Urdu Audio Deepfake Detection: A Deep Learning Approach

Muhammad Qasim Ali, Ali Raza, Abdullah Awan
GIK Institute of Engineering Sciences and Technology, Topi, Pakistan
{u2022414, u2022664, u2022274}@giki.edu.pk

***Abstract*—The rapid advancement of synthetic audio generation techniques has significantly increased the risk of audio deepfakes, posing serious challenges to information security and trustworthiness, especially for low-resource languages such as Urdu. This study focuses on developing and evaluating robust deep learning models to accurately detect synthetic Urdu speech. We propose two distinct architectures: a Convolutional Neural Network (CNN) trained on log-scaled Mel-spectrogram representations, and a Long Short-Term Memory (LSTM) network trained on Mel-Frequency Cepstral Coefficients (MFCCs). Both models are designed to differentiate between authentic (bonafide) and spoofed audio generated using state-of-the-art Text-to-Speech (TTS) systems including Tacotron and VITS. Experimental results demonstrate that the CNN model achieves a superior accuracy of approximately 89%, while the LSTM attains an accuracy near 85%, highlighting the effectiveness of spectrogram-based spatial features and temporal feature modeling, respectively. These findings establish a valuable baseline for Urdu audio deepfake detection and contribute to the broader goal of securing low-resource languages against synthetic audio threats. Future work will explore enhancements such as real-time detection capabilities, speaker- and accent-invariant models, and cross-lingual generalization using transformer-based architectures.**

***Index Terms*—Deepfake detection, Urdu language, Audio forensics, Convolutional Neural Network, Long Short-Term Memory, Mel-spectrogram, Mel-Frequency Cepstral Coefficients, Low-resource languages, Synthetic speech detection.**

I. Introduction

Audio deepfakes, which involve synthetically generated or manipulated voice recordings, have emerged as a serious threat in the digital era. Enabled by advances in neural text-to-speech (TTS) and voice conversion technologies, these artificially crafted audio clips can convincingly mimic human voices. This opens up opportunities for misuse in domains ranging from misinformation and social engineering to identity fraud, cyberbullying, and political propaganda.

While considerable progress has been made in developing detection techniques for English and other high-resource languages, low-resource languages like Urdu remain underrepresented in this domain. The linguistic, phonetic, and acoustic complexities of Urdu, combined with a scarcity of annotated datasets, make deepfake detection especially challenging. Furthermore, the sociopolitical sensitivity of Urdu-speaking regions amplifies the consequences of synthetic voice misuse.

Existing work in audio deepfake detection primarily revolves around extracting spectral or prosodic features such as Mel-Frequency Cepstral Coefficients (MFCCs), linear prediction coefficients (LPCs), and Mel-spectrograms. These features are then processed using deep learning models—particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers. However, most of these models are trained and evaluated on English-language datasets like ASVspoof and FakeAVCeleb, which do not generalize well to non-English data.

In this study, we focus on Urdu—a low-resource yet widely spoken language. We present a comparative analysis of two deep learning-based detection approaches: a CNN model trained on Mel-spectrograms, and a Long Short-Term Memory (LSTM) model trained on MFCC sequences. To enable this investigation, we utilize the CSALT Deepfake Detection Dataset, a recent contribution tailored specifically for Urdu deepfake audio analysis.

Our experiments demonstrate that both models are capable of discerning real from spoofed Urdu audio with high accuracy. The CNN model achieves approximately 89% accuracy, outperforming the LSTM model, which records an accuracy of 85%. These results underscore the feasibility of developing robust deepfake detection systems for low-resource languages and lay a foundation for future research in multilingual audio forensics, real-time detection, and adversarial robustness.

II. Related Work

The rapid advancement of generative speech technologies has prompted significant research into the detection of audio deepfakes. Most existing approaches focus on detecting artifacts introduced by synthetic voice generation models, using a combination of signal processing techniques and deep learning classifiers.

Earlier methods in audio spoofing detection relied on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and phase-based features. However, with the evolution of deep learning, there has been a shift toward end-to-end models that learn robust feature representations directly from the audio signal. Convolutional Neural Networks (CNNs), in particular, have demonstrated strong performance on spectrogram-based inputs due to their ability to capture spatial patterns and frequency-related anomalies.

Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have also been widely adopted in audio deepfake detection tasks. Their capability to model temporal dependencies makes them suitable for

analyzing the dynamic nature of speech and identifying inconsistencies in rhythm, intonation, and phoneme transitions that may arise in synthesized audio.

