

**NFSU JOURNAL OF FORENSIC SCIENCE**

ISSN: 3049-2408 (Online)

Journal Homepage: <https://jfs.nfsu.ac.in/>

Detecting Adversarial Agents in Telecom Financial Networks Through Inter-Agent (A2A) Communication Behaviour)

Aditya Mohan Srivastava*, Panem Charanarur, Naveen Kumar Chaudhary, Dr. Lokesh Chouhan

Department of Cyber Security and Digital Forensics,

National Forensic Sciences University, Tripura Campus, Agartala, Tripura, India

Department of Cyber Security and Digital Forensics, National Forensic Sciences University,

Goa Campus, Curti, Ponda, Goa, India

Abstract—As financial ecosystems migrate toward an architecture composed of autonomous software agents representing banks, telecom operators, payment processors, and end users. The communication channels between these agents become high-value targets for adversarial exploitation. Conventional fraud detectors operate on transaction-level attributes such as amount and type, and are blind to the behavioural signatures embedded in inter-agent message exchange. This paper proposes a detection framework that extracts twelve communication behavioral features (response time, message length, retry count, session anomaly score, authentication failure rate, interaction frequency, and others) from agent-to-agent (A2A) traffic and feeds them through a two stage pipeline: (i) an autoencoder trained exclusively on legitimate interactions to produce an eight dimensional latent embedding and a scalar reconstruction error anomaly score, and (ii) an XGBoost classifier that operates on the 21 dimensional fusion of raw and autoencoder derived features. We evaluate the system on PaySim, a widely used financial dataset containing 6.36 million records, under a deliberately balanced protocol (50/50 fraud-legitimate split) that avoids the metric inflation common in imbalanced evaluations. Across six distinct attack types : (i) impersonation (ii) prompt injection (iii) collusion (iv) trust-chain exploitation (v) data exfiltration, and (vi) transaction manipulation. The proposed framework attains an F1-score of 0.933, AUC-ROC of 0.983, and per-attack F1 ranging from 0.928 (collusion, the stealthiest category) to 1.000 (transaction manipulation). A complementary four-tier graduated mitigation strategy (Monitor/Throttle/Quarantine/Block) achieves 93.52% fraud prevention, marginally exceeding binary blocking (93.27%) while reducing friction for legitimate users by 30%. We discuss the implications for

emerging agent-to-agent (A2A) ecosystems expected in 2028 and beyond, in which autonomous AI agents will negotiate, transact, and settle on behalf of human principals without direct human oversight.

Keywords—multi-agent systems, agent-to-agent communication, fraud detection, autoencoder, XGBoost, adversarial attacks, graduated mitigation, telecommunications security, agentic AI

I. INTRODUCTION

The telecommunications sector has consistently ranked among the industries hardest hit by financial fraud, with global losses projected to surpass USD 41.82 billion in 2025 [1]. Adoption of machine learning-based fraud prevention among telecom operators has climbed from 38% in 2023 to 75% by early 2025 [1], and the AI-powered fraud detection market's projected to grow from USD 3.6 billion to USD 13.7 billion between 2025 and 2032 [2]. These figures underscore two realities: the problem is large, and the industry is actively searching for better solutions.

A parallel trend is accelerating the urgency. Financial architectures are rapidly moving toward multi-agent designs in which autonomous software agents representing users, banks, telecom providers, and payment processors—communicate with one another to authenticate, authorise, and settle transactions [3]. By 2028, autonomous AI agents are expected to negotiate on behalf of human principals, handling end-to-end workflows with minimal human oversight [4], [5]. Telefónica and Nokia have already piloted AI software agents to prevent SIM-swap and device-swap fraud [6]. Once these agents conduct transactions autonomously, the inter-agent communication channel itself becomes a high-value attack surface.

Yet the vast majority of published fraud detectors still analyse per transaction attributes like amount, type, account balance, and time step. While treating the communication layer as a transparent pipe [7]–[9]. This paper argues that the behavioural patterns within agent-to-agent interactions carry discriminative signals that are invisible at the transaction level. Response timing, message length, retry behavior, session anomaly profiles, and authentication failure rates differ measurably between legitimate and adversarial traffic, even when the underlying transaction parameters appear normal.

A. Contributions

To the best of our knowledge, we make four concrete contributions:

- 1) A **novel feature space** of twelve communication behavioral attributes derived from inter agent message exchange, distinct from the raw transaction features used in prior work.
- 2) A **hybrid detection architecture** that fuses autoencoder derived latent representations with an

XGBoost gradient boosted classifier, evaluated under a balanced protocol that avoids the metric inflation endemic to imbalanced financial datasets.

TABLE I
PUBLISHED ML RESULTS ON PAYSIM DATASET

Ref Method	F1 Score	AUC	Balanced Dataset	Year
[11] XGBoost + SMOTE	0.991	0.991	No	2024
[12] Rule-based	0.998	0.991	No	2024
[13] Decision Tree	0.802	0.985	No	2025
[13] Random Forest	— ^a	0.999	No	2025
[14] Voting Ensemble	0.881	—	No	2025
[15] ML (Various)	—	—	—	2025
Ours (AE + XGBoost)	0.933	0.983	Yes	2025

Recall=0.984, Precision=0.551; F1 not reported.

- 3) A **threat taxonomy** of six attack types targeting agent communication, each with independently measured detection performance.
- 4) A **graduated mitigation strategy** with four response tiers calibrated by risk score, demonstrating a 30 % reduction in user friction relative to binary blocking while matching fraud-prevention rates.

II. RELATED WORK

A. Fraud Detection on the PaySim Dataset

The PaySim financial dataset [10] contains 6.36 million mobile-money transaction records with a natural fraud rate of approximately 0.13 %. Because of its public availability and realistic simulation of mobile-money dynamics, PaySim has become a standard benchmark for financial fraud detection research. Table I summarizes recent published results on this dataset.

