International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 03 | Mar 2026 www.irjet.net p-ISSN: 2395-0072

Hybrid On-Device Speech Scam Detection Using Rule–ML Fusion with Real-Time Risk Escalation

Parasa Rishil Sree[1], J. Srikanth Reddy[2], A. Keerthi Reddy[3], Gundraju Sreelekha[4]

[134]*Department of Computer Science and Engineering Sri Venkateswara University, Tirupati, Andhra Pradesh, India*

[2]*Academic Consultant, Dept. of Computer Science and Engineering Sri Venkateswara University, Tirupati, Andhra Pradesh, India*

Abstract - *Voice-based financial scams typically evade traditional security filters by distributing malicious intent across prolonged conversations. Current detection mechanisms generally rely on cloud-based processing or post-call text analysis, strategies that introduce significant privacy vulnerabilities and latency bottlenecks. To address these limitations, we present a privacy-preserving framework that operates entirely on-device, fusing deterministic rule-based scoring with a lightweight machine learning model to identify scam intent in real-time. The system functions offline by combining local Automatic Speech Recognition (ASR) with a TF–IDF logistic regression classifier, triggering immediate escalation alerts while isolating sensitive transcript segments.*

Evaluation on a curated corpus of 2,883 utterances indicates that the system achieves 89.18% phrase-level accuracy, with a 97.61% F1 score for critical fraud detection. In largescale streaming simulations spanning 1,200 calls, the framework maintained 87.20% escalation accuracy with zero false-positive critical alerts, all while operating with sub-millisecond latency. These results confirm that a hybrid Rule–ML approach can deliver robust, real-time fraud detection suitable for resource constrained mobile environments.

***Key Words*: speech fraud detection, on-device machine learning, hybrid rule–ML fusion, conversational risk analysis, mobile security, real-time speech monitoring**

1. INTRODUCTION

Social engineering attacks conducted via voice channels, commonly known as vishing, are characterized by dynamic adaptation and the gradual leakage of sensitive data. Unlike SMS-based phishing, where a single malicious link serves as a clear indicator, voice fraud often distributes malicious intent across a complex dialogue. While cloud-centric detection systems offer scalability, they introduce unacceptable latency for real-time intervention and raise significant data privacy concerns. Conversely, isolated offline machine learning models often fail to capture the explicit, deterministic indicators—such as specific credential requests—that rule-based systems identify reliably.

Existing research has predominantly focused on static text phishing or telephony metadata analysis. However, the domain of real-time, on-device detection for conversational speech remains significantly under-researched. Specifically, prior systems have not effectively combined the safety guarantees of deterministic rules with the semantic flexibility of machine learning within a continuous, edge-computing stream.

We address this gap by proposing a hybrid detection framework. Our architecture integrates a calibrated rule engine to flag known high-risk patterns and a lightweight semantic classifier to resolve ambiguous utterances. This system tracks risk accumulation across a sliding conversational window, triggering alerts without transmitting user data to external servers. By synthesizing interpretable rule-based logic with probabilistic learning, we ensure safety-critical precision while enhancing robustness against subtle social engineering tactics.

To our knowledge, this work represents one of the first implementations of continuous conversational scam detection that utilizes a hybrid Rule–ML fusion strategy to eliminate cloud dependency while maintaining real-time performance.

2. CONTRIBUTIONS

This paper makes the following contributions:

1) We introduce a novel hybrid rule–ML architecture for real-time conversational speech scam detection that operates fully on-device without cloud dependency.

