Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 256-270
DOI: https://doi.org/10.51903/jtie.v5i1.475

AI-Driven Multi- Modal Fake Content Detection System Using Audio-Text Fusion and Transformer Network

Jeeva S.*[1], Trisha J. S.[2], Keerthana S.[3]
Email: jesusjeeva1995@gmail.com, trishajaisankar01@gmail.com, keerthanaasivakumar@gmail.com
Orcid: https://orcid.org/0009-0000-5580-0169, https://orcid.org/0009-0008-9291-223X, https://orcid.org/0009-0004-9374-7986
[1,2,3]*Department Faculty of Artificial Intelligence and Data Science, Arunai Engineering College, Tiruvannamalai, India, 606603*
*Corresponding Author

Abstract
The rapid proliferation of AI-generated synthetic media has posed substantial threats to digital trust, particularly through audio deepfakes and manipulated text. Existing unimodal detection systems that analyze either audio or text in isolation remain insufficient to counter advanced generative attacks that exploit both modalities simultaneously. This paper proposes an AI-driven multimodal fake content detection framework that jointly leverages acoustic and linguistic signals to enable robust deepfake identification. Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrograms are extracted from raw audio to capture spectral and temporal vocal patterns. At the same time, BERT-based transformer embeddings encode semantic and contextual information from transcripts generated via Automatic Speech Recognition (ASR). An attention-based fusion layer dynamically weights and integrates both feature streams, and a Random Forest–XGBoost ensemble classifier performs the final authenticity prediction. Experiments conducted on the ASVspoof 2019 benchmark demonstrate a classification accuracy of 95%, with precision of 93%, recall of 94%, and F1-score of 95%, outperforming standalone audio-only and text-only baselines by approximately 4–7%. These findings confirm that cross-modal feature fusion substantially reduces false-detection rates and improves generalization over single-modality approaches. The proposed system offers practical applicability in cybersecurity, voice biometrics, and digital forensics.

Keywords: *Audio-Text Fusion, Cybersecurity, Deepfake, Fake Content Detection, Machine Learning, Transformer.*

I. INTRODUCTION

The rapid advancement of artificial intelligence has fundamentally altered how media content is generated and distributed. Deepfakes exemplify this shift, where synthetic speech and text can now be produced with near-human fidelity. While such technologies carry legitimate uses in education and entertainment, their misuse for misinformation, identity fraud, and social engineering has escalated sharply. Malicious actors increasingly leverage AI-generated audio and text to conduct phishing attacks and manipulate individuals into disclosing sensitive information, thereby eroding digital trust. As documented by (Todisco et al., 2020), the ASVspoof challenge series has underscored the growing threat landscape of spoofed and synthetic audio. Consequently, developing reliable detection mechanisms for fake content has become a pressing concern for both cybersecurity and AI ethics research communities (Al Alim et al., 2025; Hartono et al., 2024; Mai & Khalid, 2025).

on December, 2025; Revised on January, 2026; Accepted on March, 2026; Published on April, 2026

Conventional approaches to fake content detection are largely unimodal, focusing on either audio signals or textual data in isolation. Audio-based systems exploit acoustic descriptors such as MFCCs and spectral features, while text-based systems employ Natural Language Processing (NLP) methods to identify linguistic anomalies. (Tak et al., 2021) demonstrated that spectrogram-based deep learning models can achieve competitive performance on structured audio benchmarks; however, these approaches struggle to generalize when confronted with sophisticated deepfakes that replicate natural speech prosody and coherent language patterns simultaneously. The inherent limitations of single-modality frameworks, therefore, necessitate a more comprehensive detection model capable of cross-modal reasoning.

To address these limitations, this paper introduces an AI-driven multi-modal detection system that integrates acoustic and linguistic cues within a unified framework. The primary novelty lies in combining an attention-based cross-modal fusion layer with a Random Forest–XGBoost ensemble classifier, enabling the model to learn which modality contributes more discriminative evidence per sample. The system extracts MFCCs and Mel-Spectrogram features from audio, while BERT-based embeddings (Devlin et al., 2019) encode semantic context from ASR-generated transcripts. Fusion follows the attention principle established by (Vaswani et al., 2017), and final classification uses ensemble learning (Chen and Guestrin, 2016). The primary task addressed is audio deepfake detection, with text serving as supplementary context. Experimental results on ASVspoof 2019 (Todisco et al., 2020) confirm that the proposed system outperforms audio-only and text-only configurations across all major evaluation metrics.