More recent efforts have explored the use of self-supervised learning and transformer-based architectures, which leverage large-scale pretraining to capture nuanced speech representations. These approaches show strong generalization across multiple spoofing techniques and are gaining traction in both research and real-world applications.

Several public benchmarks and datasets have been instrumental in driving research in this area. Datasets such as those released for the ASVspoof Challenges have provided standardized platforms for evaluating model performance on a variety of spoofing attacks, including text-to-speech (TTS), voice conversion, and replay attacks. Other datasets focus on multimodal deepfakes or TTS-specific scenarios, enabling researchers to train and validate models on diverse types of fake audio.

However, the majority of these datasets and models have been developed for high-resource languages like English, limiting their applicability to low-resource settings. Languages such as Urdu lack sufficient annotated data and dedicated detection frameworks, making it difficult to build accurate and robust detection systems.

To address this gap, our work uses a specialized dataset of Urdu audio deepfakes, including bonafide and spoofed samples generated using modern TTS systems such as Tacotron and VITS. By evaluating both CNN and LSTM models on this dataset, we aim to establish a baseline for future Urdu-specific research and highlight the importance of linguistic diversity in audio forensics and deepfake mitigation.

III. Methodology

A. Dataset

We utilize the CSALT Deepfake Detection Dataset (Urdu) [1], which contains both bonafide and spoofed audio samples. The spoofed samples are generated using Tacotron and VITS-based text-to-speech (TTS) systems. This dataset is one of the few publicly available resources specifically for deepfake detection research in the Urdu language.

The dataset is structured as follows:

- **Bonafide Part 1:** 708 files (1,302.66 minutes)
- **Bonafide Part 2:** 495 files (1,271.65 minutes)
- **Tacotron-generated Spoofed:** 495 files (1,061.96 minutes)
- **VITS-generated Spoofed:** 495 files (1,340.79 minutes)

Prior to feature extraction, the audio data undergoes several preprocessing steps:

- Conversion from stereo to mono audio
- Trimming of leading and trailing silence
- Amplitude normalization to standardize volume levels
- Resampling and padding to ensure consistent audio length across samples

B. Feature Extraction

Two distinct feature sets are extracted depending on the model architecture:

- **For CNN:** Log-scaled Mel-spectrograms are computed using a 25 ms window length with a 10 ms hop size. The resulting spectrograms have dimensions of 128×128.

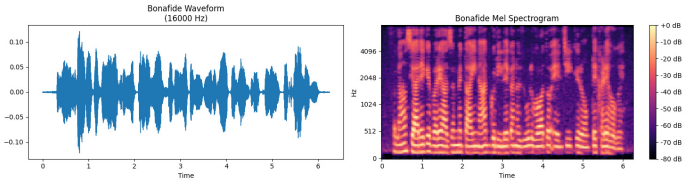

Fig. 1: Waveform and Mel-spectrogram of bonafide audio.

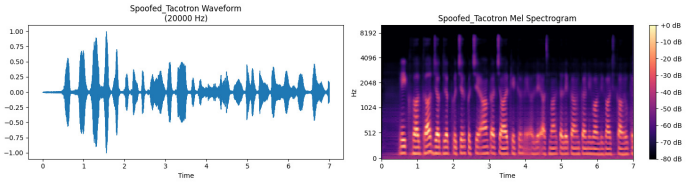

Fig. 2: Waveform and Mel-spectrogram of spoofed audio.

- **For LSTM:** 13-dimensional Mel-frequency cepstral coefficients (MFCCs) are extracted per frame. These sequences are zero-padded to maintain uniform length for batch processing.

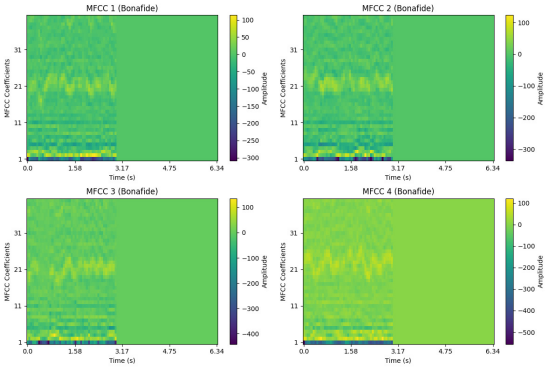

Fig. 3: MFCC coefficients extracted from bonafide audio.

