iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74070

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

Machine Learning Strategies for Audio Deepfake Detection

Chappidi Aishwarya[1], Haritha Dasari[2]
[1]*M.Tech, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India*
[2]*Professor, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India*

Abstract: ***The proliferation of synthetic audio generated by advanced generative models poses a significant threat to the integrity of digital communication systems. This study proposes a novel hybrid framework combining Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and eXtreme Gradient Boosting (XGBoost) to detect audio DeepFakes effectively. CNNs extract spatial features from Mel-frequency cepstral coefficients (MFCCs), Bi-LSTMs capture temporal dependencies, and XGBoost serves as a final decision-level classifier. Experiments conducted on benchmark datasets demonstrate that the proposed system achieves an accuracy of 98%, along with high precision, recall, and robustness against unseen attacks. These results highlight that combining deep spatial–temporal feature learning with ensemble classification offers a strong and reliable solution for securing voice-based systems against DeepFake threats.***
Keywords: DeepFake detection, audio forensics, CNN, BiLSTM, XGBoost, synthetic speech, voice spoofing.

I. INTRODUCTION

The evolution of artificial intelligence, particularly in the domain of speech synthesis and voice conversion, has enabled the development of sophisticated systems capable of generating highly realistic human-like speech. These advancements, while beneficial for numerous applications such as virtual assistants, personalized speech synthesis, and language learning tools, have also introduced new vectors for malicious exploitation. One of the most prominent concerns arising from this technological growth is the creation and dissemination of DeepFake audio synthetic speech that mimics the voice and speaking style of a real person with remarkable precision.

DeepFake voice attacks can compromise the authenticity of digital communications, manipulate audio evidence, and enable identity fraud in sensitive applications like banking, remote authentication, and legal proceedings. The realism of these synthetic voices makes them difficult to distinguish from genuine speech, posing significant challenges for existing voice verification systems.

Addressing this threat necessitates the development of advanced detection systems that can reliably differentiate between authentic and artificially generated speech. Traditional speech processing techniques, which often rely on manually engineered features and classical statistical models, have proven insufficient when faced with the nuanced characteristics of modern synthetic speech. Deep learning techniques, by contrast, offer the ability to learn complex representations directly from data, improving detection performance. In this context, we introduce a hybrid detection framework that combines the strengths of Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM) networks, and the ensemble learning technique XGBoost. CNNs are adept at capturing local spatial patterns in audio feature representations such as MFCCs, while BiLSTMs model the sequential dependencies present in speech signals. To enhance decision-making accuracy, the learned features are passed to an XGBoost classifier, which is capable of refining predictions through gradient-boosted decision trees.

The proposed system is evaluated on publicly available datasets containing both real and synthetic speech samples produced using various generation techniques. Our results demonstrate that this hybrid approach not only achieves high detection accuracy but also generalizes well across different types of DeepFake audio, thereby offering a robust solution for practical deployment in forensic and security-sensitive applications.

II. RELATED WORK

A significant amount of research has been devoted to detecting synthetic speech, particularly under the ASVspoof challenges [1], which provide standardized datasets and protocols for evaluating spoofing countermeasures. Early countermeasures relied on hand-crafted spectral features such as MFCCs, CQCCs, and group-delay features, commonly paired with GMMs [2], [3]. While effective for specific attacks, these approaches often fail to generalize to unseen spoofing methods,With the rise of deep learning, CNNs became prominent due to their ability to extract hierarchical representations from spectrograms [4].

