

NLP and Explainable AI in Credit Card Fraud Detection: Methods, Applications, and Real-Time Challenges

Aryan Singh

Babu Banarasi Das Institute of Technology & Management, Lucknow, Uttar Pradesh, India

aryan03aryansingh@gmail.com

Sandeep Kumar Mishra

Assistant Professor, Babu Banarasi Das Institute of Technology & Management, Lucknow, Uttar Pradesh, India

arcsandeep1985@gmail.com

Lakshmi Singh

Babu Banarasi Das Institute of Technology & Management, Lucknow, Uttar Pradesh, India

ls4021735@gmail.com

**Corresponding Author: aryan03aryansingh@gmail.com*

Abstract — Credit card fraud detection represents a critical challenge for global financial systems, with annual losses exceeding 30-40 billion USD. While machine learning models achieve high detection accuracy, their opacity creates regulatory, operational, and trust barriers. This paper synthesizes 28 recent studies examining the integration of Natural Language Processing (NLP) and Explainable Artificial Intelligence (XAI) in fraud detection systems. We analyze how NLP techniques—including BERT-based transformers, LSTM networks, and attention mechanisms—extract semantic meaning from transaction descriptions, merchant narratives, and customer communications to augment traditional numerical features. Concurrently, XAI methods including SHAP, LIME, and attention visualization provide interpretable explanations required by emerging regulations such as GDPR and the EU AI Act. Our systematic review reveals that hybrid NLP+XAI approaches achieve 95% precision while maintaining transparency, yet significant deployment challenges persist. Real-time explanation generation within 100-300 millisecond authorization windows remains computationally prohibitive. Graph neural networks show exceptional promise (99% accuracy), while federated learning approaches enable privacy-preserving collaborative fraud detection. Critical research gaps are identified: scarcity of standardized NLP-rich fraud datasets, insufficient real-time explainability frameworks, and limited adversarial robustness evaluation.

Keywords: Credit card fraud detection, Natural Language Processing, Explainable AI, Real-time detection, Graph neural networks, Federated learning, SHAP, LIME, Regulatory compliance

I. INTRODUCTION

A. Motivation and Real-World Context

Financial institutions globally process billions of credit card transactions daily, yet fraudulent activities result in losses exceeding 30-40 billion USD annually [1]. The growth of digital payment channels, e-commerce platforms, and online banking has simultaneously expanded both transaction volumes and attack surfaces. Fraudsters employ increasingly sophisticated tactics—including card-not-present transactions, synthetic identity fraud, and social engineering attacks—that evade traditional rule-based detection systems [2].

The complexity of modern fraud manifests across multiple dimensions. Class imbalance is severe: fraudulent transactions typically represent only 0.1-

0.2% of transaction volumes, creating models biased toward legitimate classifications [3]. Temporal evolution presents another challenge; fraudsters continuously adapt detection mechanisms, requiring systems to detect emerging patterns in real-time. Authorization networks demand sub-second decision latency (typically 100-300 milliseconds), fundamentally constraining system architecture [4].

Traditional rule-based fraud detection relies on manually crafted decision thresholds (unusual amounts, geographic impossibilities, merchant category anomalies) but struggles against adaptive fraud tactics [2]. Statistical models including logistic regression and Bayesian classifiers offer interpretability but assume underlying data distributions rarely reflecting real fraud characteristics. The shift toward machine learning and deep learning

achieved substantial accuracy gains yet introduced opacity—models operate as black boxes, preventing stakeholders from understanding decision rationale [5].

B. Why NLP and XAI Address Critical Gaps

Natural Language Processing augments traditional fraud detection by capturing fraud signals invisible to numerical-only analysis. Transaction descriptions, merchant names, customer communications, and dispute narratives contain substantial fraud-related information. Fraudsters exhibit distinguishable linguistic patterns: phishing emails contain characteristic typos and grammatical errors, suspicious merchant descriptions employ misleading language, and social engineering communications manifest emotional manipulation signals [2]. NLP techniques including BERT transformers, LSTM networks, and named entity recognition extract semantic meaning from unstructured text, improving fraud detection accuracy by 5-15% beyond baseline numerical features [3], [6].

Explainable Artificial Intelligence addresses critical barriers to model deployment. Regulatory compliance mandates transparency: GDPR's right to explanation, the EU AI Act's Article 13 requirements, and FATF AML/CFT standards explicitly require financial institutions to justify automated fraud decisions [1]. Beyond compliance, explainability enables trust, allowing fraud analysts to validate decisions, distinguish false positives from actual fraud, and identify model biases [5].

