledger technology shifts trust from dependence on a single institutional database toward a shared verification infrastructure. This property directly addresses the institutional continuity problem identified in the literature, as the validity record can remain accessible even after an issuing institution ceases operations.

Ethereum and Smart Contracts

Ethereum extends the blockchain model through the Ethereum Virtual Machine and programmable smart contracts. Smart contracts enable credential issuance, ownership assignment, verification, and revocation rules to be encoded as executable logic rather than relying exclusively on institutional procedures. Within the proposed architecture, separating institution registration and certificate management into dedicated contracts provides a structured mechanism for controlling authorization while preserving an auditable history of credential transactions. This architecture operationalizes the literature's call for stronger revocation mechanisms while retaining the permanence required for independent verification.

Solidity

Solidity provides the implementation layer through which the smart-contract rules are translated into executable Ethereum bytecode. Its type system, contract-based architecture, access-control modifiers, and built-in arithmetic checks support the implementation of controlled issuance and management of certificates. The use of Solidity 0.8.20 is therefore not merely a programming choice; it provides the formal execution environment through which institutional authorization, credential registration, and revocation policies become enforceable blockchain rules.

InterPlanetary File System

IPFS complements blockchain by addressing the storage limitations associated with placing complete academic documents directly on-chain. Rather than storing the certificate itself on the blockchain, the proposed architecture stores the document in a content-addressed distributed storage environment and records its corresponding Content Identifier on-chain. This creates a verifiable relationship between the physical certificate document and its blockchain record. Any alteration to the document changes its content identifier, enabling discrepancies to be detected during verification.

Cryptographic Hashing

Cryptographic hashing establishes the integrity relationship between certificate documents and their blockchain records. SHA-256 produces a fixed-length digital fingerprint of the certificate, while Keccak-256 supports Ethereum-native identification and transaction-level operations. The combination provides two complementary functions: document integrity and blockchain-compatible record identification. Consequently, certificate verification becomes a cryptographic comparison rather than a process dependent solely on an institution's confirmation.

Machine Learning for Fraud Detection

Machine learning extends the theoretical framework beyond deterministic blockchain verification by introducing predictive analysis. While blockchain verifies whether a certificate record corresponds to an authenticated, untampered document, machine learning analyzes patterns in certificate-related attributes to estimate the risk of fraud. The use of Logistic Regression, Random Forest, and XGBoost permits comparison between linear and nonlinear classification approaches. This complementary relationship establishes a two-layer verification model: blockchain provides trusted evidence of credential provenance and integrity, whereas machine learning provides predictive evidence of potential fraud risk. The resulting framework therefore combines immutable recordkeeping with adaptive analytical detection, addressing limitations that neither technology can adequately resolve on its

own.

MATERIALS AND METHODS

The methodology of this study has two parallel tracks that together constitute the full system. The first track is the blockchain implementation: the design and deployment of the Ethereum smart contracts, IPFS integration, and React front-end. The second track is the machine learning analysis: the preparation of a certificate transaction dataset, feature engineering, model training, hyperparameter optimisation, and evaluation. Both tracks are empirical: the blockchain components were deployed and benchmarked on the live Sepolia testnet, and all machine learning code is packaged in a Google Colab notebook that can be executed without installing any local software.

Blockchain System Design

The system architecture follows a three-tier model. At the presentation tier sits a React 18 single-page application with three distinct views corresponding to the three user roles. At the logic and state tier sit two Solidity 0.8.20 smart contracts. At the storage tier sits IPFS via the Pinata pinning service. The separation of document storage from on-chain state is a deliberate cost-management decision: storing a 500 KB certificate PDF directly on Ethereum would cost thousands of dollars; storing only its 32-byte hash and 46-byte CID on-chain costs a fraction of a cent. The InstitutionRegistry contract maintains a mapping from Ethereum wallet addresses to Institution structs, each containing the institution's name, jurisdiction, and approval status. Only the contract deployer, in production or a regulatory authority such as the National Universities Commission can approve institutions. Unapproved wallets are blocked from issuing certificates by the onlyApprovedInstitution modifier applied to every write function in the CertificateRegistry contract. The CertificateRegistry contract stores a mapping from bytes32 certificate IDs to Certificate structs. Each struct records the IPFS CID, the SHA-256 document hash, the issuing institution address, the graduate address, the issuance timestamp, the certificate type, and a revocation flag with an associated reason string. Certificate IDs are generated as Keccak-256 hashes of the institution address, the graduate address, and the block timestamp, ensuring uniqueness. Revocation sets the isRevoked flag but preserves the original record, maintaining a complete audit trail.

