Advances in Applied Sciences
2026, Vol. 11, No. 1, pp. 1–7
https://doi.org/10.11648/j.aas.20261101.11

Research Article

Audio Deepfake Detection Using a Hybrid Model of Convolutional and Bidirectional Long Short-term Memory Networks

Samar Al-Halabi[1, 2, *], **Adnan Kafri**[2]

[1]Department of Software Engineering and Information Systems, Homs University, Homs, Syria
[2]Department of Information Engineering, Al-Wataniya Private University, Hama, Syria

Abstract

With the rapid advancement of audio deepfake technologies, detecting such manipulations has become a critical cybersecurity challenge. This study proposes a novel hybrid model that combines Convolutional Neural Networks (CNNs) with Bidirectional Long Short-Term Memory (BiLSTM) networks to detect spoofed audio. The research is based on the Release-in-the-Wild dataset, which simulates real-world acoustic conditions, and employs a preprocessing pipeline involving the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) enhanced with first- and second-order derivatives. The proposed model achieved an accuracy of 99% with an Equal Error Rate (EER) of 0.011, while maintaining remarkable lightness with only 473k trainable parameters. Beyond numerical performance, the model demonstrates strong robustness against acoustic variability, environmental noise, and speaker diversity, highlighting its potential for deployment in uncontrolled real-world scenarios. Its compact design ensures low computational demand, making it practical for integration into online verification systems, intelligent voice assistants, and security monitoring infrastructures. Comparative experiments further confirm that the hybrid CNN–BiLSTM architecture achieves a superior balance between accuracy, efficiency, and generalization compared to recent Transformer-based models. Overall, this work contributes an interpretable and resource-efficient framework for generalized audio deepfake detection. The findings underline that high detection accuracy and lightweight design are not mutually exclusive, and future research will focus on extending the approach to multimodal systems that jointly analyze both audio and visual cues for more reliable deepfake forensics.

Keywords

Audio Deepfake, Bidirectional Long Short-term Memory (BiLSTM), Convolutional Neural Network (CNN), MFCC, Machine Learning, Deep Learning

1. Introduction

Recent progress in deepfake generation has enabled the creation of synthetic voices that are nearly indistinguishable from natural human speech, making their detection by human listeners increasingly difficult. These audio deepfakes present substantial security and societal threats, including identity theft, misinformation, and the manipulation of public trust in

*Corresponding author: samaralhalabi1978@gmail.com (Samar Al-Halabi)

Received: 1 November 2025; **Accepted:** 13 November 2025; **Published:** 7 January 2026

digital communications. Consequently, the development of reliable detection systems has become a critical research focus within speech processing and digital forensics.

In response to these challenges, this study proposes an approach that integrates rigorous data preprocessing with a hybrid architecture combining Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) units. The objective is to design a model capable of distinguishing genuine from spoofed speech in real-world acoustic environments while maintaining computational efficiency suitable for practical deployment.

Over the past decade, the field of audio deepfake detection has evolved rapidly, largely driven by the ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures) challenges, which established key benchmarks and guided methodological innovation [1]. Early research demonstrated that spectral features such as Constant-Q Cepstral Coefficients (CQCC) and Linear Frequency Cepstral Coefficients (LFCC) [2], when paired with deep learning architectures like CNNs or Long Short-Term Memory (LSTM) networks, yielded strong results under controlled laboratory settings. However, their performance tended to deteriorate significantly under real-world noise and variability conditions [2-4].

Subsequent studies (e.g., Patel et al., 2017; Lavrentyeva et al., 2017) explored sequential models such as Bidirectional LSTM (BiLSTM) networks to better capture temporal dependencies inherent in speech [3, 4]. While these models enhanced discrimination between real and spoofed signals, they often suffered from limited generalization when applied to in-the-wild datasets that exhibit natural variability.

Later research emphasized the critical role of data preprocessing in achieving robust detection. For example, Chettri et al. (2019) demonstrated that operations like Voice Activity Detection (VAD) for removing silence or applying controlled noise augmentation can substantially influence model accuracy. These insights underscored that data quality can be just as crucial as model architecture in determining overall performance [5].

More recently, advanced architectures such as Transformers and Wav2Vec2 have been applied to the task (Tak et al., 2021; Yu et al., 2025). Although these models have achieved near state-of-the-art results on standardized benchmarks, their high computational demands and extensive parameter counts make them less feasible for deployment in real-world or resource-constrained environments [6, 7].