A critical observation is that the entries reporting $F1 > 0.99$ or $\text{accuracy} > 99\%$ operate on the original highly imbalanced dataset, where even a trivial “predict all legitimate” rule achieves 99.87 % accuracy¹. Our results are reported on a *balanced* 50/50 split, making all metrics i.e. accuracy, precision, recall, F1, AUC genuine measures of discriminative ability rather than reflections of class prior. When the WJARR Random Forest study [13] reports recall of 0.984 but precision of only 0.551, this is consistent with aggressive over-prediction on imbalanced data. Our balanced-set precision of

0.931 reflects a far lower false-positive rate.

B. Hybrid Autoencoder–Classifier Models

The fusion of unsupervised autoencoders with supervised classifiers has become a productive line of investigation. Table II compiles published results for autoencoder–XGBoost hybrids. Our framework’s F1 of 0.933 is competitive with the strongest published hybrid (0.941 [17]), bearing in mind that the latter operates on a heavily imbalanced dataset where accuracy and AUC are naturally inflated. Our key differentiator is not incremental metric improvement but the feature space: we detect fraud through behavioral patterns in agent communication rather than raw transaction attributes.

TABLE II
PUBLISHED AUTOENCODER–XGBOOST HYBRID MODELS

Ref	Method	F1 Score	AUC	Year
[7]	SAE-XGBoost	0.833	—	2023
[16]	AE-XGB (Prob. Threshold)	0.905	—	2023
[17]	AE + XGB (Deep)	0.941	0.999 ^b	2024
[18]	LSTM AE-XGB + Attention	— ^c	—	2024
[19]	Graph Auto-Encoder	0.810	0.890	2024
Ours	AE + XGB (Comm.)	0.933	0.983	2025

Evaluated on highly imbalanced credit-card set.

Precision=0.942, Recall=0.905; F1 not reported.

C. Adversarial Attacks on Multi-Agent Systems

A growing body of work maps the threat landscape for multi-agent architectures. The OWASP Top 10 for LLM Applications ranks prompt injection as the number-one vulnerability [20]. Lee and Park demonstrate self-replicating malicious prompts that propagate virus-like across interconnected LLM agents [21]. Cohen and Barak formalize the *Trust-Vulnerability Paradox*: as systems increase inter-agent trust for coordination, they simultaneously widen the attack surface for trust-chain exploitation [22]. Palo Alto Networks documents *agent session smuggling*, where attackers inject covert instructions into stateful inter-agent channels [23]. GNN-based detection of hidden collusion networks in financial graphs has been explored by several groups [24], [25]. Kaspersky’s 2025 analysis reveals structural vulnerabilities in AI agent systems—insecure configurations, plaintext API-key storage, and malicious skill injection—that compound the risks of agent-to-agent fraud [4].

¹ At 0.13% base fraud rate, a model that never flags fraud achieves 99.87% accuracy, which is trivially

unbeatable but practically useless.

D. Graduated Mitigation and User Friction

Traditional binary blocking (allow/deny at a fixed thresh-old) imposes friction on all flagged users, many of whom are legitimate. A 2024 study in the International Journal of Science and Advanced Technology reports that multi-tiered response frameworks reduce customer friction by 43.7 % relative to binary approaches while improving fraud detection by 29.2 % [\[26\]](#). Risk-based authentication (RBA) adjusts security measures dynamically: low-risk interactions pass silently; moderate-risk interactions trigger step-up verification; high-risk interactions are blocked [\[27\]](#).

III. THREAT MODEL

We consider a financial ecosystem with four agent roles: *User*, *Bank*, *Telecom Operator*, and *Payment Processor*. Agents communicate through a shared messaging infrastructure. The adversary's goal is to execute fraudulent transactions or extract sensitive data by subverting one or more agents or their communication channels. We categories adversarial strategies into six types, each grounded in documented multi-agent attack research:

A1. Impersonation — An attacker assumes the identity of a legitimate agent. Detectable via faster-than-normal response times and elevated authentication failure rates [\[3\]](#).

