NUML International Journal of Engineering and Computing
Volume: 2 Issue: 1
https://numl.edu.pk/journals/nijec
Print ISSN: 2788-9629
E-ISSN: 2791-3465
DOI: 10.52015/nijec.v2i1.37

Deep Learning Approach for Detecting Audio Deepfakes in Urdu

Marium Mateen [a*]

[a,*] *School of Computing Department of Computer Science & Software Engineering, Jinnah University for Women Karachi, Pakistan (mariummateen12@gmail.com)*

Submitted	Revised	Published
05-June-2023	9-July-2023	26-July-2023

Abstract

The application of Deep Learning algorithms for speech synthesis has led to the widespread generation of Audio Deepfakes, which are becoming a real threat to voice interfaces. Audio Deepfakes are fake audio recordings that are difficult to differentiate from real recordings because they use AI-generated techniques to clone human voices. When prominent speakers, celebrities, and politicians are the target of Audio Deepfakes, this technology can potentially undermine public confidence and trustworthiness. Therefore, it is essential to create efficient methods and technologies to identify and stop the creation and spread of Audio Deepfakes. To address the critical issue of the widespread circulation of fake audio and to detect Audio Deepfakes, several Machine Learning and Deep Learning techniques have been developed recently. However, most such solutions have been trained using datasets in English, raising concerns about their accuracy and trustworthiness for other languages. The primary objective of this research is to develop a Deep Learning model for detecting Audio Deepfakes in Urdu. For this purpose, the deep learning model is trained using an Urdu language audio dataset. The dataset was prepared using both real and fake audio. The real Urdu audio clips were initially collected from which Deep fakes were generated with the help of the Real-Time Voice Cloning tool. Our Deep Learning-based model is built to detect Audio Deep fakes produced using imitation and synthesis techniques. According to the findings of our study, when tested and evaluated, our model obtained an accuracy of 91 percent.

Keywords: Audio DeepFakes, Machine Learning, Deep Learning, LSTM, RNN

* Corresponding Author: Marium Mateen (mariummateen12@gmail.com)

1. Introduction

Speech generation technologies based on artificial intelligence have produced several tools capable of generating fake voices that sound very similar to real voices [1]. Machine-generated voices in various forms using artificial intelligent agents have been prevalent more recently. With technology automation, people are more frequently used to control daily tasks through speech. Artificial or machine-generated voices are increasingly employed in virtual assistants such as Alexa, Google Assistant, Siri, and others. Such technologies are useful in various applications such as customer services, marketing, smart home appliances, audiobooks, etc.

Although voice generation technology has been developed for the betterment of the community, it has also been utilized maliciously to disseminate false information around the world using audio. It has sparked concern about Audio Deepfakes. An Audio Deepfake is an artificial intelligence technique that can produce highly realistic speech patterns resembling a particular person. These are usually referred to as audio manipulators, becoming available through websites, desktop machines, and mobile devices [2].

Despite the benefits of such technologies, they have become a constant threat. The enormous amount of voice recordings is broadcast online, typically on social media platforms, which makes it difficult to distinguish fakes from real ones. Hence, it has been the need of the era to authenticate any audio recordings to prevent the spreading of misinformation. Therefore, this issue has been of significant interest to the scholarly community over the years.

Many detection methods have been developed to discern fake audio files from real ones. Different ML and DL models that employ various tactics to recognize false audio have been created. This research implements a deep learning model to detect Audio DeepFakes generated from synthesized and imitation techniques. For this purpose, the generated dataset contains real audio clips in Urdu. The research has utilized Real-Time Vice Cloning Tool to generate DeepFakes of real audio clips. Deep learning models are trained on the generated Urdu Audio dataset. Later on, the performance of this model is evaluated.

Below is the summary of the major contributions of this research:

- To collect real audio clips in Urdu and pre-process them.
- To create DeepFakes of the collected real audio clips using the Real-Time Voice Cloning Tool
- Applying Deep Learning based system for detecting audio DeepFakes generated from synthesized and imitation techniques.
- Test and evaluate the performance of these models on Urdu Audio.

The rest of the paper is structured as follows. Section 2 presents a brief discussion of relevant works conducted previously. Section 3 discusses the methodologies utilized to conduct the research and section 4 shows the results of implementing those methodologies. The last section concludes the overall discussion.

2. Background And Related Work

This section explains the techniques for generating fake audio and also describes the current studies in detecting fake audio.

