AI vs. AI: Detecting Deepfake Voices

Md Rabiul Hasan
Computer Science and Engineering
University of Oulu, Finland
MdRabiul.Hasan@student.oulu.fi

Abu Taher
Computer Science and Engineering
University of Oulu, Finland
abutaher.kuet.ece@gmail.com
ataher24@student.oulu.fi

Abstract

The ongoing development of generative artificial intelligence has given the ability to clone real-time voice conversation which also raises concerns about identity misrepresentation and privacy. The misuse of deep fake's voice can impact areas like politics and social media platforms. So, detecting a fake voice in real-time is very important to reduce its effect. This study proposes a deep learning framework to correctly identify between deep fake speech and human speech by analyzing speech emotion features such as rolloff, zero crossing rate, chroma STFT and spectral centroid etc. Here, we have implemented the LSTM, CRNN, and FCNN models for real vs. fake speech classification. The overall accuracy of the LSTM, CRNN, and FCNN models is 98, 99, and 98 percent, respectively. The CRNN model outperformed compared to other models due to its affective capture of local acoustic features and temporal dynamics via CNN and RNN, which makes it more robust for fake speech detection.

1 Introduction

The rapid expansion of voice-based computer-human interfaces has made increasingly precise voice biometrics technologies necessary.Deep learning has made it possible to dramatically improve speaker verification technology's accuracy during the past ten years.The spoofing and voice impersonation capabilities of AI-based voice synthesis systems have advanced significantly at the same time.These superior text-to-speech (TTS) [18] conversion methods can outperform both humans and computerized voice verification systems.It is now essential for systems to identify logical access threats like speech synthesis and voice conversion in order to protect voice-based authentication systems against these malicious attacks.[4]

The ASVspoof1 launched in 2015 with the objective to improve voice spoofing identification defence [17].Identifying popular advanced logical speech processing and voice conversion threats was the focus of the 2015 challenge, which was centred upon unit selection, Hidden Markov models (HMM) and Gaussian mixture models.Since then, voice converter and speech synthesis systems have greatly improved in quality because to deep learning.In 2016, WaveNet [14] was proposed as the first end-to-end voice synthesizer that trained using raw audio, and it showed a mean opinion score (MOS) that was very similar to human speech. VC systems [12] and other TTS systems such as Tacotron [16] and Deep Voice [1] showed similar quality.It became more difficult to identify spoofing attacks as a result of these advancements in TTS and VC technology.

The paper's structure is as follows: Section 1 will give a overview of real and deepfake speech, while Section 2 reviews the existing techniques for categorizing speech produced by artificial intelligence. Section 3 discuss on the methodology, dataset description, and data processing techniques and Section 4 provides a graphic representation of overall performance of our suggested architecture. We concluded the report with a thorough analysis of our findings to close out our discussion.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

