NFSU JOURNAL OF CYBER SECURITY & DIGITAL FORENSICS

NFSU – Journal of Cyber Security and Digital Forensics
Volume – 5, Issue – 1, June 2026
E – ISSN – 2583-7559

Acoustic-Based Detection of AI-Generated Vishing Attacks in Telecom Environments

A Machine Learning Framework for Voice Feature Analysis and Fraudulent Call Identification

Keerthi P
Department of CSE
Christ University
Bangalore,India
pkeerthi291@gmail.com

Dr.Gokulapriya R
Department of CSE
Christ University
Bangalore,India
r.gokulapriya@christuniversity.in

***Abstract*— The proliferation of AI-generated synthetic voice and caller ID has made it all but impossible to defend against these systems with conventional security measures in telecom. In this paper, a novel supervised machine learning framework along with Convolutional Neural Network (CNN)-based deep learning method are introduced for real-time fraudulent voice communication detection based upon the acoustic and behavioural features extracted from the FOR-2sec subset of the Fake-or-Real (FoR) dataset. The features extracted were MFCCs and spectral descriptors, which were then tested against six classical classifiers: Logistic Regression, SVM, Random Forest, AdaBoost, CatBoost and XGBoost, as well as a hybrid SVM-based architecture for combining the features, and a proposed CNN model that aims to capture deep temporal and spectral patterns embedded in AI-generated speech. The CNN-based system outperforms standalone SVM with a uniform score of 0.88, while the CNN-based architecture is more competitive in terms of feature representation, by learning hierarchical acoustic representations directly from the raw spectral input with less reliance on hand-crafted feature engineering. With Explainability based on LIME, further the rationales of decisions are auditable, which are applicable for anti-spoofing guidelines by TRAI and STIR/SHAKEN standard and keep the framework technically sound and legally responsible.**

Keywords*— *Audio Forensics, Deepfake Detection, Machine Learning

I. INTRODUCTION

Voice communication is one of the most frequently used mediums to communicate, on a personal, financial and institutional level, and is therefore a common target for cybercriminals. Vishing, or voice phishing, is one of the most rapidly expanding dangers that exists today in the telecom industry. This type of phishing relies on live conversation, psychological tricks, and synthetic speech by AI to trick victims. With the addition of VoIP technology, caller ID spoofing, and generative AI, the entry level for attackers has been lowered to a new level and they can now deploy robocalls and hyper-realistic voice synthesis to scale.Now, thanks to the addition of VoIP technology, caller ID spoofing, and generative AI, the entry level for attackers has been lowered to a new level, enabling them to deploy robocalls and hyper-realistic voice synthesis at scale that traditional rule-based systems can't compete with.

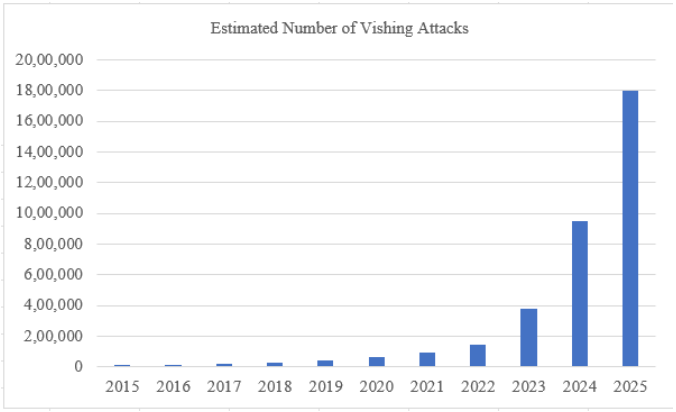


Fig 1 : Estimated Growth of Vishing Attacks (2015–2025)

Vishing attacks have surged following 2022, as illustrated in Figure 1, with the rise of mainstream AI voice synthesis tools and the ineffective response of traditional fraud prevention measures against the new vectors of AI attacks. Existing literature also shows that there is a critical need — previous research mostly focused on structured CDR data, but vishing attacks are most effectively carried out on the acoustic signal layer.

