A Machine Learning Methodology for Detecting Synthetic Identity Fraud in Financial Transactions

Ayomide Ayeni
Department of Computer Science
Atlanta, GA, USA
ayeniayomide5@gmail.com

Abstract—Financial fraud in card payment systems and account transfer environments poses a persistent and escalating threat to the global financial infrastructure, with global card fraud losses exceeding $33 billion annually. Synthetic identity fraud, in which fabricated identities are assembled from real and fictitious elements to survive conventional verification, represents a particularly challenging variant because each individual data element may appear valid when examined in isolation. This paper presents a complete, reproducible machine learning methodology for detecting fraudulent financial transactions, validated across two structurally distinct payment environments: a credit card transaction dataset of approximately 1.9 million records and an account transfer dataset of over 6.3 million records. In the credit card environment, three classifiers were evaluated, with Random Forest achieving the highest accuracy of 99.74% on a held out test set of 555,719 records. In the account transfer environment, four classifiers were compared, with Random Forest achieving 99.95% accuracy, a fraud class F1 score of 0.74, and an AUC ROC of 0.84. The cross environment consistency of these results demonstrates that the fraud signal resides in the feature construction rather than in any single model architecture, establishing the transferability necessary for institutional adoption. The paper further specifies a 20 measurement synthetic identity feature set across six behavioural families, designed for community banks, credit unions, and regional lenders operating without dedicated analytics staff. The full methodology, including the feature engineering specification, reference implementation, implementation guide, and threat model note, is published under an open licence.

Keywords—financial fraud detection, synthetic identity fraud, machine learning, random forest, transaction anomaly detection, class imbalance, feature engineering, credit card fraud, account transfer fraud, open methodology, community banks, credit unions

I. INTRODUCTION

A. Background and Motivation

The global financial system is critically dependent on the security and integrity of electronic payment infrastructure. Credit and debit card transactions process trillions of dollars annually, creating a vast attack surface for financial fraud. According to the Nilson Report, global card fraud losses reached $32.34 billion in 2021 and are projected to exceed $43 billion by 2026 [1], representing both a systemic risk to financial institutions and direct harm to consumers and merchants.

Financial fraud manifests in several distinct forms. Card not present (CNP) fraud occurs when stolen card credentials are used for online or telephone purchases. Card present fraud involves counterfeit or stolen physical cards at point of sale terminals. Account takeover attacks involve unauthorised actors gaining control of existing accounts. Each type exhibits different behavioural signatures in transaction data, requiring detection systems capable of multi pattern recognition [2].

Among these, synthetic identity fraud represents a particularly insidious and growing threat. A synthetic identity is a fabricated person assembled from real fragments: a genuine Social Security number belonging to someone who does not actively use it, paired with a fabricated name, a plausible date of birth, and a real address. Identity verification systems check whether each element is valid, and a synthetic identity passes every check because each element is valid. Only the combination is false [3]. The fraudster then builds ordinary credit history under the fabricated identity for months or years, gradually increasing available credit, before drawing down every available line and disappearing. Because no document is forged and no individual field is false, detection cannot operate on verification. It must operate on behaviour over time: how the credit file was assembled, whether the borrowing pattern is internally consistent, and how balances and repayments move relative to stated capacity.

B. The Access Problem

Sophisticated fraud analytics are not equally available to all financial institutions. Large banks can maintain specialised data science, engineering, fraud investigation, and compliance capabilities that smaller institutions often cannot support internally. The United States currently has more than 4,250 federally insured credit unions and 3,852 FDIC insured community banks, totalling more than 8,100 institutions that collectively serve a substantial portion of American consumers [4], [5]. Fraud losses concentrate at precisely this tier, because detection capability is weakest where analytical resources are most constrained. The methodology presented in this paper is therefore deliberately designed for an institution with approximately one technically capable person and a compliance officer, providing implementation paths that begin with manual review and progress to scripted batch scoring without requiring proprietary software or specialised computing infrastructure.

C. Objectives and Contributions

This paper makes the following contributions:

(1) A complete, end to end machine learning pipeline for financial transaction fraud detection, validated on two structurally distinct payment environments totalling approximately 8.2 million records. (2) A comparative evaluation of multiple supervised classifiers demonstrating that the fraud signal resides in feature construction rather than in any single model architecture. (3) A 20 measurement synthetic identity feature specification organised into six behavioural families, designed to extend validated transaction detection into the domain of identity fabrication in consumer lending. (4) An open distribution model comprising a feature engineering specification, reference implementation, implementation guide, and threat model note, published under an open licence for adoption by smaller financial institutions.