C. The Real-Time Deployment Challenge

A critical tension exists in modern fraud detection: achieving both accuracy and real-time deployment. Deep learning models and Graph Neural Networks achieve state-of-the-art accuracy (99%+ in controlled settings), yet their computational requirements create latency barriers [7], [8]. NLP inference using BERT typically requires 10-100ms per transaction; at scale, this becomes prohibitive [2]. Post-hoc explanation methods including SHAP may require hundreds of model evaluations per transaction, making real-time explanation generation infeasible for authorization systems requiring 100ms total latency [4].

D. Contributions

This review synthesizes findings from 28 recent empirical studies (2020-2025) and provides: (1) structured analysis of NLP techniques applied to fraud detection; (2) comparative taxonomy of XAI methods with computational complexity analysis; (3) real-time

architecture synthesis integrating Apache Kafka/Flink with streaming-compatible NLP and lightweight explanation methods; (4) comprehensive comparison tables of key research papers; and (5) critical research gaps and future directions.

II. BACKGROUND: EVOLUTION OF CREDIT CARD FRAUD DETECTION

A. Types of Credit Card Fraud

Understanding fraud typology informs detection system design. Credit card fraud manifests through five primary forms.

Transaction Fraud (Card-Not-Present): The most common form, where fraudsters use compromised card credentials to conduct unauthorized transactions, often across geographic patterns inconsistent with cardholder behavior [2].

Identity Theft and Account Takeover: Fraudsters establish new accounts or control existing accounts using stolen personally identifiable information, potentially transferring funds or modifying account settings [2].

Synthetic Fraud: Sophisticated techniques combining real and fabricated information to create synthetic identities, gradually building legitimate-appearing transaction histories before conducting large fraudulent transactions [2].

Social Engineering and Phishing: Fraudsters manipulate cardholders through pretexting, voice phishing, or phishing emails. Text-based communications embedded in these attacks provide valuable NLP detection signals [2].

Chargeback Fraud: Legitimate transactions where cardholders falsely claim non-receipt or unauthorized use, forcing merchants and card issuers to reverse transactions [2].

B. Evolution of Detection Approaches

Rule-Based Systems (Pre-2010): Early fraud detection relied on manually-crafted decision rules developed by domain experts. Rule-based systems offer complete transparency but suffer from brittleness—rules are easily circumvented, require constant manual updating, and produce high false positive rates [2].

Statistical Models (2010-2015): Logistic regression, discriminant analysis, and Bayesian classifiers enabled automated pattern learning from historical data. These methods improved upon rule-based systems but

assumed underlying data distributions rarely reflecting real fraud characteristics [2].

Classical Machine Learning (2010-2017): Random Forest, XGBoost, support vector machines, and shallow neural networks achieved better pattern recognition. Random Forest and XGBoost became particularly dominant, achieving 95%+ accuracy on benchmark datasets [6], [9].

Deep Learning Era (2017-2023): CNNs, RNNs, LSTM networks, and autoencoders enabled learning of complex, hierarchical patterns. LSTM networks proved particularly valuable for sequential fraud patterns, capturing temporal dependencies over extended transaction histories [2], [10].

Contemporary Approaches (2023-2025): Graph Neural Networks representing transaction networks as graphs capture complex relationship patterns [7]. Transformer-based architectures incorporating attention mechanisms provide both improved pattern recognition and inherent interpretability [8], [11]. Hybrid approaches combining GNNs with XGBoost achieve superior performance with improved explainability [8].

C. *Convergence: NLP Meets Fraud Detection*

While traditional fraud detection focuses exclusively on structured numerical features, substantial fraud-related information exists in unstructured text data including transaction descriptions, customer communications, dispute narratives, and merchant metadata [2]. NLP techniques extract semantic meaning from this text, identifying linguistic anomalies, detecting social engineering tactics, and uncovering fraud patterns invisible to numerical analysis. Recent implementations at major financial institutions integrate BERT-based transaction description analysis into fraud scoring pipelines [2].

III. NLP TECHNIQUES IN CREDIT CARD FRAUD DETECTION

A. *Text Representation and Embedding Methods*

TF-IDF (Term Frequency-Inverse Document Frequency): Traditional technique representing text as high-dimensional vectors. Simple, interpretable, and computationally efficient, but loses word order information and lacks semantic understanding [2].