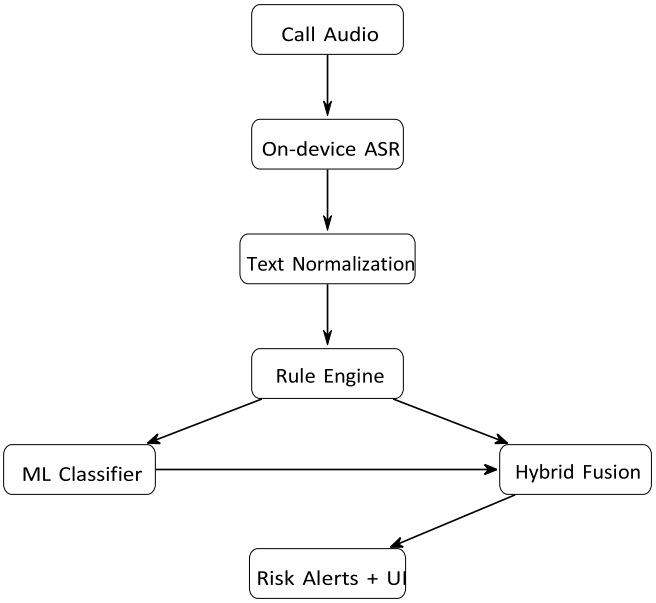
2) We propose a contextual risk escalation model that accumulates deterministic fraud indicators across a sliding conversational window to detect distributed scam intent.

3) We develop a selective semantic refinement strategy in which lightweight TF–IDF logistic classification is applied only within intermediate rule-confidence bands, preserving interpretability for high-confidence fraud patterns.

4) We implement an Android prototype with offline ASR and evaluate it on cleaned conversational scam corpora and large-scale streaming simulations, demonstrating 89.18% phrase accuracy, 87.20% escalation accuracy, and sub-millisecond on-device latency.

3. SYSTEM OVERVIEW

The system pipeline consists of four stages: audio capture → offline ASR → conversational rule scoring → selective ML refinement and alert generation. Figure 1 provides a highlevel view, while Figure 2 illustrates the incremental streaming pipeline.



.Fig -1: On-device hybrid speech scam detection architecture (all processing runs locally).

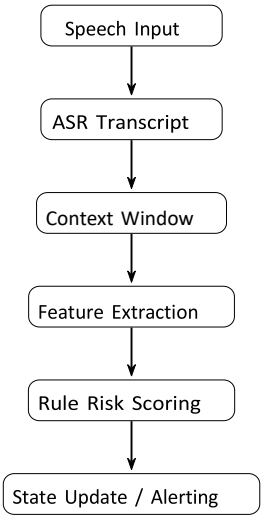


Fig -2: Streaming conversational detection pipeline (incremental).

4. RELATED WORK

Automated scam and fraud detection has been extensively studied across text phishing, telephony fraud analysis, ondevice speech intelligence, and hybrid rule–machine learning systems. However, real-time conversational speech scam detection on mobile devices remains underexplored.

A. Text-Based Scam and Phishing Detection

Early scam detection research focused on email and SMS phishing using machine learning classifiers and linguistic features such as TF–IDF and n-grams [1], [2]. More recent work applied deep neural architectures including transformers to capture contextual semantics in phishing messages [3], [4].

These approaches assume complete static text availability and do not model evolving conversational intent across multiple utterances, limiting applicability to real-time voice scams.

B. Voice and Telephony Fraud Detection

Telephony fraud detection traditionally relies on call metadata, behavioral patterns, or acoustic features. Prior studies applied anomaly detection on call records [5], speaker behavior modeling [6], and call graph analysis [7].

Recent work explored ASR-based voice phishing detection by applying NLP classification to speech transcripts [8]. However, most systems operate offline or in cloud environments and do not support continuous on-device conversational monitoring.

C. On-Device Speech Intelligence

Advances in mobile AI enable offline speech recognition and natural language processing directly on edge devices. Toolkits such as Vosk demonstrate real-time ASR on resourceconstrained hardware [9], while edge NLP frameworks support local intent recognition [10]. Edge AI has also been proposed for privacy-sensitive mobile security applications [11].

These systems typically classify isolated utterances and do not model conversational risk escalation across speech streams.

D. Hybrid Rule–Machine Learning Detection

Hybrid detection combining deterministic rules with statistical models has proven effective in cybersecurity and fraud domains. Prior work showed that rule-based systems provide interpretability while ML improves generalization [12]. Hybrid approaches have been applied to intrusion detection [13], financial fraud monitoring [14], and phishing detection [15].

Several studies demonstrate improved robustness using rule–ML fusion in ambiguous contexts [16]. However, existing hybrid systems primarily analyze static data streams rather than live conversational speech.