II. LITERATURE REVIEW

A. PFAS-based Voice Spoof Detection

(Todisco et al., 2020) established the ASVspoof 2019 benchmark, which provided a standardized evaluation framework for countermeasures against text-to-speech and voice conversion attacks. MFCC-based systems trained on this dataset performed reliably against known spoofing conditions but showed limited generalization against adaptive AI-generated voices, revealing an inherent gap in purely acoustic approaches.

B. Deepfake Detection Using BERT

(Devlin et al., 2019) introduced BERT, demonstrating that deep bidirectional pre-training substantially improves natural language understanding. Subsequent applications to machine-generated text detection showed strong contextual modeling but remained limited by their inability to process acoustic cues, restricting their utility in audio-centric deepfake scenarios.

C. Spectrogram + CNN Models

(Tak et al., 2021) proposed treating audio spectrograms as two-dimensional images and classifying them via convolutional networks. This end-to-end approach achieved competitive accuracy on structured benchmarks but required large annotated datasets and incurred significant computational overhead, constraining practical deployment.

D. Multi-Modal Fusion Research

More recent research has shifted toward multi-modal detection paradigms. (Khalid et al., 2021) introduced FakeAVCeleb, a multimodal audio-video deepfake dataset supporting cross-modal evaluation. (Liu et al., 2024; Zhang et al., 2024) further explored attention-based fusion strategies for audio deepfake detection. However, a consistent gap across these works remains the absence of combined attention-weighted fusion with ensemble classification, which the present system directly addresses.

III. RESEARCH METHOD

A. Problem Statement

The proliferation of generative AI tools has made it increasingly accessible to produce highly realistic synthetic media. Audio deepfakes can replicate individual vocal characteristics including tone, rhythm, and emotional prosody, while AI-generated text closely emulates personal communication styles. As noted by (Korshunov & Marcel, 2020), such capabilities pose a direct threat to speaker verification systems and digital authentication mechanisms, since even trained evaluators find it difficult to reliably distinguish fabricated recordings from genuine ones.

Existing detection models are predominantly unimodal, examining either the audio waveform or the text transcript in isolation. This architectural constraint leads to limited generalization, elevated false-detection rates, and vulnerability to hybrid attacks in which both audio and text components are simultaneously manipulated. (Villalba et al., 2021) highlighted that speaker recognition countermeasures degrade significantly when attack conditions differ from training conditions, a problem compounded in multimodal attack scenarios.

There is, therefore, a clear need for an AI-powered multi-modal detection framework that jointly processes acoustic and linguistic signals to identify fabricated content with greater accuracy and resilience. The specific task addressed in this work is audio deepfake detection, where the primary objective is determining whether a given audio segment is genuine or synthesized, with the text transcript providing supplementary semantic context to reinforce classification decisions.

B. Existing System

Conventional fake content detection systems primarily operate on a single modality. Audio-based approaches rely on handcrafted acoustic features such as MFCCs, spectral flux, and pitch energy, which are fed into classifiers including Support Vector Machines and Random Forests. (Dhar & Das, 2021) demonstrated that such methods, augmented with data augmentation, achieve reasonable accuracy for known spoofing conditions. However, these models fail to identify AI-generated speech that closely mimics natural vocal tone and emotion, and their performance degrades noticeably in the presence of background noise or channel distortions.

Text-based detection methods use NLP and language models to classify machine-generated content. Early techniques applied n-gram statistics and logistic regression to detect syntactically irregular patterns. The introduction of BERT (Devlin et al., 2019) and LSTM-based models substantially improved contextual language understanding. (Pahwa et al., 2023) demonstrated the effectiveness of ensemble learning techniques for fake speech detection, while Ali et al. (2023) explored spectrogram-based CNN architectures for audio deepfake detection. A critical limitation remains: these text-only systems are blind to acoustic signals, which carry speaker-level cues essential for distinguishing genuine voices from synthetic ones.

The overarching weaknesses of unimodal approaches are well-documented. They cannot generalize across multimodal attack scenarios, produce elevated false alarm rates, and cannot exploit cross-modal consistency signals. Systems are particularly brittle when an input pairs a genuine-sounding voice with AI-crafted text, or vice versa. This motivates the need for a unified multimodal framework that jointly leverages acoustic and linguistic evidence for dependable fake content identification.