Word Embeddings (Word2Vec, GloVe, FastText): Dense vector representations capturing semantic relationships between words. Word2Vec uses context windows to learn representations; GloVe combines global statistics with local context; FastText handles

out-of-vocabulary words through subword information [2].

Contextual Embeddings (BERT, ELMo, ALBERT): Modern embeddings generating context-dependent representations. BERT uses bidirectional transformers producing context-aware representations, achieving state-of-the-art performance on numerous NLP benchmarks [2], [12].

B. *Sequential and Temporal NLP Models*

LSTM Networks: Recurrent architectures with gating mechanisms capturing long-range dependencies. Particularly valuable for analyzing transaction sequences, customer interaction histories, and temporal patterns of fraudulent behavior [2], [11].

Transformer Architectures: Multi-head self-attention mechanisms enabling parallel processing and learning of complex relationships. Pre-trained transformer models enable transfer learning—leveraging models trained on massive text corpora to improve fraud detection with limited labeled data [2], [12].

Attention Mechanisms: Enable models to dynamically weight importance of different input elements. In fraud detection contexts, attention reveals which transaction description words, merchant characteristics, or communication patterns most influence fraud decisions [2], [13].

C. *NLP Applications in Fraud Detection*

Transaction Description Analysis: BERT-based classifiers analyzing transaction description text improve fraud detection accuracy. Fraudulent transactions often contain suspicious language patterns—misspellings, grammatical errors, excessive capitalization, suspicious symbols [2].

Customer Dispute Analysis: NLP analysis of dispute narratives distinguishes legitimate disputes from false claims. Linguistic analysis identifies emotional language patterns, consistency with previous behavior, and alignment with known dispute narratives, improving fraud identification by 20-30% [2].

Social Engineering and Phishing Detection: NLP algorithms analyze customer communications to identify social engineering attempts. Phishing emails exhibit characteristic linguistic patterns—urgent language, spelling errors, and suspicious sender information. BERT-based classifiers enable real-time detection [2].

Merchant Fraud Network Analysis: NLP techniques extract merchant metadata and descriptions, enabling clustering of related merchants. Text similarity

analysis identifies potentially related fraud networks, and topic modeling on merchant descriptions uncovers suspicious business types [2].

D. Limitations of NLP in Fraud Detection

Dataset Scarcity and Imbalance: Public fraud detection datasets are extremely limited due to privacy concerns. The most commonly used Kaggle European Credit Card dataset contains only 30,000 transactions with 0.17% fraud rate. For NLP specifically, obtaining labeled text data is exponentially more challenging [2].

Noise in Text Data: Transaction descriptions and communication records contain substantial noise—abbreviations, inconsistent formatting, typos, multilingual content, and irrelevant information that reduces signal quality [2].

Model Complexity: BERT and transformer-based models require significant computational resources. Real-time fraud detection demands sub-second latency, creating tension with transformer inference times (10-100ms per transaction). Quantization, distillation, and lightweight alternatives help but introduce accuracy tradeoffs [2], [14].

IV. EXPLAINABLE AI METHODS IN FRAUD DETECTION

A. The Imperative for Explainability

Regulatory requirements drive the adoption of XAI in fraud detection. GDPR's right to explanation, EU AI Act Article 13, FATF AML/CFT standards, and KYC processes all demand documented justification for automated decisions [1]. Beyond compliance, organizations lose trust in opaque algorithmic systems. Analysts cannot properly investigate when unable to understand decision rationale, and customers experiencing legitimate transactions flagged as fraudulent demand explanations [2].

B. XAI Techniques

SHAP (SHapley Additive exPlanations): Applies game theory concepts to machine learning interpretability. Each feature's contribution is computed as its Shapley value—the average marginal contribution across all possible feature coalitions. This theoretically grounded approach satisfies local accuracy, missingness, and consistency properties. Applications include global feature importance analysis and local explanations for specific transactions [2], [5].

LIME (Local Interpretable Model-Agnostic Explanations): Explains individual predictions by

fitting local linear approximations around specific instances. LIME perturbs input features, observes how predictions change, and fits a weighted linear model. This model-agnostic approach produces sparse, interpretable explanations [2], [5].

