VISHING DETECTION SYSTEM USING TEXT-BASED ANALYSIS TO PREVENT VOICE PHISHING SCAMS

Norhidayah Muhammad[1], Nor Aida Akma Alias, Siti Dhalila Mohd Satar, Nazirah Abd Hamid, Mumtazimah Mohamad

[1]Faculty of Informatic and Computing, Tembila Campus, Besut, Terengganu

ABSTRACT

Vishing or voice phishing is another form of cyber-attack that has remained a challenge for quite a while now. Scammers employ psychological manipulation to circumvent conventional security defense systems. The number of incidents involving fraud through telecommunication in Malaysia has increased tremendously, with the Malaysian Police Force reporting an increase of 47.3% in the number of cases recorded from 2022 to 2023. These fraud incidences have made substantial contributions towards the loss of RM1.2 billion according to records by Commercial Crime Investigation Department (CCID) Malaysia. In addition, statistics show that over 60% of the fraudulent cases are carried out through sophisticated social engineering techniques, which are not detected by the traditional network filters. Modern security defense systems have proved inadequate since they depend on a blacklist of numbers that is outdated. This project builds an automated system for the identification of vishing based on the analysis of two major components: acoustic features (speech sounds) and text (message content). By using genuine data collected from both YouTube and TikTok platforms, audio patterns are extracted using Mel-Frequency Cepstral Coefficients (MFCC) analysis, while transcription is performed using OpenAI Whisper for text-based analysis. There are two different classifiers of Naïve Bayes which are used independently: one classifier uses acoustic features, whereas the other uses transcriptions from the text. The combination of both classifiers is done using weights in such a way that text gets 80%, whereas acoustic features get 20% weightage. The method classifies each call as either a fraud, suspicious call, or genuine one, and its performance is measured based on parameters such as accuracy, precision, recall, and F1-Score. The findings from an experiment conducted using a sample set of 32 audio recordings (17 fraud and 15 genuine) revealed an accuracy of 96.87%, precision rate of 100%, recall of 94.12%, and F1-score of 97%. It shows the feasibility of using the developed hybrid method despite its constraints including the small amount of training data.

KEYWORDS

Vishing Detection, Mel-Frequency Cepstral Coefficients (MFCC), Naïve Bayes classifier, OpenAI Whisper, Multimodal classification, Social Engineering

1. INTRODUCTION

Vishing is a widely practiced form of cybercrime whereby the attackers pretend to be legitimate sources like the banks, government departments, and the police in their voice phishing attempts to psychologically manipulate their targets into revealing private details [1]. The conventional means for detecting phishing attacks generally involve caller ID validation and blacklisting, whereby any number or impersonation that seems suspicious is checked against an existing list of known scams; nevertheless, such techniques are always reactive since they can only work with the help of previous data, cannot detect novel phishing attacks as the scammers keep changing their phone numbers to avoid detection, and lack the capability of analyzing the content of the call in terms of language or tone [1, 3, 13].

This research introduces an intelligent system designed for the detection of fraudulent phone calls in the Malay language. This paper uses a real-world dataset consisting of fraudulent and legitimate voice recordings obtained from open sources such as YouTube and social media. This hybrid vishing detection system employs the use of MFCC-based audio features [11], Whisper automatic speech recognition model [14], Naïve Bayes text classifier [10], and a keyword scoring method. The proposed keyword scoring process classifies suspicious words according to their risk categories using information collected from Malaysian authorities' reports as well as news articles. The differential weighing of keyword risk categories is based on the idea behind feature importance ranking with Term Frequency-Inverse Document Frequency, TF-IDF methodology [15]. The hybrid fusion approach is inspired by Kim et al.'s multimodal detection methodology [13]. The novelty of the proposed system lies not in the individual techniques employed, but in the targeted adaptation to the Malay language and Malaysian vishing landscape. To the best of the authors' knowledge, this is among the first studies to develop an automated vishing detection system specifically designed for Malay language that addressing a critical gap identified in the literature [17, 21].