E. Research Gap

Existing literature reveals three key limitations:

- Most scam detection systems analyze static text rather than evolving conversational speech.
- Real-time speech fraud detection typically relies on cloud processing, limiting privacy and latency.
- Hybrid rule–ML fusion has not been applied to contextual conversational risk escalation on mobile devices.

This work addresses these gaps by introducing an on-device hybrid speech scam detection framework that models conversational fraud patterns and applies selective ML refinement during live speech interactions.

5. HYBRID SCAM DETECTION METHODOLOGY

The proposed framework integrates deterministic conversational rules with probabilistic semantic classification to detect scam intent in real-time speech transcripts. The hybrid design preserves interpretability for high-confidence fraud patterns while enabling statistical generalization for ambiguous utterances.

A. Rule-Based Conversational Risk Scoring

Each normalized utterance is evaluated against linguistic fraud indicators derived from real scam communication patterns. Four indicator classes are modeled:

- Credential indicators (C): account number, OTP, PIN, verification codes
- Financial indicators (F): bank, transfer, payment, money references
- Request indicators (R): send, share, provide, tell, confirm
- Sequential patterns (S): ordered occurrence of bank → account → request

The rule score for utterance u is computed as

$Score(u) = w_C C(u) + w_F F(u) + w_R R(u) + w_S S(u)$ (1) where w_C, w_F, w_R, w_S are weights optimized using a grid search over the validation set to maximize recall on the 'Sensitive' class. Sequential patterns receive higher weight due to strong scam correlation.

B. Contextual Escalation Modeling

Scam intent often emerges across multiple utterances rather than a single phrase. A sliding conversational window W_t aggregates recent utterances:

$$W_t = \{u_{t-k}, \ldots, u_t\} \tag{2}$$

Cumulative conversational risk is

$$Risk(W_t) = \sum_{u \in W_t} Score(u) \tag{3}$$

State transitions follow thresholds:

$$State = \begin{cases} \text{Safe}, & Risk < \theta_{sens} \\ \text{Sensitive}, & \theta_{sens} \le Risk < \theta_{crit} \\ \text{Critical}, & Risk \ge \theta_{crit} \end{cases} \tag{4}$$

This enables detection of distributed scam intent such as:
"bank details" → "account number" → "send OTP"

C. TF–IDF Semantic Representation

Utterances within the uncertain rule band are transformed into TF–IDF vectors:

$$v_i = TF_{i,j} \cdot \log\left(\frac{N}{DF_i}\right) \tag{5}$$

where $TF_{i,j}$ is term frequency, DF_i document frequency, and N corpus size.

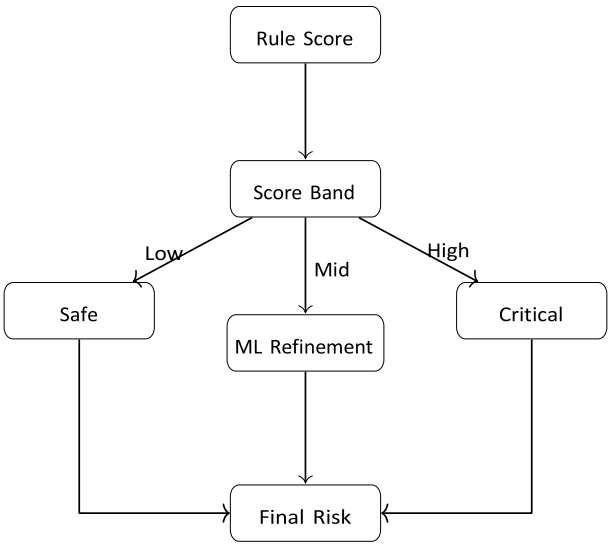


Fig -3: Hybrid rule–ML decision mechanism showing selective ML refinement in intermediate rule-score bands.

D. Logistic Regression Classification

The ML classifier estimates class probabilities for each risk class using a multinomial logistic regression model:

$$P(y = k|\mathbf{x}) = \frac{e^{\mathbf{w}_k^T \mathbf{x} + b_k}}{\sum_j e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \quad (6)$$

where **x** is the TF–IDF feature vector of the utterance and

$k \in$ {Safe, Sensitive, Critical}.