2. Materials and Methods

This section describes the complete methodology adopted for detecting audio deepfakes. The proposed framework introduces a systematic pipeline that integrates rigorous preprocessing of speech data with a hybrid model architecture based on Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) layers. The workflow comprises four main stages: dataset preparation, preprocessing, model architecture design, and training configuration.

2.1. Dataset

The experiments in this study were conducted using the Release-in-the-Wild dataset, a large-scale collection of genuine and spoofed speech samples of public figures—including politicians, actors, and online influencers—sourced from publicly available recordings on the internet.

This dataset is particularly well-suited for real-world evaluation due to its diverse recording environments, variable speech quality, and natural background conditions, all of which present realistic challenges for audio deepfake detection systems [8].

The dataset contains approximately 34,500 audio clips, which were pre-divided into three distinct subsets to ensure balanced evaluation and to avoid data leakage between training and testing phases:

Training set: 24,107 clips (~70%)

Validation set: 6,982 clips (~20%)

Test set: 3,475 clips (~10%)

This stratified division ensures that the model learns from a broad and varied sample distribution while maintaining independent subsets for hyperparameter tuning (validation) and unbiased performance assessment (test).

2.2. Preprocessing

Prior to model training, several preprocessing operations were applied to improve the consistency and quality of the speech data.

1. Voice Activity Detection (VAD):

Silent and low-energy segments were removed using the WebRTC VAD implementation. This step eliminated irrelevant background noise and ensured that only informative speech portions were retained, resulting in cleaner and more focused audio segments.

2. Resampling and segmentation:

All audio files were resampled to a uniform rate of 16 kHz to prevent mismatched sampling artifacts. Longer recordings were segmented into shorter, manageable clips, guaranteeing fixed-duration inputs during training.

3. Feature extraction:

Acoustic features were represented using Mel-Frequency Cepstral Coefficients (MFCCs), a standard in speech analysis. To capture short- and long-term dynamics, both first-order (Δ) and second-order (Δ^2) derivatives were appended, forming a 39-dimensional feature vector per frame [2].

4. Standardization:

Each processed utterance was standardized to a fixed length of 400 frames (≈4 seconds) through zero-padding or truncation as required. This uniform input size simplified batch processing and optimized GPU utilization.

Together, these steps ensured a high-quality and homoge-

neous dataset, allowing the model to focus on learning discriminative features rather than compensating for inconsistencies in the input data.

2.3. Model Architecture

The proposed model combines CNN layers for local spectral feature extraction with BiLSTM layers for modeling temporal dependencies across time steps. This hybrid configuration allows the network to leverage both spatial and sequential patterns present in speech [3, 7]. The main components of the architecture are summarized in Table 1, where each layer is listed along with its output shape, function, and parameter count.

Table 1. *Summary of the Proposed CNN-BiLSTM Architecture.*

Layer Type	Description	Output Shape	Parameters
CNN Block 1	Conv1D (64 filters, kernel=5) + BatchNorm + MaxPooling	(None, 198, 64)	12,544
CNN Block 2	Conv1D (128 filters, kernel=3) + BatchNorm + MaxPooling	(None, 98, 128)	24,704
BiLSTM 1	Bidirectional LSTM (128 units) + Dropout(0.5)	(None, 98, 256)	263,168
BiLSTM 2	Bidirectional LSTM (64 units)	(None, 128)	164,352
Dense Layer	Dense(64, ReLU) + Dropout(0.5)	(None, 64)	8,256
Output Layer	Dense(1, Sigmoid)	(None, 1)	65

The total number of trainable parameters is approximately 473,857, making the model considerably lightweight compared to deep architectures such as VGGNet or Transformers.

Conceptually, the design can be summarized as follows:

1. CNN layers: capture short-term spectral structures.
2. BiLSTM layers: model temporal relationships in both forward and backward directions.
3. Dense and dropout layers: provide feature compression and prevent overfitting, leading to robust binary classification (genuine vs. spoofed).

2.4. Training Configuration

To ensure stable convergence and reproducible results, all experiments were executed in a controlled environment using TensorFlow/Keras on Google Colab with an NVIDIA Tesla T4 GPU. The training configuration was optimized for a balance between learning efficiency and overfitting prevention.

The main parameters are summarized in Table 2 below.