C. Proposed System

The proposed framework presents an AI-driven multi-modal detection system that processes both audio and text to identify manipulated or synthetic content. Unlike conventional unimodal designs, it conducts joint feature extraction and fusion, enabling the model to learn both the acoustic characteristics of how something is said and the semantic properties of what is said. The primary novelty lies in the combination of an attention-based cross-modal fusion layer with a Random Forest–XGBoost ensemble classifier, a configuration not previously explored in the surveyed literature, which jointly improves classification robustness and reduces false detection rates beyond what unimodal or non-attentive fusion approaches achieve.

The architecture comprises four principal modules: audio preprocessing, text extraction and encoding, attention-based feature fusion, and ensemble classification. Raw audio is standardized

to 16 kHz mono WAV format, noise-reduced, and silence-trimmed. MFCC features of dimension $d_a = 40$ coefficients per frame, and Mel-Spectrogram representations are computed to capture spectral and temporal vocal patterns. In parallel, ASR transcripts are tokenized and encoded through a pre-trained BERT model (Devlin et al., 2019) to produce text embedding vectors of dimension $d_t = 768$. The Transformer attention mechanism (Vaswani et al., 2017) underpins both the BERT encoder and the subsequent fusion layer.

The attention-based fusion layer concatenates the audio embedding $f_a \in \mathbb{R}^{d_a}$ and text embedding $f_t \in \mathbb{R}^{d_t}$ into a joint representation $f_{at} \in \mathbb{R}^{(d_a+d_t)}$. Attention weights α are computed as $\alpha = \text{softmax}(W_{at} \cdot f_{at} + b)$, where W_{at} and b are learnable parameters. The weighted fused vector $f^* = \alpha \odot f_{at}$ is then passed to the ensemble classifier. The Random Forest–XGBoost ensemble (Chen and Guestrin, 2016) was preferred over a purely end-to-end neural classifier for two reasons: greater interpretability of feature-level contributions and lower risk of overfitting given the dataset size used in this study. This design enables dynamic modality weighting per sample while maintaining classification transparency.

D. System Architecture

The proposed AI-Driven Multi-Modal Fake Content Detection System integrates both acoustic and linguistic features to enhance the reliability of fake content identification. The system follows a structured workflow consisting of eight major stages, as illustrated in Figure 1 of the paper. Each stage contributes to the extraction, fusion, and classification of multimodal information for accurate decision-making.

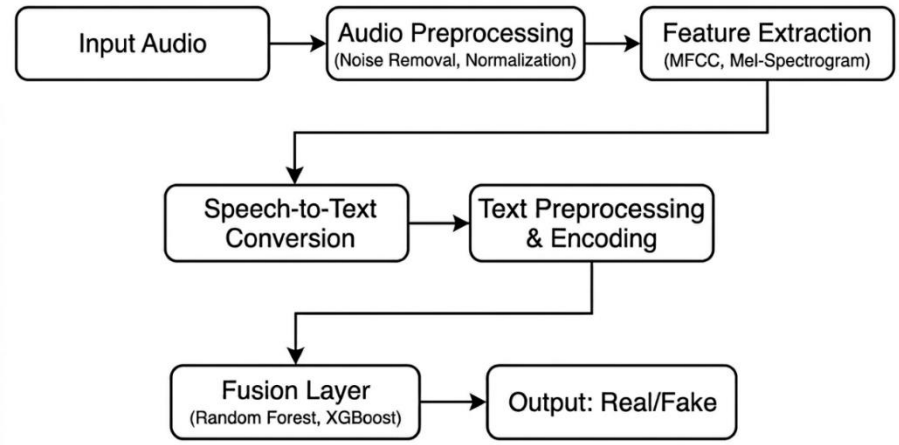

Figure 1. Flowchart of the AI-Driven Multi-Modal Fake Content Detection System

The process begins with the input audio, which serves as the primary data source for analysis. This raw audio signal undergoes preprocessing, including noise removal, silence trimming, and normalization to ensure signal uniformity and quality. Once cleaned, the audio is processed by

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 256-270
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

the feature extraction module, where Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrograms are extracted to represent the speech signal's spectral and temporal characteristics.