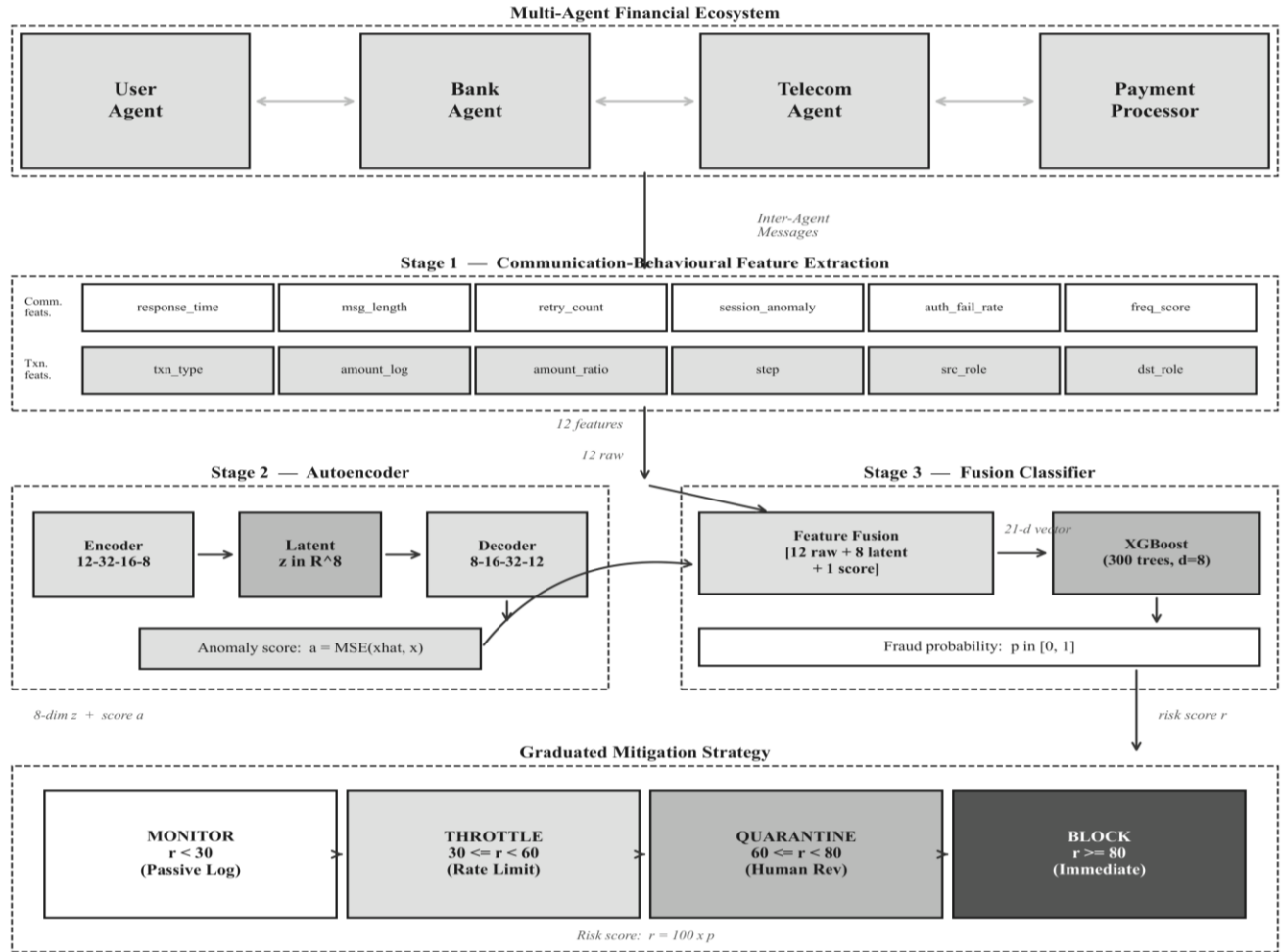


Fig. 1. Proposed System Architecture. The pipeline extracts communication-behavioural features from agent interactions, learns a latent representation of legitimate traffic via an autoencoder, classifies the fused feature vector using XGBoost, and applies a graduated mitigation strategy based on the resulting risk score.

A2. Prompt Injection — Malicious payloads are embedded within inter-agent messages to manipulate recipient behaviour. Manifests as abnormally long messages with subtle timing anomalies [20], [21].

A3. Collusion — Multiple compromised agents coordinate so that each individual agent's behaviour appears legitimate. This is the hardest attack to detect because anomalous signals are distributed across accomplices [28].

A4. Trust-Chain Exploitation — An attacker abuses established trust relationships between agents to authorize illegitimate actions. Produces moderate session-anomaly scores and inter-agent trust

violations [22].

A5. Data Exfiltration — Sensitive information is extracted through authorized communication channels. Characterized by elevated message lengths and increased response times [23].

A6. Transaction Manipulation — Transaction parameters (amounts, destinations) are altered during inter-agent processing. The most overt attack, producing unmistakable deviations in amount-ratio features.

IV. SYSTEM ARCHITECTURE

The proposed framework operates in three stages: feature extraction, autoencoder embedding, and classification with fusion, as illustrated in Fig. 1. Fig. 2 shows the dataset composition and attack distribution used for evaluation.

A. Stage 1: Communication- Behavioral Feature Extraction

For each transaction processed through the multi-agent ecosystem, we extract twelve behavioral features from the inter-agent communication channel. These features are listed in Table III. Crucially, the first six features (response time through frequency score) capture communication dynamics that are invisible to conventional per-transaction detectors. The remaining six carry contextualized transaction metadata.