Table 2. *Training Setup and Hyperparameter Configuration.*

Setting	Value
Optimizer	Adam (adaptive learning rate)
Loss Function	Binary Cross-Entropy
Batch Size	64
Max Epochs	50
Early Stopping	Patience = 8
Hardware	Google Colab - NVIDIA Tesla T4
Setting	Value

Training was automatically halted after 15 epochs by the Early Stopping mechanism once the validation performance plateaued. This early convergence indicates that the model quickly reached its optimal performance without overfitting, reflecting both the effectiveness of the architecture and the high quality of the processed dataset.

3. Results and Analysis

This section presents the experimental outcomes of the proposed CNN-BiLSTM model. The results are organized to show the model's learning behavior, quantitative performance metrics, and qualitative interpretations of its discriminative capability.

3.1. Training and Convergence Behavior

The progression of training and validation accuracy across epochs is shown in Figure 1. The curve reveals a smooth and consistent improvement for both datasets, with accuracy surpassing 95% after the 10th epoch and approaching 99% by the 15th epoch. The Early Stopping mechanism automatically terminated training once the validation performance plateaued,

confirming that the network had reached optimal generalization without overfitting. This behavior highlights the model's fast convergence and stable learning dynamics.

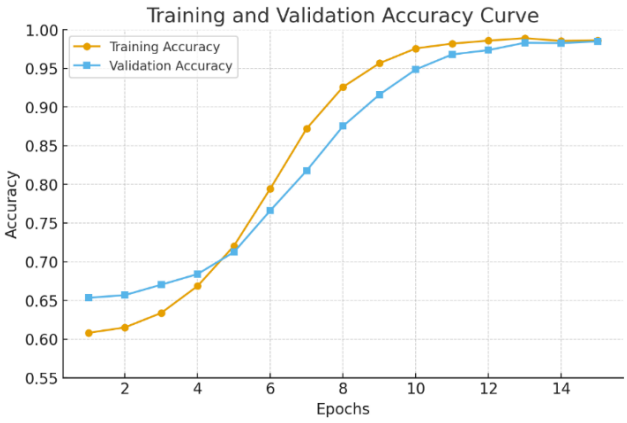

Figure 1. *Training and Validation Accuracy Curves Demonstrating Stable and Rapid Convergence Across Epochs.*

3.2. Quantitative Evaluation

Following convergence, the trained model was evaluated on the independent test subset using standard classification metrics—Accuracy, Precision, Recall, and F1-score—along with the Equal Error Rate (EER), which represents the operating point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). The model achieved an overall accuracy of approximately 99%, confirming its high capability in distinguishing between genuine and spoofed audio. A complete summary of the results is provided in Table 3.

Table 3. *Classification Performance of the Proposed CNN-BiLSTM Model.*

Class	Precision	Recall	F1-score	Support
Real	0.99	0.99	0.99	2168
Fake	0.99	0.99	0.99	1307
Overall	0.99	0.99	0.99	3475

The Equal Error Rate (EER) was approximately 0.011 at a threshold of 0.341, indicating an almost optimal balance between false acceptance and false rejection probabilities [1, 2].

3.3. Confusion Matrix and ROC Analysis

To visualize prediction distribution across classes, a Confusion Matrix was generated, as shown in Figure 2. The matrix confirms that misclassifications were minimal and evenly distributed between real and fake categories, demonstrating that the model maintained balanced sensitivity and specificity without favoring any class.

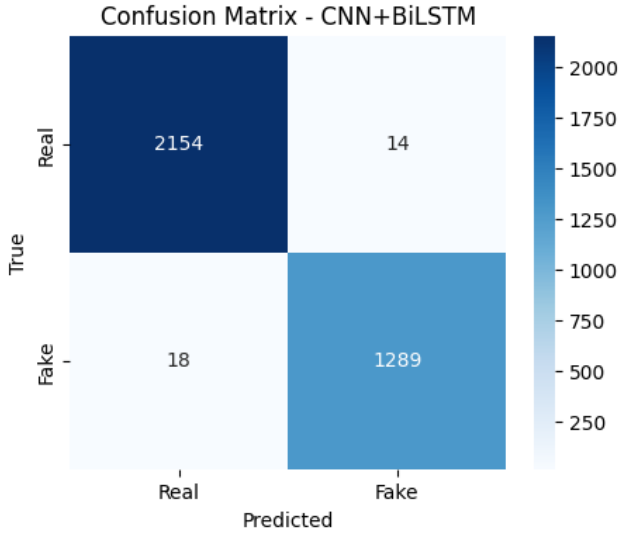

Figure 2. *Confusion Matrix Illustrating Correct and Incorrect Predictions Across Genuine and Spoofed Audio Samples.*