C. Model Architectures

1) CNN Model: The CNN architecture comprises two convolutional layers with 32 and 64 filters, respectively. Each convolutional layer is followed by ReLU activations and 2×2 max pooling. After convolutional blocks, a dropout layer is applied to reduce overfitting, followed by two fully connected (dense) layers. The final output layer uses a sigmoid activation function for binary classification between bonafide and spoofed audio.

2) LSTM Model: The LSTM-based model contains a single LSTM layer with 128 units to capture temporal dependencies in the MFCC sequences. This is followed by dropout and fully connected layers. The final classification layer uses a sigmoid activation function. Input sequences are pre-padded to maintain consistent length during training and inference.

IV. Training and Evaluation

A. Training Configuration

Both models use binary crossentropy loss and the Adam optimizer with a learning rate of 0.001. Batch size is 32, and early stopping is employed with a patience of 3 to prevent overfitting.

B. Evaluation Metrics

To comprehensively evaluate the performance of our deepfake audio detection models, we employ several standard metrics:

- **Accuracy:** The ratio of correctly classified samples (both bonafide and spoofed) to the total number of samples. It provides a general overview of model performance.
- **Precision:** The proportion of true positive predictions among all positive predictions made by the model. High precision indicates a low false positive rate, which is critical to avoid misclassifying bonafide audio as spoofed.
- **Recall:** Also known as sensitivity, it measures the proportion of actual positive samples correctly identified by the model. A high recall ensures that most spoofed audios are detected.
- **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is imbalanced.
- **Confusion Matrix:** A tabular visualization of true positives, true negatives, false positives, and false negatives that offers detailed insight into classification errors.
- **Training and Validation Curves:** Loss and accuracy plots over epochs for both training and validation sets are analyzed to monitor model convergence, detect overfitting, and assess generalization capability.

TABLE I: Performance comparison of CNN and LSTM models on the CSALT Urdu Deepfake Dataset

Model	Accuracy	Precision	Recall	F1-Score
CNN	89%	88%	90%	89%
LSTM	85%	84%	86%	85%

From Table I, it is evident that the CNN model outperforms the LSTM model across all metrics, particularly in terms of accuracy and recall. The superior performance of the CNN can be attributed to its ability to effectively capture spatial patterns in the log-scaled Mel-spectrogram inputs. Meanwhile, the LSTM, which processes sequential MFCC features, shows competitive but slightly lower performance, likely due to challenges in modeling temporal dependencies over variable-length sequences.

Additionally, the training and validation loss/accuracy curves (not shown here) demonstrate stable convergence for both models, with the CNN exhibiting faster convergence and less overfitting, as indicated by closer training and validation curves.

Overall, these metrics and visualizations provide a robust framework for assessing the models' ability to distinguish bonafide and spoofed audio in the Urdu deepfake detection task.

V. Results and Visualizations

Figures 4 and 5 illustrate the training and validation loss and accuracy curves for the CNN and LSTM models, respectively.

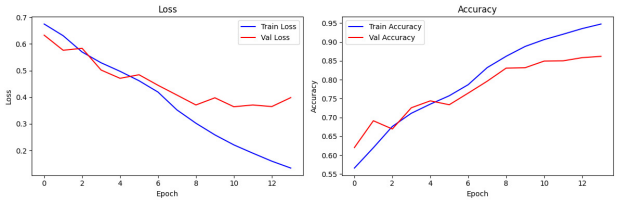

Fig. 4: Training and validation loss and accuracy curves for the CNN model. The model shows steady convergence with minimal overfitting.

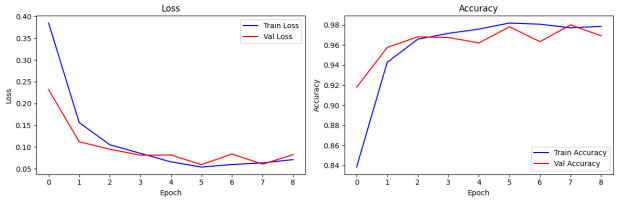

Fig. 5: Training and validation loss and accuracy curves for the LSTM model. Although the model converges, it exhibits slightly higher variance between training and validation metrics compared to CNN.

The CNN model demonstrates faster and more stable convergence, with the training and validation curves closely aligned, indicating effective generalization. In contrast, the LSTM model, while also converging, shows some divergence between training and validation performance, suggesting potential overfitting or sensitivity to input sequence length variations.

These results, combined with the quantitative evaluation metrics, reinforce the effectiveness of CNNs in capturing spatial features from Mel-spectrograms for the task of Urdu deepfake audio detection.