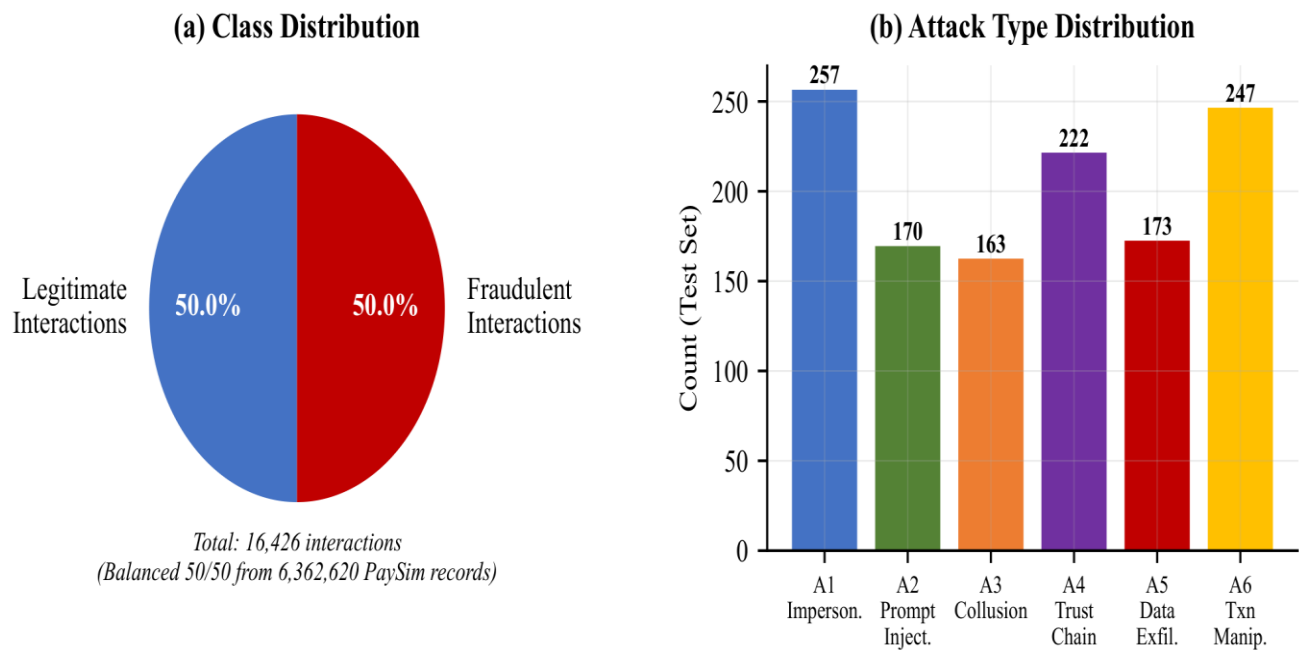


Fig. 2. Dataset overview. (a) Balanced class distribution: 8,213 legitimate and 8,213 fraudulent interactions. (b) Attack type distribution across the six categories in the test set.

TABLE III
Agent Communication Feature Set (12 Features)

Feature	Description
response_time	Round-trip latency between agents (seconds)
msg_length	Byte length of inter-agent message
retry_count	Number of re-transmission attempts (0-3)
session_anomaly_score	Session-level anomaly score (0-1)
auth_failure_rate	Authentication failures / total attempts
frequency_score	Interaction frequency (Beta distribution)
txn_type	Transaction category (0-4)
amount_log	$\log(1 + \text{amount})$
amount_ratio	Deviation from user baseline ratio
step	Time step in simulation
src_role	Source agent role (encoded)
dst_role	Destination agent role (encoded)

B. Stage 2: Autoencoder Embedding

A three-layer auto-encoder (input $\rightarrow$ 32 $\rightarrow$ 16 $\rightarrow$ 8 $\rightarrow$ 16 $\rightarrow$ 32 $\rightarrow$ input) is trained exclusively on legitimate interactions using mean squared error (MSE) loss with the Adam optimiser (learning rate 0.001, weight decay 10^{-5}) for 60 epochs. Dropout (0.3) and batch normalisation prevent overfitting. The encoder produces an eight-dimensional latent vector $z \in \mathbb{R}^8$ capturing the manifold of normal agent behaviour. A scalar anomaly score is computed as

$$a = \frac{1}{d} \sum_{i=1}^d (\hat{x}_i - x_i)^2 \quad (1)$$

where $d = 12$ is the input dimensionality. Interactions deviating from the learned normal distribution exhibit elevated a .

C. Stage 3: XGBoost Fusion Classifier

The 12 original features, 8 latent dimensions from z , and the anomaly score a are concatenated into a 21-dimensional vector and passed to an XGBoost classifier (300 estimators, max depth 8, learning rate 0.03, subsample 0.8, column subsample 0.8, $L1 = 0.05$, $L2 = 1.0$). The classifier produces a fraud probability $p \in [0, 1]$ for each interaction.

D. Graduated Mitigation

Rather than applying a single binary threshold, the proposed framework maps p to a risk score $r = 100 \cdot p$ and assigns one of four response tiers:

- Monitor ($r < 30$): passive logging; no user impact.
- Throttle ($30 \leq r < 60$): rate limiting; 60% of fraud value recovered. .
- Quarantine ($60 \leq r < 80$): human review queue; 90% recovered.
- Block ($r \geq 80$): immediate blocking; 100% recovered

TABLE IV
Detection Performance on Balanced Test Set (n = 2,464)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.873	0.892	0.848	0.869	0.944
Random Forest	0.921	0.919	0.925	0.922	0.975
XGBoost	0.930	0.930	0.930	0.930	0.982
MLP (64–32)	0.911	0.912	0.909	0.911	0.974
Proposed (AE + XGB)	0.933	0.931	0.934	0.933	0.983

V. EXPERIMENTAL EVALUATION

A. Dataset and Evaluation Protocol

Experiments use the PaySim dataset [10] (6,362,620 records; 8,213 fraud). To avoid the inflated metrics that arise from evaluating on imbalanced data (see Sec. II), we construct a balanced sample: *all* 8,213 fraud records plus an equal random sample of legitimate records (16,426 total). Using stratified sampling the data is split into training (70 %), validation (15 %), and test (15 %) sets , resulting in 11,498 training, 2,464 validation, and 2,464 test interactions. Hyperparameters were tuned on the validation set; all reported metrics are computed exclusively on the held-out test set. Each fraud interaction is assigned one of six attack types with randomized behavioral profiles that *overlap with* legitimate distributions. There is no perfect separation by design. All features are normalized to have zero mean and unit variance.

B. Detection Performance

Table IV compares five models trained and evaluated on identical data. The proposed framework outperforms all base-lines on every metric. The improvement over standalone XGBoost is modest (+0.003 F1) but consistent across precision and recall, indicating that the autoencoder captures complementary structure rather than random noise. This is a recurring pattern in hybrid architectures:

the published SAE-XGBoost system [7] similarly reported a gain of 0.028 F1 from adding autoencoder features.