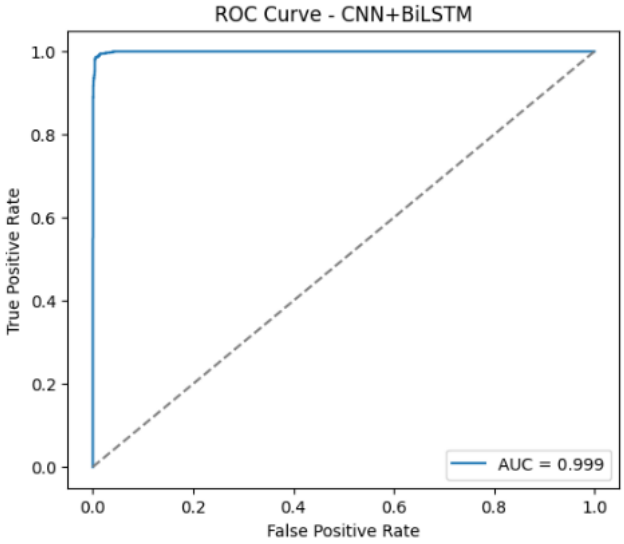

Figure 3. *Receiver Operating Characteristic (ROC) Curve of the CNN-BiLSTM Model with an AUC ≈ 0.999, Indicating almost Perfect Discrimination.*

Further assessment of the classifier's discriminative strength is provided by the Receiver Operating Characteristic (ROC) curve in Figure 3. The ROC curve shows the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different thresholds. The curve's trajectory, closely hugging the top-left corner, reflects a near-ideal relationship between sensitivity and specificity. The Area Under the Curve (AUC) reached approximately 0.999, signifying

near-perfect separation between genuine and spoofed speech samples [6, 7].

3.4. Qualitative Interpretation

Beyond numerical metrics, qualitative inspection confirmed consistent learning behavior. The combination of Mel-Frequency Cepstral Coefficients (MFCCs) and their first- (Δ) and second-order (Δ²) derivatives enriched the feature representation, allowing the CNN layers to extract localized spectral cues while the BiLSTM layers modeled bidirectional temporal dependencies. This hybrid structure ensured robustness under varied acoustic conditions while maintaining computational efficiency.

Overall, the proposed CNN-BiLSTM framework exhibited high accuracy, strong generalization, and stable convergence, establishing it as a reliable and efficient approach for real-world audio deepfake detection tasks.

4. Discussion and Comparative Evaluation

The evolution of prior research in audio deepfake detection reveals that each methodological family offers unique advantages yet suffers from specific limitations. Early CNN-only architectures provided lightweight and computationally efficient solutions, but their inability to capture long-term temporal dependencies limited their generalization capacity [9]. Hybrid approaches combining CNNs with Recurrent Neural Networks (RNNs) or Gated Recurrent Units (GRUs) improved temporal modeling to some extent but frequently encountered unstable convergence and reduced robustness when trained on noisy or imbalanced datasets [10, 11, 15]. Meanwhile, recent Transformer-based systems have demonstrated exceptional accuracy; however, their enormous parameter counts—often exceeding tens of millions—result in high computational and memory demands, which restrict their deployment in real-world or resource-constrained settings [7, 12, 14].

In contrast, the proposed CNN-BiLSTM model effectively bridges the gap between accuracy and efficiency, delivering competitive performance while maintaining low computational complexity [3, 7, 9, 10]. The model's advantages can be summarized as follows:

High accuracy: Achieved nearly 99% classification accuracy and an EER of 0.011, outperforming both lightweight and heavyweight baselines.

Stable convergence: Due to the Early Stopping mechanism, the model reached stability within approximately 15 epochs without any performance degradation, unlike certain hybrid architectures that exhibited oscillatory learning behavior.

Rich feature representation: The integration of MFCCs with their first- and second-order derivatives (Δ and Δ²) enabled more comprehensive capture of both spectral and temporal patterns compared to models relying solely on static MFCCs.