There are three major types of manipulation techniques: imitation-based, synthetic-based, and replay-based that are used to generate fake audios which are discussed below:

2.1 Imitation-based

A technique for changing speech (confidential sound) to appear like another voice (intended sound), with the main goal of preserving the secrecy of the secret audio. There are several techniques to mimic audio, including requesting an individual with a voice similar to the speaker to do so. Deepfake audios may be produced using certain masking algorithms, such as Efficient Wavelength Mask. Target and original audio with comparable features will be captured specifically. The target audio will then be created utilizing an imitation generation approach, which will create a new voice that is the false one, using the signal of the original audio. As a result, it is challenging for people to distinguish between false and genuine sounds produced using this technique [3].

2.2 Synthetic-based

It comprises three segments: a text analysis model, an acoustic model, and a vocoder, and its goal is to convert text into appropriate and lifelike voice in real time [4]. Two essential measures must be taken to produce artificial Deepfake audio. In the beginning, clear and organized raw audio and a transcript of the audio speech should be gathered. Second, a synthetic audio generation model must be created utilizing training data from the Text to Speech model. The most realistic audio may be produced using the well-known model generation methods Tactoran 2, Deep Voice 3, and Fast Speech 2 [5,6].

2.3 Replay-based

Replay attacks are described as playing back a speaker's audio recording. Cut-and-paste and far-field attacks are the two categories of detection attacks [7]. A far-field microphone recording of the target played back on a phone handset with a speaker is used as the test section in far-field detection replay attacks. To simulate the sentence required by a text-dependent system, a recording must be constructed by cutting and pasting brief recordings [7]. The emphasis of this work is imitation-based, and synthetic-based DeepFake Audio Detection, replay-based detection will not be covered.

2.4 DeepFake Detection Mechanism

The necessity of tools for recognizing such Deepfakes in various dialects and languages has risen due to the availability of AI-based technologies for producing Deepfakes. This section presents the Deepfake Detection techniques currently available for voices fabricated using imitation- and synthetic-based spoofing methods.

Most of the research in this area uses the ASVspoof dataset [8]. It focuses on its Logical Access partition, which raises concerns about the algorithms' ability to identify Deep Fakes when presented with voices from people with various accents and linguistic backgrounds. Iqbal et al. [9] presented a novel approach for detecting Deepfake audio using feature engineering and machine learning techniques. The proposed approach utilized a set of audio features, including Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and Spectral Contrast (SC), to capture the unique characteristics of human speech. These features are then fed into a Random

Forest classifier, trained to distinguish between real and Deepfake audio. The authors conducted experiments on a dataset of real and Deepfake audio recordings and achieved an accuracy of 93.3%. The proposed approach offers a promising solution for detecting Deepfake audio and can be extended to other languages and dialects. Hamza et al. [10] propose a novel technique for detecting Deepfake audio using MFCC features and machine learning. The authors report achieving a high accuracy rate of over 90% in their experimental results. However, more research and testing are required to evaluate this approach's effectiveness in real-world applications fully.

Camacho et al. [11] suggested a convolutional neural network (CNN)-based model. In this model, audio is first translated to scatter plot pictures of nearby samples before being fed as CNN input. The model is trained and evaluated using the Fake or Real (FoR) dataset [12]. The model claims 88.9% accuracy. Using data from several generation techniques during training, the suggested model solved the generalization issue of DL-based models. Yet, its performance lagged below that of other models in the literature. A novel audio feature descriptor known as ELTP-LFCC [11] was created by T. Arif et al. based on a Local Ternary Pattern (ELTP) and Linear Frequency Cepstral Coefficients (LFCC). The Deep Bidirectional Long-Term Memory (DBiLSTM) network was employed in conjunction with this descriptor to improve the model's resilience and ability to identify bogus audio in various indoor and outdoor ambient settings. The developed model was evaluated using artificially false audio and replicated from the ASVspoof 2019 dataset. According to the experiment results, the model scored better with the audio synthetic dataset (with 0.74% EER) than it did with samples based on imitations (33.28% EER).

Wang et al. [13] created the Deep-Sonar DNN model to represent the neurological activities of speaker recognition (SR) systems versus artificial intelligence (AI)-generated bogus sounds. The classification of this model is based on the Layer-wise neuron behaviors. On the voices of English speakers from the FoR dataset [11], the suggested model attained a detection rate of 98.1% with an EER of around 2% [13]. However, real-world noise has a significant impact on DeepSonar's performance. Another Deep learning-based model known as Deep4Snet [14] was constructed. The suggested model used a 2D CNN model to show the audio dataset (histogram). The model has a 98.5% success rate in distinguishing between imitation and artificial sounds. The performance of Deep4SNet, however, was affected by the data translation process and was not scalable. Zhenchun Lei et al. [15] presented a 1-D CNN and Siamese CNN to identify fraudulent audio. While the Siamese CNN was based on two trained GMM models, the 1-D CNN used speech log probabilities as input. The Siamese CNN concatenated two identical 1-D CNNs utilizing a fully connected and SoftMax output layers. It featured two identical 1-D CNNs. The proposed Siamese CNN beat the GMM and 1-D CNN by improving the min-tDCF and Equal Error Rate (EER) by around 55% when employing the LFCC features, according to tests conducted on the ASVspoof 2019 dataset.