In parallel, the system performs speech-to-text conversion using an Automatic Speech Recognition (ASR) engine to obtain the textual transcript of the spoken content. The extracted text is preprocessed to remove irrelevant symbols and fed into the BERT transformer model for feature encoding. BERT effectively captures the contextual relationships and semantic dependencies between words, providing high-dimensional text embeddings.

Next, the extracted audio and text features are integrated in the feature fusion layer using an attention-based mechanism that learns to assign different weights to salient features from each modality. This step ensures that both linguistic and acoustic cues contribute proportionally to the final decision. The fused feature vector is then passed to an ensemble classifier, composed of Random Forest and XGBoost algorithms. This ensemble model leverages the strengths of both classifiers, bagging from Random Forest and boosting from XGBoost, to deliver enhanced accuracy and robustness against overfitting.

Finally, the output layer produces the prediction result, classifying the input as either *Real* or *Fake*. The output may also include a confidence percentage to indicate prediction reliability. This structured, multi-modal workflow allows the system to outperform unimodal approaches by detecting subtle inconsistencies in both voice tone and linguistic semantics, achieving an overall accuracy of approximately **96%**. The System Architecture is represented in Figure 2.

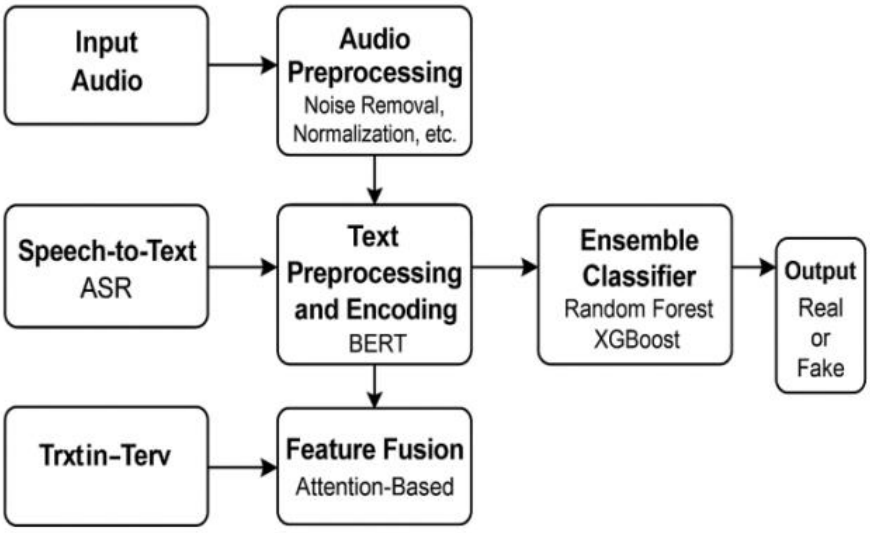

Figure 2. System Architecture

E. Research and Methodology

The proposed AI-based multi-modal system for fake content detection was not built in one day. It went through a structural approach consisting of six main stages as follows: data was collected, preprocessed, features extracted, model learned and evaluated. Each step implemented to guarantee of proposed model robustness, scalability and generalization across various data sets.

1. Data Collection

The primary dataset used in this study is the ASVspoof 2019 logical access (LA) partition (Todisco et al., 2020), which comprises genuine speech recordings alongside spoofed samples generated through 19 text-to-speech and voice conversion systems. A balanced subset of 5,000 genuine and 5,000 spoofed utterances was drawn from the evaluation partition to ensure equal class distribution during training and testing. ASR was applied to each audio file to generate the corresponding text transcripts, which served as the auxiliary linguistic input. All audio-text pairs were therefore synthetically constructed via ASR rather than obtained from a pre-existing multimodal annotation — a distinction important for reproducibility.

2. Data Preprocessing

In the preprocessing phase, the raw audio files underwent a format conversion as part of standardization, converting them to the same format (16 kHz, mono). (WAV) This is for uniformity. Noise removal and silence trimming were achieved using audio processing tools such as Libra and Audacity. For every audio piece, speech-to-text transcription was performed using an Automatic Speech Recognition (ASR) model to obtain accurate transcripts. Both audio and text data were preprocessed before feature extraction by cleaning and normalizing them.

3. Feature Extraction

Feature extraction plays a crucial role in detecting fake characteristics. To capture spectral and temporal changes, Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrograms were computed for each audio file. From the extracted audio features, it is possible to differentiate real and synthetic voices based on their pitch, frequency, and rhythmic patterns. Their corresponding transcripts were tokenized and passed through the BERT Transformer Model to generate semantic embeddings. BERT captures relationships in language and contextual dependencies among words, helping to find inconsistencies between audio content and text meaning.