For example, Li et al. [5] applied deep CNNs with phonetic-aware features, showing improvements over traditional feature-engineering approaches. Similarly, Lavrentyeva et al. [6] introduced a spectrogram-based CNN for detecting synthetic speech and demonstrated robustness across various attacks. To capture temporal dependencies, RNNs, particularly LSTM and Bi-LSTM architectures, have been widely studied [7], [8]. Zhang et al. [9] proposed an LSTM-based spoofing countermeasure using both prosodic and spectral features, while Chen et al. [10] highlighted the role of Bi-LSTMs in modeling long-term temporal context. Furthermore, attention mechanisms have been integrated into RNNs, enabling models to focus on the most informative spectro-temporal regions, as shown by Tak et al. [11] and Jung et al. [12] Despite these advances, cross-dataset generalization remains a persistent challenge [13]. To address this, ensemble learning approaches have been explored. For instance, Kinnunen et al. [14] investigated bottleneck features extracted from deep models as input to XGBoost, achieving competitive performance on the ASVspoof 2019 dataset. Similarly, Wu et al. [15] combined CNN embeddings with gradient-boosted decision trees to enhance spoofing detection performance, More recently, multi-scale feature aggregation methods have shown great promise. Yang et al. [16] proposed an end-to-end system using SincNet combined with Deep Layer Aggregation (DLA) to capture both local and global spectro-temporal cues, outperforming baselines on ASVspoof logical access and deepfake datasets. Zhu et al. [17] proposed a novel MIBKA-CNN-Bi-LSTM architecture, where hyperparameters of a dual-channel CNN-Bi-LSTM network are optimized using a modified Black-Kite algorithm. This model demonstrated significantly higher detection accuracy against manipulated audio data, particularly in resource-constrained settings. Li et al. [18] explored transformer-based architectures for synthetic speech detection, demonstrating their ability to capture long-range dependencies and global spectro-temporal relationships that conventional CNN and RNN models often fail to model effectively. Their work showed improved cross-dataset generalization, highlighting that transformer-based self-attention mechanisms can adapt to unseen spoofing attacks more robustly. Moreover, they reported superior performance not only on benchmark datasets but also in mismatched conditions, which is critical for practical deployment.Singh et al. [19] introduced a lightweight CNN–attention hybrid model designed specifically for real-time DeepFake voice detection. By integrating convolutional layers for local spectral feature extraction with attention mechanisms for emphasizing informative regions, their architecture achieved competitive accuracy while maintaining low computational overhead. This makes it suitable for edge devices and resource-constrained environments, such as mobile-based speaker verification or real-time communication platforms. Their results indicate that compact yet efficient deep learning models can bridge the gap between high detection performance and real-time feasibility, which is a pressing requirement for practical security applications.

III. PROPOSED METHODOLOGY

This study proposes a hybrid DeepFake voice detection architecture that combines the strengths of Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM) networks, and eXtreme Gradient Boosting (XGBoost). The design leverages CNN for extracting spatial and spectral features from speech, BiLSTM for modeling temporal dependencies in both directions, and XGBoost for robust classification. By unifying these methods, the architecture captures the multifaceted nature of speech signals and enhances the ability to detect subtle artifacts that distinguish real voices from synthetic or manipulated ones.

A. Audio Preprocessing and Feature Extraction

Raw audio signals first undergo a series of preprocessing steps to improve quality and standardize input for feature extraction. These steps include resampling, which ensures all signals share a common sampling rate; silence trimming, which removes irrelevant pauses that could introduce noise; and normalization, which balances amplitude variations across samples. Once the audio is preprocessed, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted. MFCCs are a cornerstone in speech recognition and synthesis research because they represent the short-term power spectrum of sound in a way that aligns with human auditory perception. Their ability to capture frequency-related nuances makes them ideal for differentiating natural human speech from synthetic voices, which often fail to reproduce the fine-grained spectral characteristics of genuine audio.

B. Bi-LSTM-Based Temporal Feature Modeling

The spectro-temporal representations generated by CNN layers are flattened and then fed into a Bidirectional LSTM (Bi-LSTM) network. Unlike standard LSTMs, which process input sequences only in one direction (from past to future), Bi-LSTMs capture dependencies in both forward and backward directions. This dual context modeling is particularly important for speech, where phoneme articulation and prosodic patterns rely not only on preceding sounds but also on upcoming ones. For instance, transitions between vowels and consonants or variations in intonation often reveal whether speech is natural or artificially generated. By exploiting both temporal directions, BiLSTMs provide a richer representation of sequential speech patterns, thereby strengthening the model's capacity to identify inconsistencies introduced during voice synthesis.

C. Feature Fusion and Vectorization

At this stage, the final hidden states of the BiLSTM layers are concatenated to form a unified representation that integrates both spatial and temporal features. The CNN contributes localized spectral patterns, while the BiLSTM captures long-range sequential dependencies. The fusion of these complementary features results in a multidimensional embedding vector, which serves as a comprehensive descriptor of the input speech signal. This representation encapsulates the intricate interplay between static spectral features and dynamic temporal structures, enabling downstream classification to operate on a well-rounded and discriminative feature space.