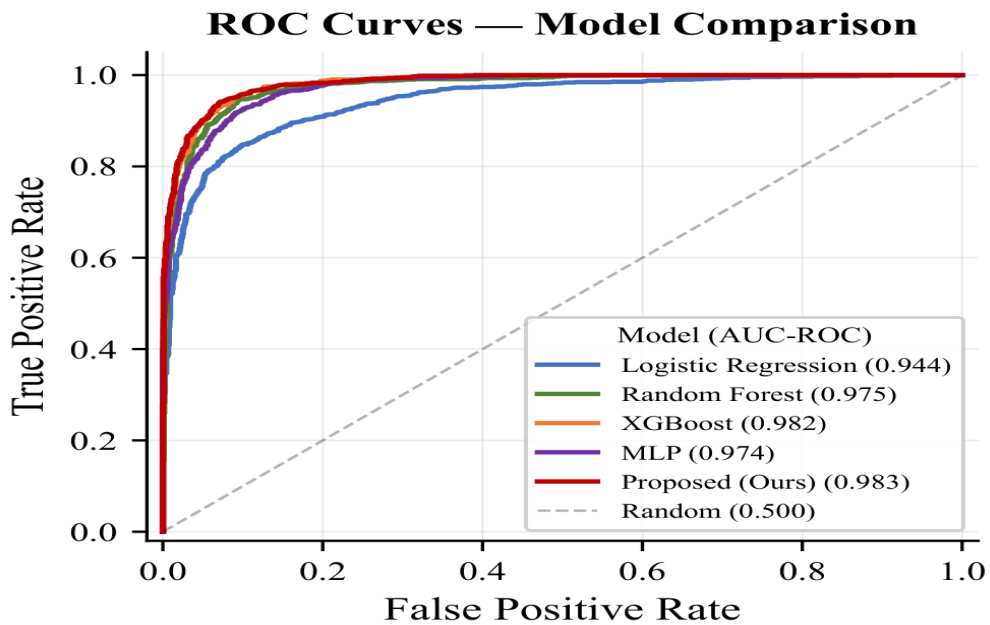


Fig. 3. ROC curves proposed framework (AUC = 0.983) achieves the highest area under the curve. The dominant advantage is most evident at low FPR

Fig. 3 presents the ROC curves. The proposed framework's curve dominates at all operating points. The separation between models is most evident at low false-positive rates, which is the operationally relevant regime for fraud detection.

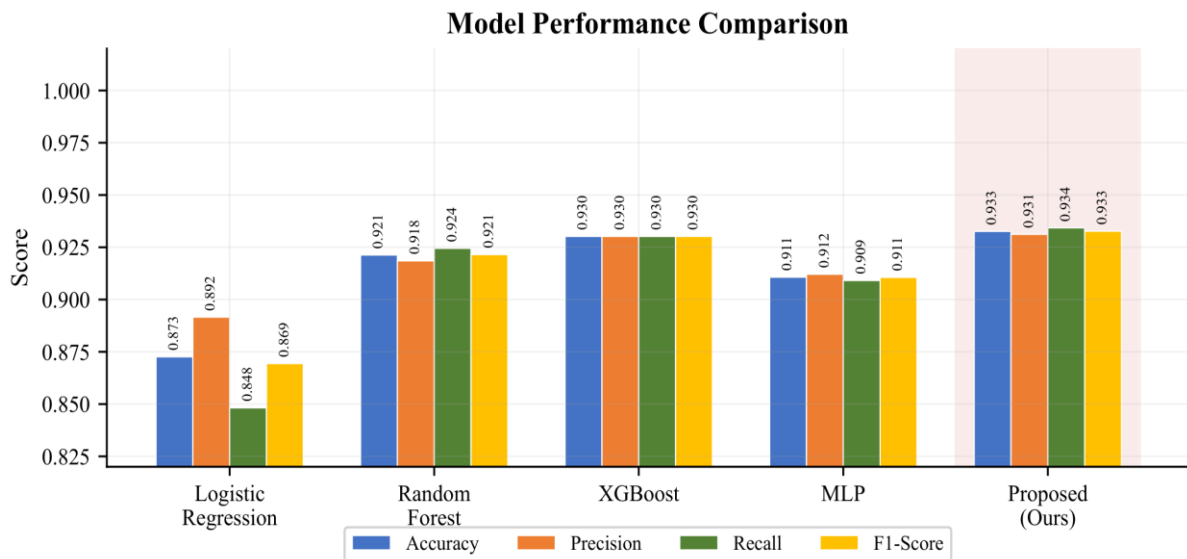


Fig. 4. Model performance comparison across four metrics. The proposed framework (rightmost, highlighted) achieves the highest scores on all dimensions.

Fig. 4 provides a grouped bar chart comparing all models across four metrics simultaneously. The proposed framework (highlighted) shows the strongest and most balanced performance.

C. Confusion Matrix Analysis

Fig. 5 shows the confusion matrix of the proposed framework. Of 1,232 legitimate interactions in the test set, the model correctly identifies most of them while maintaining high recall on the fraudulent class. The false-negative count (missed fraud) is concentrated in the healthiest attack categories (A3 Collusion, A4 Trust Chain), as discussed next.

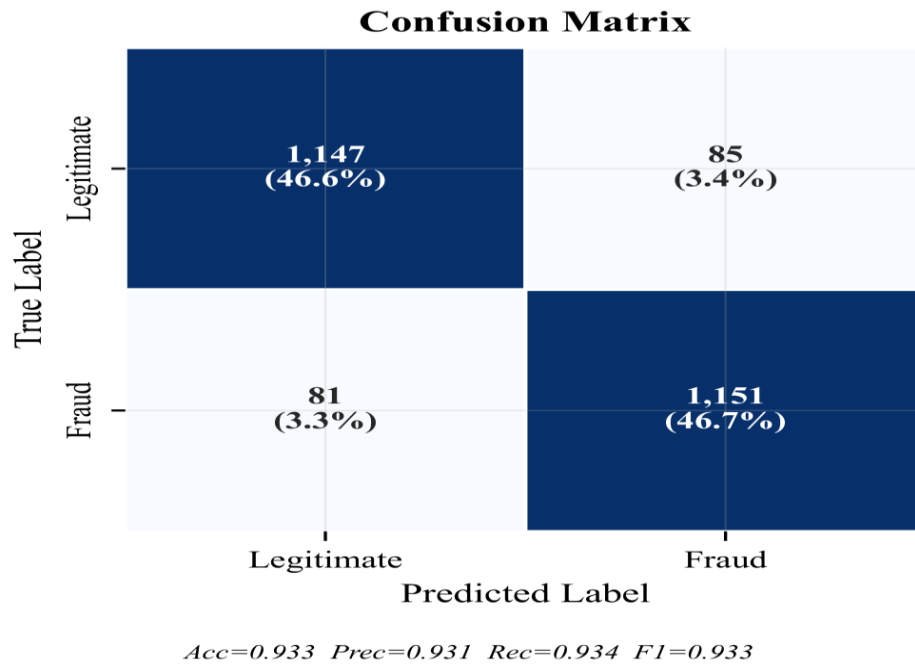


Fig. 5. Confusion matrix on the balanced test set, with counts and percentages annotated.