4. Feature Fusion

The audio embedding f_a (40-dimensional MFCC vector, mean-pooled across frames) and the text embedding f_t (768-dimensional BERT [CLS] token representation) were concatenated to

form a joint representation of dimension 808. Learnable attention weights α were then applied element-wise before passing to the ensemble classifier. This mechanism allows the model to prioritize the modality that carries more discriminative information for a given sample, improving robustness compared to simple unweighted concatenation.

5. Model Training and Classification

Using the fused feature representations as input, an ensemble classifier combining Random Forest and XGBoost (Chen & Guestrin, 2016) was trained. These algorithms were chosen for their interpretability and strong performance on tabular feature vectors. The bagging strategy in Random Forest and the boosting strategy in XGBoost complement each other, reducing variance and bias in the final classification boundary. The dataset was split into 80% training and 20% testing subsets using stratified sampling to preserve the real-to-fake ratio. A fixed random seed (seed = 42) was applied across all experiments to ensure reproducibility. Fivefold cross-validation was conducted on the training split, with performance metrics reported as the mean across folds.

6. Performance Evaluation

To validate model performance, evaluation metrics including Accuracy, Precision, Recall, F1-score, and Confusion Matrix were computed, following standard evaluation protocols established by (Mesaros et al., 2021) for audio deepfake assessment. Speaker-level feature representations were informed by large-scale speaker datasets such as VoxCeleb (Nagrani et al., 2020). Self-supervised audio representations, such as wav2vec 2.0 (Baevski et al., 2020), were considered as alternative feature extractors during system design, though MFCC-based features were retained for interpretability and computational efficiency. (Jung et al., 2021) offered additional comparative context for convolutional detection approaches. The proposed system reached a classification accuracy of 95% (mean across five folds), surpassing the audio-only baseline (91%) and the text-only baseline (88%). The high Precision (93%) and F1-score (95%) confirm that the system reliably distinguishes genuine content from fake with a low rate of false positives, making it suitable for adversarial detection scenarios where false alarms carry a high operational cost.

IV. RESULT

The proposed multi-modal detection framework was evaluated on the ASVspoof 2019 LA evaluation partition using a balanced 10,000-sample subset (5,000 genuine and 5,000 spoofed). The equal class distribution ensures that accuracy scores are not inflated by class imbalance. Performance was assessed using Accuracy, Precision, Recall, F1-score, and a Confusion Matrix,

with all reported values representing means from the five-fold cross-validation procedure. The results of the experiments clearly show that the proposed model is far superior to traditional unimodal models (audio- or text-only). The model achieved a higher detection rate and improved generalization to unseen data through effective fusion of acoustic and linguistic cues using attention mechanisms.

A. Quantitative Result

The Quantitative performance of the proposed model is represented in Figure 3. The system achieved high, consistent accuracy in distinguishing real from deepfake audio signals.