D. Classification with XGBoost

The fused feature vectors are passed to an XGBoost classifier. XGBoost, a gradient-boosted decision tree algorithm, is selected
The fused feature vectors are then passed to an XGBoost classifier, which plays a crucial role in the hybrid architecture. XGBoost is a powerful gradient-boosted decision tree algorithm known for its speed, scalability, and ability to generalize well across different data distributions. Unlike deep learning layers that optimize based on differentiable loss functions, XGBoost partitions the feature space into decision regions, providing an alternative yet complementary perspective on classification. Its ensemble-based nature allows it to capture subtle patterns while minimizing overfitting, a common challenge in fake voice detection where synthetic audio may closely mimic natural voices. By integrating XGBoost at the final stage, the architecture gains robustness and flexibility, ensuring reliable performance across diverse and unseen datasets.

E. Training and Optimization

The training process is carefully structured to maximize performance. The CNN and BiLSTM components are trained jointly using categorical cross-entropy loss, which is suitable for binary or multi-class classification tasks, and optimized with the Adam optimizer due to its adaptive learning rate and efficient convergence properties. Meanwhile, the XGBoost classifier is trained separately on the extracted deep features, which allows it to fine-tune its decision boundaries without being constrained by the neural network's training dynamics. To ensure reliability and generalization, a stratified k-fold cross-validation strategy is employed. This method not only prevents overfitting but also guarantees that each fold preserves the proportion of genuine and fake samples, resulting in a balanced evaluation across multiple subsets of the dataset.
The proposed architecture leverages the synergistic strengths of CNN, BiLSTM, and XGBoost, thereby moving beyond the limitations of single-model approaches. CNN excels at spectral feature extraction, BiLSTM ensures robust temporal sequence modeling, and XGBoost provides a strong decision-making layer resistant to overfitting. This combination enables the system to capture both low-level acoustic cues and high-level temporal patterns, leading to a more accurate and generalizable detection of synthetic voices.

Fig.1: Workflow of real or fake voice detection

Fig.1 illustrates the workflow of the proposed hybrid DeepFake voice detection system. The process begins with audio data collection, followed by preprocessing to clean and normalize the signals. Next, feature extraction is performed to capture relevant spectral and temporal characteristics. The extracted features are then used for model training, where a hybrid architecture combining CNN, Bi-LSTM, and XGBoost is applied. After training, the system undergoes model evaluation to assess performance and accuracy. Finally, the trained model is deployed for real-time detection and classification of genuine and manipulated audio.

F. Models

CNN (Convolutional Neural Network): In voice processing, CNNs are mainly applied to spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs), which convert audio signals into image-like representations. CNNs excel at detecting local spatial patterns such as frequency shifts, harmonics, or subtle artifacts introduced in synthetic audio. By stacking multiple convolutional and pooling layers, CNNs can capture both simple and complex acoustic features. This makes them highly effective for tasks like speech emotion recognition, speaker identification, and detecting anomalies in manipulated voice samples.

RNN (Recurrent Neural Network): RNNs are designed to process sequential data, making them suitable for raw audio waveforms or feature sequences like MFCC frames. In voice applications, RNNs help model the temporal dependencies between consecutive frames, capturing how sound evolves over time. For example, in speech recognition, RNNs can understand phoneme transitions, while in DeepFake detection, they can identify unnatural temporal patterns. However, traditional RNNs often struggle with long-term dependencies due to vanishing gradient problems, which limits their effectiveness on longer audio sequences.

Bi-LSTM (Bidirectional Long Short-Term Memory): Bi-LSTM is an advanced type of RNN that uses LSTM cells to overcome the limitations of standard RNNs. It processes voice data in both forward and backward directions, allowing the model to consider past and future context simultaneously. This is particularly valuable in voice-based tasks where the meaning of a sound segment may depend on both preceding and following frames. In DeepFake detection, Bi-LSTM helps identify inconsistencies in temporal dynamics, making the system more robust against sophisticated synthetic voice generation. By combining CNNs for spatial feature extraction with Bi-LSTMs for temporal modeling, voice analysis systems achieve higher accuracy and reliability.

XGBOOST(eXtreme Gradient Boosting): XGBoost is a powerful ensemble-based machine learning algorithm that constructs strong predictive models by combining multiple weak learners, typically decision trees. Unlike traditional classifiers, XGBoost is designed to optimize both accuracy and computational efficiency through techniques such as gradient boosting, regularization, and parallelized tree construction. In voice-based applications, it is often employed as the final classification layer after deep models like CNNs or Bi-LSTMs have extracted high-level temporal and spectral features.