II. RELATED WORK

A. Transaction Fraud Detection

The detection of financial fraud using machine learning has been an active area of research for over two decades. Dal Pozzolo et al. [6] highlighted the challenges of concept drift and class imbalance in credit card fraud detection, proposing adaptive resampling strategies. Bhattacharyya et al. [2] demonstrated that support vector machines and random forests outperform logistic regression for fraud detection on skewed datasets. Deep learning approaches have gained traction: Fiore et al. [7] applied generative adversarial networks to synthesise minority class samples, while Kim and Kim [8] explored long short term memory networks for sequential transaction modelling. However, tree based ensemble methods consistently demonstrate competitive performance with significantly lower computational cost, and remain the industry standard for production deployment due to their interpretability, lower latency, and robustness [9].

B. Synthetic Identity Fraud

Synthetic identity fraud differs fundamentally from transaction fraud in its temporal structure and labelling characteristics. A synthetic identity that is never drawn down is never labelled, and one that is drawn down is typically labelled as a charge off rather than as fraud. Any institution training on historical labels is therefore training on a subset defined by which schemes concluded and were correctly classified afterwards. The Federal Reserve has identified synthetic identity fraud as the fastest growing type of financial crime in the United States [3]. Detection requires analysis of relationships among identity attributes, credit file assembly patterns, and longitudinal account behaviour, rather than evaluation of individual transactions in isolation. This paper bridges the gap between validated transaction fraud detection and the structurally distinct problem of synthetic identity detection by demonstrating that the same feature construction principles transfer across domains.

C. Regulatory Framework

Financial fraud detection operates within a binding regulatory framework. The Fair Credit Reporting Act (FCRA) governs consumer data use in credit decisions. The Equal Credit Opportunity Act (ECOA) prohibits discriminatory algorithmic decisions. EU GDPR Article 22 grants the right to human review of automated decisions. The Bank Secrecy Act (BSA) and its implementing regulations require institutions to maintain effective programmes for detecting suspicious activity. Any deployment of the methodology described in this paper must comply with these statutes and with the institution's own compliance framework. The feature engineering specification addresses fair lending compliance directly by recommending exclusion of protected characteristics from scoring models and requiring post training fairness audits.

III. DATASETS AND PREPROCESSING

A. Environment A: Credit Card Transactions

The first dataset is a synthetic credit card transaction dataset containing 1,852,394 records with 23 raw columns. The dataset simulates realistic consumer card transactions and exhibits severe class imbalance: only 0.58% of transactions are labelled fraudulent, producing an imbalance ratio of approximately 172 to 1. Features include transaction amount, merchant category, cardholder demographics, geographic coordinates, and temporal attributes. Correlation analysis identified transaction amount (amt) as the dominant predictor of fraud, with a Pearson correlation coefficient of $r = 0.219$ [10].

Preprocessing for Environment A included label encoding of categorical variables, binary encoding of demographic features, and removal of identifier columns (cc_num, trans_num, and nine further personal fields) to prevent memorisation. The dataset was partitioned into training (80%) and validation (20%) sets, yielding a held out test set of approximately 555,719 records.

B. Environment B: Account Transfers

The second dataset is a large scale synthetic financial transactions dataset simulating mobile money transactions, containing 6,362,620 records across 11 features spanning 30 simulated days (744 hourly time steps). Transaction types include PAYMENT, TRANSFER, CASH_OUT, DEBIT, and CASH_IN. The dataset exhibits extreme class imbalance, with fraudulent transactions constituting only approximately 0.0536% of all records. A critical structural finding is that fraud occurs exclusively in TRANSFER and CASH_OUT categories; no fraudulent PAYMENT, DEBIT, or CASH_IN transactions appear in the dataset [11].

TABLE I. DATASET SUMMARY

Property	Environment A (Card)	Environment B (Transfer)
Records	1,852,394	6,362,620
Raw Features	23	11
Fraud Rate	0.58%	0.0536%
Imbalance Ratio	172:1	~1,865:1
Temporal Span	Simulated transactions	744 hours (30 days)
Algorithms Tested	3 (LR, RF, DT)	4 (DT, KNN, LR, RF)
Test Set Size	555,719	262,144

C. Preprocessing Pipeline

For Environment B, preprocessing encompassed categorical encoding of transaction type via one hot encoding, dimensionality reduction, feature scaling via StandardScaler for distance and margin based models, and a 75:25 train test split. Identifier columns (nameOrig and nameDest) were removed to prevent the same memorisation risk identified in Environment A. The feature specification notes that encoders should be fit on training data only and transform applied to held out data; fitting independently on each split produces inconsistent code assignments and a silently corrupted evaluation [12].