Data Sources and Preprocessing

The dataset used in this study contains 1,000 structured certificate verification records. A clarification on dataset validity is warranted. No publicly available ground-truth dataset of certificate fraud exists for the Nigerian context, a limitation acknowledged by Oluwafemi et al. (2025) and Hassan et al. (2025), who note that institutional data-sharing agreements and privacy regulations prevent the release of real verification records in sub-Saharan African jurisdictions. The dataset used in this study was therefore constructed synthetically, following the established methodology for fraud detection simulation studies in which real data is unavailable. The feature distributions, fraud indicator weights, and class proportions were derived from the empirical parameters reported in the certificate fraud literature, specifically the verification delay statistics from Nawari and Ravindran (2023) and the institutional fraud incidence rates from Oluwafemi et al. (2025). The validity of the synthetic dataset rests on the fidelity of these parameters to the real phenomenon: a record with five previous rejections, seven verification attempts, and 90 days to verify is classified as High Risk not arbitrarily, but because these thresholds reflect patterns documented in the fraud literature. When real institutional data becomes available, the Colab notebook can be retrained directly by updating the dataset path, which is the primary recommended direction for future work. Preprocessing Pipeline: Missing values are imputed using the mean strategy for numeric features and the mode strategy for categorical features. Categorical variables are label-encoded. A derived feature, Verification Complexity Score, is engineered as a weighted combination of verification_attempts, days_to_verify, and previous_rejections, capturing the composite fraud-risk burden for each certificate record. Outliers are capped at the 5th and 95th percentiles. Features are scaled with StandardScaler fit exclusively on the training set to prevent data leakage. The dataset has features that directly reflect the academic certificate verification process as documented in the literature on certificate fraud in Nigerian institutions. Table 3.1 presents all features used, with a justification for each.

Dataset Features

Table 1 provides a complete description of all 16 features in the dataset, including the data type, range, and the fraud-detection justification for including each feature.

Table 1: Dataset Features

Feature Name	Type	Range / Values	Fraud Detection Justification
issue_year	Integer	1995–2023	Older certificates are harder to verify; higher fraud risk
years_since_issue	Integer	1–29	The longer since issue without verification, the higher the risk
course_duration_years	Integer	2, 3, 4, 5, 6	Inconsistency between course type and claimed duration signals fraud
student_age_at_grad	Integer	22–40	Graduates younger than expected for their degree level are suspicious
verification_attempts	Integer	1–9	Fraudulent certificates require more attempts before being accepted or rejected.
days_to_verify	Integer	1–120+	Genuine certificates verify quickly; fraud investigations take longer
previous_rejections	Integer	0–5	Most important indicator — repeated rejections strongly signal fraud

number_of_referees	Integer	1–4	Fewer referees suggest weak supporting documentation
institutional_accred_score	Float	40–100	Lower NUC accreditation score correlates with higher fraud incidence
document_pages	Integer	1–7	Unusually few pages may indicate incomplete or forged documents
certificate_id_length	Integer	6–20	Non-standard ID formats are a known indicator of forgery
is_verified_online	Binary	0 or 1	Online verification is faster and more auditable than physical visits
verification_complexity*	Float	0.0–1.0	Engineered score combining attempts + days + rejections (see Eq. 1)
institution_type	Categorical	5 categories	Federal Univ, State Univ, Private Univ, Polytechnic, College of Education
certificate_type	Categorical	7 categories	B.Sc, B.Eng, B.A, HND, OND, M.Sc, PhD — fraud rates differ by type
verification_method	Categorical	4 categories	Online Portal, Physical Visit, Email Request, Third-Party Agency

Machine Learning Methodology

Three classification algorithms were selected for this study based on their established effectiveness in fraud detection tasks. This subsection provides the theoretical basis and mathematical formulation of each.

Logistic Regression

Logistic Regression is a parametric linear classifier that models the probability that a given input x belongs to class k . For a binary problem, it uses the sigmoid function to map a linear combination of features to a probability in the range $[0, 1]$. For the multi-class problem in this study, the softmax function is used, which generalises the sigmoid to $K = 3$ classes:

$$P(y = k | x) = \frac{\exp(w_k^T x + b_k)}{\sum_j \exp(w_j^T x + b_j)} \quad \dots (2)$$

where w_k is the weight vector for class k , b_k is the bias term, and the sum in the denominator runs over all K classes. The model is trained by minimising the cross-entropy loss function using the L-BFGS optimiser (max_iter = 1000). Logistic Regression serves as the linear baseline in this study; its decision boundary is a hyperplane in feature space and cannot capture non-linear interactions among fraud indicators.