2. RELATED WORK

2.1. Existing Vishing Detection Approaches

Existing vishing detection approaches can be broadly categorised into text-based, audio-based, and multimodal methods, each with their own strengths and limitations. Text-based systems relying on blacklisted phone numbers and keyword identification are inherently reactive, they can only flag previously reported threats and fail to detect new attacks as scammers regularly modify their scripts [17]. Text classifiers are also heavily dependent on speech-to-text transcription accuracy, background noise and poor call quality introduce errors that degrade classification performance [18]. Furthermore, adversarial scripts generated by large language models have been shown to significantly reduce the detection performance of conventional machine learning classifiers [19].

Audio-based methods relying solely on MFCC or spectrogram features have demonstrated reduced robustness against modern voice conversion technologies producing highly natural-sounding synthetic speech [13]. Sumbiri and Jonathan [20] demonstrated that despite achieving high accuracy, the SVM classifier showed an extremely low ROC AUC of 0.19, indicating poor class differentiation, highlighting that acoustic features alone are insufficient for reliable detection. Although multimodal systems combining text and audio have shown improved performance [13], most existing research focuses on English or Korean. The system by Kim et al. [13] was trained and evaluated exclusively on Korean datasets, limiting its generalizability. Languages such as Bahasa Melayu remain critically underrepresented [21], despite the language being the primary communication medium for Malaysian users, particularly elderly individuals who are frequent vishing targets.

2.2. Reference System Analysis

Kim et al. [13] proposed a multimodal voice phishing detection system integrating KoBERT-based text classification with a CNN-BiLSTM model for MFCC-based synthetic speech detection. As illustrated in Figure 1, the system applies speaker diarization to separate speakers and performs parallel text and audio analysis before fusing results using an 80:20 weighted average (text:audio), achieving an F1-score of 0.994 on Korean datasets [13]. Their empirical finding that text-dominant fusion configurations produce more stable results than voice-dominant configurations directly inform the fusion strategy adopted in the present study. The key

distinction of the proposed system is its adaptation to the Malay language context using computationally lighter Naïve Bayes classifiers and a domain-specific keyword scoring mechanism grounded in documented Malaysian vishing cases [4, 5, 6, 7], which is absent in the reference system.

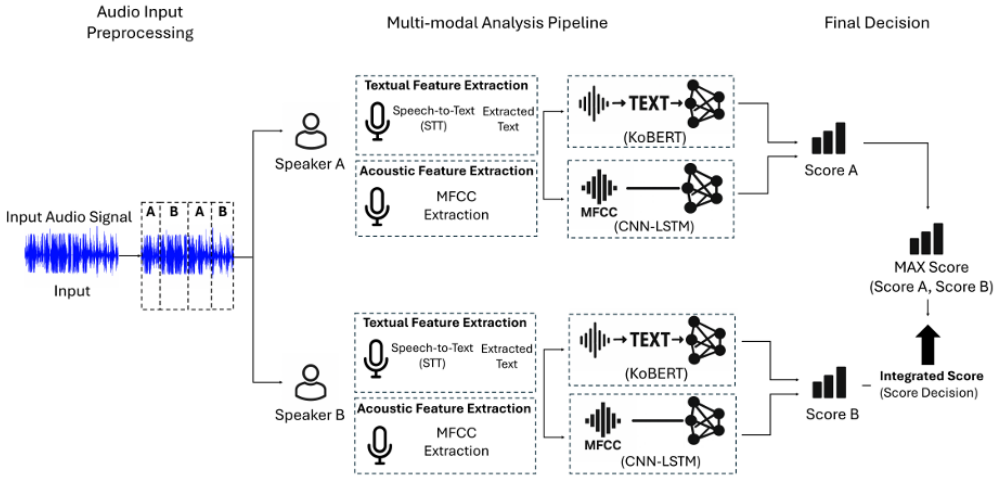

Figure 1. Overall architecture of the proposed multimodal voice phishing detection system by Kim et al. [13]. The system applies speaker diarization to separate speakers and then performs parallel analysis using a text-based and an audio-based detection model

3. METHODOLOGY

3.1. Framework

The vishing detection system under consideration adopts the pipeline approach which is characterized by six distinct modules: (i) audio preprocessing and MFCC features extraction, (ii) Audio Naïve Bayes classifier, (iii) speech to text transcriptions through OpenAI Whisper, (iv) text preprocessing, (v) keyword scoring and Text Naïve Bayes classifier, and (vi) fusion and decision making. Figure 2 depicts the overall structure of the system architecture.