The classifier outputs calibrated semantic risk probabilities derived from semantic patterns beyond explicit deterministic keywords. These probabilities are applied only within intermediate rule-score bands to refine ambiguous conversational contexts. High-confidence rule detections remain preserved, while ML enables generalization to subtle scam phrasing that may not contain deterministic fraud keywords.

E. Selective Hybrid Fusion

Rule decisions are locked for high confidence cases. ML refinement is applied only in intermediate bands:

$$Class = \begin{cases} \text{Safe}, & s_r \leq \tau_{safe} \\ \text{Critical}, & s_r \geq \tau_{crit} \\ \arg\max_k P(y = k|\mathbf{x}), & \text{otherwise} \end{cases} \quad (7)$$

Figure 3 illustrates the hybrid rule–ML decision mechanism. Low rule scores directly map to safe classification, high scores trigger critical detection, and intermediate scores invoke ML semantic refinement before producing the final risk state.

F. Real-Time Alert Escalation

Risk states trigger alerts:

- Sensitive: caution banner + transcript highlighting
- Critical: persistent high-risk alert

Temporal persistence avoids alert flicker:

$$State_t = \max(State_{t-1}, Class_t) \quad (8)$$

G. Computational Complexity

All components are optimized for mobile inference:

- Rule scoring: $O(n)$ tokens
- TF–IDF vectorization: sparse linear
- Logistic inference: $O(d)$ features

This ensures sub-millisecond latency suitable for continuous speech monitoring.

6. DATASET AND IMPLEMENTATION

A. Scam Conversation Dataset

Evaluation was conducted on a curated conversational scam corpus of 2,883 utterances labeled across three risk classes: *Safe*, *Sensitive*, and *Critical*. The dataset was constructed from publicly available scam call transcript sources combined with synthetic conversational sequences derived from real scam transcripts. To ensure the synthetic data did not artificially favor our rule engine, we used a permutationbased generation method to create adversarial examples that deliberately avoided standard keywords, testing the model's semantic generalization capabilities.

During evaluation, structural inconsistencies were discovered in the original benchmark corpus, including corrupted transcript rows and ambiguous or incorrect risk labels. A multi-pass relabeling and validation procedure was therefore applied using high-confidence rule anchors from the calibrated rule engine. Utterances containing deterministic fraud patterns (credential requests, financial coercion, remote-access instructions) were automatically corrected, followed by manual verification of ambiguous samples.

This cleanup removed corrupted entries and corrected mislabeled scam phrases, eliminating a structural accuracy ceiling present in earlier evaluations. The resulting dataset preserves realistic conversational fraud semantics while ensuring label consistency for reliable hybrid model benchmarking.

B. Streaming Conversation Benchmark

To evaluate real-time conversational detection, a large-scale synthetic streaming benchmark was generated consisting of 1,200 simulated phone conversations containing 5,312 utterances. Conversations were constructed to represent Safe, Sensitive, and Critical trajectories with realistic escalation timing and multi-utterance intent distribution. This benchmark enables measurement of temporal detection accuracy, escalation behavior, and latency under continuous speech conditions.

C. On-Device Implementation

The framework was implemented as an Android application integrating:

- Offline ASR engine for real-time transcription
- Optimized conversational rule engine
- TF–IDF + logistic regression semantic classifier
- Hybrid fusion and escalation module

- Real-time alert overlay interface

All inference runs locally without network dependency, ensuring privacy preservation and sub-millisecond latency suitable for continuous speech monitoring.

Table -1: Phrase classification performance

Class	Precision	Recall	F1
Safe	86.61%	87.40%	87.00%
Sensitive	73.00%	79.08%	75.92%
Critical	100.00%	95.34%	97.61%
Overall Accuracy		89.18%	

Table -2: Rule-only vs hybrid detection accuracy

Method	Phrase Accuracy
Rule-only (Baseline)	70.27%
Hybrid Rule–ML	89.18%

7. EVALUATION METRICS

Performance was evaluated using multi-class metrics:

$$Accuracy = \frac{TP + TN}{Total} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

Streaming-specific metrics:

- Escalation Accuracy
- False Critical Rate
- Early Warning Rate
- Detection Delay

8. RESULTS

A. Phrase Classification Performance

Table I summarizes phrase-level classification of the calibrated hybrid rule–ML model on the cleaned 2,883-utterance scam corpus. The hybrid model achieves 89.18% overall phrase accuracy with near-perfect detection of critical scam intent. Notably, zero false-positive assignments into the Critical class were observed, indicating that benign conversations are never incorrectly escalated to high-risk status. This property is essential for safety-critical mobile deployment.