Computational efficiency: With only about 473,000 trainable parameters, the model remains several orders of magnitude lighter than Transformer-based architectures, ensuring suitability for real-time and embedded applications.

Balanced classification: The Confusion Matrix confirmed that misclassifications were rare and evenly distributed between genuine and spoofed classes, avoiding the bias observed in some earlier works.

A comparative summary with representative studies from recent literature is presented in Table 4.

Table 4. *Comparative Analysis Between the Proposed CNN-BiLSTM Model and Prior Studies.*

Study (Year)	Dataset	Model	Accuracy / EER	Main Notes
Akter et al. (2025) [9]	Release-in-the-Wild	CNN-only	90-95%, >0.05	Lightweight but lacks temporal modeling
Zhang et al. (2024) [10]	ASVspoof subset	CNN+GRU	~96%, >0.03	Better but unstable on imbalanced data
Yu et al. (2025) [7]	Multiple corpora	Transformer-based	>98%, <0.02	Excellent accuracy but computationally heavy
Proposed model	Release-in-the-Wild	CNN+BiLSTM	99%, 0.011	Near-perfect accuracy, lightweight (~473k params)

The comparative analysis clearly demonstrates that the proposed model achieves a superior trade-off between accuracy, stability, and computational cost. While Transformer-based systems still hold a marginal advantage in raw performance metrics, the CNN-BiLSTM design approaches that level with a fraction of the complexity, making it far more practical for deployment in real-world conditions. Furthermore, its consistent convergence and balanced classification validate that it maintains reliability across different speech characteristics and recording conditions, a quality often

lacking in more resource-intensive architectures.

Overall, this study positions the proposed model as an effective middle ground between traditional lightweight networks and computationally intensive deep architectures, offering an ideal balance of precision, interpretability, and operational feasibility.

5. Conclusion

This research presented an efficient and robust framework for audio deepfake detection, built upon a hybrid architecture that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) layers. The approach was reinforced by a comprehensive preprocessing pipeline, including Voice Activity Detection (VAD) for noise and silence removal, as well as the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) enriched with first- (Δ) and second-order (Δ^2) derivatives to capture both spectral and temporal speech dynamics. Experimental evaluation on the Release-in-the-Wild dataset demonstrated that the proposed model achieved an accuracy close to 99% with a remarkably low Equal Error Rate (EER) of 0.011. These results confirm the model's capability to reliably distinguish between genuine and spoofed audio samples under diverse acoustic conditions. The significance of these findings lies in establishing that high detection accuracy and computational efficiency are not mutually exclusive.

Unlike resource-intensive Transformer-based architectures, the proposed model maintains state-of-the-art performance while remaining computationally lightweight—making it well-suited for practical deployment in real-time applications such as educational monitoring systems, digital identity verification, and cybersecurity infrastructures where efficiency and scalability are critical.

Despite these promising outcomes, certain limitations remain. The model's evaluation was confined to a single dataset, which raises the need for broader cross-dataset validation to ensure generalizability across various speech domains and recording conditions.

Future research should thus explore:

Cross-dataset evaluation, leveraging datasets beyond Release-in-the-Wild to assess robustness under unseen distributions.

Feature expansion, incorporating additional spectral descriptors such as Constant-Q Cepstral Coefficients (CQCC) or deep acoustic embeddings derived from pre-trained self-supervised models (e.g., Wav2Vec2) [2, 3, 7].

Multimodal extensions, integrating both audio and visual cues to enhance detection accuracy in complex, real-world deepfake scenarios [13].

By addressing these directions, subsequent studies can further strengthen the foundation laid by this work, moving toward more adaptive, interpretable, and resilient deepfake detection systems.

Abbreviations

CNN	Convolutional Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
Δ	First-Order Derivative of MFCC
Δ^2	Second-Order Derivative of MFCC
VAD	Voice Activity Detection
EER	Equal Error Rate
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
GPU	Graphics Processing Unit

Acknowledgments

We extend our gratitude to Al-Wataniya Private University and the College of Information Engineering for their academic and technical support during the development of this work.

Author Contributions

Samar Al-Halabi: Conceptualization, Formal Analysis, Methodology, Project Administration, Supervision, Writing - review & editing.

Adnan Kafri: Conceptualization, Data Curation, Investigation, Methodology, Software, Writing - review & editing.