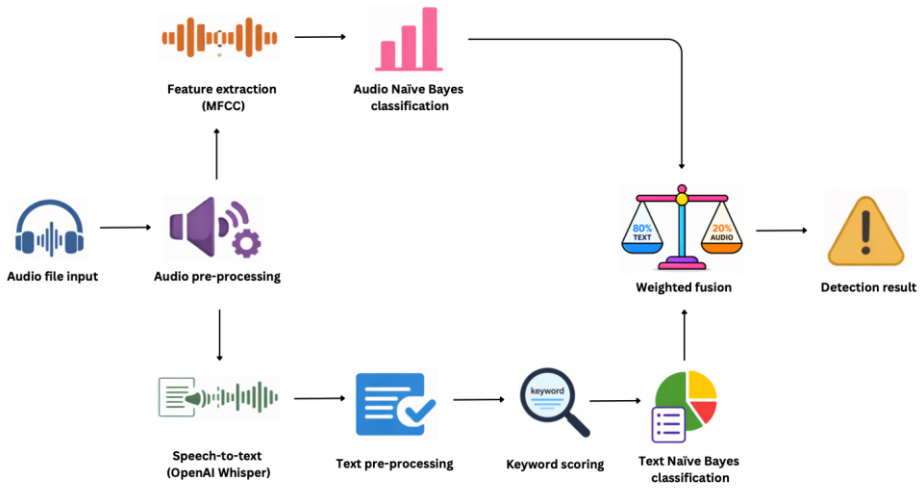

Figure 2. Framework of vishing detection system

3.2. MFCC Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) have been extensively applied in the field of speech analysis since they contain perceptually relevant information about human speech, such as tone color and spectral envelope shape [11]. The extraction of MFCCs is performed on each audio sample via the librosa module. For each audio signal, thirteen MFCC coefficients are generated, which is a common approach in speech feature extraction [11, 12]. Furthermore, the mean and standard deviation values of each coefficient in all frames are computed, leading to a 26-dimensional vector that describes the spectral and temporal features of the speaker's voice [13]. According to Kim et al., MFCCs offer the best trade-off between performance and efficiency for acoustic representation when performing vishing identification tasks [13]. MP3 audio samples are transformed to WAV format through the pydub module before extracting their MFCC coefficients.

3.3. Speech-to-Text Transcription

The process of converting the audio content to Malay language text is done through OpenAI Whisper (Small model). Whisper is a massive weakly supervised automatic speech recognition model pre-trained on 680,000 hours of multilingual and multitask data [14], which performs effectively on low-resource languages, such as Malay [14]. According to Radford et al. [14], the training method of Whisper allows for effective speech recognition in a variety of languages and acoustic conditions. The conversion is made using the language argument set to 'ms', ensuring the correct language conversion. The selection of Whisper to convert the audio content follows the methodology presented by Kim et al. [13].

3.4. Text Preprocessing

Prior to passing the transcribed text to the text classifier, the transcribed text goes through three stages of preprocessing according to the standard text preprocessing practices in NLP [16]. Text preprocessing is an essential process that helps reduce noise in text and improve feature representations in NLP pipelines [16]. To begin with, all characters in the transcribed text are made lowercase. Secondly, the text is cleaned of punctuations using regular expressions, as punctuation does not play much semantic role in fraudulent texts [16]. Lastly, the text is cleaned of stopwords in the Malay language (examples include 'yang', 'dan', 'di', 'ke', 'dari'). Elimination of stopwords is important to eliminate redundant high-frequency words from texts that have little or no discriminative value [16]. The processed text is fed into the TF-IDF Vectorizer which gives high importance to words that have better discriminating capacity [15].