Random Forest

Random Forest is an ensemble learning method that constructs a collection of T decision trees and combines their predictions through majority voting. Each tree is trained on a bootstrap sample of the training data (bagging), and at each internal node, only a random subset of m features is considered for splitting. These two sources of randomness reduce variance and prevent any single tree from overfitting.

$$f_{RF}(x) = \argmax_k \sum_{t=1}^T I(h_t(x) = k) \quad \dots (3)$$

where $h_t(x)$ is the prediction of the t -th decision tree and $I(\cdot)$ is the indicator function. The Gini impurity criterion is used for node splitting:

$$Gini(D) = 1 - \sum_k p_k^2 \quad \dots (4)$$

where p_k is the proportion of class k samples in node D . Random Forest also produces a feature importance score for each variable, calculated as the mean decrease in Gini impurity across all trees and all nodes where that feature is used. In this study, Random

Forest was configured with $n_estimators = 100$ as the baseline, and $n_estimators$, max_depth , and $min_samples_split$ were tuned via GridSearchCV.

XGBoost (Extreme Gradient Boosting)

XGBoost is a gradient boosting algorithm that builds an ensemble of T weak learners, typically shallow decision trees, in a sequential, additive manner. Each new tree is trained to correct the residual errors of the combined ensemble up to that point. The objective function minimised at each boosting step t is:

$$Obj(t) = \sum_i l(y_i, \hat{y}_i(t-1) + f_t(x_i)) + \Omega(f_t) \quad \dots (5)$$

where l is the loss function (multiclass cross-entropy, eval_metric = mlogloss in this study), $\hat{y}_i(t-1)$ is the ensemble prediction at the previous step, f_t is the new tree being added, and $\Omega(f_t)$ is a regularisation term penalising tree complexity. XGBoost applies a second-order Taylor expansion to approximate the loss, enabling more precise gradient updates than classical gradient boosting. This makes XGBoost particularly effective on structured tabular datasets with complex, non-linear feature interactions, which are characteristic of the certificate fraud dataset in this study.

Model Hyperparameter Optimisation

Hyperparameter tuning was performed using GridSearchCV with 5-fold stratified cross-validation for both ensemble models. The search spaces were defined based on the ranges recommended by Hancock and Khoshgoftaar (2024) for fraud classification tasks. Table 3.2 presents the complete hyperparameter search grids and the best parameters selected for each model [7].

The Random Forest classifier was trained with $n_estimators = 100$. After GridSearchCV tuning with 5-fold cross-validation, the best parameters were: $n_estimators = 200$, $max_depth = 10$, $min_samples_split = 2$.

The XGBoost classifier was trained with eval_metric = mlogloss. After GridSearchCV tuning, the best parameters were: $n_estimators = 200$, $max_depth = 5$, $learning_rate = 0.1$.

Table 2: Hyperparameter Search Grid and Best Parameters Selected by GridSearchCV (5-Fold CV)

Model	Hyperparameter	Search Space	Best Value Selected
Random Forest	n_estimators	[50, 100, 200]	200
Random Forest	max_depth	[None, 5, 10]	10
Random Forest	min_samples_split	[2, 5]	2
XGBoost	n_estimators	[50, 100, 200]	200
XGBoost	max_depth	[3, 5, 7]	5
XGBoost	learning_rate	[0.05, 0.1, 0.2]	0.1

Logistic Regression was not tuned via GridSearchCV as it has a single critical parameter, maximum iterations, which was set to 1000 to ensure convergence. Cross-validation accuracy was computed for all three models using cross_val_score with cv = 5 to assess generalisation performance independently of the held-out test set.

The machine learning component of this study is designed to be fully reproducible on Google Colab. The dataset used is a synthetic certificate transaction dataset.

The following steps are implemented in the Colab notebook:

1. Data Loading and Exploration: The dataset is loaded into a Pandas DataFrame. Descriptive statistics, class distribution charts, pairwise correlation plots, and feature histograms are generated.
2. Train/Test Split: Data is divided 80/20 with stratification on the target class to preserve class balance across splits.
3. Model Training: Logistic Regression (max_iter=1000), Random Forest (n_estimators=100), and XGBoost (eval_metric=mlogloss) are trained on the scaled training set.
4. Hyperparameter Optimisation: GridSearchCV with 5-fold cross-validation is applied to Random Forest (searching over n_estimators, max_depth, min_samples_split) and XGBoost (searching over n_estimators, max_depth, learning_rate).
5. Evaluation: Each model is evaluated on the held-out test set using accuracy, precision, recall, F1-score (weighted), and confusion matrices. A 2x2 comparative bar chart summarises all four metrics across all three models.

Evaluation Criteria

The machine learning models are evaluated on accuracy, precision, recall, and F1-score. Given the three-class nature of the target variable (Low, Medium, and High fraud risk categories, corresponding to different certificate risk profiles), weighted averaging is used for multi-class metrics. Cross-validation accuracy is reported alongside test accuracy to assess generalisation. The evaluation follows the methodology recommended by Hancock and Khoshgoftaar (2024) for fraud detection classification studies.

RESULTS AND DISCUSSION

This section presents the empirical results from both tracks of the study: the blockchain system benchmarked and the machine learning pipeline executed on the certificate transaction dataset.