The results are compared with Logistic Regression, SVM, Random Forest, AdaBoost, CatBoost, XGBoost, a hybrid SVM framework, and the proposed CNN model, and their outcomes are in line with the TRAI anti-spoofing directives and STIR/SHAKEN standard

II. Literature Review

Pietri et al. [1] carried out a wide analytical mapping of the spam, scam and robocall threats in 5G and AI-enabled telecommunication environments and recorded the economic impacts and regulatory vulnerabilities across international jurisdictions such as India. Although the study is a comprehensive identification of AI-based voice impersonation as an ever-growing threat-creating study, its analytical focus is strictly descriptive in nature - it does not offer nor confirm any automated detection engine, making it more a threat characterization study than a mitigation framework. The following basic limitation is the direct motivation of real-time machine learning detection system which has been proposed in the present paper.

Tripathy et al. [2] discussed the problem of class imbalance in detecting telecom fraud by resampling the skewed class data using SMOTE-based resampling and classical classifier and ensemble classifier on real world skewed datasets. Though the study has shown that resampling greatly enhanced the minority classes detection, its complete reliance on structured call detail records as the critical methodological limitation of the study because it does not mention the acoustic signal analysis, AI-generated synthetic speech, and vishing detection voice-level at all. This gap is especially notable considering that the current vishing attacks are at the level of the acoustic layer, rather than the transactional layer, which implies that there is a fundamental disconnection between the dataset options of existing studies and the facts of the reality of telecom voice fraud.

Khatun et al. [3] compared supervised classifiers such as the Random Forest, SVM, and XGBoost on structured telecom call records and found ensemble superiority of the ensemble in high-dimensional non-linear feature spaces than the performance of linear classifiers on the same data sets. The fact that the study has been limited in terms of transactional and behavioral metadata, and that there has been no integration of acoustic feature extraction in the study limits one to pose a critical question about whether the findings of the study can be generalized to the voice-based fraud detection scenarios. The fact that the ensemble methods outperform the standalone classifiers in general, is directly refuted by the experimental results of the current paper, wherein the standalone SVM proved to be a better tool in general.

Wang et al. [4] presented a challenge-response authentication scheme named Caller ID Verification (CIV), which is a PKI-free authentication system that deals with caller ID spoofing, in heterogeneous networks such as VoIP, cellular, and SS7-based systems. Although CIV is a significant improvement in both protocol-level identity verification and in general network authentication, it only deals with the network authentication layer, and offers no protection against AI-generated voice synthesis or acoustics-level fraud, which is a significant blind spot since an authenticated caller ID does not imply protection against the deployment of a cloned or synthesized voice during the call itself. This architectural shortcoming highlights the need to have such an acoustic-layer detection as a complement to the current framework, which it directly offers.

Krasic and Celar [5] compared machine learning classifiers to models trained on live telecom operator data, and found that ensemble techniques trained on re-sampled all-SMOTE data, consistently outperformed their counterparts trained on raw imbalanced data. Although its methodological approach is sound in its own right, the fact that the study is entirely limited to the study of structured call detail records and that the analysis of the threat of AI-generated vishing is conducted entirely in the feature space of transactional features that both this and the Tripathy et al. study exclusively explore. The fact that this restriction is common to several studies indicates that there is a gap in the telecom fraud detection literature that is systemic.

This earlier research on deepfake speech detection by [6] identified a methodological flaw in the area in the form of the lack of a standardized and reproducible evaluation framework but proposed a cross-dataset benchmarking framework to allow fair comparison of models. Importantly, although models have been shown to perform reasonably well under controlled environments, performance declines significantly when the models are exposed to new synthesis methods - a generalization failure, which has both direct and serious implications about the real world deployment of telecom fraud, wherein the attackers continue to evolve their voice synthesis methods beyond any fixed training distribution. This fragility compromises the feasibility of detection models that are not specifically trained to make adversarial generalization.