3.5. Dual Naïve Bayes Classification

The proposed system uses two separate Naïve Bayes classifiers that run on different modalities, as per the dual modality analysis technique described by Kim et al. [13]. The Naïve Bayes classifier works using Bayes' theorem under the condition that features are independent given the class label [10]. Even though this is a restrictive assumption, Naïve Bayes classifiers perform efficiently in a broad range of applications and especially in scenarios where training data is scarce [10].

Model 1 is the Audio Naïve Bayes, which is a Gaussian Naïve Bayes (GaussianNB) that uses the 26-dimensional MFCC feature vectors as its inputs. GaussianNB is used since the MFCC is a continuous feature vector that can contain negative values, and the Gaussian likelihood assumption models continuous features under the assumption that the features are normally

distributed in every class [10]. Model 2 is the Text Naïve Bayes, which is a Multinomial Naïve Bayes (MultinomialNB) that uses TF-IDF weighting of processed Malay transcripts. The TF-IDF scoring method emphasizes terms that are unique to certain documents while de-emphasizing terms that have high occurrence frequency [15]. The output of both classifiers is the probability score for each class (1 for fraud; 0 for non-fraud), presented as a percentage from 0 to 100.

3.6. Tiered Keyword Scoring Mechanism

Apart from the Naïve Bayes classifiers, the system employs a keyword scoring module that analyzes the transcription of the audio for pre-defined keywords considered suspicious. In contrast to the system employed by Kim et al. [13], which is entirely dependent on the use of deep learning classifiers, the proposed system uses the additional feature of keyword scoring based on the documented vishing methods employed in Malaysia [4, 5, 6, 7]. The various categories of the keywords were defined on the basis of the following information:

The high-risk keywords (+3 score) are 'otp', 'kata laluan', 'nombor pin', 'saman tertunggak', 'akaun disekat', and 'kes jenayah'. These were gathered from the official warning statements provided by the National Scam Response Center (NSRC). NSRC specifically emphasizes that legitimate agencies will never ask for financial account details such as passwords, PIN, TAC, or OTP from the general public [5]. The LHDN stated that the vishing syndicate informs the victim that his/her financial account will be segregated under the Anti-Money Laundering Act and that he/she is associated with criminal activities [4].

The medium risk keywords (score +2) are "disekat," "mahkamah," "lhdn," "sprm," "saman," and "bungkusan." They were found using the Macau Scam reported by Astro Awani, where the scam victim was told that there was a parcel (bungkusan) with contraband goods that was registered in her name, and she would face legal prosecution (mahkamah) for that [6]. The BSN stated that impersonation of court personnel, LHDN, and SPRM is a typical modus operandi used in Malaysia's vishing attacks [7].

Keywords with low risks (score +1) consist of general finance words like "bank," "akaun," "bayaran," "segera," and "polis," which show up in scam contexts but could also be used in honest communications [7]. It is theoretically reasonable for there to be a difference in the weights of the keywords because of how features work in the TF-IDF model, where more specific terms receive higher weights [15]. Table 1 presents the keyword categories along with their sources.

Table 1. Keyword Tier Categories and Sources

Tier	Score	Example Keywords	Source
High Risk	+3	otp, kata laluan, nombor pin, akaun disekat, kes jenayah	NSRC [5], LHDN [4]
Medium Risk	+2	mahkamah, bungkusan, lhdn, sprm, saman	BERNAMA [6], BSN [7]
Low Risk	+1	bank, akaun, bayaran, segera, polis	BSN [7], LHDN [4]

3.7. Weighted Fusion and Verdict Generation

The overall score for final classification is obtained using the weighted sum of the text-based and audio-based Naïve Bayes confidence score using the late-fusion approach as discussed by Kim et al. [13]. Kim et al. [13] experimentally observed that a 80:20 ratio between text and audio offers the most consistent performance in terms of:

$$\text{Final Score} = (0.8 \times \text{Text Score}) + (0.2 \times \text{Audio Score})$$

If the value of the score is above 70, the call is considered fraudulent; if the value of the score is above 50 or the score of the keyword is above the suspicious value (2), the call is considered suspicious, otherwise it is a legitimate call. The threshold value of 70 was chosen after Kim et al. [13] and their multimodal fusion approach, where they have set the exact threshold. The ratio between the textual information and the acoustic information is set as 80:20 since the textual information is more reliable than the acoustic one, due to the lack of big training data in the acoustic domain as found by Kim et al. [13].