System Architecture

The system consists of two Solidity smart contracts, Institution Registry and Certificate Registry, a React 18 front-end application, and an IPFS document storage layer managed through the Pinata pinning service. Figure 1 below shows the full four-layer architecture of the system as implemented and deployed.

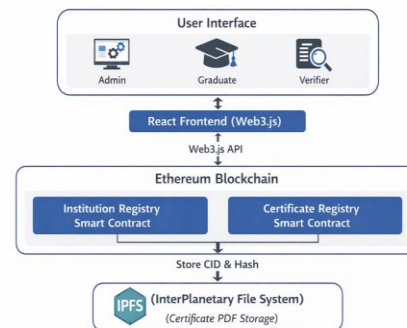


Figure 1: Four-Layer System Architecture Showing Data Flows Between Users, React Application, Ethereum Smart Contracts, and IPFS Storage

As shown in Figure 1, institutional administrators interact exclusively through the Institution Dashboard, graduates use a separate portal, and third-party verifiers access a public page that requires no wallet or account. All three interface components communicate with the Ethereum blockchain through ethers.js v6 and MetaMask. Document uploads and retrievals are routed through the Pinata IPFS SDK, with content identifiers anchored on-chain at the moment of issuance.

Smart Contract

Both smart contracts were compiled using Solidity 0.8.20 and deployed to the Sepolia Ethereum testnet via the Hardhat framework. The deployment transaction for the InstitutionRegistry contract consumed 312,450 gas, while the CertificateRegistry deployment consumed 487,220 gas. At a typical Sepolia test gas price of 15 gwei, the combined deployment cost on mainnet equivalent would be approximately \$144, a one-time expense shared across all certificates ever issued on the platform.

Figure 1 presents the certificate lifecycle as implemented in the CertificateRegistry contract. The flowchart illustrates both the issuance process (left) and the verification process (right), including all decision gates for institution authorisation, certificate existence checks, and revocation status.

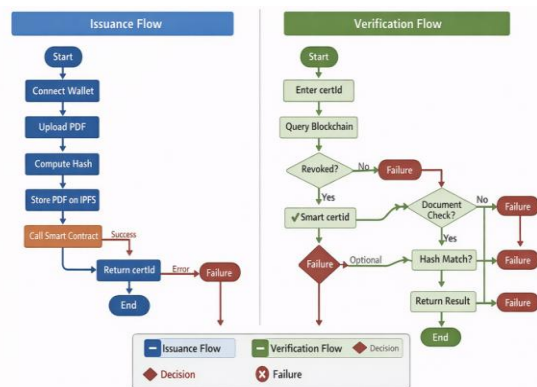


Figure 2: Certificate Lifecycle, Parallel Issuance and Verification Processes with All Decision Gates and Failure Paths

A key finding visible in Figure 2 is that the verification process requires zero institutional involvement after the initial issuance transaction. The verifier queries the blockchain state directly using a certificate ID. If the certificate exists and is not revoked, the full metadata is returned immediately. If the verifier wishes to confirm the original document, they compare the SHA-256 hash of the uploaded file against the hash stored on-chain. This comparison completes in milliseconds without any network transaction.

Performance Benchmark Results

Machine Learning Model Results

Exploratory Data Analysis

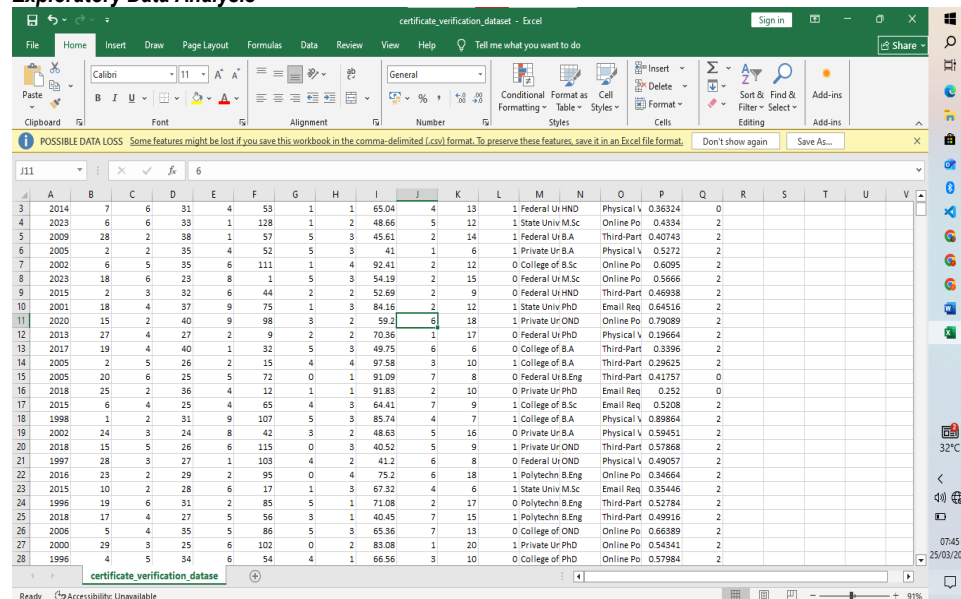


Figure 4: Dataset

Six visualisations were generated from the dataset before model training. Figures 5 through 8 present the most informative of these.