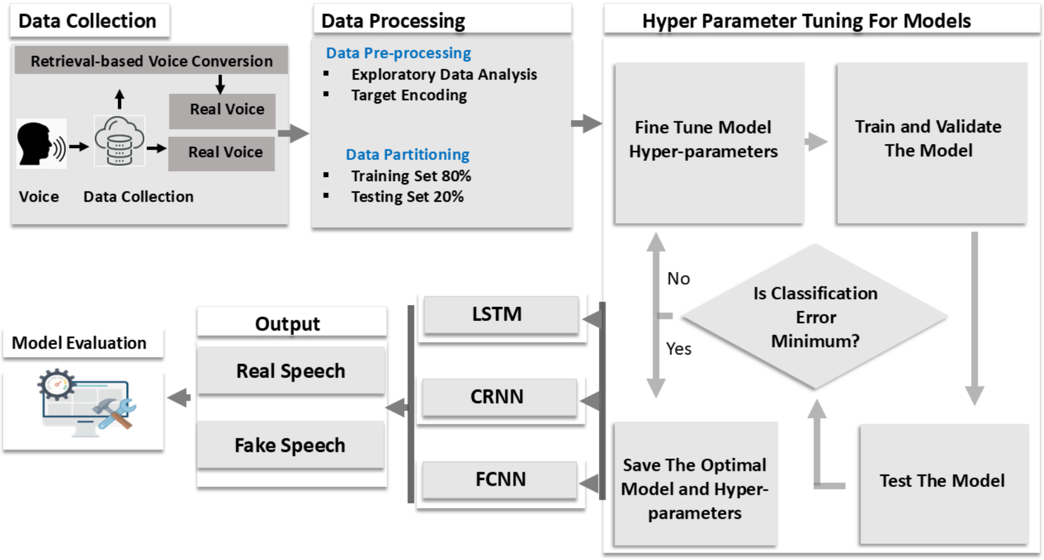

Figure 1: Methodology diagram

2 Literature Review

To detect AI generated speech Galina et al.[10] used ASVspoof 2019 dataset, where they used LFCC, CQT, FFT techniques for feature extraction and an LCNN model for classification with A-SoftMax as the loss function. A similar dataset was used by Guang et al.[9] with raw audio as a feature for the ResNet model. In 2021, Xin Wanget al.[15] proposed ensemble-based LCNN and LCNN-LSTM models for synthetic speech detection. They applied LSB, SPEC, and LFCC as features with MSE for P2SGrad as a loss function. Xinhui Chen et al.[6] presented Variants of the ECAPA-TDNN model with LFCC as feature extraction technique, where they used data compression techniques like MP3, ACC, Landlie, cellular, and VoiP for data augmentation. Later, a similar dataset and data extraction technique was used by Tianxiang Chen et al.[5] for automatic speaker verification utilizing ResNet, MLP, and SWA models with large margin cosine and cross entropy as loss functions. Combining LFCC and MFCC data extraction techniques Joaquın C ́ aceres et al.[3] proposed the TDNN and RawNet2 model for deep fake speech detection. Anton Tomilov et al. [13] proposed MSTFT feature extraction based Resnet18, LCNN, Sinc+CNN models for AI speech detection by combining data augmentation technique like Mixup,FIR filters. In 2024, Yujie Yang et al.[19] used both ASVspoof 2019 and ASVspoof 2021 dataset for deep fake audio detection by combining XLS-R, Hubert, WavLM as encoders and ResNet as decoder.

3 Proposed Method

The fundamental objective of this project is to design a framework to correctly identify real voices and AI-generated voices by utilizing LSTM, CRNN and FCNN models. Firstly, the real voices have been collected, and then, using retrieval-based voice conversion, fake speech has been generated. In the data processing stage, we have performed exploratory data analysis and target encoding. Then we have tuned the hyperparameters of the LSTM, CRNN, and FCNN models shown in Table 2. Finally, we have evaluated our models based on performance.

3.1 Dataset Description

Real audio recordings from eight famous people are included in the DEEP-VOICE dataset. [2] as well as versions of these samples that have been Retrieval-based Voice Conversion (RVC)-converted into each other's voices. Any background noise was eliminated before conversion, and then the original background noises were re-incorporated into the DeepFake versions.There are two formats