4. DATASET

Since there are few available and open-source audio datasets for fraud calls made in Malay language, an exclusive dataset was developed for this research purpose. The dataset is composed of a total of 32 audio samples: 17 audio samples of scam calls and 15 audio samples of genuine calls. Scam audio samples were collected from actual vishing cases prevalent in Malaysia, such as impersonating government institutions (LHDN, IPD), delivery companies (Pos Laju), and telecommunications firms (Telekom Malaysia). This follows previous research findings on vishing tactics in Malaysia [4, 6, 7]. Legitimate samples include phone recordings of regular phone calls.

It should be noted that the number of samples for this experiment, 32 samples, is relatively low compared to other experiments using similar methodologies. For instance, Kim et al. [13] worked on over 25,000 samples and Lee and Park [8] made use of thousands of transcribed audio clips. This is mainly because the number of available samples in the current research is rather limited since there are only a few sources of public vishing materials in the Malay language makes it impossible to collect more data.

5. RESULTS

5.1. Evaluation Metrics

The performance of the proposed model will be measured through four evaluation criteria: accuracy, precision, recall, and F1 score. These criteria will be determined based on the confusion matrix that divides prediction results into four categories, namely TP, TN, FP, and FN.
Accuracy measures the proportion of correctly classified instances out of all instances:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision measures the proportion of correctly detected fraudulent calls among all calls predicted as fraudulent:

$$\text{Precision} = TP / (TP + FP)$$

Recall measures the proportion of actual fraudulent calls that were successfully detected:

$$\text{Recall} = TP / (TP + FN)$$

F1-Score is the harmonic mean of precision and recall, providing a balanced measure particularly useful for imbalanced datasets [Basit et al., 2021]:

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Precision is particularly important in vishing detection to minimize false alarms on legitimate calls, while recall ensures that fraudulent calls are not missed, both have direct implications for user safety and system reliability [13].

5.2. Confusion Matrix

Table 2 presents the confusion matrix obtained from evaluating the proposed system on the full dataset of 32 audio samples. The confusion matrix shows that the system correctly identified all 15 legitimate calls (TN = 15) with zero false positives and successfully detected 16 out of 17 fraudulent calls (TP = 16). The single false negative represents one fraudulent call misclassified as legitimate, while the absence of false positives confirms that the system does not incorrectly flag genuine calls as fraud, a desirable property for real-world deployment.

Table 2. Confusion Matrix

	Predicted Fraud	**Predicted Legitimate**
Actual Fraud	16 (TP)	1 (FN)
Actual Legitimate	0 (FP)	15 (TN)

5.3. Performance Calculation

Based on the confusion matrix values (TP = 16, TN = 15, FP = 0, FN = 1), the performance metrics are calculated as follows:
Accuracy = (16 + 15) / (16 + 15 + 0 + 1) = 31 / 32 = 0.9687 = 96.87%
Precision = 16 / (16 + 0) = 16 / 16 = 1.00 = 100%
Recall = 16 / (16 + 1) = 16 / 17 = 0.9412 = 94.12%
F1-Score = 2 × (1.00 × 0.9412) / (1.00 + 0.9412) = 2 × (0.9412 / 1.9412) = 0.9700 = 97.00%
Table 3 summarises the classification performance of the proposed system. The 100% precision rate stands out the most due to the lack of false positives from the process. The 94.12% recall rate means that there was one instance of fraud missed due to transcription errors made by the Whisper model that led to the failure to detect suspicious keywords. The high F1-score of 97.00% demonstrates that there is a good tradeoff between precision and recall rates.