In [7], the survey classified AI-based deepfake detection into feature-based, machine learning and deep learning paradigm, with dominance in capturing spatial and temporal audio manipulations. Although the review rightly identifies explainable AI as a critical future point of need, real-time deployment, and adaptive learning, it falls short of making any deployable framework, to meet these requirements simultaneously, defaulting to a prescriptive agenda, instead of an operational contribution. The current paper is a direct answer to this agenda as it combines LIME-based explainability with a real-time classification pipeline in a telecom-specific fraud detection environment.

Panda [8] explored the transformation of social engineering and phishing attacks by conducting a systematic review of

attack taxonomies and defense mechanisms and found that multi-layered behavioral and technical defense mechanisms are more effective than single-point defense mechanisms. Although the study offers a very good theoretical background in which to understand the psychological mechanics behind vishing, its purely conceptual framework of defence - that is, one that does not have an automated, platform-integrated detection tool as part of its operational base - limits its functional usefulness. The lack of a system of detection developed empirically and validated is a significant weakness, especially because of the recognition of AI-assisted impersonation as a rapidly advancing threat that must be addressed urgently with the help of technical means.

Scanlon et al. [9] tested the utility of ChatGPT in multiple digital forensic use cases, and found it had potential in generating and summarizing forensic evidence, but it also had severe limitations, such as hallucinated forensic evidence, inability to test evidence, and unacceptability of AI-generated conclusions in legal proceedings. Although the research does make a significant contribution towards establishing the limits of the applicability of the LLM in the field of forensic science, it also reveals one important gap: the lack of domain-specific AI forensic tools that can provide automated analysis, and structured output that is legally admissible. The significance of explainable mechanisms of decision-making, such as the LIME-based audit trail, which is already part of the current framework, strengthens this gap and, therefore, justifies the importance of explainable mechanisms of decision-making, including the LIME-based audit trail, which is already part of the current framework.

III. Dataset Description

The FOR-2sec set is uniformly segmented 2-second real human speech and 2-second fake speech samples generated by the AI author using synthetic speech. The data set includes recordings from several speakers in different acoustic conditions, allowing some degree of speaker and environment diversity which facilitates generalization in trained classifiers. All the audio samples are presented in normalized WAV format with a fixed sampling rate, so that features can be extracted in the same way throughout the whole corpus.The data set consists of two main categories: real and fake. The real class is made up of actual voices recorded by humans, and the fake class is made up of AI-generated fake voices using a variety of state-of-the-art TTS and voice cloning systems. The binary classification approach neatly aligns with the fraud detection goal of the proposed framework, namely identifying real voices in the network as opposed to those generated by AI for vishing attacks.

IV. Threat Taxonomy of Telecom-Enabled Voice Fraud

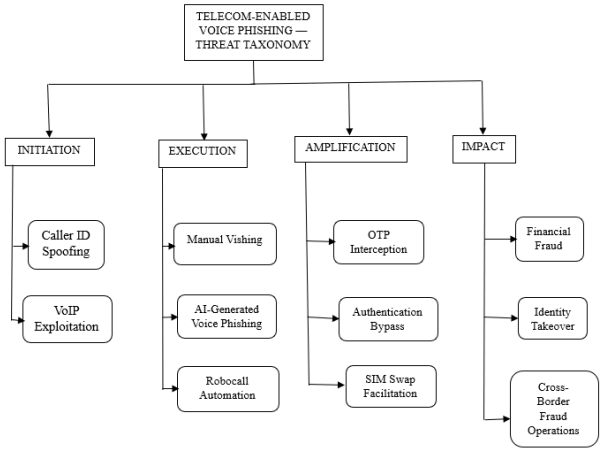


Fig 2: Threat Taxonomy of Telecom-Enabled Voice Phishing Attacks

The growing sophistication of vishing attacks requires a formalized type of classification of the threat environment to be able to establish both technical detection mechanisms and regulatory responses. Figure 3 illustrates a hierarchical threat taxonomy of telecom-enabled voice phishing that outlines the attack life cycle on four connected layers: initiation, execution, amplification and impact.