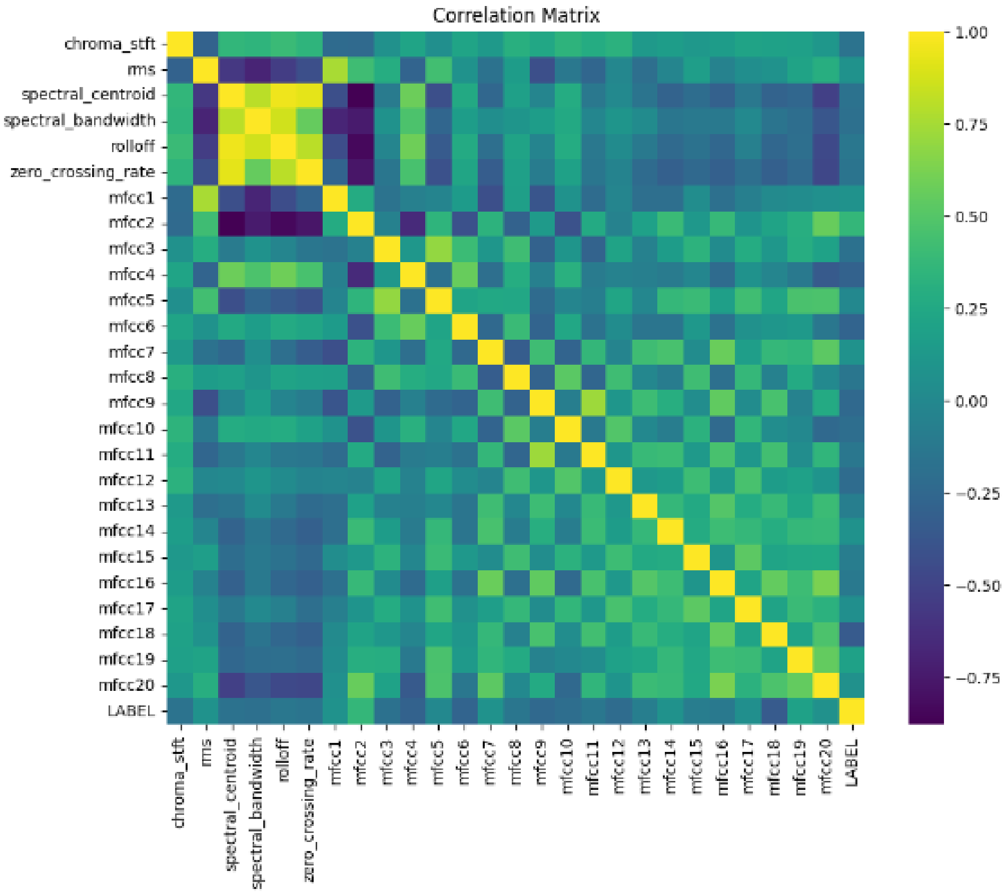

Figure 2: Correlation matrix of the features

for the dataset. Raw audio files in the first format, "AUDIO," are separated into "REAL" and "FAKE" categories. Both the original speaker and the target voice are indicated by the file names; for example, "Obama-to-Biden" indicates that the speech of Barack Obama was modified to sound like Joe Biden. The analysis in the study below was based on the second format, "DATASET-balanced.csv," which comprises audio attributes that have been extracted from one-second segments and balanced using random sampling.

3.2 Data Processing

3.2.1 Exploratory Data Analysis

We have generated a heat map of correlation matrix to facilitate the selection of relevant features for deep learning models. To discover emotional characteristics in the dataset, we have conducted a t-test on each feature. Features that showed statistically notable variations (p-value < 0.05) were found, offering insight into the possible expressions of emotional changes in audio recordings. Box plots were used to better depict these important aspects, giving us an understanding of the distributions and variations in emotional cues between actual and synthetic speech.

3.3 Model Description

3.3.1 LSTM

First developed by Graves and Schmidhuber[7] , the popular LSTM configurationalso referred to as the "vanilla LSTM"—is used as a benchmark for evaluating other variations. Using full-gradient training throughout, this version combines the original LSTM model with improvements made by

Table 1: Significant features related to speech emotion

Feature	P value
chroma stft	1.255482e-67
rms	6.812161e-15
spectral centroid	5.306269e-74
spectral bandwidth	3.118181e-96
rolloff	2.509645e-49
zero crossing rate	3.000349e-65
mfcc1	6.695210e-16
mfcc2	0.000000e+00