Table 3. Performance of the Proposed System

Metric	**Value**
Accuracy	96.87%
Precision	100.00%
Recall	94.12%
F1-Score	97.00%

5.4. Comparative Analysis

Table 4 presents a comparative analysis between the proposed system and the reference multimodal vishing detection system by Kim et al. [13]. As shown in Table 5, the proposed system achieves competitive results compared to Kim et al. [13] despite significant differences in classifier complexity and dataset scale. The 2.9% accuracy gap (96.87% vs 99.90%) and 2.4% F1-score gap (97.00% vs 99.40%) are attributable to the use of Naïve Bayes versus KoBERT and the substantially smaller training dataset (32 vs 25,000+ samples). Notably, the proposed system achieves a higher precision (100% vs 99.80%), indicating fewer false alarms on legitimate calls. The proposed system also introduces a domain-specific tiered keyword scoring mechanism grounded in Malaysian vishing modus operandi [4, 5, 6, 7], which is absent in the reference system.

Table 4. Comparison with Reference System

Aspect	**Proposed System**	**Kim et al.**
Language	Bahasa Melayu	Korean
Text Classifier	Multinomial Naïve Bayes [10]	KoBERT (Transformer) [13]
Audio Classifier	Gaussian Naïve Bayes (MFCC) [10, 11]	CNN-BiLSTM (MFCC) [13]
Fusion Strategy	Weighted Average (80:20) [13]	Weighted Average (80:20) [13]
Keyword Scoring	Yes (3-tier) [4, 5, 6, 7]	No
Speaker Diarization	No	Yes [13]
Dataset Size	32 samples	25,000+ samples [13]
Accuracy	96.87%	99.90% [13]
Precision	100.00%	99.80% [13]
Recall	94.12%	98.20% [13]
F1-Score	97.00%	99.40% [13]

5.5. Result Discussion

The proposed model recorded 96.87% accuracy, 100% precision, 94.12% recall, and 97.00% F1 score. The precision of 100% is very important here because this shows that there were no false positives, which means all calls that were identified as legitimate were indeed legitimate. In real-world deployment, false alarms on legitimate calls erode user trust and reduce system adoption [13], making high precision a critical property for any vishing detection system. The single misclassification involved one fraudulent call incorrectly classified as legitimate (false negative), likely caused by transcription inaccuracies from the Whisper model that resulted in key suspicious keywords being mistranslated and subsequently undetected by both the keyword scoring mechanism and the Naïve Bayes classifier.

Compared to existing literature, the proposed system demonstrates competitive performance relative to its design constraints. Lee and Park [8] achieved strong results on Korean vishing data using machine learning-based text classification, while Moussavou Boussougou and Park [9] reported improved performance using a CNN-BiLSTM hybrid model. Both studies, however, benefited from substantially larger datasets and focused on languages with more abundant training resources. The proposed system differs in its explicit focus on the Malay language, a domain critically underrepresented in existing vishing detection research [17, 21]. The tiered keyword scoring mechanism, grounded in documented Malaysian vishing modus operandi [4, 5, 6, 7], introduces a domain-specific layer absent in the reference system [13], providing targeted detection capability for Malaysian fraud scenarios involving LHDN impersonation, Macau Scam tactics, and courier-based fraud.

The performance gap relative to Kim et al. [13] is 99.90% accuracy versus 96.87% is primarily attributable to two factors. First, Kim et al. employed KoBERT, a transformer-based model pretrained on large-scale Korean corpora, which captures deeper contextual semantic relationships through self-attention mechanisms compared to the Multinomial Naïve Bayes classifier [13]. Second, the reference system trained on over 25,000 samples provides substantially greater statistical coverage of vishing patterns. The Naïve Bayes assumption of feature independence [10] further limits the proposed system's ability to model complex inter-feature relationships. Despite these limitations, the 80:20 weighted fusion strategy adopted from Kim et al. [13] proved effective, consistent with their empirical finding that text-dominant fusion configurations produce more stable results than voice-dominant configurations, particularly when audio training data is limited.

Several opportunities for further improvement are identified. First, expanding the dataset through collection of real-world Malay-language vishing recordings or LLM-based data augmentation would substantially improve model generalizability. Kim et al. [13] demonstrated that ChatGPT-based scenario generation can effectively diversify training data while preserving phishing intent. Second, incorporating speaker diarization as implemented by Kim et al. [13] would enable independent analysis of each speaker's utterances in multi-party calls. Third, replacing the Naïve Bayes classifiers with more expressive models such as fine-tuned multilingual BERT variants or CNN-BiLSTM architectures could close the performance gap with state-of-the-art systems. Finally, a proper train-test split evaluation with a larger dataset would provide more reliable and generalizable performance estimates beyond the in-sample testing conducted in this study.