Extraction of the main feature representation, the Mel-Frequency Cepstral Coefficients (MFCCs) was performed over the 40 coefficients of 2-second standardized audio segments of the FOR-2sec dataset with a 25ms Hamming window, 10ms frame shift, and 512-point Fast Fourier Transform (FFT). MFCCs were chosen for their proven ability to capture the spectral distortions caused by AI voice synthesis systems, such as formant violations, phase variations, and unnatural intonation patterns, and for their computational efficiency, which is important for real-time use in telecom applications. In addition to MFCC, three types of spectral complementary descriptors were extracted: spectral centroid (the frequency at which the maximum amount of the signal's energy occurs), spectral bandwidth (the spread of energy over various frequency bands), and spectral roll-off (the frequency below which 85% of the total energy is calculated). Further, temporal statistics such as the zero-crossing rate and the root mean square energy were extracted to include irregularities in rhythm and prosody that are typical of robocalls and synthetic speech. All features were combined into a single feature vector, standard scalar normalized to remove scale dependency between different types of telecom audio sources, and fed to the classification pipeline..

NFSU JOURNAL OF CYBER SECURITY & DIGITAL FORENSICS

NFSU – Journal of Cyber Security and Digital Forensics
Volume – 5, Issue – 1, June 2026
E – ISSN – 2583-7559

A. Layer 1 — Initiation: Entry Vectors

The vishing attack lifecycle is always initiated using one of the two major entry vectors. The initial and most common is the spoofing of the number being called when the attackers use their tricks to alter the number being shown on the receiving end to mirror an authoritative entity - such as a scheduled commercial bank, government ministry or the telecom service provider. This method would directly compromise the inherent trust that users have on the systems of caller identification, which leverages that most telecommunication networks lack a strong call authentication system. The second entry point is the use of Voice over Internet Protocol (VoIP) services, whereby attackers can conceal their geographic location, use anonymity, and run extensive calling campaigns at a low cost in particular structural weaknesses of contemporary telecommunications systems.

B. Layer 2 — Execution: Attack Modalities

After the entry vector is in place, vishing attacks are implemented in three different modalities. Manual vishing is another form of fraud using human-operated calls and backed by a written social engineering conversation that is used to generate a sense of urgency, fake authority, or fake distress. More and more, though, there is the deployment of AI-based voice phishing, where generative deep learning models are used to produce voice that sounds like a trusted person family member, bank official, government representative, etc, and to some extent sounds real enough that it is highly challenging to detect using the unaided human ear. Robocall automation is a third modality, which is a combination of both automated mass-dialling infrastructure and pre-recorded or AI-synthesised vishing messages, allowing fraudulent campaigns to call thousands of targets at once with minimum human effort.

C. Layer 3 — Amplification: Credential Exploitation

The implementation layer directly drives into an amplification stage where mined out information is weaponized. OTP interception is the first-order amplification vector The most imminent example is the live chat of a vishing attack, where the attackers attempt to social-engineer their unsuspecting victims into revealing One-Time Passwords in real time, by-passing the security measure that multi-factor authentication is meant to offer. This helps bypass authentication, whereby, extracted credentials are exploited to obtain unauthorized access to banking websites, telecommunication accounts, and government services websites. Furthermore, personal data obtained in the course of vishing calls is also used more and more to make a SIM swap request, where attackers provide the victim with their information to a telecommunication company and get an actual replacement SIM card - thus being able to redirect all further OTPs and calls to an attacker-controlled device.

D. Layer 4 — Impact: Downstream Consequences

The implications of a successful vishing attack are downstream and work in three dimensions. On the personal level, the victims are defrauded financially in the form of unauthorised banking procedures, diversion of UPI payments, and compromise of digital wallets. On the identity level, vishing, OTP interception and SIM swapping off-the-counter result in total identity takeover - a state of affairs whereby the attacker obtains complete control over the victim's online and monetary identity. On a systemic level, most modern vishing activities are organized at the cross-jurisdictional level, and this is a cross-border activity of fraud on the basis of regulatory fragmentation between national telecommunications regulatory bodies and an obstacle to traditional law enforcement tools.