For the categorization of audio DeepFakes, Chintha et al. [16] created two brand-new models based on convolution RNN. Five layers of recovered audio signals are used in the first model, Convolution Recurrent Neural Network Spoof (CRNN-Spoof), which is fed into a bidirectional LSTM network to forecast false sounds. The second one, the Wide Inception Residual Network Spoof (WIRE-Net-Spoof), employs a function called weighted negative log-likelihood and a different training procedure. In the ASV spoof challenge 2019 dataset, the CRNN-Spoof technique outperformed the WIRE-Net-Spoof method by 0.132% of the Tandem Decision Cost Function (t-DCF) with a 4.27% EER. This study's utilization of several layers and convolutional networks resulted in managerial complexity, one of its limitations.

Alzantot et al. [17] emphasized the requirement for creating a residual CNN-based system for Audio DeepFake detection. This model aims to extract three significant characteristics from the input—MFCC and constant Q cepstral coefficients. (CQCC), and STFT. To calculate the Counter Major (CM) score of the counterfeit audio. A high CM score establishes the authenticity of the audio, whereas a low CM score raises doubts about its originality. The suggested approach improved the CM rate in two matrices of t-DCF (0.1569) and EER (6.02) by 71% and 75%, respectively, showing encouraging results. However, more research is still required because of the generalization mistakes in the suggested system. Table 1 summarizes the contribution of various research studies in detecting the generation of Audio Deepfakes.

Table 1: Summary of Literature Review

Year	Reference	Language	DeepFake	Model	Features	Dataset	Limitations
2022	Iqbal et al. [9]	English	Synthesizes	SVM, MLP, DT, LR NB and XGB	MFCC, spectral roll-off, spectral centroid, spectral contrast, spectral bandwidth, and zero crossing rate	FoR [12]	Relatively small dataset size and a lack of testing against advanced and realistic deepfake audio techniques.
2022	Hamza et al. [10]	English	Synthesizes	SVM, MLP, DT, XGB, ETC, LR, GNB, AB, GB, LDA and QDA	MFCC	FoR [12]	Relatively small dataset, which may limit the generalizability of the findings
2021	S. Camacho [11]	English	Synthesizes	CNN	Scatter Plots	FoR [12]	It performed worse than the conventional DL techniques, and the model required additional training.
2021	T. Arif et al. [18]	English	Imitated + Synthesized	DBiLSTM	ELTP-LFCC	ASV Spoof 2019 [8]	Gives poor performance on imitated samples.

2021	M. Ballesteros et al. [14]	English + French + Spanish +Portuguese + Tagalog	Imitated + Synthesized	Deep4Snet	Histogram, Spectrogram, Time domain waveform	H-Voice [15]	The data transformation procedure had an impact on the model, and it is non-scalable.
2020	Wang et al. [13]	Chinese + English	Synthesized	DeepSonar	Raw neuron, Activated neuron High-dimensional. data visualization of MFCC	FoR [12]	highly influenced by sounds in the actual world
2020	Zhenchun Lei [15]	English	Synthesized	CNN and Siamese CNN	CQCC, LFCC	ASV Spoof 2019 [8]	The models only function well with LFCC and are not resilient to other characteristics.
2020	Chintha et al. [13]	English	Synthesized	CRNN-Spoof and WIRE-Net-Spoof	CQCC and MFCC	ASV Spoof 2019 [8]	The suggested model is complex and includes several layers and convolutional networks.
2019	Alzantot et al. [17]	English	Synthesized	Residual CNN	MFCC, CQCC, STFT	ASV Spoof 2019 [8]	The model cannot be applied to new threats due to its extreme overfitting using synthetic data.

3. Methodology

3.1 Data Creation

The dataset created for this study comprised both real and fake audio. These audio clips are gathered from several publicly available sources, including ted talks, informative videos, and online lectures in Urdu. It is gender and racial unbiased as the dataset contains real audio clips of 10 males and ten females of different age groups. The overall dataset has 400 real and fake audio clips involving 20 subjects. The final dataset created for this research is available at [19]

3.2 Data Pre-processing

Since the obtained audio samples within the dataset consist of different durations, making this dataset effective for fake speech detection is necessary. The gathered audio clips of variable length duration were pre-processed by splitting them into clips of 1-minute duration. The research has utilized length normalization to guarantee that each sample has the same length.