6. CONCLUSIONS

The proposed study involved the creation of a combined approach to vishing detection in the Malay language that combines MFCC audio features, OpenAI Whisper transcriptions from speech to text, text preprocessing, two-stage classification using Naive Bayes algorithms, as well as a keyword score based on the known methodology of vishing in Malaysia. The method uses the weighted approach (80%/20%, text/audio) inspired by the reference paper [13] and has achieved 96.87% accuracy, 100% precision, 94.12% recall, and 97% F1 score on 32 Malay audio clips. The primary contribution of this work is the adaptation of established multimodal detection techniques to the underserved Malay-language domain, incorporating a keyword scoring mechanism grounded in real documented Malaysian vishing modus operandi [4, 5, 6, 7] is a domain-specific element absent in existing systems [13].

Though the findings are encouraging, there are a few limitations associated with the study. Firstly, the size of the dataset, containing 32 instances, is much smaller when compared to those in other similar studies [13]. Secondly, the evaluation was performed on the entire dataset without any division into training and testing phases, resulting in overly positive performance metrics. Lastly, the use of the Naïve Bayes algorithm assumes that features are independent of one another [10], and the lack of speaker diarization further restricts the capabilities of the proposed system when compared to the reference system [13].

REFERENCES

[1] S. Ashfaq, P. Chandre, S. Pathan, U. Mande, M. Nimbalkar, and P. Mahalle, "Defending against vishing attacks: A comprehensive review for prevention and mitigation techniques," in Cyber Security and Digital Forensics, REDCYSEC 2023, Lecture Notes in Networks and Systems, vol. 896, Springer, Singapore, 2024, pp. 411–422.

[2] Sinar Harian, "Malaysia catat 2.98 juta panggilan scam, kerugian cecah RM1.57 bilion," 2025. [Online]. Available: https://www.sinarharian.com.my/article/715683

[3] Astro Awani, "107,716 kes penipuan dalam talian dikesan dari 2020 hingga 2023," 2024. [Online]. Available: https://www.astroawani.com/video/video-terkini-x7sio1/penipuan-atas-talian-20202023-babitkan-kerugian-rm32bilion-x8tszck

[4] Lembaga Hasil Dalam Negeri Malaysia (LHDN), "Peringatan penipuan: Modus operandi penipuan mengaitkan nama LHDNM," 2024. [Online]. Available: https://www.hasil.gov.my/peringatan-penipuan/

[5] National Scam Response Centre (NSRC), "Maklumat lanjut mengenai pusat respons penipuan kebangsaan," 2023. [Online]. Available: https://nfcc.jpm.gov.my/index.php/ms/nsrc

[6] Astro Awani, "Akibat panik, penjawat awam rugi RM429,000 ditipu Macau Scam," 2022. [Online]. Available: https://www.astroawani.com/berita-malaysia/akibat-panik-penjawat-awam-rugi-rm429000-ditipu-macau-scam-378000