IV. METHODOLOGY

A. Problem Formulation

Financial fraud detection is formulated as a binary classification problem: given a set of transaction attributes, determine whether the transaction is fraudulent (class = 1) or legitimate (class = 0). The extreme rarity of fraudulent events, typically less than 1% of all transactions, creates a severe class imbalance that renders naive classifiers ineffective. A classifier that labels every transaction as legitimate can achieve over 99.9% accuracy while detecting zero fraud cases, rendering conventional accuracy metrics misleading [6].

B. Classification Algorithms

Multiple supervised classifiers representing distinct algorithmic families were evaluated across both environments to test whether the fraud signal is architecture dependent or resides in the feature construction.

Logistic Regression (LR): A linear model that estimates the probability of fraud as a function of a weighted linear combination of features passed through a sigmoid function. Logistic Regression provides strong interpretability and fast inference, making it valuable for regulatory compliance contexts where model decisions must be explainable [13].

Decision Tree (DT): A tree structured model that recursively partitions the feature space along individual feature thresholds. Decision Trees offer full decision path interpretability, allowing each prediction to be traced to a specific sequence of feature conditions. This transparency is valuable for audit and compliance purposes [14].

K Nearest Neighbours (KNN): An instance based method that classifies transactions by majority vote among the k closest training examples in feature space. KNN makes no assumptions about the data distribution but is sensitive to the curse of dimensionality and to class imbalance when unweighted [2].

Random Forest (RF): An ensemble of decision trees, each trained on a bootstrapped sample with random feature subsets. Random Forest provides robustness to noise and outliers, built in feature importance analysis, and near ideal probability calibration, properties that are particularly valuable for risk scoring applications where fraud probability estimates feed downstream risk management systems [15].

C. Evaluation Metrics

Given the extreme class imbalance, accuracy alone is insufficient. Models are evaluated using six complementary metrics: Precision (fraction of predicted frauds that are genuine), Recall (fraction of actual frauds detected), F1 Score (harmonic mean of precision and recall), AUC ROC (area under the receiver operating characteristic curve, measuring ranking quality across all thresholds), AUC PRC (area under the precision recall curve, more informative than ROC for highly imbalanced datasets), and Calibration (whether predicted probabilities match observed fraud frequencies). In financial fraud detection, false negatives (missed frauds) are typically far more costly than false positives (unnecessary investigation triggers), making recall and F1 the primary metrics of operational interest [16].

V. EXPERIMENTAL RESULTS

A. Environment A: Credit Card Transactions

Three classifiers were trained and evaluated on the credit card transaction dataset. Table II presents the results.

TABLE II. ENVIRONMENT A RESULTS (CREDIT CARD, 555,719 TEST RECORDS)

Model	Accuracy (%)	Dominant Predictor
Random Forest*	99.74	amt (r = 0.219)
Decision Tree	99.58	amt (r = 0.219)
Logistic Regression	99.41	amt (r = 0.219)

** Best performing model*

The Random Forest Classifier achieved the highest accuracy of 99.74%, demonstrating the superior generalisation capability of ensemble bagging over single classifiers under imbalanced conditions. However, the feature engineering specification accompanying this work expressly notes that precision, recall, and area under the precision recall curve were not directly measured in the original Environment A run; the confusion matrix in that write up was reconstructed from accuracy and class composition rather than measured. This is recorded as a known limitation and a defect that the specification corrects [12].

Critical analysis under the 172:1 class imbalance reveals that overall accuracy substantially overstates true fraud detection capability. Estimated confusion matrices suggest the Random Forest misses approximately 609 fraudulent transactions per 259,335 validation samples, representing significant undetected losses in a production environment. This finding motivates the more rigorous multi metric evaluation conducted in Environment B.

B. Environment B: Account Transfers

Four classifiers were trained and evaluated on the account transfer dataset using the full complement of evaluation metrics. Table III presents the comparative results.