3.3 Generating DeepFakes

The dataset creation process is incomplete without generating fake clips for Deepfake detection. Hence, fake clips are generated from the Deepfake generation process for this purpose. The research has used real-time voice cloning (RTVC) to generate fake audio. It is a tool that uses transfer learning to create voice clones.

3.4 Model Development

The RNN and Long Short-Term Memory (LSTM) network architecture was used to create the deep learning model. LSTM is used to handle long-term dependencies and address the vanishing gradient problem. The model was trained on a dataset including real and fake audio clips.

3.5 Model Evaluation

The model's performance was evaluated utilizing different types of audio clips. The evaluation set contained 200 real audio samples and 200 Deepfakes generated via the RTVC tool. The model was evaluated based on its ability to distinguish audio recordings as real or fake properly.

4. Experiments

The experiments start with loading dataset files. These audio files are pre-processed and converted to waveform. The pre-processing operations are applied to generate the waveform. The waveform is resampled to match the data

hyperparameters. After the audio files are converted into waves, they are passed to compute embedding for a single utterance. These utterances are then sliced into partial. Such that if the utterance covers the large range, it is sliced into partial. These split utterances are then forwarded to the model to generate spectrograms from waves.

To train the model, we randomly selected 160 audio clips from the set of real audios, and the remaining are left for testing. The model consists of three layers. The first layer is based on the LSTM network. It is a variation of Recurrent Neural Networks (RNNs) capable of learning long-term dependencies. It is specifically used in sequence prediction problems. The second layer of the model is based on Linear; the last layer uses the ReLu function for activation.

The hyperparameters used in the model included a mel_window_length of 25ms, mel_window_step of 10ms, and mel_n_channels of 40 for the Mel-filterbank. The audio hyperparameters included a partials_n_frames of 160 (1600ms) and a sampling_rate of 16,000. Voice Activation Detection (VAD) hyperparameters consisted of a vad_window_length of 30ms, vad_moving_average_width of 8, and vad_max_silence_length of 6. The audio volume normalization was set to audio_norm_target_dBFS of -30. The model hyperparameters included model_hidden_size of 256, model_embedding_size of 256, and model_num_layers of 3.

5. Results

To compute the scores, research has converted audio clips to spectrograms. The score of each clip is then compared to the ground truth value of 0.84. If the computed score is 0.84 or above, the model predicts the clip is real, and if the score is less than 0.84, it predicts the clip to be fake

The evaluation is based on the metrics, including precision, recall, and f1 score, presented in Table 2. The model achieved an accuracy of 91%. The confusion matrix represents the performance of a binary classification model, indicating the number of correct and incorrect predictions made by the model. In this case, the model has correctly labeled 230 out of 240 real clips and 172 out of 200 fake clips, while it has made ten incorrect predictions for real clips and 28 incorrect predictions for fake clips, as shown in Figure 1.

The machine learning model was trained on a desktop computer with an Intel Core i5-1155G7 CPU @ 2.50 GHz with four cores and 16 GB of RAM. We used an Intel(R) Iris(R) Xe Graphics with 8 GB of GPU memory to train the model.

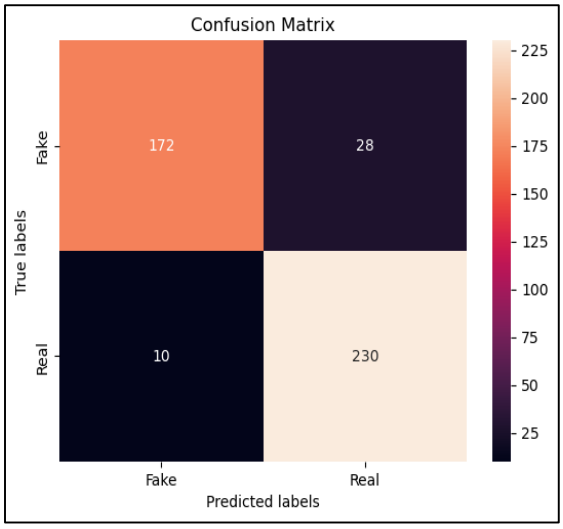

Figure 1: Confusion matrix

Table 2: Results

	Precision Score	Recall Score	F1 Score	Accuracy Score
1	0.8914	0.9583	0.9236	0.9136