V. METHODOLOGY

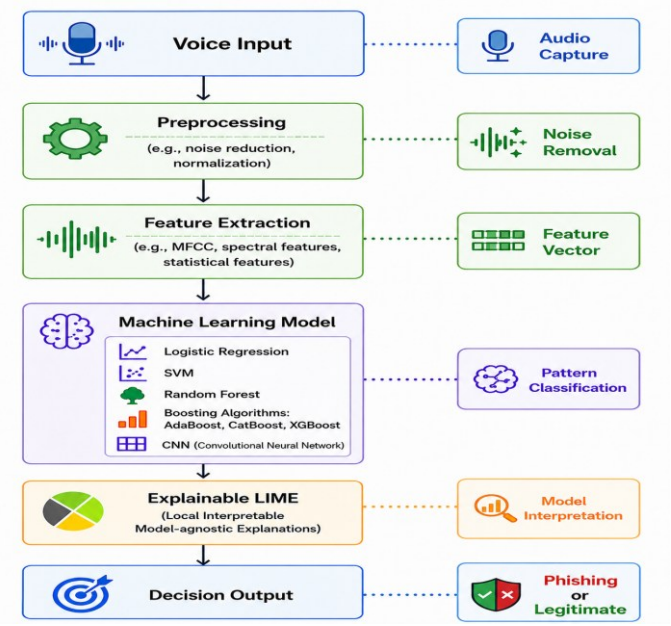


Fig 3: System architecture

The recommended supervised machine learning pipeline had mixed results for each of the classifiers, with SVM being the most balanced and most successful with an accuracy of 0.88 on all of the evaluation measures, and being the best one to build an effective decision boundary in high dimensional acoustic feature space, which is a key requirement for operational capability in real time telecom fraud detection. Random Forest had the best recall value of 0.88 and the comparatively lowest precision value of 0.72, thus a higher

false positive rate. In today's high-risk business of fraud prevention from the telecom service providers, however, there's always a compromise that can be made for that, but it may mean a greater cost of operation in implementing the solution over a large scale because of the false alarms generated.

The result showed the overall accuracy of Logistic Regression was 0.81, which again demonstrated the inherent limitation of linear classifiers in modelling the non-linear acoustic relationships, and also supported the discriminative power of the feature set for systems with limited computation power. However, the results of the hybrid Gradient Boosting-SVM architecture showed comparable performance metrics, with a recall of 0.88, precision of 0.70 and F1-score of 0.78, suggesting that the complexity of the architecture is not always proportional to the improvement in detection.

Additionally, a Convolutional Neural Network (CNN) was integrated to enable deep feature learning by automatically learning spatial and hierarchical patterns from acoustic representations like MFCC-based feature maps and spectrals features. CNN significantly improved the capacity to recognize the subtle patterns in voices, indicators of voice spoofing, and complex voice variations that occur with AI-generated and fraudulent speech, while traditional machine learning models focused mainly on handcrafted feature vectors. For optimal generalization, however, CNN needs to have more computational resources and larger data sets, which is important for real-time telecom deployments where performance is a critical factor.

VI. Results

A. Comparision Of Ml Models

SVM was the best performer with a balanced accuracy of 0.88 on all metrics, surpassing the performance of Random Forest that had a high recall but a high false positive on the other and Logistic Regression that was limited by the linear nature. It is noteworthy that the hybrid Gradient Boosting-SVM model gave a lower F1-score of 0.78 which did not offer any significant improvement over the standalone SVM and indicated that the architecture complexity did not increase the performance proportionally with the current dataset size.