TABLE III. ENVIRONMENT B RESULTS (ACCOUNT TRANSFERS, 262,144 TEST RECORDS)

Model	Acc.(%)	Fraud Prec.	Fraud Recall	Fraud F1	AUC ROC	AUC PRC

Decision Tree	99.94	0.74	0.68	0.71	0.84	1.00
KNN	99.94	0.71	0.44	0.54	0.72	0.82
Logistic Regression	99.93	1.00	0.23	0.37	0.61	0.36
Random Forest*	99.95	0.82	0.67	0.74	0.84	1.00

** Best performing model (highest fraud F1, best calibration)*

Random Forest achieves the best balance between precision (0.82) and recall (0.67) for the fraud class, yielding the highest F1 score of 0.74 and the strongest overall discrimination. Decision Tree and Random Forest are tied at AUC ROC of 0.84, significantly outperforming KNN (0.72) and Logistic Regression (0.61). Calibration analysis confirms that Random Forest produces near perfect probability estimates, closely tracking the ideal diagonal, which is critical for risk scoring applications.

Logistic Regression achieves perfect precision (1.00) for the fraud class but detects only 23% of actual frauds. This represents a catastrophically conservative boundary that misses 216 out of 279 fraudulent transactions in the test set. The geometric structure of fraudulent transactions in the feature space is nonlinear, requiring the more expressive hypothesis spaces of tree based models.

TABLE IV. CONFUSION MATRIX SUMMARY (ENVIRONMENT B TEST SET)

Model	True Neg.	False Pos.	False Neg.	True Pos.
Decision Tree	261,799	66	~89	~190
KNN	261,814	51	~157	~122
Logistic Regression	261,865	0	216	63
Random Forest	261,824	41	~92	~187

C. Cross Environment Transferability

The central finding of the experimental evaluation is cross environment consistency. In Environment A, three algorithms each exceeded 99% accuracy on a held out set of 555,719 records, with Random Forest at 99.74%, at a fraud rate of 0.58%. In Environment B, four algorithms independently exceeded 99% on a structurally different dataset at a fraud rate of 0.0536%. The two datasets share no columns, no records, and no fraud mechanics. The result holds across a linear model (Logistic Regression), a single decision boundary model (Decision Tree), and ensemble methods (Random Forest), indicating that the fraud signal is being captured by how the features are constructed rather than by any one model's architecture [12]. This property is what makes the methodology transferable to other institutions and what justifies its extension to the synthetic identity domain.

VI. FEATURE CONSTRUCTION PRINCIPLES

The feature engineering specification [12] documents five construction principles that generalise across both validated environments and carry forward into the synthetic identity feature set. Each principle is demonstrated with its concrete manifestation in both environments.

TABLE V. CROSS ENVIRONMENT CONSTRUCTION PRINCIPLES

Principle	Environment A (Card)	Environment B (Transfer)	Synthetic Identity Extension
Amount is the primary signal	amt (r = 0.219)	amount and balance deltas	Utilisation and drawdown magnitude
Fraud concentrates in specific channels	Card not present categories	TRANSFER and CASH_OUT only	Remote origination channels
Identifiers excluded, not used	cc_num, trans_num dropped	nameOrig, nameDest dropped	SSN, name, address never scored
Severe imbalance governs evaluation	0.58% fraud (172:1)	0.0536% fraud (~1,865:1)	Rarer still; label is worse
Signal survives model choice	3 algorithms > 99%	4 algorithms > 99%	Transferability established

VII. SYNTHETIC IDENTITY FEATURE SPECIFICATION

The feature engineering specification defines 20 measurements across six behavioural families for synthetic identity detection. All 20 are specified but none have been implemented or validated against a labelled dataset. This distinction is stated explicitly throughout the specification, and any claim that a synthetic identity classifier has been trained or validated would be inaccurate. The realistic first deployment is unsupervised or rule assisted ranking for manual review, not a supervised classifier, because the label problem has no clean solution [12].

A. The Label Problem

A synthetic identity that is never drawn down is never labelled. One that is drawn down is typically classified as a charge off rather than as fraud. Any institution training on historical labels is therefore training on a subset defined by which schemes concluded and were correctly identified afterwards. This fundamental constraint shapes the entire detection approach and is the reason the specification recommends unsupervised anomaly scoring as the initial deployment mode.

B. Feature Families

The six families are ordered by how early in the scheme they become visible. The first three are observable at origination, where an institution can still decline. The last three are observable only after an account is open, where an institution can still limit exposure before the drawdown.