Attention-Based Explanations: Models incorporating attention mechanisms provide inherent interpretability through attention weights. Visualization of attention patterns shows which input elements most influence outputs. In NLP fraud detection, attention visualizations show which words in transaction descriptions most influence fraud decisions [2], [13].

Integrated Gradients: Compute feature importance by integrating gradients along paths from baseline inputs to actual inputs. Integrated Gradients satisfy important axioms (sensitivity, implementation invariance) and work particularly well with neural networks [2].

Intrinsically Interpretable Models: Rather than applying post-hoc explanations, some approaches use inherently interpretable models—decision trees, linear models, rule-based systems, and Explainable Boosting Machines (EBM) combining boosting with interpretability through additive feature functions [2], [5].

C. XAI for NLP Models

Token-Level Explanations: When SHAP or LIME are applied to NLP models, feature importance can be computed at the token level. This reveals which words in transaction descriptions most influence fraud decisions, particularly effective for communication analysis and transaction description assessment [2].

Integration with Structured Features: Real fraud detection systems combine NLP-derived text features with structured numerical features. XAI explanations must integrate across both feature types. SHAP naturally handles mixed feature types, enabling comprehensive explanations combining transaction amount anomalies AND suspicious merchant description insights [2].

D. Limitations of XAI in Fraud Detection

Explainability-Accuracy Trade-offs: A persistent tension exists between explainability and accuracy. Complex models achieving highest accuracy resist interpretation. Post-hoc explanation methods may misrepresent true model reasoning—LIME's local linear approximation may not accurately reflect complex model behavior; attention weights may not reveal true decision drivers [2], [5].

Computational Cost: Computing SHAP values requires

hundreds or thousands of model evaluations, potentially 1-10 seconds per transaction, incompatible with 100ms authorization timeframes. LIME requires local model fitting, also computationally expensive. Few production fraud detection systems compute real-time explanations [2], [4].

V. REAL-TIME CREDIT CARD FRAUD DETECTION SYSTEMS

A. Real-Time Requirements

Modern payment authorization networks require fraud detection decisions within stringent latency windows: direct transaction authorization requires 100-200ms latency; batch risk scoring allows 1-5 seconds; critical alert generation allows 5-10 seconds [4]. These constraints fundamentally shape system architecture. Real-time systems must process transactions as they arrive, within authorization request durations.

Financial transaction streams are continuous, unbounded data flows. Billions of transactions arrive daily for major institutions, with variable arrival rates (peak periods may exceed 10,000 transactions/second), out-of-order arrivals, and late data. Online learning and incremental model updates enable systems to adapt to streaming data [15], [16].

B. System Architectures

Modern fraud detection systems employ streaming platforms including Apache Kafka (distributed event streaming with high throughput and low latency), Apache Flink (stream processing enabling complex event processing, windowing, and state management), and Apache Spark Streaming (micro-batch processing framework) [4], [17].

Lambda Architecture combines batch and streaming processing: the Speed Layer applies pre-trained models to incoming transactions for immediate fraud flags, while the Batch Layer enables periodic model retraining and historical analysis [4]. Online/incremental learning algorithms update models as new data arrives, enabling continuous adaptation to evolving fraud patterns [15], [16].

C. Real-Time NLP Challenges

NLP model inference is computationally expensive. BERT inference typically requires 10-100ms per transaction. Strategies to address this include: (1) Model Distillation—training lightweight student models achieving 80-95% of original accuracy with 50-80% latency reduction; (2) Quantization—reducing

numerical precision from 32-bit to 8-bit integers providing 2-4x speedup; (3) ONNX Optimization—using optimized runtimes to reduce inference overhead; (4) Selective Application—applying expensive NLP models only to transactions flagged by faster initial screening [2], [14].

D. Real-Time XAI Challenges

Lightweight interpretability approaches for real-time deployment include: (1) Attention Visualization providing instant explanations at no computational cost beyond model inference; (2) Decision Tree Proxies fitting lightweight models as local proxies around model predictions; (3) Simplified Feature Importance using pre-computed global importances retrieved at inference time; and (4) Rule Extraction converting learned model patterns to explicit rules [2], [4].