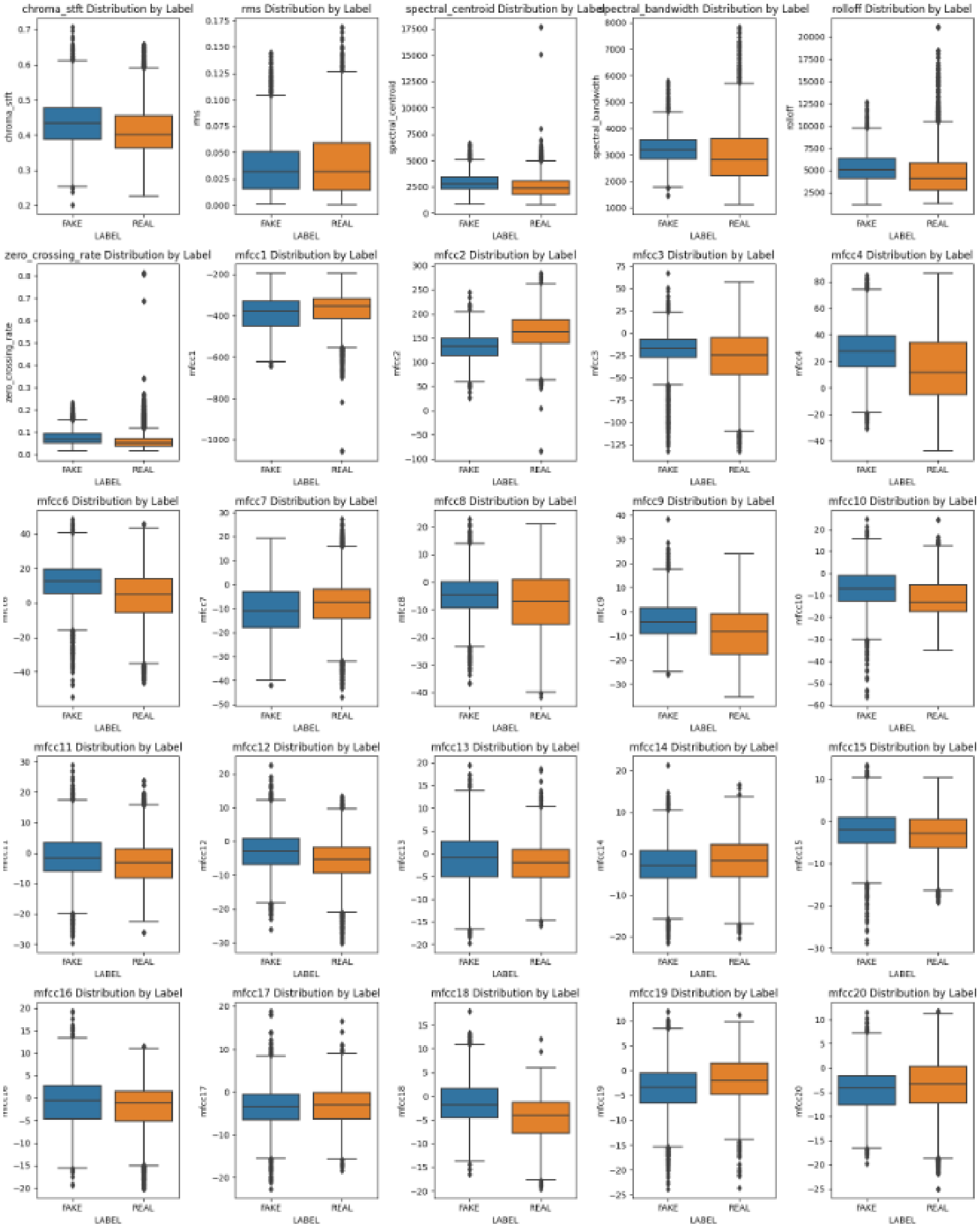

Figure 3: Box plot of features related to speech emotion

Gers[8] and Schmidhuber and Gers et al. . The vanilla LSTM design consists of a input block, a memory cell , a final activation function, peephole connections, and fundamental three gates: input,output and forget. The output block is also continually fed back into the input block and every gates.

3.3.2 CRNN

This study's CRNN model integrates recurrent and convolutional layers for effective sequence modeling. When 1D convolutional layers first extract local patterns, it employs max-pooling and dropout to reduce dimensionality and prevent overfitting. To recognize temporal dependencies in the data, an LSTM layer comes after these layers. The output of fully connected layers is subsequently flattened by the model, which also includes ReLU activation for further regularization. The sigmoid activation function in the last dense layer is used for binary categorization. Adam optimizer has been used for model optimization and a loss function of binary cross-entropy for calculating the loss during training.

3.3.3 FCNN

FCNN model , also known as a multilayer perceptron (MLP), where in single layer all neuron is linked all other neurons in the layer underneath it.Due to its fully connected structure, FCNNs may perform complex data transformations employing multiple layers of nonlinear operations. Typical FCNN component consists of an input layer, single or multiple hidden layers, and finally an output layer. Weighted sums and activation functions are applied by neurons in each layer to enhance the learned representations [11].

Table 2: Hyperparameters of LSTM, CRNN, and FCNN models.

Hyperparameter	LSTM	CRNN	FCNN
Rate of Learning	0.0001	0.001	0.001
Batch Size	32	32	32
Number of Layers	2	4	3
Rate of Droupout	0.3	0.3	0.5
Optimizer	Adam	Adam	Adam
Epochs	100	50	100
Activation Function	ReLU	ReLU	ReLU

4 Results Analysis and Discussions

The efficacy of deep learning models for real vs. fake speech classification can be evaluated from their evaluation metrics. For better understanding, the overall evaluation metrics has been structured in three phases. Phase 1 demonstrates the single model outputs; phase 2 represents the performance comparison of models; and phase 3 outlines the limitations and future scope for this study.