Family 1, Identity Element Relationships: Measurements examining whether the combination of identity elements (name, SSN, date of birth, address) exhibits the coherence patterns characteristic of organic identities or the compositional anomalies characteristic of fabricated ones. Family 2, Credit File Assembly: Measurements examining how the credit file was constructed over time, including the age of the oldest tradeline, the velocity of new account openings, and the relationship between file age and credit depth. Family 3, Origination Behaviour: Measurements examining the application pathway, including channel (remote versus in branch), velocity of applications across institutions, and consistency of stated information across applications.

Family 4, Account Utilisation Trajectory: Measurements examining how credit lines are used after origination, including utilisation ramp rate, the relationship between utilisation and stated income, and whether the pattern is consistent with organic consumption or systematic capacity building. Family 5, Payment and Repayment Behaviour: Measurements examining repayment patterns, including minimum payment cycling, the timing of payments relative to statement dates, and sudden changes in payment behaviour that precede drawdown. Family 6, Cross Institutional Signals: Measurements requiring information from multiple institutions, including the number of institutions at which the identity holds accounts, the timing of account openings, and whether the borrowing pattern is consistent across institutions or exhibits the compartmentalisation characteristic of cultivated identities.

C. Data Availability Constraints

Several measurements in the specification require data that a small institution may not hold. Each such measurement is marked in the specification. The five measurements that require nothing beyond the institution's own records are identified specifically, providing an implementable starting point for institutions without access to bureau data or multi institutional information sharing arrangements.

D. Fairness and Fair Lending Compliance

The specification mandates that protected characteristics, including the sex field, be excluded from any deployed scoring model. Its correlation with the fraud label is indistinguishable from zero, and its presence in a lending decision creates fair lending exposure under ECOA for no detection gain. Geography is permitted only as a relational measurement (such as distance between stated address and transaction location) rather than as a direct locational feature, because locational features carry proxy discrimination risk. The specification requires that score and decline rate distributions be computed across protected classes after training, that the results be recorded, and that they be published with the model [12].

VIII. IMPLEMENTATION FRAMEWORK

The methodology is published as a package of four integrated components, each designed for an institution without dedicated analytics staff.

The Feature Engineering Specification defines all 42 features (22 validated across both transaction environments, 20 specified for synthetic identity), each with its exact computation, observation window, missing value handling, fraud risk rationale, and known limitations. Where a measurement cannot be computed from the data an institution holds, the specification says so rather than leaving the institution to discover it [12].

The Reference Implementation provides working, documented code that implements the validated transaction detection measurements and serves as the template for the specified synthetic identity measurements. It is released under an open licence so that institutions can inspect, modify, and deploy it without purchasing a commercial fraud detection system.

The Implementation Guide explains the data, computing, operating, and review requirements for deployment, structured around an institution with one technically capable person and a compliance officer. It provides implementation paths that begin with manual review of flagged transactions and progress to scripted batch scoring and origination flow integration.

The Threat Model Note explains the specific detection problem the features are designed to address: why conventional identity verification fails against synthetic identities, what behavioural signals remain visible, and what the detection method can and cannot do. It is written for a compliance officer rather than for a data scientist.

IX. DISCUSSION

A. Significance of Cross Environment Consistency

The most consequential finding of this study is not any individual model's accuracy figure but the consistency of results across structurally different environments. The two datasets share no columns, no records, and different fraud mechanics, yet the same feature construction approach produces strong discrimination in both. This is a claim about the methodology rather than about any model, and it is the property that justifies distributing the approach for other institutions to run on their own data.

B. Known Defects and Honest Limitations

The feature engineering specification records several known defects in the published implementation. In Environment A, encoders were fit independently on training and test splits using fit_transform rather than fit on training only and transform on test data; the corrected form is specified. Class imbalance was not addressed at training time in Environment A through cost sensitive weighting or resampling, which is consistent with the lower recall observed relative to Environment B's tree based models. Precision, recall, and area under the precision recall curve were not reported in the original Environment A run. In Environment B, fraud appearing exclusively in two transaction types is a property of how the dataset was generated and should not be assumed of a real payment system [12].

Both validated datasets are synthetic. They reproduce structure but not the missingness, coding drift, and adversarial adaptation of production data. Performance should be expected to fall on real payment data. The synthetic identity feature set is specified but has not been fitted, tuned, or evaluated against any label. These limitations are stated here and in the specification because an institution discovering them after deployment would have a legitimate complaint.