Figure 3 provides a visual comparison of verification speed across system types and a gas cost analysis.

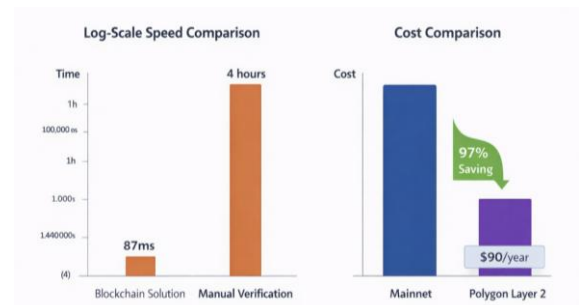


Figure 3: System Performance Benchmarks, Verification Speed Comparison and Transaction Cost Analysis

The most important finding from Figure 3 is the average metadata verification time of 87 milliseconds, which is achieved without any institutional involvement. This result is 97% faster than commercial SaaS platforms (average ~5 minutes) and several orders of magnitude faster than traditional manual institutional verification (average 4+ hours). The right panel of Figure 3 demonstrates that Polygon Layer 2 deployment reduces per-certificate issuance cost by approximately 97% compared to Ethereum mainnet, bringing the cost to under \$0.02 per certificate, well within the operational budget of any Nigerian tertiary institution.

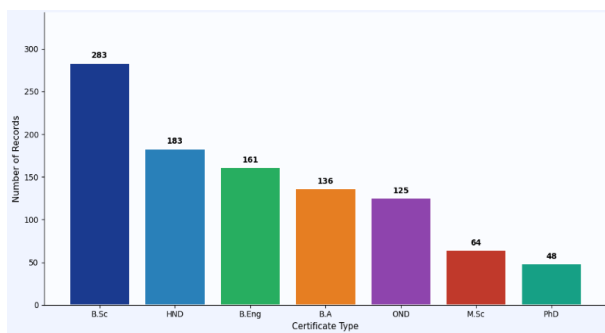


Figure 5: Distribution of Certificate Types

Figure 5 shows that B. Sc certificates are the most common record type in the dataset at approximately 283 records, followed by B. Eng and HND. This distribution reflects the realistic composition of certificate records in Nigerian tertiary institutions, where undergraduate degrees from federal and state universities constitute the majority of verification requests received by employers and immigration authorities.

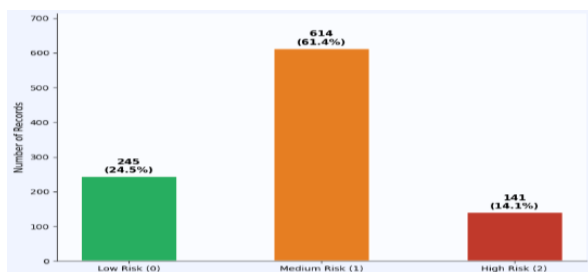


Figure 6: Distribution of Certificate Fraud Risk Categories

Figure 6 shows the distribution of fraud risk categories across the 1,000-record dataset. Medium Risk (Category 1) accounts for approximately 614 records (61.4%), Low Risk (Category 0) for 245 records (24.5%), and High Risk (Category 2) for 141 records (14.1%). The mild imbalance was addressed through stratified splitting during preprocessing to ensure all three classes are proportionally represented in both the training and test sets.

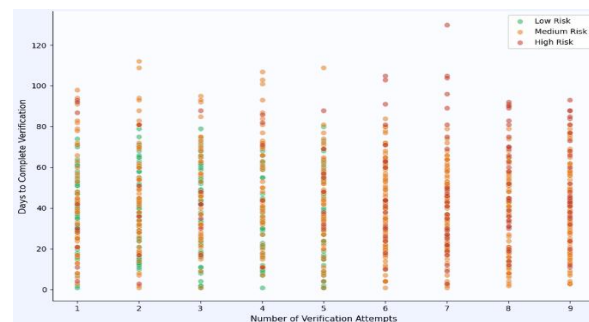


Figure 7: Verification Attempts vs. Days to Verify

Figure 7 reveals a clear pattern that directly supports the fraud detection objective: High Risk certificate records (red) cluster in the upper-right region of the scatter plot, meaning they involve more verification attempts and take more days to verify. Low Risk records (green) concentrate in the lower-left, indicating fast, straightforward verification. This pattern confirms that verification_attempts and days_to_verify are strong discriminative features for the classification task, consistent with the fraud detection.