4.1 Phase 1: Models Performance

4.1.1 LSTM

In training phases of LSTM, we have used a batch size of 32 with 0.0001 learning rate. The model has been trained on a number of epochs but we have got a better accuracy of 98% with the epoch number of 100, where each epoch consists of 295 steps. 80% of data used for model training and 20% for testing. The LSTM model has shown better performance in predicting the true fake speech of 1176, where true real speech was 1143. The score of falsely positive and falsely negative rate was 20 and 17, respectively shown in fig.1 .

4.1.2 CRNN

We have used the same batch size and learning rate for CRNN model training, where the epoch number was 50. We applied Adam optimizer for the model compilation and binary cross entropy for

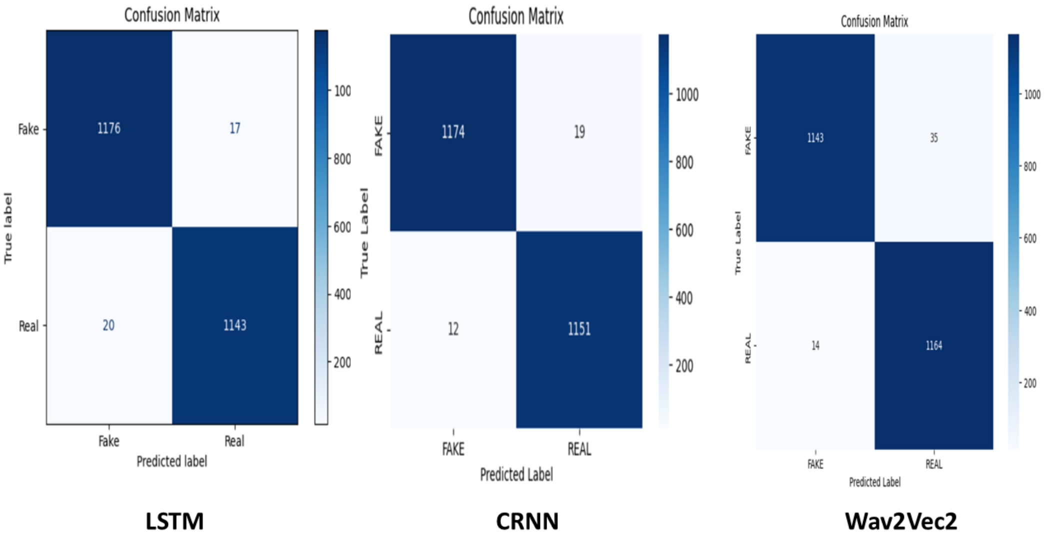

Figure 4: Models confusion metrics

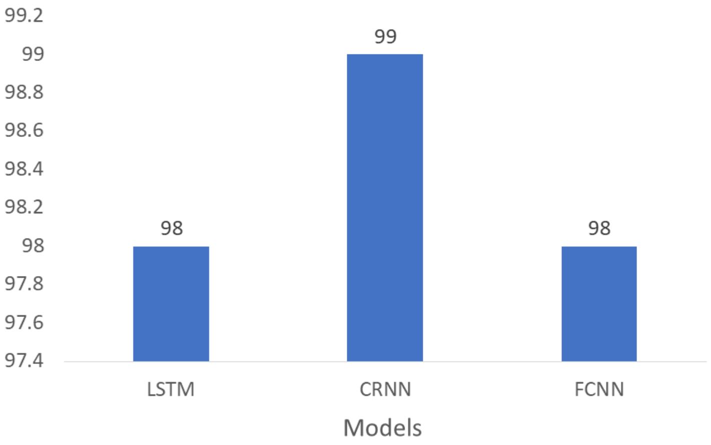

Figure 5: Models accuracy comparison