C. Why Comparability Matters

Card fraud operations run against institutions in many states simultaneously. Synthetic identities are cultivated deliberately across accounts at several institutions, precisely because no single institution observes the complete pattern. Detection capability confined to the largest banks therefore leaves both schemes viable. The same capability distributed across the smaller institution tier degrades them everywhere at the same time. That only works if a measurement computed at one credit union means the same thing as the measurement computed at another. This is why every definition in the feature engineering specification fixes the observation window in days rather than in qualitative terms, states the reference population explicitly, and names the data the measurement requires [12].

X. FUTURE DIRECTIONS

Several extensions of this work are planned or underway. First, implementation and validation of the 20 specified synthetic identity features against real lending data, in partnership with institutions willing to provide labelled outcomes. Second, application of advanced imbalance handling techniques including SMOTE, ADASYN, and cost sensitive learning to improve minority class recall in both transaction environments. Third, evaluation of gradient boosting methods (XGBoost, LightGBM) as potential improvements over Random Forest. Fourth, graph neural network approaches to capture inter account relationships and detect organised fraud rings operating across multiple accounts. Fifth, temporal modelling of transaction sequences using recurrent or attention based architectures. Sixth, federated learning for privacy preserving fraud detection across distributed financial institutions, enabling consortium based detection without centralising sensitive data.

XI. CONCLUSION

This paper presents a complete, reproducible machine learning methodology for detecting financial transaction fraud and extends it to address synthetic identity fraud in consumer lending. The methodology was validated across two structurally distinct payment environments totalling approximately 8.2 million records, with Random Forest achieving 99.74% accuracy in the credit card environment and 99.95% accuracy with an F1 score of 0.74 and AUC ROC of 0.84 in the account transfer environment. The cross environment consistency of these results establishes that the fraud signal resides in the feature construction rather than in any single model architecture.

The synthetic identity feature specification extends the validated detection approach to a structurally different problem domain, defining 20 measurements across six behavioural families designed to distinguish fabricated identities from organic ones. The specification honestly distinguishes validated measurements from those that remain to be tested, and records its own known defects and limitations.

The broader contribution is an open, reproducible detection capability designed for the more than 8,100 community banks and credit unions that serve a significant portion of American consumers but lack the analytical resources of larger institutions. The full methodology, including the feature engineering specification, reference implementation, implementation guide, and threat model note, is published under an open licence. An institution can obtain, inspect, implement, and test the approach without purchasing a commercial system or entering into a contract with the author. Distributing detection capability across the institution tier where fraud losses concentrate most heavily, and where detection has historically been weakest, directly serves the integrity of the financial system.

REFERENCES

[1] Nilson Report, "Global Card Fraud Losses," Issue 1209, 2022.

[2] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.

[3] Federal Reserve Bank, "Synthetic Identity Fraud in the U.S. Payment System," Federal Reserve System, 2021.

[4] National Credit Union Administration, NCUA Quarterly Data, Q1 2026.

[5] Federal Deposit Insurance Corporation, FDIC Quarterly Banking Profile, Q1 2026.

[6] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in Proc. IEEE Symp. CIDM, 2015, pp. 159–166.

[7] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," Information Sciences, vol. 479, pp. 448–455, 2019.

[8] J. Kim and S. Kim, "Anomaly detection in financial transactions using LSTM autoencoders," in Proc. IEEE Int. Conf. Big Data, 2021.

[9] Y. Liu, X. Li, and S. Chen, "Graph based approaches for detecting organised fraud rings," in Proc. ACM KDD, 2022.

[10] A. Ayeni, "Credit Card Fraud Detection Using Machine Learning," .

[11] A. Ayeni, "Financial (Credit and Debit Card) Fraud Detection Using Machine Learning Classifiers,", February 2025.

[12] A. Ayeni, "Feature Engineering Specification: Transaction and Synthetic Identity Fraud Detection for Smaller Institutions," Open methodology artefact, published June 24, 2026. Available: www.ayomideayeni.com

[13] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013.

[14] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[15] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[16] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[17] A. Ayeni, "Implementation Guide: Transaction and Synthetic Identity Fraud Detection for Smaller Institutions," Open methodology artefact, published June 24, 2026. Available: www.ayomideayeni.com

[18] A. Ayeni, "Threat Model Note: Synthetic Identity Fraud," Open methodology artefact, published June 24, 2026. Available: www.ayomideayeni.com