VI. Challenges and Limitations

Despite the encouraging results, several challenges and limitations were encountered during this study:

- **Limited Dataset Availability:** There is a scarcity of publicly available Urdu audio deepfake datasets, which restricts model training and evaluation diversity.

- **Accent and Speaker Diversity:** Variability in regional accents and speaker characteristics affects the model's ability to generalize across different audio samples.
- **Quality of Spoofed Samples:** Some spoofed audio clips contain low synthesis quality or artifacts, which can bias the model toward detecting obvious flaws rather than subtle deepfake cues.
- **Real-time Detection Latency:** The computational complexity of the models, particularly LSTM processing of sequential features, leads to latency issues that challenge real-time deployment scenarios.

Addressing these limitations in future work through expanded datasets, data augmentation, and optimization for faster inference will be crucial to improving the robustness and practicality of Urdu deepfake audio detection systems.

VII. Conclusion and Future Work

In this study, we presented a comparative analysis of CNN and LSTM architectures for the task of Urdu deepfake audio detection. Our results demonstrate that CNN models outperform LSTM counterparts, primarily due to their superior ability to capture robust spatial features from Mel-spectrogram representations. This highlights the effectiveness of convolutional approaches for audio forensics in low-resource languages such as Urdu.

These findings establish a strong foundation for further research in deepfake detection within resource-constrained settings. Moving forward, future work will focus on the following directions:

- Expanding the size and diversity of publicly available Urdu deepfake audio datasets to improve model generalization.
- Exploring advanced feature representations by integrating wav2vec 2.0 and transformer-based architectures, which have shown promise in various speech processing tasks.
- Developing lightweight, computationally efficient models capable of real-time deepfake detection for practical deployment.
- Investigating the cross-lingual transferability of models trained on Urdu data to other low-resource languages, enhancing the adaptability of the proposed techniques.

Additionally, it is important to consider the ethical implications of deepfake detection technologies. While these systems can help mitigate malicious uses of synthetic media, such as misinformation and fraud, care must be taken to ensure they do not inadvertently discriminate against certain accents or speakers. Transparent evaluation and inclusive dataset curation are essential to foster fairness and trustworthiness in audio forensics. Ultimately, advances in this field will contribute to safer digital communication environments and protect individuals from audio-based deception.

Addressing these areas will be crucial for advancing the state-of-the-art in audio deepfake detection, particularly for underrepresented languages and real-world applications.

References

[1] A. Author *et al.*, "Deepfake Defense: Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset," in *Proc. ACL*, 2024.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, et al., "Tacotron: Towards End-to-End Speech Synthesis," *Proc. Interspeech*, 2017.

[3] K. Kim, J. Lee, and W. Kim, "VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," *Proc. ICML*, 2021.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[5] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[6] J. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018.

[7] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE ICASSP*, 2013.

[8] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE ICASSP*, 2013.

[9] N. Evans, S. Marcel, and H. Wang, "Audio Forensics: New Approaches to Old Problems," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 110–116, 2015.

[10] N. Chen, Y. Qian, and K. Yu, "Transformer-based end-to-end speech recognition," *Proc. Interspeech*, 2020.

[11] T. Kinnunen, H. Delgado, et al., "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," *arXiv preprint arXiv:1904.05441*, 2019.

[12] Z. Li, S. Fang, and H. Lin, "Audio Deepfake Detection Using Self-Adaptive Neural Networks," *IEEE Access*, vol. 9, pp. 104672–104681, 2021.

[13] J. Ko, J. Park, and S. Kang, "Audio Data Augmentation for Deep Learning Based Environmental Sound Classification," *Proc. ICASSP*, 2018.

[14] M. Sandler, A. Howard, M. Zhu, et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. CVPR*, 2018.

[15] S. Chen, J. Xie, et al., "Self-Attention Networks for Audio Classification," *Proc. ICASSP*, 2020.

[16] D. Snyder, D. Garcia-Romero, et al., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *Proc. ICASSP*, 2018.

[17] W. Wu, Y. Sun, and Y. Qian, "Robust Audio Deepfake Detection via Multi-Scale Feature Fusion," *IEEE Transactions on Information Forensics and Security*, 2023.

[18] S. Chakraborty, P. K. Ghosh, et al., "End-to-End Deepfake Audio Detection Using Raw Waveforms," *Proc. Interspeech*, 2022.

[19] F. Li, C. Yang, et al., "Multimodal Deepfake Detection Using Audio-Visual Consistency," *IEEE Transactions on Multimedia*, 2022.

[20] J. Shen, R. Pang, et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *Proc. ICASSP*, 2018.