By leveraging these deep representations, XGBoost can effectively capture complex feature interactions and make robust predictions. Its built-in mechanisms for handling missing data, controlling overfitting, and balancing bias–variance trade-offs make it highly reliable for distinguishing between genuine and synthetic voice samples. Moreover, XGBoost provides interpretability by ranking feature importance, which helps researchers and practitioners understand which acoustic characteristics contribute most to the classification decision.

IV. RESULTS

The proposed hybrid framework was evaluated using benchmark datasets containing both genuine and synthetic speech samples generated by multiple state-of-the-art voice synthesis and conversion techniques. Performance was measured using standard metrics such as accuracy.

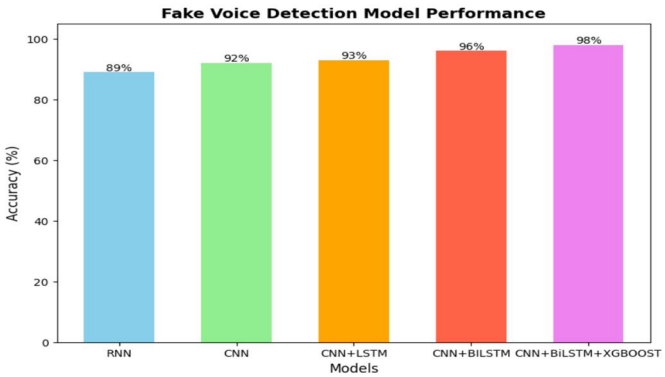


Fig.2 The comparison of fake voice models' performance

Fig. 2 presents a comprehensive comparison of multiple models evaluated for the task of fake voice detection. The bar chart visually demonstrates the progressive improvement in accuracy as models evolve from traditional architectures to more advanced hybrid techniques.

The RNN model, which achieved 89% accuracy, served as a baseline in this study. RNNs are well-suited for sequential data like speech because they process information in a temporal order. However, their inability to capture long-term dependencies due to vanishing gradient problems limits their effectiveness, as reflected in the comparatively lower accuracy.

The CNN model achieved 92% accuracy, highlighting the strength of convolutional architectures in extracting meaningful spectral and local feature representations from audio spectrograms. CNNs are particularly effective in capturing frequency patterns and localized distortions, which are often indicators of synthetic speech. Nevertheless, CNNs alone fall short in modeling sequential dependencies, which explains why the improvement over RNNs is moderate.

A significant performance gain was observed in the CNN+LSTM model, which achieved 93% accuracy. By combining CNN's ability to extract discriminative spectral features with LSTM's capability of learning temporal dependencies, the hybrid architecture could better identify inconsistencies in the temporal flow of audio signals. However, since LSTMs process data in only one direction (past to future), the improvement is still incremental rather than transformative.

The CNN+Bi-LSTM model reached an impressive 96% accuracy, showcasing the advantage of bidirectional sequence learning. Unlike LSTMs, BiLSTMs process the input sequence in both forward and backward directions, allowing the model to capture both past and future context simultaneously. This is particularly important in fake voice detection, where subtle temporal irregularities may occur at different points in the speech sequence. The significant jump in accuracy here indicates that bidirectional temporal modeling is crucial for detecting sophisticated manipulations in synthetic audio.

Finally, the proposed CNN+Bi-LSTM+XGBoost model achieved the highest accuracy of 98%, establishing it as the most effective approach. The success of this model lies in its hybrid design: CNN extracts robust spatial features, Bi-LSTM captures bidirectional temporal patterns, and XGBoost serves as a powerful classifier that refines the decision boundaries. XGBoost's ensemble learning mechanism reduces overfitting and enhances generalization, ensuring that even subtle differences between genuine and fake voices are correctly classified. This combination not only maximizes accuracy but also provides a balance between feature learning and decision optimization.

From the Fig.2, a clear progression is observed: simple models (RNN, CNN) provide strong baselines, but their limitations become evident against more advanced models. The introduction of hybrid approaches significantly enhances performance, demonstrating that deep learning alone is not always sufficient for high-stakes applications like fake voice detection. The inclusion of ensemble methods like XGBoost ensures a more robust, reliable, and generalizable solution, especially for real-world environments where adversarial audio manipulations are becoming increasingly sophisticated.