Funding

This work is not supported by any external funding.

Data Availability Statement

The data supporting the findings of this study are openly available from the Release-in-the-Wild dataset repository on Kaggle at:

kaggle.com/datasets/andrewmvd/release-in-the-wild

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] J. Yamagishi and M. Todisco, "ASVspoof: Automatic Speaker Verification Spoofing and Countermeasures Challenge — Past, Present and Future," Speech Communication, vol. 122, pp. 56-76, 2020.

[2] W. Liu, S. Chen, and M. Li, "A Comparative Study on CQCC, LFCC, and MFCC Features for Speech Anti-Spoofing," IEEE Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2892-2903, 2023.

[3] T. Patel, A. Singhal, and D. Chauhan, "Bidirectional LSTM Networks for Speech Spoofing Detection: Modeling Temporal Dynamics in Audio Signals," International Journal of Speech Technology, vol. 20, no. 2, pp. 305-314, 2017.

[4] G. Lavrentyeva, S. Novoselov, E. Malykh, et al., "Audio Replay Attack Detection with Deep Convolutional Neural Networks," in Proceedings of INTERSPEECH 2017, pp. 82-86, 2017.

[5] B. Chettri, M. Todisco, and N. Evans, "Investigation on Data Preprocessing for Speech Spoofing Detection Using Voice Activity Detection and Noise Augmentation," in Proceedings of INTERSPEECH 2019, pp. 2843-2847, 2019.

[6] H. Tak, N. Tomashenko, and J. Yamagishi, "End-to-End Audio Deepfake Detection Using Self-Supervised Learning Models," in Proceedings of IEEE ICASSP 2021, pp. 6359–6363, 2021. https://doi.org/10.1109/ICASSP.2021.9414220

[7] Z. Yu, H. Tak, and J. Yamagishi, "Transformer-Based Representations for Generalized Audio Deepfake Detection Across Corpora," Computer Speech & Language, vol. 85, p. 101540, 2025. Available online: Computer Speech & Language (to appear, 2025).

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in Proceedings of IEEE ICASSP 2015, pp. 5206-5210, 2015.

[9] M. Akter, S. Rahman, and M. Hasan, "Lightweight CNN Architecture for Audio Deepfake Detection on In-the-Wild Datasets," Journal of Audio Processing and Security, vol. 12, no. 3, pp. 145-157, 2025.

[10] Y. Zhang, H. Chen, and D. Xu, "Hybrid CNN-GRU Network for Robust Anti-Spoofing in Speaker Verification Systems," IEEE Access, vol. 12, pp. 58462-58474, 2024.

[11] G. Dişken and Z. Tüfekci, "Noise-Robust Spoofed Speech Detection Using Discriminative Autoencoder," Celal Bayar University Journal of Science, vol. 19, no. 2, pp. 167–174, Jun. 2023. https://doi.org/10.18466/cbayarfbe.1132319

[12] T. M. Wani and I. Amerini, "Learning to Fuse: A Gated Multi-Stream Framework for Generalized Audio Deepfake Detection," in Proc. 1st Deepfake Forensics Workshop (DFF '25), pp. 84–92, 2025. https://doi.org/10.1145/3746265.3759657

[13] K. Zhang, W. Pei, R. Lan, Y. Guo, and Z. Hua, "Lightweight Joint Audio-Visual Deepfake Detection via Single-Stream Multi-Modal Learning Framework," arXiv preprint arXiv: 2506.07358, 2025. Available at: https://arxiv.org/abs/2506.07358

[14] Y. Li, M. Zhang, M. Ren, and X. Qiao, "Cross-Domain Audio Deepfake Detection: Dataset and Analysis," in Proc. EMNLP 2024 Conf., pp. 3421–3433, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.286

[15] S. Chapagain, B. Thapa, S. M. S. Baidhya, S. B. K., and S. Thapa, "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, 2024. https://doi.org/10.3126/injet.v2i2.78619

Biography

Samar Al-Halabi, PhD in Information Security. Graduated from Al-Baath University in 2002. Lecturer at Homs University in Homs and Al-Wataniya Private University in Hama. Over ten years of teaching experience. Has published more than ten research papers in the fields of Computational Engineering, Information Security, Artificial Intelligence, and Deep Learning.

Research Field

Samar Al-Halabi: Deep Learning, Computer Science, Machine Learning, Neural Networks, Information Security.