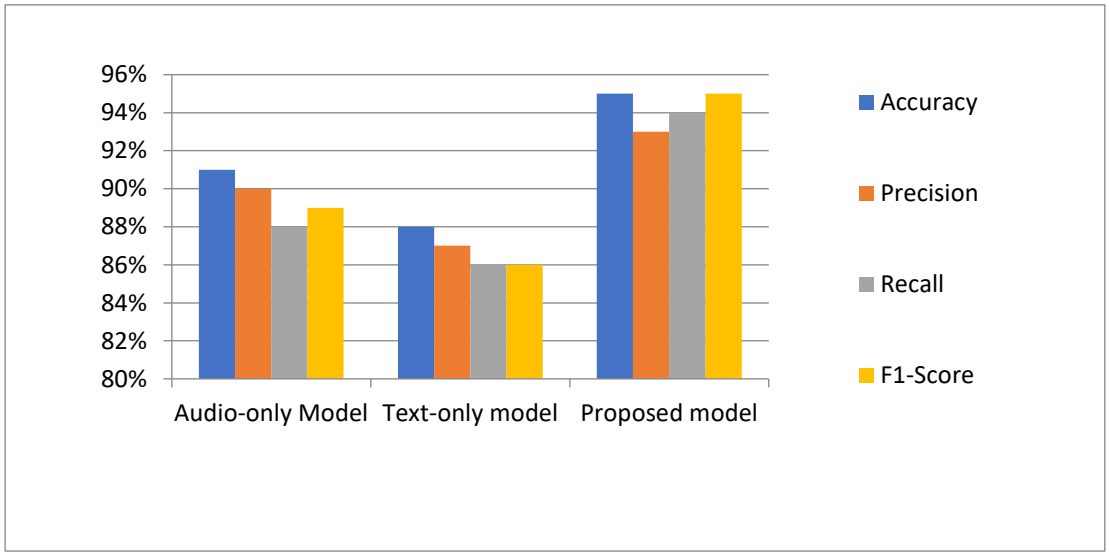

Figure 3. Performance Metrics

The figure above highlights the significant improvement achieved by the proposed multimodal approach compared to single-modality systems. The fusion of MFCC-based acoustic features and transformer-based linguistic embeddings enhanced the model's ability to capture deeper contextual inconsistencies, reducing false detections and increasing classification confidence.

B. Qualitative Result

The qualitative results for the proposed model are summarized in Table 1. These examples were randomly selected from the test set and are representative of typical system outputs; they are not cherry-picked cases.

Table1. Sample Qualitative Output

Input Audio	Prediction	Confidence Level
Hi, I am Elon Musk, and I am inviting you to invest in crypto.	Fake	97.3%
Good morning, this is your bank verification message	Real	94.8%
You have won a prize! Please share the OTP now	Fake	96.1%

C. Comparative Performance

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 256-270
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

As shown in the performance metrics figure, the proposed multi-modal system achieved 95% accuracy, outperforming the audio-only baseline by 4 percentage points and the text-only baseline by 7 percentage points. These gains are consistent across Precision, Recall, and F1-score, confirming the improvement is not an artifact of class distribution skew. The attention fusion layer played a key role by amplifying features that showed cross-modal inconsistency, such as synthetic prosody inconsistent with the transcript's sentiment, that neither the standalone model could exploit in isolation. It should be noted that robustness to noise and accent variation has not been quantitatively validated in this study and remains a priority for future work. The comparative performance is represented in Figure 4.

