TRUETONE

TRUETONE

REAL OR FAKE SPEECH

BY LAMPS

2025

Introduction

“Audio deepfakes created by modern Text-To-Speech(TTS) and voice-conversion models can mimic human voices so convincingly that they enable fraud and disinformation. Our project presents a compact deep learning system trained on both real and synthetic speech to detect the subtle spectral and temporal artifacts in fake audio, providing a fast, reliable tool for voice authentication and digital forensics.”

Problem Statement

TRUETONE

02

SECURITY & TRUST RISKS

Malicious actors can weaponize fake audio for disinformation campaigns, identity fraud, voice phishing (vishing), and reputational harm.

04

Human Detection Limits

Audio deepfakes enable disinformation campaigns, financial fraud, and reputation attacks by exploiting the trust placed in spoken words.

01

RISE OF AUDIO DEEPFAKES

Advances in text-to-speech (TTS) and voice conversion models can produce highly realistic synthetic speech that is often indistinguishable from genuine human voices.

03

DETECTION CHALLENGE

Subtle spectral & temporal artifacts in synthesized speech require specialized deep-learning methods to reliably distinguish real vs. fake audio.

TRUETONE

Literature Survey

IEEE 2024 [2]: Built an ensemble of CNN, CRNN, RNN, ResNet, EfficientNet, and Whisper-embedding models on ASVspoof 2019-LA, yielding EER = 0.03 and AUC = 0.994 .

Elsevier 2023 [4]: Compared VGG-16 and a custom CNN across MFCC, mel-spectrum, chromagram, and spectrogram features; the custom CNN consistently outper--formed VGG-16

IEEE 2023 [5]: Tested CNN, BiLSTM, TCN, and STN architectures on ASVspoof 2021, ADD 2023, and In-the-Wild datasets, achieving up to 94.33% accuracy

IEEE 2022 [7]: Applied a Multi-Feature Audio Authenticity Network on In-the-Wild and Fake-or-Real datasets, reaching 98.93% (In-the-Wild) and 94.47% accuracy

IEEE 2025 [1]: Employed a hybrid Wav2Vec2 + AASIST model on the SingFake singing-voice dataset, achieving an equal-error-rate (EER) of 8.23% .

- Elsevier 2024 [3]: Evaluated various DL classifiers under adversarial attack on ASVspoof 2019-LA; focused on robustness, with significant performance degrada -tion under attack

IEEE 2023 [6]: Used an MFCC-based CNN–LSTM model on ASVspoof 2019, reporting 88% accuracy

Elsevier 2022 [8] : Evaluated CNNs and RNNs on ASVspoof 2019, achieving 96% accuracy

TRUETONE

METHODOLOGY

PREPROCESSING

- File Type Validation and Analysis
- Standardization
- Resampling
- Duration Normalization
- RMS Volume Normalization
- Mono Conversion

01

FEATURE EXTRACTION

- 64-band log-Mel Spectrogram
- Log scaling (dB)

02

MODEL EXPLORATION

- DNN
- CNN
- BiLSTM
- Hybrid CNN + BiLSTM

03

EVALUATION & TUNING

- Accuracy, AUC
- Loss
- Dropout, hyperparams

04

DEPLOYMENT & FUTURE WORK

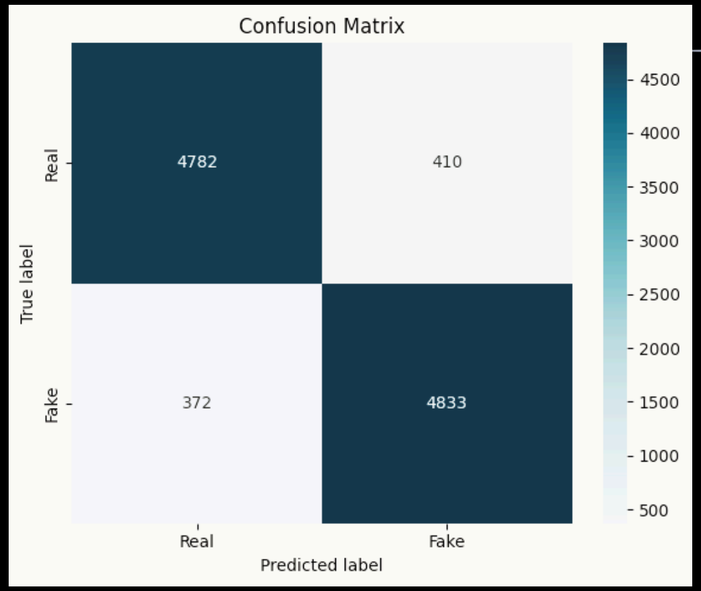
- Real-time Detection
- Edge devices
- Multilingual

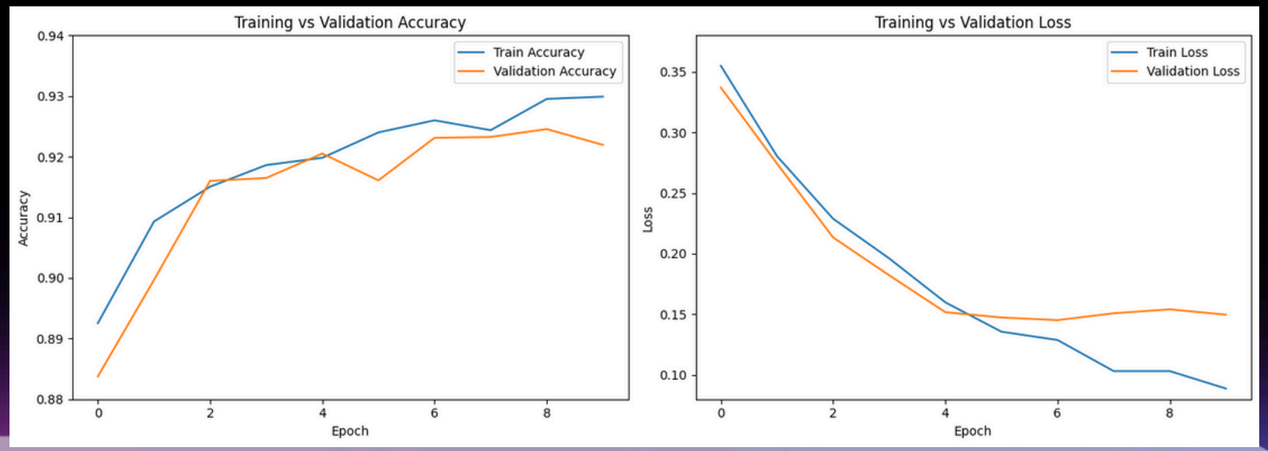
05

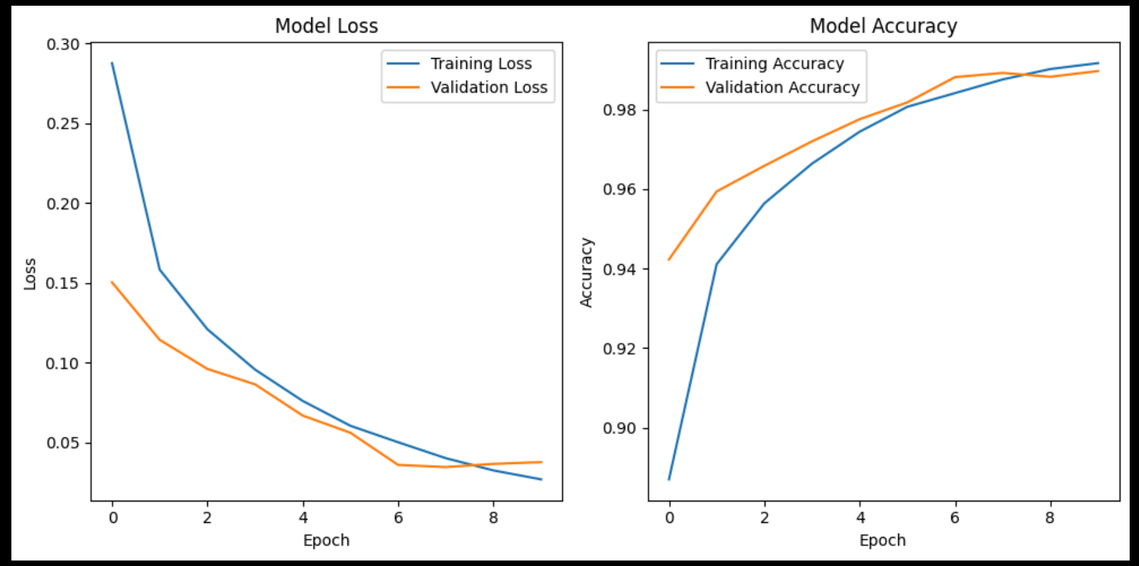
TRUETONE

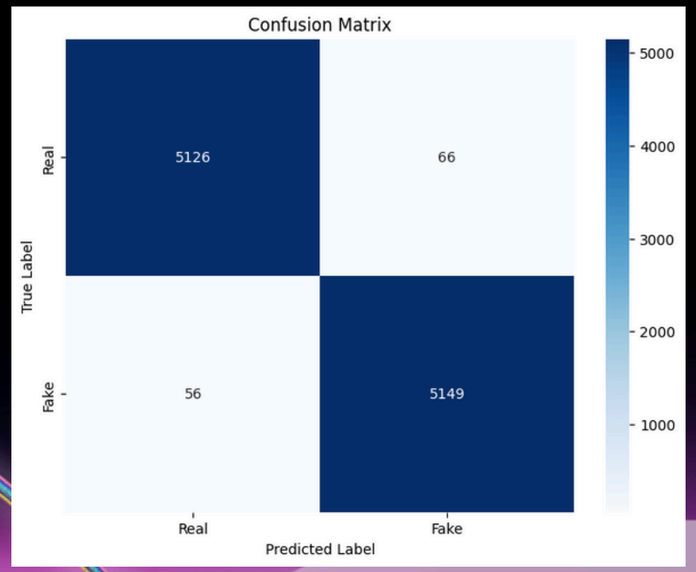
DNN

The DNN begins by flattening the input data to convert it into a one-dimensional vector. It then passes through two dense (fully connected) layers with 128 and 64 neurons respectively, each activated by the ReLU function and followed by dropout layers to prevent overfitting. The final layer uses a sigmoid activation to output probabilities for binary classification. The model is optimized using the Adam optimizer with binary cross-entropy as the loss function, and accuracy as the evaluation metric.









TRUETONE

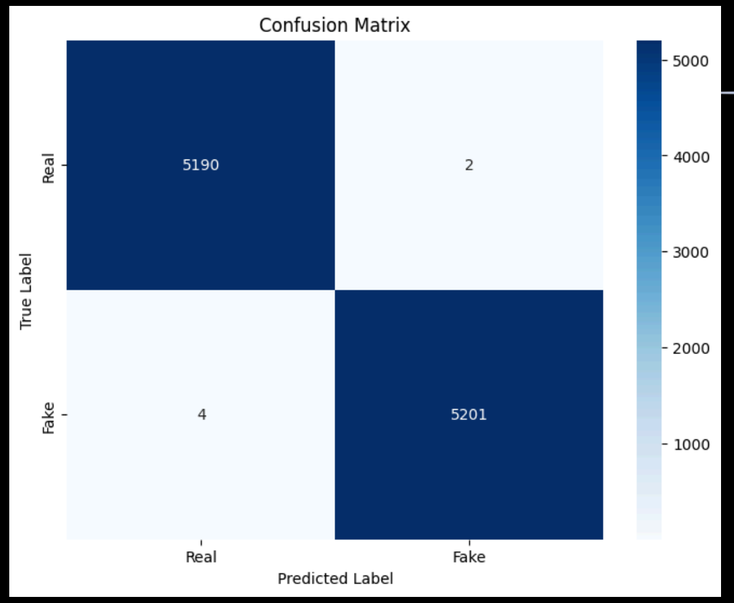
BiLSTM

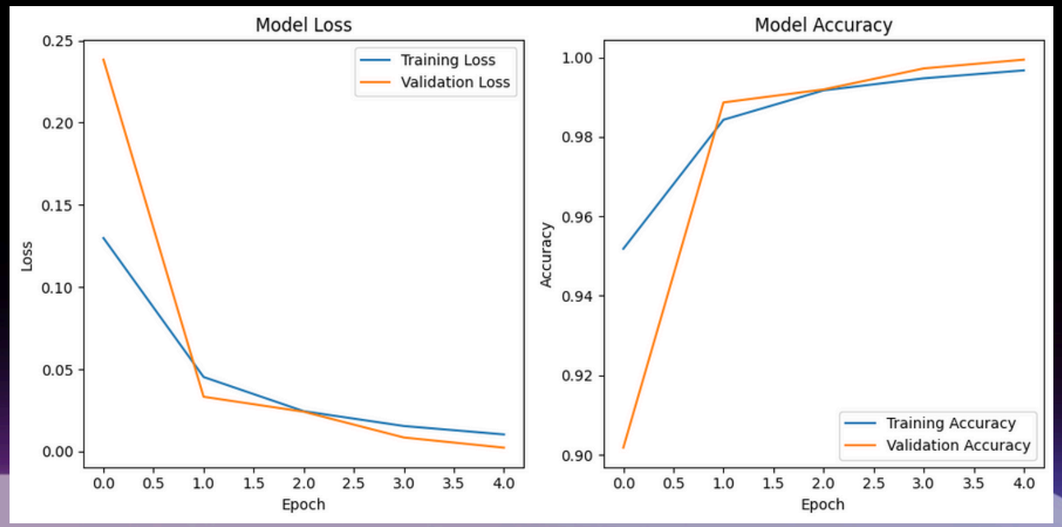
A stacked BiLSTM architecture is employed to effectively learn temporal patterns from the audio features. The first BiLSTM layer (with 128 units) returns full sequences to retain temporal depth, followed by a second BiLSTM layer (64 units) for deeper sequence abstraction. Subsequent dense layers with ReLU activations and progressive dropout help in refining learned representations and reducing overfitting. The final output layer uses a sigmoid activation function to perform binary classification.

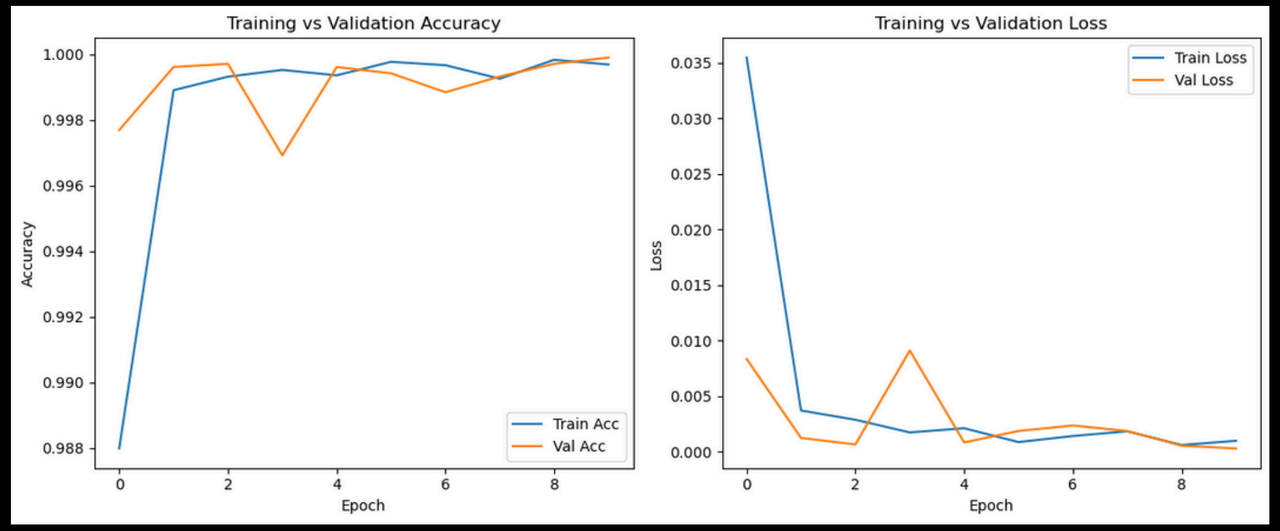
TRUETONE

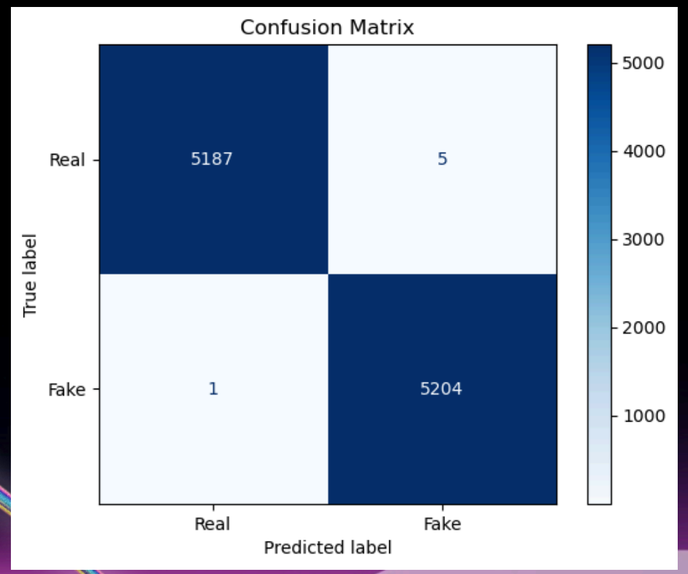
CNN

The CNN architecture has been designed to classify audio inputs by progressively extracting and refining features through three convolutional blocks with increasing filter depths (32, 64, and 128). Each block is followed by batch normalization, max pooling, and dropout (0.5) to enhance training stability and reduce overfitting. The extracted features are then passed through fully connected layers, with additional regularization, culminating in a sigmoid-activated output neuron for binary classification.









HYBRID

CNN and BiLSTM

A stacked BiLSTM architecture is employed to effectively learn temporal patterns from the audio features. The first BiLSTM layer (with 128 units) returns full sequences to retain temporal depth, followed by a second BiLSTM layer (64 units) for deeper sequence abstraction. Subsequent dense layers with ReLU activations and progressive dropout help in refining learned representations and reducing overfitting. The final output layer uses a sigmoid activation function to perform binary classification.

TRUETONE

Performance Benchmarks

Model	Test Accuracy	Test AUC	Test Loss
DNN	92.3%	0.955	0.148
BiLSTM	98.83%	0.998	0.0387
CNN	99.80%	0.9984	0.0121
Hybrid	99.94%	0.9997	0.0021

OUR HYBRID CNN + BILSTM MODEL OUTPERFORMS THE OTHERS BECAUSE IT UNITES THE STRENGTHS OF CONVOLUTIONAL FEATURE LEARNING (FOR FINE-GRAINED SPECTRAL CUES) AND RECURRENT SEQUENCE MODELING (FOR TEMPORAL DYNAMICS). THIS DUAL CAPACITY ALLOWS IT TO DETECT THE MOST SUBTLE INCONSISTENCIES IN FAKE SPEECH, RESULTING IN SUPERIOR ACCURACY (99.94%), AUC (0.9997), AND MINIMAL LOSS (0.0021).

TRUETONE

Future Prospects

Enterprise-Grade

- **Scalable**: Handles high-volume audio streams
- **Robust**: Withstands adversarial audio manipulations
- **User-Friendly**: Intuitive interface for all users

Real-Time & Edge-Optimized

- **Streamline pipelines:** for live audio monitoring with minimal latency
- **Tailor lightweight models:** for on-device inference and low-power hardware

Integrated & Adaptive

- **Partner with media platforms & SDK providers:** for seamless deployment
- **Leverage transfer learning and continual learning:** to stay ahead of novel deep-fake methods

Conclusion

The rapid evolution of AI-generated content has opened up new vistas in speech synthesis, but it has also introduced some extremely serious problems–chief among them the threat of audio deep fakes. This project aimed to counter that threat by creating a deep learning-based system that can detect synthetic speech with reliability. By trying out different architectures of neural networks from CNNs to BiLSTMs, and even a CNN-BiLSTM hybrid model we were able to achieve astaggering accuracy of 99.94%.

TRUETONE

References

[1]AUDIO FEATURES INVESTIGATION FOR SINGING VOICE DEEPFAKE DETECTION
[2]DEEPFAKE AUDIO DETECTION USING SPECTROGRAM-BASED FEATURE AND ENSEMBLE OF DEEP LEARNING MODELS
[3]AUDIO-DEEPFAKE DETECTION: ADVERSARIAL ATTACKS AND COUNTERMEASURES
[4]THE EFFECT OF DEEP LEARNING METHODS ON DEEPFAKE AUDIO DETECTION FOR DIGITAL INVESTIGATION
[5]AUDIO DEEPFAKE APPROACHES
[6]UNMASKING THE TRUTH: A DEEP LEARNING APPROACH TO DETECTING DEEPFAKE AUDIO THROUGH MFCC FEATURES
[7]DEEPFAKE AUDIO DETECTION VIA MFCC FEATURES USING MACHINE LEARNING
[8] MOHAMMED ABDELDAYEM. THE FAKE-OR-REAL DATASET. KAGGLE. 2023.
[9] KOMINEK, J., & BLACK, A. W. (2004). THE CMU ARCTIC SPEECH DATABASES.
[10] ITO, K. (2017). THE LJ SPEECH DATASET.
[11] VOXFORGE. (2006). FREE SPEECH CORPUS.

TRUETONE

THANK YOU!