6. Conclusion

The research has emphasized the importance of DeepFake detection, which has become the most growing prevalence. Since Urdu is the widely spoken language in South Asia, there is a lack of research for detecting Deepfake in this language. The study has proposed a Deep Learning approach by identifying DeepFake audio using a data set that contains Urdu audio clips. The data set consisting of original audio clips is pre-processed and trained. After which Audio Deepfakes are generated of these audio clips. The model has achieved an accuracy of 91%. To improve the performance and effectiveness of the model, we intend to extend this research work by enhancing the dataset in terms of the number and type of Urdu language audios. Building comprehensive and diverse datasets having multiple language samples is crucial for training and evaluating deep fake detection models.

The deep fake detection domain has many prospective paths for research. Exploring transformers, Autoencoders, attention models, and many other emerging techniques can be exploited to identify the deepfakes. With the advancement in deep learning, detecting audio deepfakes is getting more challenging. Moreover, more research is needed to develop robust detection methods to maintain the trustworthiness of audio content and to safeguard against its potential misuse.

References

[1] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Algorithms, vol. 15, no. 5, 2022, doi: 10.3390/a15050155.

[2] N. Diakopoulos and D. Johnson, "Anticipating and addressing the ethical implications of deepfakes in the context of elections," New Media Soc, vol. 23, no. 7, pp. 2072–2098, 2021, doi: 10.1177/1461444820925811.

[3] D. Rodríguez-Ortega, Y., Ballesteros, D.M. and Renza, "A Machine Learning Model to Detect Fake Voice," in Third International Conference, ICAI 2020, Ota, Nigeria, October 29–31, 2020, Springer International Publishing, 2020, pp. 3–13. doi: 10.1007/978-3-030-61702-8_1.

[4] Y. Ning, S. He, Z. Wu, C. Xing, and L. J. Zhang, "Review of deep learning based speech synthesis," Applied Sciences (Switzerland), vol. 9, no. 19, pp. 1–16, 2019, doi: 10.3390/app9194050.

[5] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," Adv Neural Inf Process Syst, vol. 32, no. NeurIPS, 2019.

[6] S. Ö. Arık and J. Miller, "Deep voice: Real-time neural text-to-speech," in International Conference on machine learning, 2021, pp. 195--204.

[7] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating Replay Attacks Against Voice Assistants," Proc ACM Interact Mob Wearable Ubiquitous Technol, vol. 3, no. 3, pp. 1–26, 2019, doi: 10.1145/3351258.

[8] M. Todisco et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", doi: 10.7488/ds/2555.

[9] F. Iqbal, A. Abbasi, A. R. Javed, Z. Jalil, and J. Al-Karaki, "Deepfake Audio Detection via Feature Engineering and Machine Learning," CEUR Workshop Proc, vol. 3318, 2022.

[10] A. Hamza et al., "Deepfake Audio Detection via MFCC features using Machine Learning," IEEE Access, vol. 10, no. December, pp. 134018–134028, 2022, doi: 10.1109/ACCESS.2022.3231480.

[11] S. Camacho, D. M. Ballesteros, and D. Renza, "Fake Speech Recognition Using Deep Learning," in Communications in Computer and Information Science, Springer, 2021, pp. 38–48. doi: 10.1007/978-3-030-86702-7_4.

[12] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in 2019 10th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019, 2019, pp. 1–10. doi: 10.1109/SPED.2019.8906599.

[13] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, "Semantic Similarity Metrics for Evaluating Source Code Summarization," in IEEE International Conference on Program Comprehension, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

[14] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4SNet: deep learning for fake speech classification," Expert Syst Appl, vol. 184, no. November 2019, p. 115465, 2021, doi: 10.1016/j.eswa.2021.115465.

[15] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection," Proceedings of the Annual Conference of the International Speech

Communication Association, INTERSPEECH, vol. 2020-Octob, pp. 1116–1120, 2020, doi: 10.21437/Interspeech.2020-2723.

[16] A. Chintha et al., "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection," vol. 14, no. 5, pp. 1024–1037, 2020.

[17] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," no. December, 2019.

[18] T. Arif, A. L. I. Javed, M. Alhameed, F. Jeribi, and A. L. I. Tahir, "Voice Spoofing Countermeasure for Logical Access Attacks Detection," IEEE Access, vol. 9, pp. 162857–162868, 2021, doi: 10.1109/ACCESS.2021.3133134.

[19] Marium Mateen, "Urdu VRF (Voice Real-or-Fake) Audio Corpus," 2023.