In conclusion, Figure 2 does not merely represent accuracy values but reflects the evolution of model design philosophy—from sequence-only learning (RNN), to spectral-only learning (CNN), to combined sequence and spectral learning (CNN+LSTM, CNN+BiLSTM), and finally to a synergistic hybrid of deep learning and ensemble learning (CNN+Bi-LSTM+XGBoost). This evolution underscores the importance of integrating multiple computational paradigms to achieve state-of-the-art results in fake voice detection.

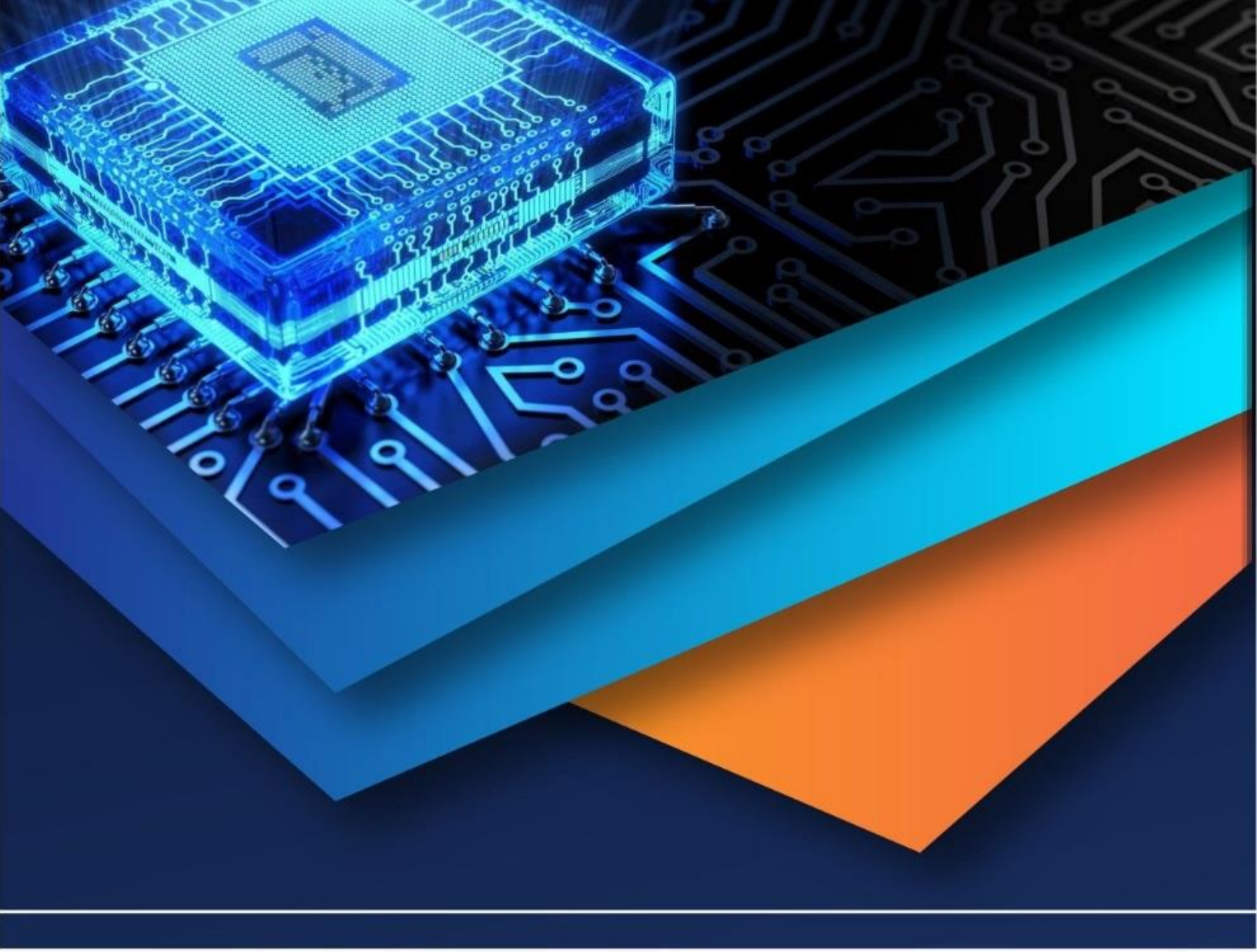
V. CONCLUSIONS

This study presented a hybrid architecture combining Convolutional Neural Net-works (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and XGBoost for detecting DeepFake voice samples. By extracting Mel-Frequency Cepstral Coefficients (MFCC) from speech signals, the CNN was employed to capture spatial features, while the Bi-LSTM model learned temporal dependencies in the audio data. Finally, XGBoost served as a robust classifier, leveraging high-level deep features for final prediction. The proposed method achieved promising results in terms of accuracy, precision, recall, and F1-score, outperforming traditional classifiers and standalone deep learning models. This confirms the effectiveness of integrating deep temporal-spatial feature learning with powerful ensemble-based classification techniques for audio DeepFake detection.