[7] Bank Simpanan Nasional (BSN), "What is a Macau Scam?" 2024. [Online]. Available: https://bsn.com.my/page/macau-scam
[8] M. Lee and E. Park, "Real-time Korean voice phishing detection based on machine learning approaches," Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 8173–8184, 2023.
[9] M. K. Moussavou Boussougou and D. J. Park, "Attention-based 1D CNN-BiLSTM hybrid model enhanced with FastText word embedding for Korean voice phishing detection," Mathematics, vol. 11, no. 14, p. 3217, 2023.
[10] I. Rish, "An empirical study of the Naïve Bayes classifier," in Proc. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, no. 22, 2001, pp. 41–46.
[11] B. Elizalde and D. Emmanouilidou, "Detection of robocall and spam calls using acoustic features of incoming voicemails," Proceedings of Meetings on Acoustics, vol. 45, p. 060004, 2021.
[12] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio deepfake detection: What has been achieved and what lies ahead," Sensors, vol. 25, p. 1989, 2025.
[13] J. Kim, S. Gu, Y. Kim, S. Lee, and C. Kang, "A multimodal voice phishing detection system integrating text and audio analysis," Applied Sciences, vol. 15, no. 20, p. 11170, 2025.
[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.
[15] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing & Management, vol. 39, no. 1, pp. 45–65, 2003.
[16] E. Loper and S. Bird, "NLTK: The natural language toolkit," in Proc. ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, 2002, pp. 63–70.
[17] A. E. Chichwadia and N. Mpekoa, "A machine learning algorithm to detect and prevent smishing and vishing," 2024.
[18] I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," Procedia Computer Science, vol. 184, pp. 853–858, 2021.
[19] W. Li, S. Manickam, Y. Chong, and S. Karuppayah, "Talking like a phisher: LLM-based attacks on voice phishing classifiers," arXiv preprint arXiv:2507.16291, 2025. [Online]. Available: http://arxiv.org/abs/2507.16291
[20] Z. Sumbiri and D. Jonathan, "Identifying and evaluating the best machine learning predictive models for detecting voice (phone-call) vishing attacks on MoMo users in real time," EdinBurg Peer Reviewed Journals and Books Publishers Journal of Information and Technology, vol. 5, no. 6, 2025.
[21] H. Cho and M. Seo, "Towards reliable and practical phishing detection," 2025.

AUTHORS

Norhidayah Muhammad received her Ph.D. from Universiti Teknologi Mara (UITM), her M.Sc. from Universiti Malaysia Pahang (UMP) and her BIT from Universiti Malaysia Terengganu (UMT). She is a lecturer at Universiti Sultan Zainal Abidin, specializing in software project management, cryptography and cryptanalysis (including key management and random number generation), data hiding and steganography, data security and digital security and privacy, and data encryotion.

Nazirah Abd Hamid is a lecturer in University Sultan Zainal Abidin, Terengganu, Malaysia. She received her BIT from Universiti Utara Malaysia (UUM), her M.Sc. from Universiti Teknologi Malaysia (UTM) and Ph.D. from Universiti Teknikal Malaysia Melaka (UTeM). She is specializing in biometrics security system, security services (including digital forensic, steganography, network security, and public key infrastructure and biometrics.

Siti Dhalila Mohd Satar received her BIT from Universiti Kebangsaan Malaysia (UKM), her M.Sc. in Universiti Teknologi Malaysia (UTM) and her Ph.D. from Universiti Putra Malaysia (UPM). She is a lecturer and also academic program

coordinator (PPA) Bachelor of Computer Science (computer network security with honours) in Universiti Sultan Zainal Abidin. She is specializing in authentication system and security services (including digital forensic, steganography, network security, and public key infrastructure and biometrics).

Mumtazimah Mohamad was born in Terengganu, Malaysia. She received a bachelor's degree in information technology from Universiti Kebangsaan Malaysia, in 2000, an M.Sc. degree in computer science from Universiti Putra Malaysia, and a Ph.D. degree in computer science from Universiti Malaysia Terengganu in 2014. She was a Junior Lecturer in 2000. Currently, she is an Associate Professor with the Department of Computer Science, Faculty of Informatics and Computing (FIK), Universiti Sultan Zainal Abidin, Terengganu, Malaysia. She has published over 50 research articles in peer-reviewed journals, book, chapters, and proceedings. She has been appointed as a reviewer and technical committee for numerous conferences and journals and has worked as a researcher in several nationally funded Research and Development projects. Her research interests include pattern recognition, machine learning, artificial intelligence, and parallel processing.

Nor Aida Akma Alias is a Final Year Student pursuing her B.Sc. in Computer Science (Computer Network Security) at Universiti Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia. Her academic background covers cybersecurity disciplines including penetration testing, cryptography, computer forensics, supported by a strong foundation in artificial intelligence and have hands-on exposure to IoT. Her final year research focuses on developing an automated vishing detection system using MFCC feature extraction and Naïve Bayes classification for Malay language.