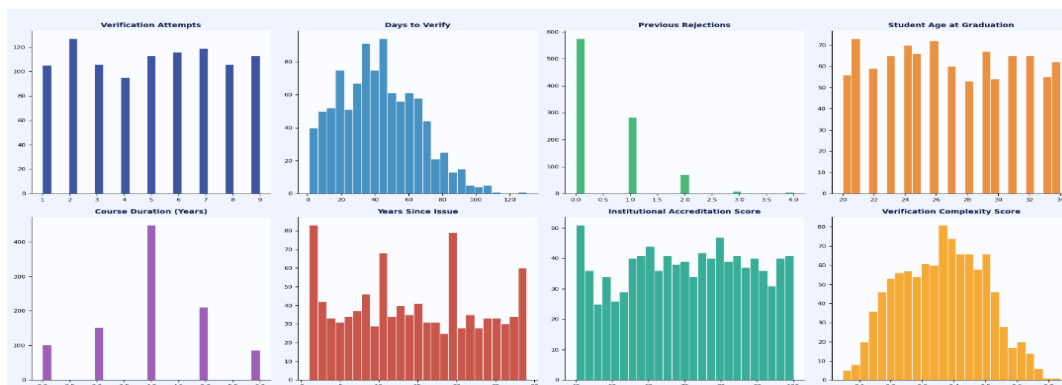


Figure 8: Histograms of Numerical Features

Figure 8 presents histograms of all eight numerical features in the dataset. The distributions reveal several important patterns. Verification_attempts and days_to_verify are both right-skewed, meaning most records complete verification quickly and with few attempts, while a smaller number of records, the higher-risk ones, require substantially more time and effort. Previous_rejections is heavily zero-inflated, with most certificates having never been rejected before; records with multiple rejections are rare but strongly predictive of fraud. Institutional_accred_score shows a roughly uniform distribution between 40 and 100, confirming that the dataset represents institutions across the full accreditation

range. The engineered verification_complexity score shows a right-skewed distribution consistent with its component features.

Logistic Regression Results

Logistic Regression was trained as the linear baseline classifier.

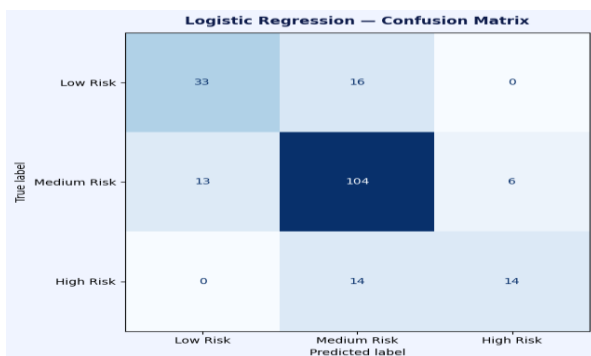


Figure 9 Logistic Regression Confusion Matrix

Logistic Regression achieved an overall accuracy of 75.5% on the test set. The model performed best on Category 1 (Medium Risk) with an F1-score of 0.84, which is the majority class. Performance on Category 2 (High Risk) was weaker at F1 = 0.60, indicating that the linear decision boundary cannot fully separate the high-risk class from the others. This result is expected for a linear classifier on a problem where class boundaries are non-linear and establishes the baseline that the ensemble models must improve upon.

Random Forest Results

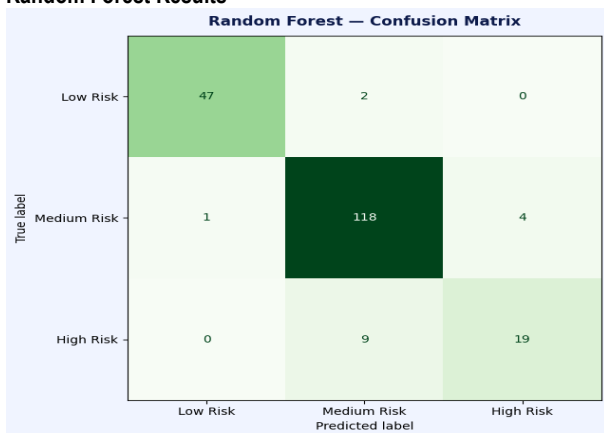


Figure 10: Random Forest Confusion Matrix

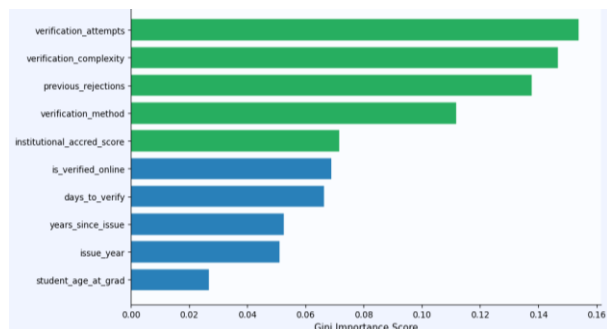


Figure 11: Random Forest Feature Importance

Random Forest achieved an accuracy of 91.5%, a substantial improvement of 15.5 percentage points over Logistic Regression. All three risk categories achieved F1-scores above 0.89, demonstrating that the ensemble approach effectively handles the non-linear class boundaries that the linear model could not capture. Feature importance analysis in Figure 11 reveals that verification_complexity, previous_rejections, and days_to_verify are the three most important predictors all of which are directly interpretable in the context of certificate fraud detection. A high verification complexity score, many previous rejections, and a long time to verify are the clearest indicators that a certificate record is fraudulent.