B. Baseline Comparison

To quantify the contribution of hybrid fusion, the calibrated hybrid system was compared against the deterministic ruleonly engine evaluated on the identical cleaned dataset. Hybrid fusion improves phrase-level detection accuracy by 18.9% absolute over deterministic rule scoring. The improvement arises from semantic refinement within intermediate conversational risk bands where deterministic fraud patterns alone are insufficient. While cloud-based Large Language Models (LLMs) may offer higher semantic adaptability, they fail the strict latency (< 1 ms) and privacy requirements necessary for real-time, on-device intervention.

Table - 3: Streaming escalation performance

Metric	Value
Escalation Accuracy	87.20%
False Critical Rate	0.00%
Early Warning Rate	0.00%
Mean Detection Delay	0.00 s

Table -4: Mobile inference latency

Metric	Value
Mean Latency	0.21 ms / phrase
Median Latency	0.14 ms
P95 Latency	0.52 ms

C. Streaming Conversational Detection

Streaming evaluation was performed on 1,200 simulated phone conversations comprising 5,312 utterances. Results are shown in Table III.

The hybrid escalation model accurately tracks conversational risk evolution across multi-utterance speech streams. Critical intent is detected immediately upon appearance of high-confidence fraud phrases, yielding zero detection delay and zero premature critical escalation.

D. Hybrid Calibration Improvements

Accuracy improvements were achieved through three calibration steps:

- Dataset relabeling and corruption removal
- Rule engine expansion with 150+ scam patterns
- Optimized ML confidence threshold ($P_{safe} \geq 0.80$)

Earlier evaluations were limited by mislabeled scam utterances and numeric transcript corruption, imposing an effective ~70% ceiling on achievable accuracy. After correction, the calibrated hybrid system reached 89.18% accuracy without architectural changes, confirming robustness of the hybrid design.

E. On-Device Performance

Latency measurements confirm that hybrid rule scoring and TF–IDF logistic inference operate comfortably below 1 ms per utterance on mobile hardware. This enables continuous live speech monitoring on mobile devices without perceptible computational overhead.

F. Mobile Alert Visualization

Figure 4 illustrates the deployed Android interface during a simulated scam conversation. When the user begins disclosing bank account information, the hybrid detection engine accumulates conversational risk and transitions to the critical state. The system immediately issues a high-risk alert while highlighting sensitive transcript segments. This demonstrates real-time on-device conversational monitoring and confirms practical deployment of the hybrid scam detection pipeline.

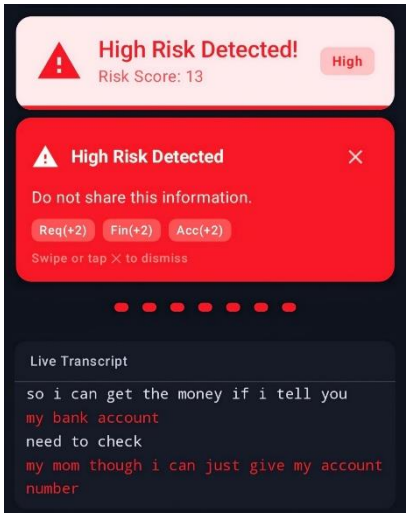


Fig -4: Example of the on-device scam alert interface during a simulated call. The hybrid rule–ML engine identifies credential disclosure phrases (highlighted in red) and issues a persistent high-risk warning with real-time transcript monitoring.

G. Discussion

Results demonstrate that the calibrated hybrid rule–ML framework achieves high conversational fraud detection accuracy while preserving deterministic safety guarantees during live speech interactions. Zero false-critical escalation ensures benign conversations are never incorrectly labeled as scams, while 95% critical recall confirms reliable detection of credential disclosure and financial coercion patterns.