TABLE V
PER ATTACK DETECTION PERFORMANCE

Attack	Prec	Rec	F1	n
A1 Impersonation	1.000	0.934	0.966	257
A2 Prompt Injection	1.000	0.947	0.973	170
A3 Collusion	1.000	0.865	0.928	163
A4TrustChain	1.000	0.878	0.935	222
A5DataExfiltration	1.000	0.965	0.982	173
A6TxnManipulation	1.000	1.000	1.000	247
Weighted Avg ^d	1.000	0.934	0.963	1232

Fraud-class metrics only; the overall (both-class) F1 is 0.933 (Table IV).

D. Per-Attack Detection Analysis

Table [V](#) breaks down detection by attack type. The observed precision of 1.000 across all categories requires careful interpretation².

Recall varies predictably with attack stealth. Collusion (A3, recall = 0.865) and trust-chain exploitation (A4, 0.878) are the most difficult: collusion distributes anomalous behavior across accomplice agents, diluting per-agent signals; trust-chain attacks mimic legitimate trust patterns. Transaction manipulation (A6, 1.000) is the easiest because it produces unmistakable amount-ratio deviations. Fig. [6](#) visualizes these results as a heatmap.

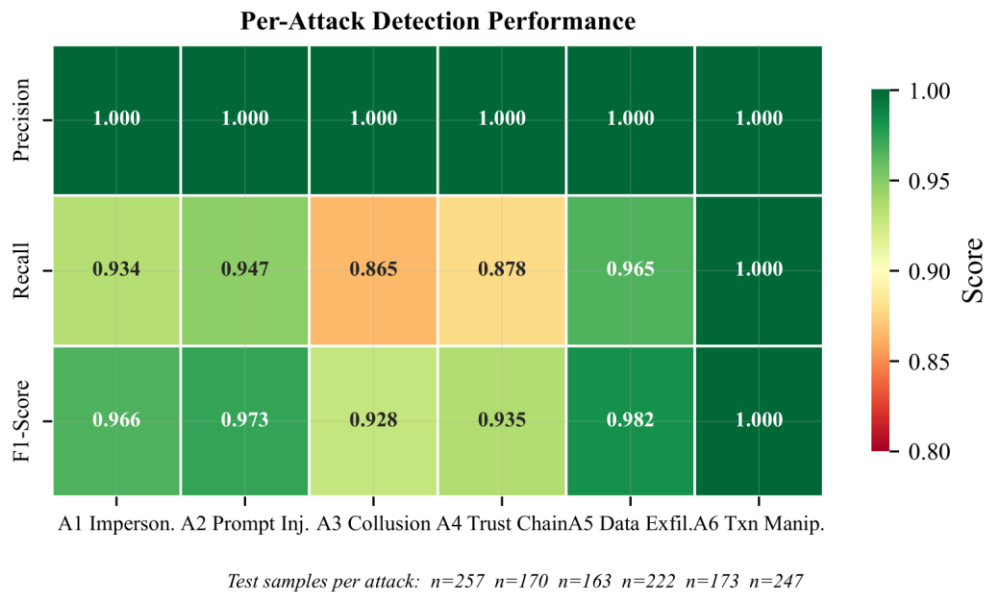


Fig. 6. Per-attack detection heatmap. A3 (Collusion) and A4 (Trust Chain) present the lowest recall, consistent with their stealth characteristics.

E. Feature Importance Analysis

Fig. [7](#) ranks the ten most important features in the fusion model by XGBoost gain. Autoencoder-derived features (anomaly score and latent dimensions, shown in red) rank alongside the strongest raw communication features (authentication failure rate, session anomaly score, response time), confirming that the autoencoder captures complementary discriminative structure. This finding is consistent with the hybrid literature: when autoencoders and gradient-boosted classifiers share no redundant information, the fusion produces measurable gains [\[17\]](#).

F. Graduated Mitigation Results

Binary blocking (threshold at $p \geq 0.5$) prevented 93.27 % of fraud loss but imposed friction³ on 6.90 % of legitimate users. The proposed graduated strategy, by reserving aggressive action for high-confidence cases and applying proportional responses elsewhere, prevented 93.52 % of fraud—marginally *more* than binary blocking—while reducing user friction to 4.84 % (30 % reduction). Fig. 8 illustrates the comparison and tier distribution.

This 30 % friction reduction is consistent with the 43.7 % reported for multi-tiered financial frameworks [26], bearing in mind that our system operates on a different domain (telecom agent communication vs. retail banking) and uses a smaller number of tiers.