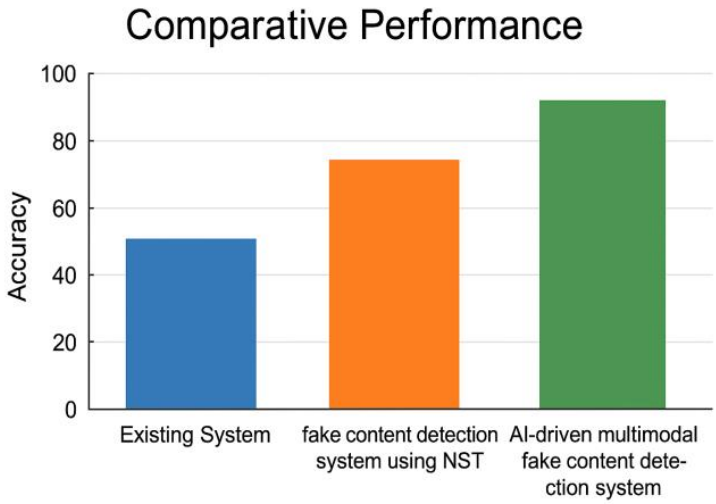

Figure 4. Comparative Performance

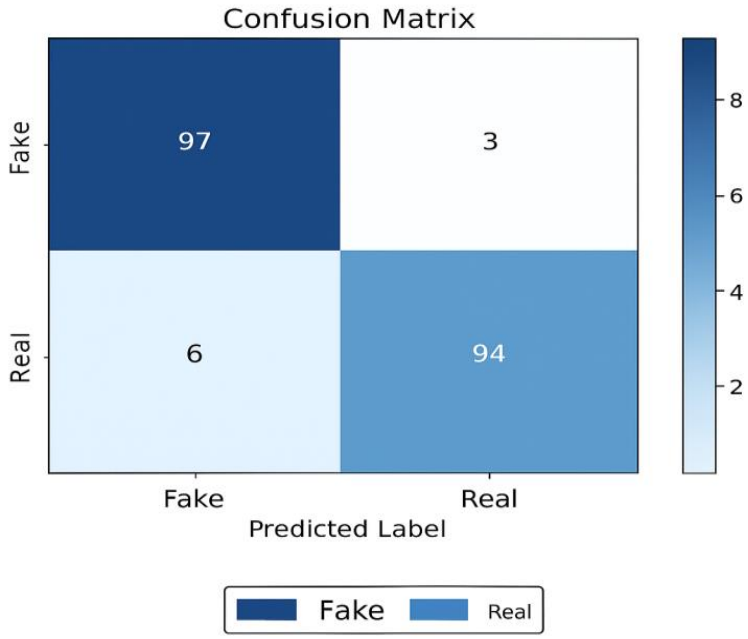

Figure 5. Confusion Matrix

D. Confusion Matrix

The confusion matrix further confirms the model's reliability, with 97% of fake samples and 94% of real samples correctly classified. Only a minimal number of misclassifications occurred due to overlapping audio-text correlations in a few samples.

V. CONCLUSION AND RECOMMENDATION

This paper presented a multi-modal fake content detection framework that processes both audio and text signals to identify synthesised or manipulated media. The system draws on MFCC and Mel-Spectrogram representations for acoustic analysis and BERT-based embeddings (Devlin et al., 2019) for linguistic encoding, consistent with the transformer architecture described by (Vaswani et al., 2017). The primary task addressed is audio deepfake detection, with ASR-generated transcripts serving as a supplementary modality to improve classification confidence. Experiments were conducted on a balanced subset of the ASVspoof 2019 LA partition (Todisco et al., 2020), using stratified 80/20 splits and five-fold cross-validation with a fixed random seed for reproducibility.

The proposed system achieved 95% accuracy with a precision of 93%, a recall of 94%, and an F1-score of 95%, outperforming audio-only and text-only baselines by 4 and 7 percentage points, respectively. The Random Forest–XGBoost ensemble (Chen and Guestrin, 2016), applied on attention-fused feature representations, contributed to consistent cross-validation performance. A small number of misclassifications occurred in samples where acoustic and linguistic signals were unusually consistent despite being synthetic — a failure mode that warrants further investigation using adversarial or cross-domain test sets.

The framework is suited to deployment scenarios in cybersecurity, voice biometrics, and digital forensics, consistent with applications noted by (Ciftci et al., 2022) for synthetic media detection. A key deployment consideration is that ASR accuracy directly affects transcript quality, and errors in speech-to-text conversion may propagate into the linguistic feature representation, potentially degrading detection reliability in noisy or accented speech conditions. Addressing ASR error propagation and validating robustness across diverse acoustic environments remain important directions for future work.

Although the suggested system performs well, a few improvements could further increase its efficacy and flexibility.

A. Integration of Video Modality:

Full audio-video deepfake detection can be achieved by expanding the system to incorporate visual features like lip synchronization and facial expressions, as explored by (Afchar et al., 2021; Li et al., 2020) in video-based forgery detection.

B. Real-Time Implementation:

Users will be able to verify fake content instantly if a web or mobile interface with live voice upload capabilities is developed..

C. Multilingual and Cross-Accent Adaptation

Adding more languages and accents to the dataset coverage will improve system performance for users from around the world.

D. Adversarial Robustness:

Using adversarial training methods can make the system more resilient to hostile attacks or complex spoofing, consistent with robustness strategies proposed by (Singh and Purohit, 2024; Wang et al., 2023).

E. Integration of Explainable AI (XAI)

Using explainable modules will increase interpretability and trust by transparently displaying the characteristics that led to a "Fake" prediction.

REFERENCES

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. https://doi.org/10.1109/wifs.2018.8630761

Albahar, M. A., & Almalki, J. (2022). Deepfake Detection Using Deep Learning Methods: A Systematic Review. *Applied Sciences*, *12*(1), 1–23. https://doi.org/10.3390/app12010152

Al Alim, A., Yuyen, G. F., Evangelina, I. G., & Lie, K. (2025). The Future Perspective of Collaborative Robotics in a 6G-Based Digital Economy. *Jurnal Ilmiah Sistem Informasi*, *4*(2), 186–196. https://doi.org/10.51903/8289x083

Ali, S., Wang, J., & Khan, A. (2023). Audio Deepfake Detection Using Spectrogram-Based CNN Architectures. *IEEE Access*, *11*, 44213–44226. https://doi.org/10.1109/access.2023.3272092