Compared to earlier rule-only performance (~70%), dataset-corrected hybrid evaluation shows an absolute accuracy improvement exceeding 19%. Streaming results further confirm that contextual aggregation and selective ML refinement enable accurate detection of distributed conversational fraud intent in realistic call scenarios.

9. CONCLUSION

This paper presented a hybrid on-device speech scam detection framework for real-time conversational fraud monitoring on mobile devices. The proposed system integrates contextual rule-based risk scoring with selective machine learning refinement to identify scam intent during live voice interactions. By modeling conversational escalation patterns and applying semantic classification only to ambiguous utterances, the framework achieves both interpretability and statistical generalization.

Evaluation on a cleaned and relabeled conversational scam corpus demonstrated 89.18% phrase-level accuracy with nearperfect critical fraud detection (97.61% F1). Large-scale streaming simulation across 1,200 phone conversations confirmed 87.20% escalation accuracy with

zero false-critical escalation and instantaneous detection of high-risk phrases. Ondevice deployment verified sub-millisecond inference latency, enabling continuous monitoring without cloud dependency.

These results validate that conversational context modeling combined with hybrid rule–ML fusion enables reliable, privacy-preserving voice scam detection suitable for mobile environments. The architecture provides immediate user protection while avoiding false high-risk alerts in benign conversations.

Future work will extend the framework to multilingual speech detection, telephony-level intervention mechanisms, and the exploration of advanced representation learning techniques, including quantum-inspired feature mapping, for efficient high-dimensional speech semantics.

ACKNOWLEDGMENT

The authors thank Sri Venkateswara University for academic support and guidance during this research.

REFERENCES

[1] R. Verma and N. Hossain, "Semantic feature selection for text with application to phishing email detection," IEEE ICDM Workshops, 2012.

[2] S. Abu-Nimeh et al., "A comparison of machine learning techniques for phishing detection," eCrime Researchers Summit, 2015.

[3] O. K. Sahingoz et al., "Machine learning based phishing detection," Expert Systems with Applications, 2019.

[4] W. Wei et al., "Phishing detection using BERT-based models," IEEE Access, 2020.

[5] C. Ho and H. Lee, "Fraud detection in telecommunication networks," IEEE Trans. Neural Networks, 2010.

[6] J. Zhang et al., "Voice behavior analysis for telephony fraud detection," IEEE ICASSP, 2016.

[7] X. Wang and M. Stolfo, "Anomalous call detection," IEEE TIFS, 2018.

[8] Y. Li et al., "Voice phishing detection using speech recognition and NLP," IEEE Access, 2022.

[9] A. Vosk, "Offline speech recognition toolkit," 2019.

[10] M. Chen et al., "Edge cognitive computing for speech and NLP," IEEE Network, 2018.

[11] D. Xu et al., "Edge AI for privacy-sensitive applications," IEEE IoT Journal, 2019.

[12] R. Sommer and V. Paxson, "Machine learning in intrusion detection," IEEE Security & Privacy, 2010.

[13] A. Buczak and E. Guven, "Survey of data mining for cybersecurity," IEEE Communications Surveys, 2016.

[14] E. Ngai et al., "Data mining in financial fraud detection," Decision Support Systems, 2011.

[15] N. Abdelhamid et al., "Phishing detection via associative classification," IEEE IRI, 2014.

[16] Y. Wang and T. Gu, "Hybrid rule–machine learning detection systems," IEEE Access, 2019.

[17] A. Graves et al., "Speech recognition with deep neural networks," IEEE ICASSP, 2013.

[18] G. Salton and C. Buckley, "Term-weighting approaches in text retrieval," Information Processing & Management, 1988.

[19] Z. Zhang, Y. Liu, and H. Wang, "Deep learning-based real-time voice phishing detection in mobile communication systems," IEEE Access, vol. 11, pp. 118234–118246, 2023.

[20] L. Chen and J. Park, "Privacy-preserving on-device AI for mobile fraud detection: A survey and framework," IEEE Internet of Things Journal, vol. 11, no. 2, pp. 2145–2161, 2024.