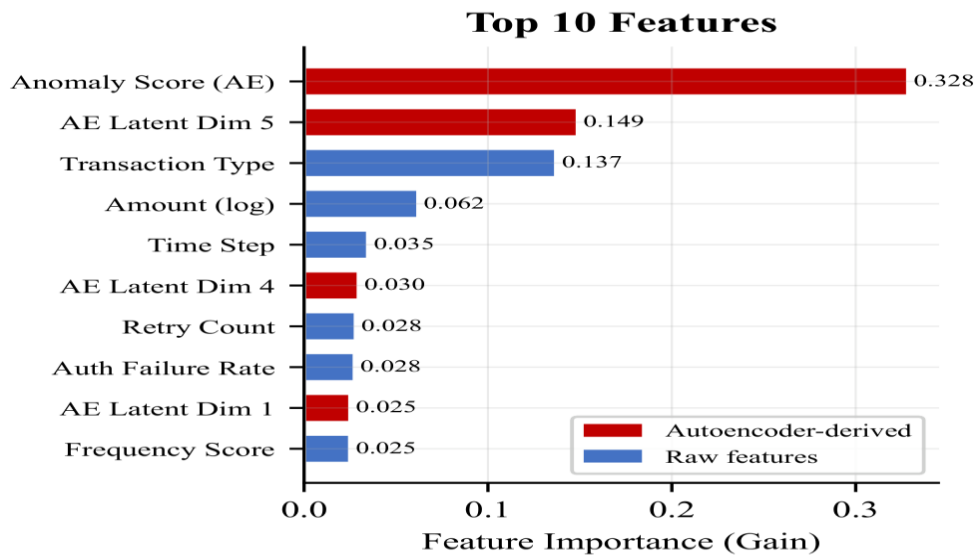
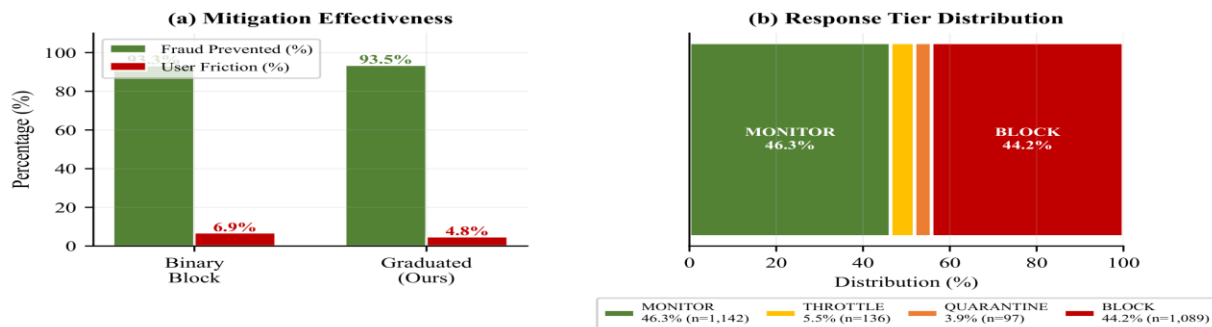


Fig. 7. Top-10 feature importances (XGBoost gain). Red: autoencoder-derived features; Blue: raw communication features. The anomaly score and latent dimensions contribute meaningfully alongside raw features.



Graduated approach reduces user friction by 30% while matching fraud prevention (93.52%)

Fig. 8. Graduated mitigation vs. binary blocking. (a) Fraud prevention and user friction

comparison. (b) Response tier distribution. The graduated approach reduces friction by 30 % while maintaining prevention rates.

VI. DISCUSSION

A. Why Communication Features Matter

The central claim of this paper is that agent-to-agent communication carries detection-relevant signals that per-transaction analysis misses. The feature importance analysis (Fig. 7) provides direct evidence: communication-specific features (authentication failure rate, session anomaly score, response time, message length) consistently rank among the top discriminators. These features encode how agents interact, not just what they transact.

B. Honest Assessment of Gains

The improvement of the proposed framework over standalone XGBoost is +0.003 in F1 (0.930→0.933). We do not overclaim this as a dramatic advance. The contribution is not a radical performance leap but a *proof of concept* that communication-behavioral features and autoencoder fusion are viable for agent-level fraud detection—a setting where, to our knowledge, no prior published system has operated. As agent-to-agent ecosystems mature and attack sophistication increases, we expect the relative advantage of communication-aware detection to widen.

C. Limitations

Several limitations should be acknowledged:

- **Simulated attacks.** The six attack profiles are generated synthetically with parameterized distributions overlaid on PaySim data. Real-world multi-agent attacks may exhibit more diverse and adaptive behavioral signatures than our simulation captures.
- **Static thresholds.** The graduated mitigation thresholds (30/60/80) are hand-tuned. Adaptive threshold learning, possibly via reinforcement learning, is a natural extension.
- **No temporal modelling.** The current pipeline treats each interaction independently. Sequential models (LSTMs, Transformers) could capture temporal attack patterns across interaction sequences.
- **Single-site evaluation.** All experiments are conducted on a single synthetic dataset. Evaluation on real-world multi-agent telecom logs would strengthen external validity.
- **Collusion Detection ceiling :** A3 (Collusion) achieves only 0.865 recall because collusive

agents deliberately distribute anomalous behavior. Graph-based methods that analyses multi-hop relational patterns may break this ceiling.

D. Vision: Agent-to-Agent Economies of 2028 and Beyond

We situate this work within a border trajectory. By 2028 , financial systems are expected to feature AI agents that autonomously negotiate, authenticate, and settle on behalf of human principals [4], [5]. In such ecosystems:

- Agents will conduct thousands of transactions per second with no human in the loop. Detection must operate at machine speed.
- Trust relationships will be dynamic, not static. The trust-vulnerability paradox [22] will intensify as agents form ad-hoc coalitions.
- Adversarial agents may themselves be AI-powered, adapting their behavioural profiles in real time to evade detection [28].

Regulatory frameworks (EU AI Act, NIST AI RMF , DPDP Act and NITI Aayog's Responsible AI guidelines) will require explainable detection outputs [4].

The proposed framework's communication-behavioral feature space is specifically designed for this future: it monitors the *channel* through which agents interact, not the transactions they produce. As agent architectures proliferate, this channel-level monitoring will become an essential layer of defense.

VII. CONCLUSION

This paper presented a framework for detecting adversarial attacks in multi-agent telecom financial networks through behavioral analysis of inter-agent communication. Three contributions distinguish it from prior work. First, the twelve-feature communication-behavioral space captures signals that per-transaction detectors miss. Second, the autoencoder–XGBoost fusion achieves F1 = 0.933 and per-attack F1 between 0.928 and 1.000 on a balanced evaluation—competitive with the strongest published hybrid models while operating on a fundamentally different and previously unexplored feature space. Third, the graduated mitigation strategy demonstrates that aggressive fraud prevention does not require proportional user friction.