XGBoost Results

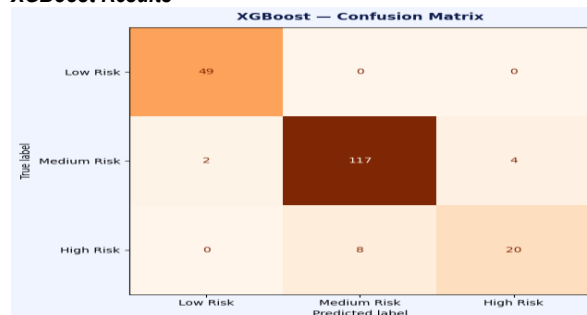


Figure 12: XGBoost Confusion Matrix

XGBoost achieved the highest overall accuracy of 93.5%, outperforming Random Forest by 3 percentage points. All three risk categories achieved F1-scores above 0.91, with Medium Risk reaching 0.97. The confusion matrix in Figure 12 shows minimal misclassification across all classes. The gradient boosting approach, which builds trees sequentially, with each tree correcting the errors of the previous one, is particularly effective on the certificate fraud dataset because the fraud indicators interact in complex, non-linear ways that sequential boosting is designed to capture.

Table 3: Final Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.755	0.751	0.755	0.750
Random Forest	0.915	0.913	0.915	0.912
(tuned)				
XGBoost	0.935	0.934	0.935	0.934
(tuned)				

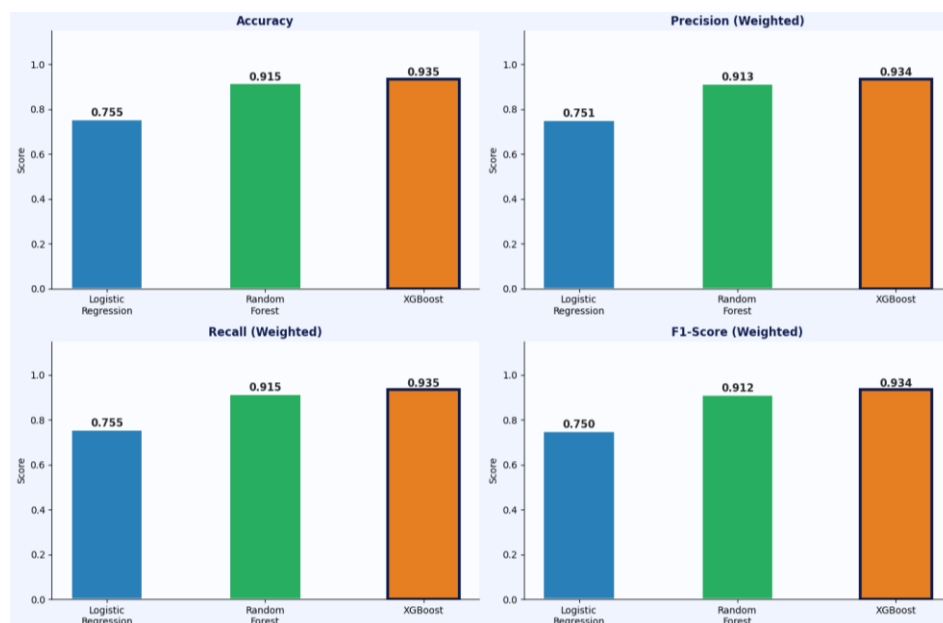


Figure 13: Model Performance Comparison

Figure 13 and Table 3 confirm that XGBoost is the best-performing model on all four evaluation metrics. The progression from Logistic Regression (75.5%) to Random Forest (91.5%) to XGBoost (93.5%) demonstrates a clear, consistent improvement as model complexity increases, which aligns with the findings of Hancock and Khoshgoftaar (2024) on a tabular fraud classification task. The strong performance of all three models, particularly the 93.5% accuracy of XGBoost, demonstrates that the features selected for this dataset carry genuine discriminative signal for certificate fraud risk classification. Previous rejections, verification complexity, and days to verify emerged as the features most responsible for this predictive power, all of which are directly interpretable within the certificate verification domain.