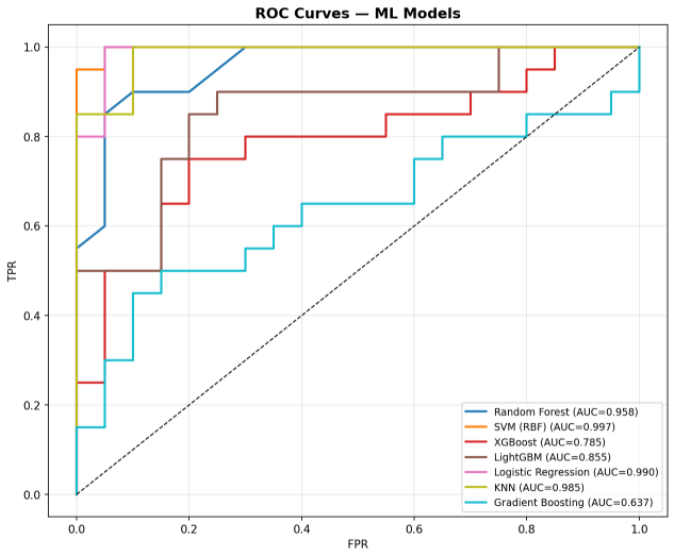


Fig 4: ROC Curve Of ML Model

VII. Conclusion

The study presents an enhanced telecom vishing detection system comprising classical machine learning classifiers along with a CNN-based deep learning framework and tested on FOR-2sec subset of Fake-or-Real dataset. Although the CNN model achieved the best classification performance with an accuracy score of 1.00 by automatically learning the hierarchical features from Mel-spectrogram representations, its unstable validation curve suggests overfitting as a result of the small size of the dataset and should be treated with care. LIME-based explainability also generates auditable decision rationales that can assist in the prosecution of fraud and regulatory compliance per TRAI anti-spoofing guidelines and the STIR/SHAKEN standard. Future work will involve expanding the training set to larger and more diverse corpora, adopting LSTM and Transformer networks, data augmentation and dropout regularization techniques, and incorporating continual learning techniques for ongoing detection performance against changing voice synthesis techniques.

Acknowledgment

We thank Christ (Deemed to be university) for providing necessary arrangement and facility for carrying the research work.

References

[1] M. Pietri et al., “Telecom spam and scams in the 5G and AI era: Analyzing economic implications, technical challenges and global regulatory efforts,” International Journal of Information Security, Springer Nature, 2025.

[2] J. Tripathy and S. Subudhi, "Effect of data imbalance in telecom fraud," in Proc. Int. Conf. Computational Intelligence in Pattern Recognition (CIPR 2024), Springer, Singapore, 2025..

[3] I Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. Discover Computing, 28(1). https://doi.org/10.1007/s10791-025-09550-0

[4] H. Khatun, A. Khan, R. Mehera, and R. K. Pal, "Supervised machine-learning methods and its application on automated telecom fraud detection," in Proc. Int. Conf. Smart Systems and Wireless Communication (SSWC 2024), Springer, Singapore, 2025

[5] T., Breitinger, F., & Baier, H. (2025). Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases. Forensic Science International Digital Investigation, 52, 301882. https://doi.org/10.1016/j.fsidi.2025.301882

[6] Firc, A., Malinka, K., & Hanáček, P. (2025). Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors. Cybersecurity, 8(1). https://doi.org/10.1186/s42400-024-00346-1

[7] Pontorno, O., Guarnera, L., & Battiato, S. (2025). DeepFeatureX-SN: Generalization of deepfake detection via contrastive learning. Multimedia Tools and Applications, 84(39), 47721–47740. https://doi.org/10.1007/s11042-025-21060-1

[8] Panda, S. P. (2025). The evolution and defense against social engineering and phishing attacks. International Journal of Science and Research (IJSR), 14(5), 397–408. https://doi.org/10.21275/sr25504223645

[9] Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J., & Sheppard, J. (2023). ChatGPT for digital forensic investigation: The good, the bad, and the unknown. Forensic Science International Digital Investigation, 46, 301609. https://doi.o

[10] S. Wang et al., "Spoofing against spoofing: toward caller ID verification in heterogeneous telecommunication systems," ACM Transactions on Privacy and Security, 2023.

[11] M. Abdeldayem, "The Fake-or-Real (FoR) Dataset (deepfake audio)," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-datasetM. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.