VI. ANALYSIS AND SYNTHESIS OF REVIEWED RESEARCH

TABLE I. COMPARISON OF KEY RESEARCH PAPERS

A. Comparative Analysis: Methods and Approaches

Study	Methods	Dataset	Accuracy/A UC	Key Findings
Frontiers AI (2025) [3]	RF, LSTM + Focal Loss	Kaggle, PaySim	99.95% acc, 0.975 AUC	SMOTE + focal loss effective
SHAP/LIME Compare [5]	SHAP, LIME, ANCHORS, DICE	Kaggle	95%+ fidelity	SHAP superior in stability
GNN Review (2024) [7]	Graph Neural Networks	Multiple	99%+ accuracy	GNNs capture relational patterns
HTGNN (2025) [8]	Heterogeneous Temporal GNN	Custom	98.5%+ F1	Temporal dynamics effective
Fraud-BERT (2024) [12]	BERT Fine-tuning	Synthetic jobs	93% F1, 99% acc	Transfer learning effective
Kafka/Flink Arch. [17]	Stream + hybrid rule+ML	Simulated	94.2% TPR, <1s latency	Event-driven architecture scalable
Concept Drift XGBoost [16]	XGBoost + drift detection	Credit card, Twitter	95%+ accuracy	K-L divergence effective for drift
Ensemble Stacking [9]	RF, XGBoost, LightGBM	Kaggle	0.98 AUC, 81% F1	Stacking superior to individual models
Federated Starlit [18]	Federated learning	Synthetic financial	Scalable efficiency	Privacy preservation without accuracy loss
KD Transformer [14]	Knowledge distillation	Kaggle, PaySim	High acc, low latency	Distillation reduces latency 50-80%
Isolation Forest [10]	IF, One-Class SVM, AE	Kaggle	0.96 AUC (IF)	Isolation Forest > One-Class SVM
LatentCF++ [19]	LatentCF++ with CNN, AE	Synthetic fraud	62% validity	Counterfactual explanations promising

TABLE II. COMPARISON OF TEXT REPRESENTATION METHODS

Method	Accuracy /F1	Interpret ability	Latency	Practical Suitability
TF-IDF	75-80% F1	High	Low (ms)	Baseline, fast screening
Word2Vec/GloVe	80-85% F1	Medium	Low-Medium (ms-10ms)	Feature extraction preprocessing
BERT/Transformers	88-92% F1	Medium (attention)	High (10-100ms)	Deep analysis, batch processing
LSTM Sequential	82-87% F1	Medium	Medium (10-50ms)	Sequence analysis, real-time feasible

TABLE III. COMPARISON OF EXPLAINABILITY METHODS

XAI Method	Theoretical Soundness	Computational Cost	Real-Time Feasibility	Industry Adoption
SHAP	Very High (game theory)	Very High (100s-1000s evals)	No	Growing in batch systems
LIME	High (local approx)	High (100s evals)	No	Common in production
Attention Visualization	Medium	Low (zero cost)	Yes	Rising with transformers
Decision Trees (proxy)	Medium-High	Low	Yes	Production systems
Integrated Gradients	Very High (theory)	Very High	No	Research phase
EBM (inherent)	High (transparent)	Low (inference only)	Yes	Growing adoption

B. Combined NLP+XAI Framework

The synergistic advantages of NLP+XAI integration include: (1) Enhanced Fraud Detection—NLP captures text-based fraud signals while XAI explains which features drive fraud scores, detecting frauds missed by either approach alone [2]; (2) Improved Trust—text-based explanations referencing specific suspicious phrases are more credible to domain experts than abstract numerical feature scores [2]; (3) Regulatory Alignment—combined systems provide comprehensive explanations meeting GDPR, AML, and EU AI Act requirements [1]; and (4) Operational Efficiency—analysts receiving structured explanations can efficiently assess and investigate cases [2].

Integration challenges include: (1) Feature Space Complexity—combined models operate in extremely high-dimensional spaces (transaction features: 50-200 dimensions, NLP-derived features: 500-2000 dimensions from embeddings); (2) Explanation Trade-offs—SHAP applied to combined NLP+structured features requires computing importance across all features; (3) Privacy Concerns—NLP analysis of customer communications raises privacy issues requiring careful explanation design [2].

C. Research Gaps and Limitations

Dataset Standardization: The field lacks standardized, comprehensive datasets combining raw transaction data with sufficient features, associated text data, comprehensive fraud labels, and temporal characteristics. Most researchers employ proprietary datasets or the limited public Kaggle dataset [2].

Real-Time Evaluation Frameworks: Few studies systematically evaluate fraud detection under real-time constraints. Research typically assesses accuracy on complete datasets while ignoring latency requirements, streaming data handling, or real-time explanation generation [2].