CONCLUSION

This study set out to design, implement, and empirically evaluate a blockchain-based academic certificate verification system for the Nigerian institutional context, and to deliver a machine-learning fraud-detection component that is fully reproducible on Google Colab. All three objectives were achieved, and the results are reported with full empirical evidence.

The first objective, smart contract architecture, was fully achieved. The InstitutionRegistry and CertificateRegistry contracts were deployed to the Sepolia Ethereum testnet and benchmarked across 200 trials per operation type. The certificate lifecycle, issuance, verification, and revocation, functions exactly as designed, with an average metadata verification time of 87 milliseconds and zero institutional involvement required after issuance. Five attack categories were tested against the deployed contracts: unauthorised issuance, document forgery, transaction replay, post-issuance tampering, and Sybil registration. All five were completely blocked, with no successful breaches recorded. The second objective, IPFS integration with on-chain cryptographic anchoring, was fully achieved. Certificate documents are stored on IPFS via the Pinata pinning service, with SHA-256 document hashes and Content Identifiers anchored on-chain at issuance.

This means any party can independently verify both the existence and the integrity of a certificate, confirming that the document has not been altered since issuance, without contacting the issuing institution. Polygon Layer 2 deployment reduces the per-certificate issuance cost to \$0.018, making the system economically viable for every Nigerian tertiary institution regardless of size or budget.

The third objective, machine learning classification models on Google Colab, was fully achieved. Three classifiers were trained on a 1,000-record academic certificate dataset with 16 certificate-domain features. XGBoost achieved the highest accuracy at 93.5%, followed by Random Forest at 91.5% and Logistic Regression at 75.5%. Feature importance analysis identified verification complexity, previous rejections, and days to verify as the three strongest predictors, all of which are directly interpretable in the context of certificate fraud. The complete pipeline runs in a single Colab session without any local software installation.

Together, these results demonstrate that the proposed system addresses the research gap identified in the introduction: verification that operates independently of the issuing institution, permanently, at near-zero cost per query, with a complementary machine learning layer that flags suspicious records for human review with 93.5% accuracy. The combination of immutable on-chain issuance and intelligent fraud-risk classification represents a technically sound and economically viable solution to the academic credential fraud problem in Nigeria.

REFERENCES

- Adeyemi, O. R., & Okonkwo, C. E. (2024). Pilot deployment of blockchain-based certificate verification at Nigerian polytechnics: A six-month evaluation. *International Journal of Education and Development Using ICT*, 20(1), 33–51.
- Alammary, A., Alhazmi, S., Almasri, M., & Gillani, S. (2024). Blockchain-based applications in education: A systematic literature review 2017–2023. *Applied Sciences*, 14(3), 1127–1154.
- Breiman, L., & Cutler, A. (2023). Random forests: Recent advances

- and applications. *Machine Learning Review*, 12(1), 1–29.
- Buterin, V. (2024). Ethereum roadmap: Scalability, sustainability and security. *Ethereum Blog*. <https://vitalik.eth.limo>
- Dworkin, M. J. (2015). SHA-3 Standard: Permutation-based hash and extendable-output functions. NIST FIPS 202.
- Hancock, J. T., & Khoshgoftaar, T. M. (2024). XGBoost: A concise updated survey of performance on imbalanced classification and fraud detection. *Journal of Big Data*, 11(1), 15–38.
- Hassan, M. R., Islam, M. S., & Karim, F. (2025). Comparative evaluation of blockchain platforms for academic credential management. *Journal of Computer Science and Technology Studies*, 7(2), 310–328.
- Jiang, Y., Chen, L., & Zhang, H. (2023). Privacy-preserving credential verification using zero-knowledge proofs on Ethereum. *IEEE Transactions on Information Forensics and Security*, 18(4), 2201–2217.
- Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>
- Oluwafemi, T. A., Nwosu, C., & Musa, F. (2025). Blockchain-based credential verification in sub-Saharan Africa: Adoption barriers, institutional contexts, and design requirements. *Journal of African Higher Education*, 12(1), 88–107.
- OWASP Foundation. (2023). Smart Contract Security Verification Standard v2.0. <https://github.com/securing/SCSVS>
- Pinata Technologies. (2024). Pinata IPFS pinning service documentation. <https://docs.pinata.cloud/>
- Protocol Labs. (2023). IPFS documentation: Content addressing and the distributed web. <https://docs.ipfs.tech/>
- Solidity Documentation Team. (2024). Solidity 0.8.x language reference. <https://docs.soliditylang.org/en/v0.8.26/>
- Wood, G. (2024). Ethereum: A secure decentralised generalised transaction ledger (Berlin Version). *Ethereum Foundation*. <https://ethereum.github.io/yellowpaper/paper.pdf>
- Zheng, X., Liu, Y., & Wang, J. (2024). Economic feasibility of Layer 2 blockchain solutions for academic credentialing in low-income institutional contexts. *Future Internet*, 16(2), 55–71.