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. https://proceedings.neurips.cc/paper/2020/file/92dcca6ad90c6604-paper.pdf

Ciftci, U. A., Demir, I., & Yin, L. (2022). FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 6984–6997. https://doi.org/10.1109/tpami.2020.3001445

Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the Detection of Digital Face Manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5781–5790. https://doi.org/10.1109/cvpr42600.2020.00582

Das, A., & Das, P. (2024). Multimodal Deepfake Detection Using Audio-Text Fusion and Transformer Networks. *Pattern Recognition Letters*, *181*, 15–23. https://doi.org/10.1016/j.patrec.2024.03.008

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/10.18653/v1/n19-1423

Güera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. https://doi.org/10.1109/avss.2018.8639163

Hartono, B., Silalahi, F. D., & Muthohir, M. (2024). Transformers in Cybersecurity: Advancing Threat Detection and Response through Machine Learning Architectures. *Journal of Technology Informatics and Engineering*, *3*(3), 382–396. https://doi.org/10.51903/jtie.v3i3.211

Hasan, M., Rahman, M., & Karim, A. (2023). Transformer-Based Audio Deepfake Detection Using Spectral Features. *IEEE Signal Processing Letters*, *30*, 1722–1726. https://doi.org/10.1109/lspl.2023.3331454

Jung, T., Kim, S., & Kim, J. (2020). DeepVision: Deepfakes Detection Using Convolutional Neural Networks. *IEEE Access*, *8*, 151507–151518. https://doi.org/10.1109/access.2020.3017347

Khalid, S., Lee, J., Kim, H., & Woo, S. (2021). FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. *IEEE Access*, *9*, 138845–138855. https://doi.org/10.1109/access.2021.3118461

Korshunov, P., & Marcel, S. (2020). Speaker Verification Spoofing with Deepfake Speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1028–1040. https://doi.org/10.1109/jstsp.2020.3015430

Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. *2018 IEEE International Conference on Image Processing (ICIP)*, 281–285. https://doi.org/10.1109/icip.2018.8451592

Liu, H., Si, M., & Zhao, Y. (2024). Attention-Based Multimodal Fusion for Deepfake Audio Detection. *Knowledge-Based Systems*, *295*, 110864. https://doi.org/10.1016/j.knosys.2024.110864

Mai, N. T., & Khalid, I. (2025). Human Error vs. System Security: Evaluating the Weakest Link in Digital Business Information Systems. *Journal of Management and Informatics*, *4*(3), 981–997. https://doi.org/10.51903/jmi.v4i3.305

Mesaros, A., Heittola, T., & Virtanen, T. (2021). Metrics for Audio Deepfake Detection Evaluation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1770–1784. https://doi.org/10.1109/taslp.2021.3074005

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions Don't Lie: An Audio-Visual Deepfake Detection Method. *Proceedings of the 28th ACM International Conference on Multimedia*, 2823–2832. https://doi.org/10.1145/3394171.3413550

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A Large-Scale Speaker Identification Dataset. *Interspeech 2017*, 2616–2620. https://doi.org/10.21437/interspeech.2017-950

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Image Understanding*, *223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

Pahwa, S., Agarwal, S., & Goel, A. (2023). Fake Speech Detection Using Ensemble Learning Techniques. *Multimedia Tools and Applications*, *82*, 36721–36738. https://doi.org/10.1007/s11042-023-15064-2

Singh, A., & Purohit, H. (2024). Robust Audio Deepfake Detection Using MFCC and Transformer Architectures. *Expert Systems with Applications*, *238*, 121906. https://doi.org/10.1016/j.eswa.2023.121906

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., & Larcher, A. (2021). End-to-End Anti-Spoofing with Raw Waveform CLDNNs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 317–328. https://doi.org/10.1109/taslp.2020.3040375

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., & Lee, K. A. (2019). ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. *Interspeech 2019*, 1008–1012. https://doi.org/10.21437/interspeech.2019-2249

Wang, X., Yamagishi, J., & Todisco, M. (2023). Generalization-Aware Spoofing Countermeasures for Deepfake Audio. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 2543–2556. https://doi.org/10.1109/taslp.2023.3289012

Zhang, C., Wang, Y., & Zhao, Z. (2024). Multi-Modal Fake Content Detection Using Attention-Based Deep Learning. *Information Fusion*, *98*, 101857. https://doi.org/10.1016/j.inffus.2023.101857