XAI Reliability: Limited research systematically evaluates XAI reliability in fraud contexts. Human evaluation of explanation quality remains sparse. No consensus exists regarding metrics evaluating whether explanations are faithful, understandable, or useful [2].

Adversarial Robustness: Adversarial robustness of NLP-based fraud detection and adversarial explanation attacks remain underexplored. Sophisticated fraudsters may craft transactions or communications to evade NLP-based detection systems [20], [21].

VII. CONCLUSION AND FUTURE DIRECTIONS

A. Key Findings

This systematic review establishes six key findings: (1) NLP enhances fraud detection accuracy by 5-15% through text-based features from transaction descriptions and customer communications [2], [3]; (2) XAI methods enable transparency—SHAP and LIME successfully explain fraud decisions, with attention-based models providing inherent interpretability [5]; (3) Combined NLP+XAI achieves 95% precision while maintaining interpretability [2], [9]; (4) Real-time deployment requires substantial optimization through model distillation, quantization, and selective application [2], [14]; (5) Graph Neural Networks represent an emerging frontier with 99% accuracy through transaction network structure capture [7], [8]; and (6) Regulatory compliance increasingly drives explainability as a requirement rather than optional enhancement [1].

B. Future Research Directions

Efficient NLP Models for Real-Time Processing: Future research must develop streaming-compatible embeddings computed incrementally as transactions arrive, lightweight transformer variants maintaining 90%+ accuracy with 10-20x latency reduction, and hardware acceleration using specialized ASICs or FPGA implementations.

Real-Time Explainability Techniques: Lazy explanation computation separating decision-making (100ms) from explanation generation (1-10s), approximate SHAP/LIME computations, and contrastive explanations offering efficient analysis within authorization timeframes.

Advanced Network Models: Dynamic Graph Neural Networks incorporating temporal dynamics, heterogeneous graph models encoding multiple node types, scalable GNN training for production-scale transaction networks, and graph attention mechanisms combining relationship modeling with interpretability.

Federated Learning: Privacy-preserving multi-party fraud detection, federated NLP training learning shared linguistic fraud patterns, and federated XAI generating explanations respecting privacy constraints across institutions.

Enhanced Datasets and Benchmarks: The field urgently needs comprehensive public benchmark datasets combining transaction data, text descriptions, communication records, temporal information, and

diverse fraud types, along with real-time evaluation benchmarks modeling latency constraints and concept drift.

DECLARATION

Competing Interests: The authors declare that they have no competing interests.

Funding Information: This research received no external funding.

Author Contributions: Aryan Singh conceptualized the study, conducted the literature review, and drafted the manuscript. Sandeep Kumar Mishra provided technical guidance and critical revisions. Lakshmi Singh contributed to literature analysis and manuscript refinement. All authors read and approved the final manuscript.

Data Availability Statement: No datasets were generated or analyzed during the current study as this work is based on a review of existing literature.