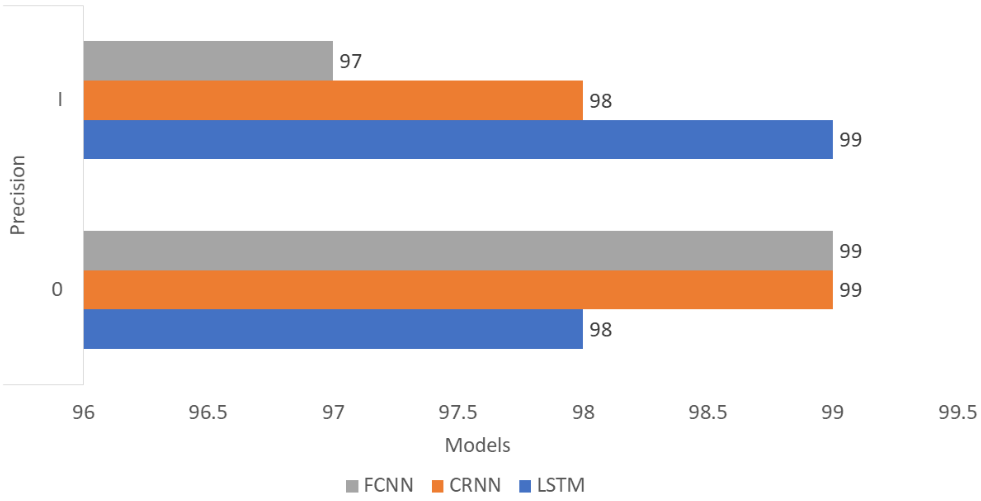

Figure 6: Models precision comparison

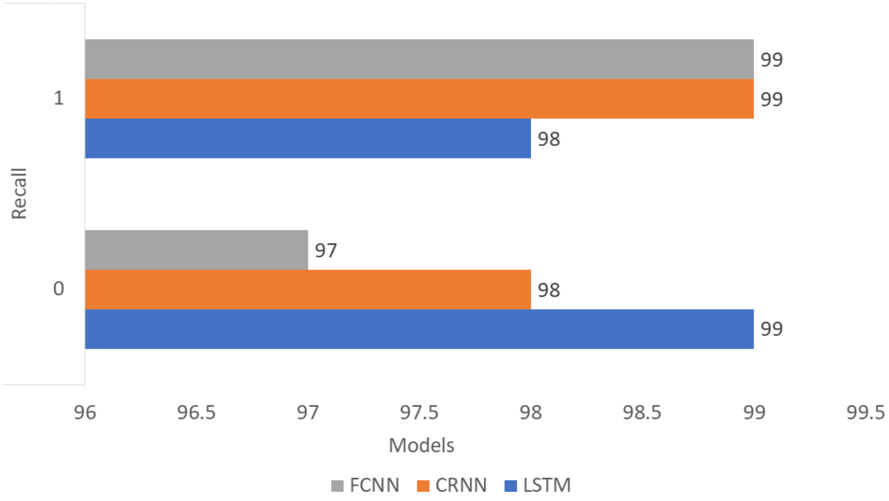

Figure 7: Models recall comparison

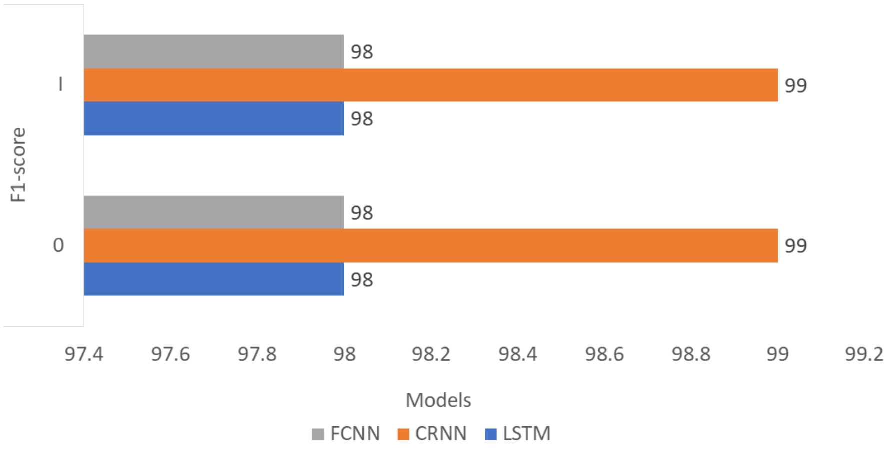

Figure 8: Models f1-score comparison

measuring the loss during training. The overall accuracy we have got is 99%, where other metrics like precision score were 99% and 98% for fake and real speech classes. The model has correctly identified the true fake and true real speech of 1174 and 1151, where the false positive and false negative were 12 and 19 shown in fig. 1.

4.1.3 FCNN

The Fully Connected Neural Network has been trained on 100 epochs with 0.001 learning rate , where the Adam optimizer has been used for gradient-based optimization to minimize the loss. The overall accuracy of the model is 98% with true fake and true real speech detection scores of 1143 and 1164. The model has misclassified the real speech as fake speech with a score of 14.

4.2 Phase 2: Model's Performance Comparison

Table 3: Model's Performance Comparison

Model	Accuracy (%)	Precision (%)		Recall (%)		F1-score (%)	
		0	1	0	1	0	1
LSTM	98	98	99	99	98	98	98
CRNN	99	99	98	98	99	99	99
FCNN	98	99	97	97	99	98	98

Here, in the table Fake speech=0, Real speech=1

Table 2 describes the overall performance of LSTM, CRNN and FCNN models. The maximum accuracy of 99% we have obtained from the CRNN model compared with the LSTM and FCNN models is 98% and 98% accuracy. The CRNN model combines both CNN and RNN and has the benefits of capturing local acoustic features through CNN and temporal dynamics through RNN, which is very efficient for speech emotion recognition and fake speech detection.

5 Conclusion

In this project we have developed a framework for correctly classifying AI-generated speech and real speech. Our approach leverages exploratory data analysis techniques with speech emotion analysis. For speech emotion analysis, throughout the extracted feature we have performed t-test and feature values less than 0.05 have been considered the most important features related to speech emotion and distinguishing between real and fake speech. We have applied three different deep learning model architectures—LSTM, CRNN, and FCNN—for real vs. fake speech classification. After hyperparameter tuning and model optimization, the CRNN model has given the highest performance of accuracy of 99%, where precision values were 99% and 98% for real and fake speech. As CRNN combines both local acoustic features and temporal features through CNN and RNN, that's why it overperformed compared to other models.

6 Contribution Distribution

Md Rabiul Hasan System architecture design, data processing and feature extraction, FCNN model training and tuning, as well as report drafting and finalization.

Abu Taher Speech emotion analysis, LSTM and CRNN model training and tuning, and contributed to report drafting and finalization.

References

[1] Sercan Ö Arık et al. "Deep voice: Real-time neural text-to-speech". In: *International conference on machine learning*. PMLR. 2017, pp. 195–204.

[2] Jordan J Bird and Ahmad Lotfi. "Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion". In: *arXiv preprint arXiv:2308.12734* (2023).

[3] Joaquın Cáceres et al. "The Biometric Vox system for the ASVspoof 2021 challenge". In: *Proc. ASVspoof2021 Workshop*. 2021.

[4] Tianxiang Chen et al. "Generalization of Audio Deepfake Detection." In: *Odyssey*. 2020, pp. 132–137.

[5] Tianxiang Chen et al. "Pindrop labs' submission to the ASVspoof 2021 challenge". In: *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (2021), pp. 89–93.

[6] Xinhui Chen et al. "UR channel-robust synthetic speech detection system for ASVspoof 2021". In: *arXiv preprint arXiv:2107.12018* (2021).

[7] Alex Graves and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural networks* 18.5-6 (2005), pp. 602–610.

[8] S Hochreiter. "Long Short-term Memory". In: *Neural Computation MIT-Press* (1997).

[9] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. "Towards end-to-end synthetic speech detection". In: *IEEE Signal Processing Letters* 28 (2021), pp. 1265–1269.

[10] Galina Lavrentyeva et al. "STC antispoofing systems for the ASVspoof2019 challenge". In: *arXiv preprint arXiv:1904.05576* (2019).

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[12] Jaime Lorenzo-Trueba et al. "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods". In: *arXiv preprint arXiv:1804.04262* (2018).

[13] Anton Tomilov et al. "STC antispoofing systems for the ASVspoof2021 challenge". In: *Proc. ASVspoof 2021 Workshop*. 2021, pp. 61–67.

[14] Aaron Van Den Oord et al. "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499* 12 (2016).

[15] Xin Wang and Junich Yamagishi. "A comparative study on recent neural spoofing countermeasures for synthetic speech detection". In: *arXiv preprint arXiv:2103.11326* (2021).

[16] Yuxuan Wang et al. "Tacotron: Towards end-to-end speech synthesis". In: *arXiv preprint arXiv:1703.10135* (2017).

[17] Shitao Weng et al. "The sysu system for the interspeech 2015 automatic speaker verification spoofing and countermeasures challenge". In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 152–155.

[18] Ziwei Yan, Yanjie Zhao, and Haoyu Wang. "VoiceWukong: Benchmarking Deepfake Voice Detection". In: *arXiv preprint arXiv:2409.06348* (2024).

[19] Yujie Yang et al. “A robust audio deepfake detection system via multi-view feature”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 13131–13135.