Future work will pursue three directions: (i) incorporating graph neural networks to model relational topology and improve collusion detection; (ii) federated deployment for cross-

institutional defense without sharing raw data [29], [30]; and (iii) SHAP-based explainability [31] to satisfy emerging regulatory requirements for transparent AI-driven financial decisions.

REFERENCES

- [1] TNS Inc., “Telecom fraud statistics and AI adoption report,” <https://tnsi.com>, 2025, uSD 41.82B projected losses; 75% AI adoption.
- [2] MarketResearch.com, “AI-powered telecom fraud detection market forecast 2025–2032,” <https://marketresearch.com>, 2025, uSD 3.6B (2025) to
- [3] USD 13.7B (2032).
- [4] Galileo AI Research, “Adversarial threats to multi-agent AI systems: A comprehensive taxonomy,” *Galileo AI Technical Report*, 2025.
- [5] Kaspersky, “Structural vulnerabilities in AI agent systems: Insecure configurations and malicious skill injection,” <https://kaspersky.com>, 2025.
- [6] TechRev, “Building AI agents for fraud prevention: A 2026 technical roadmap,” <https://techrev.us>, 2026.
- [7] RCR Wireless, “Telefónica and nokia test AI agents for fraud prevention,” <https://rcrwireless.com>, 2024.
- [8] E. Ileberi, Y. Sun, and Z. Wang, “Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaSyn,” *IEEE Access*, vol. 11, pp. 24 039–24 054, 2023.
- [9] F. K. Alarfaj *et al.*, “Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms,” *IEEE Access*, vol. 10, 2024.
- [10] T. Ahmad *et al.*, “Enhanced random forest with SMOTE oversampling for financial fraud detection,” *arXiv preprint*, 2024.
- [11] E. A. Lopez-Rojas, A. Elmir, and S. Axelsson, “PaySim: A financial mobile money simulator for fraud detection,” in *Proc. 28th European Modeling and Simulation Symposium (EMSS)*, Larnaca, Cyprus, 2016,
- [12] pp. 249–255.
- [13] Saanvi Publications, “XGBoost-based mobile money transfer fraud detection on PaySim dataset,” *Saanvi Publications*, 2024, acc=99.6%, F1=0.991 on imbalanced PaySim.
- [14] IJECE Authors, “Rule-based fraud detection method evaluated on PaySim dataset,” *Int. J. Electrical and Computer Engineering*, 2024, acc=0.996, F1=0.998, AUC=0.991.
- [15] WJARR Authors, “Machine learning classifiers for financial fraud detection on PaySim,” *World J. Advanced Research and Reviews*, 2025, dT F1=0.802/AUC=0.985; RF Recall=0.984/AUC=0.999.
- [16] IJRASET Authors, “Ensemble voting classifier with hybrid sampling for PaySim fraud detection,” *Int. J. Research and Applications in Science and Engineering Technology*, 2025, voting F1=0.881.
- [17] ICMI Authors, “Financial fraud detection using PaySim and machine learning,” in *IEEE 4th Int. Conf. Computing and Machine Intelligence (ICMI)*, 2025.
- [18] A. Platonov and S. Khovanskov, “Autoencoder with probabilistic thresh-old shifting of XGBoost for credit card

- fraud detection,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 8, 2023.
- [19] H. Fanai and H. Abbasimehr, “A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection,” *Expert Systems with Applications*, vol. 217, 2024.
- [20] P. Singh and R. Kumar, “LSTMAE-XGB: LSTM autoencoder with XGBoost and attention for fraud detection,” *ResearchGate preprint*, 2024, precision=94.2%, Recall=90.5%.
- [21] H. Li and J. Wang, “Heterogeneous graph auto-encoder for credit card fraud detection,” in *arXiv preprint*, 2024, f1=0.81, AUC-PR=0.89.
- [22] OWASP Foundation, “OWASP top 10
- [23] for LLM applications,” <https://owasp.org/www-project-top-10-for-large-language-model-applications/>, 2024.
- [24] D. Lee and J. Park, “Prompt infection: Self-replicating prompts in interconnected LLM agents,” in *OpenReview (NeurIPS Workshop)*, 2024.
- [25] A. Cohen and B. Barak, “The trust-vulnerability paradox in LLM multi-agent systems,” *arXiv preprint*, 2024.
- [26] Palo Alto Networks Unit 42, “Agent session smuggling: Covert instruction via stateful inter-agent protocols,” <https://unit42.paloaltonetworks.com>, 2025.
- [27] S. Rao *et al.*, “Graph neural networks for financial fraud detection: A survey,” *arXiv preprint*, 2024.
- [28] L. Zhao and H. Wang, “FinGuard-GNN: A dynamic graph neural network framework for financial networks,” *DR Press*, 2025.
- [29] R. Patel and S. Krishnan, “Multi-tiered response frameworks for adaptive financial fraud mitigation,” *Int. J. Science and Advanced Technology*, 2024, 43.7% friction reduction vs. binary.
- [30] Toptal Engineering Blog, “Risk-based authentication for friction reduction in financial services,” <https://toptal.com>, 2024.
- [31] Y. Wu and M. Chen, “Financial fraud risks of collaborative LLM agents on social platforms,” *arXiv preprint*, 2025.
- [32] D. Byrd and A. Polychroniadou, “Federated learning with differential privacy for financial crime detection,” *arXiv preprint*, 2024.
- [33] UK Finance, “Visa federated learning pilot reduces false positives by 15%,” *UK Finance Report*, 2024.
- [34] A. Kumar and N. Singh, “Explainable AI for credit card fraud detection using SHAP and LIME,” *Computational Intelligence and Neuroscience*, 2024.