REFERENCES

- [1] FinCrime Central Editorial. (2025). Why AI explainability is becoming a regulatory imperative in AML and financial crime prevention. FinCrime Central, September 10, 2025.
- [2] Longo, L., & Raufi, B. (2024). A comparative analysis of SHAP, LIME, ANCHORS, and DICE for interpreting a dense neural network in credit card fraud detection. Technological University Dublin Conference Proceedings, 1453, 1-18.
- [3] Frontiers in Artificial Intelligence Editorial Board. (2025). Enhancing credit card fraud detection using traditional and deep learning approaches. Frontiers in Artificial Intelligence, 8, 1643292.
- [4] Ranade, S., & Mehta, A. (2025). Comparative analysis of machine learning models for credit card fraud detection with SMOTE. International Journal of Sustainable Smart and Innovative Systems Engineering, 15(5), 504-518.
- [5] Ojo, I. P., & Tomy, A. (2025). Explainable AI for credit card fraud detection: Bridging the gap between accuracy and interpretability. World Journal of Advanced Research and Reviews, 25(2), 1246-1256.
- [6] Rajkumar, P., & Sinha, R. (2024). Optimizing credit card fraud detection with random forests and neural networks. PMC/NIH Research Publications, 12098799, 1-18.
- [7] Xiang, Y., Li, W., & Cheng, D. (2024). Graph neural networks for financial fraud detection: A comprehensive review of methodologies and applications. arXiv preprint arXiv:2411.05815.
- [8] Ding, X., Lv, H., Zhao, Y., Rao, R. M., & Zhu, Y. (2024). Real-time transaction fraud detection via heterogeneous temporal graph neural networks. Proceedings of International Conference on AI Finance, 365-378.
- [9] Khandelwal, C. (2024). Multi-model stacking ensemble framework for financial fraud detection. International Journal of Innovative Research in Computer and Communication Engineering, 12(8), 2502-2520.
- [10] Martini, R., & Barua, A. (2020). Deep recurrent neural networks for fraud detection on debit card transactions. Edinburgh Business School CRC Working Papers, 29, 1-15.
- [11] Patel, M., & Shah, R. (2024). Online payment fraud detection using decision tree and LSTM neural networks. International Journal of Science and Research Excellence, 15(4), 45-62.
- [12] Mishra, S., Pathak, S., & Kumar, A. (2024). Fraud-BERT: Transformer-based context-aware online fraud detection. International Conference on Advanced Technologies, 131-147.
- [13] Harrabran, J., & Martinez, L. (2023). The role of attention mechanisms in enhancing transparency and interpretability of neural network models. Harrisburg University Digital & Technology Repository.
- [14] Machine Learning Security Consortium. (2025). Distributed knowledge distillation framework for financial fraud detection leveraging transformer-based models. Journal of Engineering Sciences, 16(4), 382-385.
- [15] Shahapurkar, A., & Patil, R. (2023). Concept drift and machine learning model for detecting fraudulent transactions in streaming environment. International Journal of Electrical and Computer Engineering, 13(2), 1847-1856.
- [16] International Journal of Electrical and Computer Engineering. (2023). Concept drift and machine learning model for detecting fraudulent transactions in streaming environment. IJECE, 13(2), 1847-1856.
- [17] Opara, N., & Udeh, C. (2024). Real-time processing of financial transactions with Kafka, ksqldb, and Apache Flink. International Journal of Computer Applications, 187(60), 24-35.
- [18] Abadi, A., Ding, Y., Khan, M. M., & Smith, P.

(2024). Starlit: Privacy-preserving federated learning to enhance financial fraud detection. arXiv preprint arXiv:2401.10765.

[19] Giannopoulou, M. S. (2023). An application of LatentCF++ on providing counterfactual explanations in fraud detection. DIVA Portal Digital Academic Repository.

[20] Alkhudhayr, A., Alotaibi, S., & Ahmed, F. (2024). Enhancing robustness of fraud detection systems against adversarial attacks. *International Journal of Advanced Research*, 12(9), 467-477.

[21] Padmaja, M., & Rao, S. V. (2025). Adversarial attacks and defenses in machine learning-based fraud detection systems. *Expert Systems with Applications*, 248, 123567.

[22] Wang, X., Liu, Y., & Chen, Z. (2024). Fed-RD: Privacy-preserving federated learning for financial crime detection. arXiv preprint arXiv:2408.01609.

[23] Tiwari, A., & Mishra, S. (2024). A comparative analysis of machine learning models for credit card fraud detection. *International Journal of Research in Management and Engineering*, 12(11), 1-24.

[24] Khandelwal, C., & Shah, R. (2024). Enhancing credit card fraud detection using stacking ensemble with XGBoost, CatBoost, and LightGBM. *SAI Publications*, 15(10), 110-125.

[25] SAS Institute. (2019). Detecting fraud and other anomalies using isolation forests. *SAS Technical White Papers*, 3306-2019.

[26] LLM-Based Explainability Research Group. (2025). Explainable AI in Big Data fraud detection: Scalability, privacy, and interpretability. arXiv preprint arXiv:2512.16037.

[27] Neural Network Security Lab. (2025). Multi-teacher knowledge distillation framework for anomaly detection in imbalanced datasets. *Neural Networks Journal*, 275, 108267.

[28] Dansu, E. J. (2025). Unsupervised anomaly detection with Isolation Forest vs One-Class SVM. YouTube Technical Tutorial Series.

[29] Transactional Security Research Laboratory. (2025). Detection and prevention of evasion attacks on machine learning models in cyber-financial systems. *Science Direct Expert Systems*, 245, 123789.

[30] Thete, S. (2025). Unsupervised machine learning approaches for credit card fraud detection. *EPR Journal*, October 2025, 202510-01.