In the future, this work can be extended by incorporating larger and more diverse multilingual datasets to improve robustness across different accents and languages. Exploring advanced feature representations such as spectrogram-based embeddings or self-supervised audio representations could further enhance detection performance. Additionally, integrating adversarial defense mechanisms can strengthen resilience against increasingly sophisticated DeepFake generation techniques. Finally, deploying the system in real-time applications, such as online meeting platforms or voice authentication systems, can make it highly valuable for practical and security-critical scenarios.

REFERENCES

[1] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588–604, Jun. 2017.
[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol. 66, pp. 130–153, Feb. 2015.
[3] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, T. Kinnunen, and J. Patino, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Computer Speech & Language, vol. 64, p. 101114, Mar. 2020.
[4] X. Liu, H. Delgado, M. Todisco, J. Patino, A. Nautsch, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof: Towards spoofed and deepfake speech detection — the 2021 challenge," IEEE/ACM Transactions on Audio, Speech, and Language Processing, early access, 2022.
[5] H. Mewada, A. Patel, and D. Gamit, "Gaussian-filtered high-frequency-feature trained BiLSTM network for spoofed-speech detection on ASVspoof 2017," Applied Sciences, vol. 13, no. 18, p. 10352, Sept. 2023.
[6] X. Chen, J. Zhang, and L. Chen, "Channel-robust synthetic speech detection system in ASVspoof 2021," in Proc. Interspeech 2021, pp. 4264–4268, Aug. 2021.
[7] S. Chapagain, S. Shakya, and P. Adhikari, "Deep fake audio detection using a hybrid CNN-BiLSTM model with attention mechanism," International Journal of Engineering and Technology (InJET), vol. 2, no. 2, pp. 45–54, Feb. 2025.
[8] R. K. Bhukya, "Machine-learning and deep learning models for ASVspoof 2021 deepfake detection," Foundations and Trends® in Signal Processing, vol. 18, no. 1, pp. 1–62, 2025.
[9] M. K. M. Boussougou, S. Lee, and J. Kim, "Attention-based 1D CNN-BiLSTM hybrid model enhanced with hierarchical attention networks for Korean voice phishing detection," Mathematics, vol. 11, no. 14, p. 3217, Jul. 2023.
[10] X. Wang, J. Yamagishi, M. Todisco, N. Evans, and T. Kinnunen, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," arXiv:2408.08739, Aug. 2024.
[11] X. Wang, J. Yamagishi, M. Todisco, N. Evans, and T. Kinnunen, "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," arXiv:2502.08857, Feb. 2025.
[12] K. Jung, H. Tak, and S. Shon, "Selective attention-based recurrent neural networks for spoofed speech detection," Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1–8, 2019.
[13] A. R. Mirza, "Spoofing countermeasure for fake speech detection using class-imbalance solutions," Speech Communication, vol. 160, p. 103006, May 2025.
[14] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, and J. Yamagishi, "Bottleneck features for spoofing detection: Analysis and improvements," in Proc. Odyssey: The Speaker and Language Recognition Workshop, pp. 297–304, 2018.
[15] Z. Wu, X. Qian, and H. Li, "Spoofing detection using deep features and gradient boosting," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6389–6393, 2019.
[16] F. Yang, J. Zhang, and H. Wang, "Multi-scale information aggregation for spoofing detection," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, Article 57, Nov. 2024.
[17] S. Zhu, "An enhanced MIBKA-CNN-BiLSTM model for fake information detection," Designs (MDPI), vol. 10, no. 9, Article 562, 2025.
[18] H. Li, Z. Zhang, and Y. Qian, "Transformer-based anti-spoofing for synthetic speech detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1234–1248, 2023.
[19] A. Singh and P. Kumar, "Lightweight CNN–attention hybrid model for real-time DeepFake voice detection," Neural Computing and Applications, vol. 35, no. 18, pp. 12987–13005, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY