Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

# Audio Shield-Deep Fake Audio Detection

Ammanamanchi Sravya Sree
Computer Science and Engineering
Geethanjali College of Engineering and Technology
Hyderabad, Telangana

Baddam Vaishali Reddy
Computer Science and Engineering
Geethanjali college of Engineering and Technology
Hyderabad, Telangana

Teak Shashank
Computer Science and Engineering
Geethanjali College of Engineering and Technology
Hyderabad, Telangana

K. Durga Kalyani
Assistant Professor, Computer Science and Engineering
Geethanjali college of Engineering and Technology
Hyderabad, Telangana

***Abstract*** **- Advancements in artificial intelligence and deep learning have enabled highly realistic speech synthesis and voice cloning systems, creating both beneficial applications and serious security risks such as impersonation, fraud, and misinformation. This paper presents *AudioShield*, a deep learning-based framework for detecting synthetic and manipulated audio recordings. The proposed system incorporates audio preprocessing, MFCC and spectrogram-based feature extraction, and a hybrid CNN–BiLSTM architecture to capture spatial spectral artifacts and temporal inconsistencies in speech signals. The CNN component learns discriminative spectral patterns, while the BiLSTM models sequential dependencies to improve detection robustness. Additionally, the framework provides confidence scores and visual explanations to enhance interpretability and user trust. Experimental results demonstrate improved accuracy and generalization compared to standalone deep learning models. The proposed approach offers a scalable and reliable solution for cybersecurity, digital forensics, media authentication, and secure digital communication systems.**

***Index Terms*** **- Deepfake Audio Detection, Speech Forensics,Convolutional Neural Networks, Bidirectional LSTM, Audio Signal Processing, Artificial Intelligence, Cybersecurity, Media Authentication**

## I. INTRODUCTION

Artificial intelligence is reshaping speech generation technologies at an incredible rate. Neural voice cloning, deep generative audio synthesis, and contemporary text-to-speech systems are some of the ways the revolution is happening. Major fresh deep learning model developments like generative adversarial networks(GANs), neural vocoders, and speech models based on transformer are making it possible for machines to create speech records that are nearly human in terms of nature and expressiveness while at the same time imitating human vocal features such as tone pitch accent, rhythm, and emotion. As a result, the technology has been put to various good uses: virtual assistants, speech impairment accessibility solutions, automated customer support systems, and personalized digital communication platforms are just a few ways to mention the positive impact of these tools. At the same time however the identical technologies pose serious security, ethical, and social risks.

Deepfake audio is the speech that has been artificially generated or tampered with to closely resemble and mimic the voice of a real person. Unscrupulous individuals may make use of such synthetic audio to conduct various criminal activities such as identity theft, financial fraud, misinformation campaigns, social engineering, and the creation of fake digital evidence. The growing availability of open-source voice cloning software and speech data that is publicly accessible has drastically reduced the level of technical expertise one needs to produce highly convincing deepfake audios, successively leading to raised cybersecurity risks. Voice biometric authentication systems and platforms for digital communication, for instance, are among the most susceptible to being tricked or spoofed by one that is the product of a deep neural network capable of prose synthesis.

Conventional audio authentication and spoof detection technologies mainly depend on manually-designed acoustic features and statistical signal-processing methods, like pitch analysis, spectral distribution measurements, and energy-based thresholds. Although helpful for the detection of the earlier forms of audio manipulation, these methods have a hard time identifying AI-generated speech, which is with very few perceptual artifacts and can closely resemble genuine human recordings. Therefore, the demand for the detection systems of high-level and complex representations automatically learning from the speech data by the ”intelligent” systems capabilities is on the rise. Deep learning methods have the potential to address this issue as they can model the spectral features and temporal dependencies in the audio signals for the purpose of detecting very small differences that may have been introduced by the speech synthesis process.

To tackle the downsides, the present study introduces AudioShield, a hybrid deepfake audio detection framework combining the strengths of both Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. The framework is designed to improve



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

reliability, accuracy, and interpretability in secure digital communication scenarios. CNN part extracts spatial features from spectrograms to spot spectral artifacts of synthetic audios; at the same time, the BiLSTM network captures the bidirectional temporal dependencies in speech, which help to detect the speech dynamics and sequential inconsistencies. In other words, integrating spectral and temporal learning mechanisms in a single architecture, AudioShield is able to deliver high detection capability alongside transparency via the interpretable results, thereby it can be a candidate for cybersecurity, digital forensics, and media authentication facilities that one can trust

## II. LITERATURE REVIEW

In recent years, the study of deepfake audio detection has progressively concentrated on machine learning and deep learning techniques to distinguish artificially produced or doctored speech signals. Initially, the detection techniques were based on hand crafted acoustic features like Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), spectral centroid,and zero-crossing rate. These features were then modeled with traditional machine learning classifiers such as Support Vector Machines (SVM) and Decision Trees. Although such methods yielded some degree of success in the beginning, they faltered when confronted with more sophisticated speech generation systems which could create highly realistic audio with very few perceptual differences.

Thanks to the progress in deep learning, using spectrogram-based representations is one of the most common ways to do deepfake detection. Spectralgrams are a form of visual representation that shows time and frequency of the audio signal, which is why neural nets can decipher complex spectral patterns typical for synthetic speech production. CNNs have exhibited great ability to extract spatial frequency features by detecting abnormal harmonics, frequency smoothing effects, and spectral regularities introduced by neural vocoders and voice cloning systems. Such models learn hierarchical feature representations automatically, so they do not depend on human-engineered features.

However, speech signals, by their very nature, have sequential as well as temporal dependencies that are not entirely captured even with convolutional operations. To get over this problem, recurrent neural networks, especially Long Short-Term Memory (LSTM) networks, have been used to represent the temporal aspects of the speech phenomena like phoneme transitions, rhythm changes, and energy variations. Thanks to their memory gating mechanisms, LSTM networks are capable of preserving contextual information which results in an enhanced ability to detect inconsistencies that are generally present in tampered auditory recordings. Therefore, CNNRNN combination models which are able to extract spectral features and at the same time capture the temporal sequences have become a powerful and widely adopted technology.

Nevertheless, major advance notwithstanding, there are still a number of challenges hanging on in the deepfake audio detection systems that already exist. A lot of the existing models do not generalize well when they are tested on unseen datasets or new spoofing methods because they get overfitted to the specific training distributions, among other things. Besides, deep learning models are often computationally intensive and require long training periods not allowing ones in real-time settings that have limited resources. Furthermore, the lack of interpretability is a critical problem as many systems only provide binary classification outputs without giving users meaningful explanations or visualization support.

The AudioShield system that we're proposing will solve these problem in a modular and interpretable detection pipeline. Our feature extraction mechanism combines MFCC and spectrogram-based which together with a hybrid CNN-BiLSTM architecture enables the system to detect both spectral artifacts and temporal irregularities that are typical of fake speech. Besides, AudioShield produces explainable outputs, such as confidence scores and visual analysis, which raise the level of transparency, make the system easier to use and build users' trust in the deepfake detection results in cybersecurity, digital forensics, and media verification scenarios

## III. EXISTING SYSTEM

Mostly, current deepfake audio detection systems are based on handcrafted acoustic features merged with the traditional machine learning classifiers like Support Vector Machines (SVM), Random Forests, and Gaussian Mixture Models. These methods rely on the use of the human-engineered descriptors, such as MFCC, pitch-related features, spectral contrast, and energy based parameters, to discriminate between genuine and synthetic speech signals. Even though these techniques can perform quite well in controlled conditions, their success heavily relies on the quality of feature design and domain-specific expertise making them less versatile to the changing speech synthesis technologies.

Recent deep learning techniques have brought about the usage of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for feature learning automation and temporal modeling. CNN-based methods are very good at capturing spatial patterns in spectrogram representations which help in detecting very subtle spectral artifacts that are typical of audio generation. Similarly, LSTM models look at the temporal dependencies in speech sequences which help in spotting unnatural rhythm, prosody and transition inconsistencies. Hybrid CNNLSTM architectures integrate spectral and temporal analysis for even more accurate detection.

Even with these upgrades, deep learning methods currently applied still encounter a number of practical hurdles. A lot of models are only able to generalize to a limited extent when they meet new datasets or new types of voice cloning that have not been seen before. Besides, very deep networks demand so much in computational resources that they become very costly not only in training but also in inference, thereby making real-time deployment a problem. The absence of interpretability is yet another big fault. Most of them just give a classification result without providing users with a way to see or make sense of what detection decisions are being made.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Such limitations emphasize the requirement for a scalable, efficient, and interpretable deepfake audio detection frame- work. Inspired by these issues, the system we propose is designed to combine powerful feature extraction, hybrid deep learning architectures, and explainable result generation in order to enhance the reliability, transparency, and practical usability aspects of cybersecurity and digital forensic appli- cations in the real world.

## IV. PROPOSED SYSTEM

The planned AudioShield system is a complete end- to- end intelligent deepfake audio detection platform that inte- grates user interaction, automated audio processing, hybrid deep learning analysis, and explainable result visualization. While traditional detection systems only focus on backend classification, the proposed framework offers a full workflow starting from secure user access to forensic-level interpretation of detection results. The system architecture runs via several interconnected modules, each corresponding to the functional outputs depicted in the system interface screens. The entire workflow guarantees usability security scalability, and trust- worthy deepfake detection performance.

### *A. User Authentication and System Access*

The doorstep of AudioShield platform is secured by user authentication through modules i.e. registration and login. Users landing on the home page can access the system features such as project overview, account creation, login, and audio analysis. At the point of registration, users are asked for credentials which, are not only confirmed but also securely archived using authentication protocols. Similarly, the module of login checks the user authenticity and only then allows access to the system's resources thus, not only preventing illicit usage but also handling securely uploaded data of the audio. Following authentication, the users are taken to their respective personalized dashboards.

### *B. Dashboard and Analysis Management*

The dashboard acts as the main control panel for users to monitor the system in real-time or review the summary report of activities performed by the system. Users get to see total number of processed audio files, history of detection results, confidence scores of predictions, along with visual summaries of acoustic features including pitch contours and energy changes, all of which can be browsed along with the corresponding timestamps of analyses. This single location makes it possible for users to follow the trail of past predic- tions easily and do side-by-side comparison of various audio analyses in the most user-friendly way, which can greatly help in case of forensic investigation and research evaluation.

### *C. Audio Upload and Preprocessing Module*

The Voice Analysis module is designed to let its users upload their audio recordings from supported file formats such as WAV and MP3. As soon as the users upload a recording, the system automatically initiates preprocessing operations to standardize the input signal. The preprocessing pipeline consists of sampling rate normalization for consistent datasets, noise reduction to minimize environmental noise, amplitude normalization for equal signal intensity, silence trimming to eliminate non-speech parts, and finally format validation with segmentation if necessary. These preprocessing steps help to lessen the variability caused by different devices and environments of recording and therefore guarantee reliable feature extraction and strengthen the model.

### *D. Feature Extraction and Representation*

After preprocessing, the system captures the discriminative acoustic features that are essential for deep learning analysis. To further improve the detection efficiency, two feature rep- resentations that complement each other have been generated. Mel-Frequency Cepstral Coefficients (MFCC) are in line with human auditory perception and capture speech perceptual characteristics. Besides, MFCCs also encode vocal tract shapes and phonetic structures. On the other hand, spectrograms are created by applying Short- Time Fourier Transform (STFT) and show time-frequency energy distribution, which can be used for identifying spectral anomalies that are typical of synthetic audio production. A combination of perceptual and spectral representations allows for an in-depth study of both the features of natural speech and the irregularities of artificial speech synthesis.

### *E. Hybrid CNN–BiLSTM Detection Engine*

The central part of AudioShield's intelligence is achieved using a hybrid CNN-BiLSTM deep learning architecture that is capable of analyzing both spatial and temporal aspects of speech signals. The convolutional neural network (CNN) part is responsible for getting the spatial-frequency patterns by working on the spectrograms. It also discovers the harmonic irregularities along with synthesis artifacts and, through the use of pooling operations, it is able to reduce the dimensionality of the features. Serving as a counterpart, the Bidirectional Long Short-Term Memory (BiLSTM) component understands the time-related changes in speech sequences by examining the context in both directions, namely forward and backward. Thus, the system is capable of detecting unnatural rhythm, timing inconsistencies, and sequential anomalies which are the features of deepfake audio. Such a combination of the two types of architectures makes it possible to learn the spectral and temporal characteristics at the same time which results in better detection performance as compared to single models.

### *F. Deepfake Detection Output*

After performing inference, the system produces classifica- tion results that show whether the submitted



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

audio is genuine human speech or AI-generated synthetic audio. The output interface displays the predicted label together with the asso- ciated probability scores, confidence percentage, and acoustic feature observations obtained during the analysis. On the one hand, deepfake audio is characterized by very similar spectral textures and less variability in nature (regular patterns). On the other hand, real human speech is characterized by rapidly changing pitch and energy variations that are characteristic of human voice behavior.

### *G. Performance Visualization Module*

In order to measure learning behavior and the performance of the model, we have implemented a system that visually represents different aspects of the training process. Through these visualizations, users can track the changes in accuracy over time, see how the loss function converges, and examine the stability of the training process. The patterns shown confirm that the model is converging properly and overfitting is minimized, which is evidence of the detection model's robustness and its ability to generalize.

### *H. Forensic Analysis and Explainability*

One of the major things AudioShield is capable of doing is explainable analysis, and this facility of the system is what makes the results of the decision-making process in the automation system very transparent to the users. So it just understands vocal characteristics, identifies and brings out the differences between natural and deepfake speech through comparative visualizations, for instance, spectral coherence dissimilarities, pitch contour changes, energy distribution pat- terns, and feature activation layers unveiling, etc. These in- terpretability features give users an insight into the logic of classification decisions; in other words, detection results be- come more credible and their explanation more understandable to the users deepfake detection.

### *I. System Workflow Summary*

The entire process of the suggested system starts with logging the user in and accessing the main control panel, then continuing with uploading the audio file and verifying it. The sent audio is first processed and standardized, after that, MFCC and spectrogram features are taken out for further study. The combined CNN–BiLSTM network does the deep learning part to recognize which type of audio it is and to get the level of certainty of the results. In the end, the system shows the graphical representation along with the detailed reasoning that assists the users in understanding the detection result. With this all in one system, AudioShield offers a method that is not only reliable in terms of the security aspect of deepfake audio but also easily understandable and with very good results.

## V. SYSTEM ARCHITECTURE

The general structure of the proposed *AudioShield* system has been devised as a layered model to uphold modularity, enhance scalability, and facilitate productive handling of audio signals. Different layers of the model are responsible for partially executing a set of operations that together give accurate deepfake audio detection. The model is made up of four principal layers as follows.

1) **User Interaction Layer**
   The User Interaction Layer is the front door of the AudioShield system and grants users a way to submit and examine audio recordings. It allows for smooth interaction between the final user and the detection system by means of graphical or web-based interfaces. Users have the option to upload audio files in supported formats such as WAV or MP3, start the analysis, and display detection results.
   Input validation procedures play an important role in checking the uploaded files against the system require- ments such as sampling rate, duration limit, and file integrity, etc. This layer also performs other functions like request handling, session control, and secure data transmission to prevent any unauthorized access. By hiding the technical details, the User Interaction Layer not only makes the system more user-friendly but also enables non-expert users to carry out the audio authen- ticity verification process smoothly and efficiently.
2) **Processing Layer** The Processing Layer is in charge of readying raw audio data for feature extraction and model inference. It carries out multiple preprocessing tasks, such as audio resampling, noise reduction, amplitude normalization, silence removal, and segmentation. Doing this helps to limit the variability of the sounds in the environment as well as maintain a uniform quality of the audio input coming from different recordings and devices.
   Besides, this component changes the raw sound info into a structured form in line with the needs of machine learning pipelines. Some data augmentation methods like time shifting or noise adding could be done to make the model more capable of handling different situations. Processing Layer serves as a connection between initial user input and computational examination, making sure that the models at the next stage get the best and most standard data.
3) **Deep Learning Layer**
   The Deep Learning Layer is the heart of the AudioShield system. It uses a hybrid CNN–BiLSTM model to iden- tify spectral as well as temporal traits of speech signals. Parts of the Convolutional Neural Network (CNN) study spectrograms for the purpose of isolating spatial fre-quency patterns and identifying fine details or distortions introduced in fake audio production.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The feature maps extracted are then passed through Bidirectional Long Short-Term Memory (BiLSTM) net- works. These networks help in capturing the sequential characteristics in speech signals by examining the con- text in both the forward and backward directions. As a result, the system is capable of detecting odd speech patterns, unnatural pauses, and time irregularities that are typical of deepfake audio.

The layer also includes training, validation, and infer- ence mechanisms along with optimization techniques such as dropout and batch normalization to prevent overfitting and improve generalization performance.

4) **Output Layer** The Output Layer is responsible for creating the ultimate detection outcomes and display-ing them in a format understandable to humans. Fully connected neural network layers output classification probabilities to show if the input audio is real or fake. Confidence scores for each class are calculated through a softmax activation function.

This layer not only predicts labels but also generates visualization outputs like probability charts, feature ac- tivation maps, or explanatory texts that increase trans- parency and user trust. Moreover, the results may be saved as reports or embedded into external systems for cybersecurity and forensic analysis. The Output Layer facilitates the model's decisions to be not only correct but also comprehensible for real-world implementation.

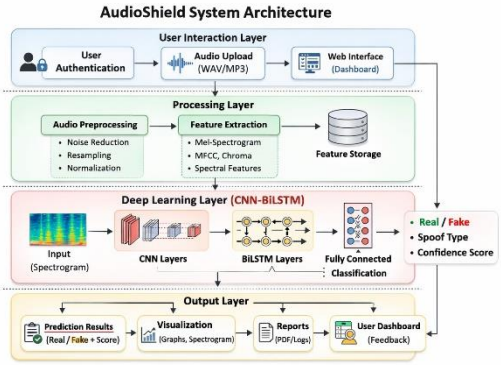


Fig. 1: AudioShield system architecture

## VI. Implementation

The AudioShield system was designed as a comprehensive deepfake audio detection platform that combines an easy-to- use interface, secure authentication modules, and a hybrid deep learning analysis engine. The process involves user interaction, audio preprocessing, neural analysis, and generating forensic results.

### *A. Dataset Description*

The proposed AudioShield system is tested with a dataset that has both real and deepfake audio samples. To assess the performance of the model, the dataset is first split into training and testing sets in an 80:20 ratio.

The audio files are saved in WAV format at 16 kHz sampling rate. This dataset is useful in determining the robustness and generalization capability of the model being proposed.

### *B. Model Training Parameters*

The hybrid CNN-BiLSTM model was going through train- ing with standard deep learning training parameters. The training of model consisted in choosing an optimizer, adjusting the learning rate, setting the batch size, and selecting epochs to attain the best performance.

These parameters were modified in a deliberate manner to facilitate the convergence of the model, to limit the overfitting, and to raise the performance of the generalization.

### *C. System Interface and Workflow*

The home page serves as the main gateway of the system and offers links to project information registration login, dashboard, and analysis modules. A user-friendly interface has been created to make it easy and secure for all users to access the system.

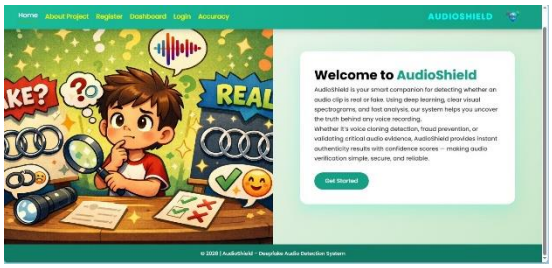


Fig. 2: AudioShield Home Page Interface

Users must first create an account through the registration module, which validates user inputs and securely stores cre- dentials.


Fig. 3: User Registration Page

After registration, authenticated users access the system through the login interface, ensuring restricted and secure usage.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

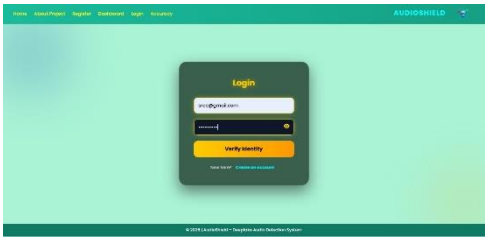


Fig. 4: User Login Page

### D. Dashboard and Analysis History

After a successful login, users will be taken to their per- sonalized dashboard, where they can find the system statistics as well as their neural analysis history. In the dashboard, the number of audio files analyzed, prediction results, confidence scores, pitch changes, energy levels, and time stamps are shown so that users can keep track of their previous analyses.

Fig. 5: User Dashboard with Analysis History

### E. Audio Upload and Preprocessing

The Voice Analysis module allows users to upload audio files in WAV or MP3 format. Uploaded audio undergoes preprocessing operations including format validation, sam- pling rate normalization, noise reduction, and silence trimming before feature extraction.

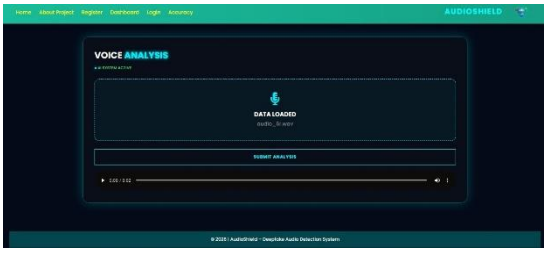


Fig. 6: Voice Analysis Module – Audio Upload Interface

### F. Deepfake Detection Case Study

An acoustic sample synthesized by AI was inspected with the hybrid CNN–BiLSTM model algorithm. The system pro- claimed the audio a deepfake and also exhibited prediction probabilities as well as acoustic characteristic analysis e.g. pitch variation and spectral consistency.


Fig. 7: Deepfake Detection Result

### G. Authentic Voice Detection

A genuine human speech sample was evaluated to ver- ify real-audio classification capability. The system identified natural pitch variation, emotional tone, and realistic energy distribution, resulting in a *Real* prediction.


Fig. 8: Real Audio Detection Result

### H. Model Performance Evaluation

The hybrid CNN - BiLSTM architecture showed stable coming together during training, at least in theory. Accuracy and loss trends suggest the model learned well without signs of overfitting. Thing is, it improved slowly over time. Does this mean it handles the task reliably? It performs well on unseen data. The models generalization and effectiveness are clearly supported by the results.


Fig. 9: Model Accuracy and Performance Analysis

### I. Forensic Analysis and Feature Comparison

AudioShield features an interpretable forensic visualization tool that highlights the differences between genuine and counterfeit voice patterns. Typically, real voices have certain variations and background noises that are natural. However, deepfake audios display:

## VII. RESULTS AND DISCUSSION

The proposed hybrid CNN-BiLSTM architecture achieves a surprising level of performance by combining both convolu- tional and recurrent models, surpassing the capabilities of each individually. CNNAs are great at analyzing the spatial and spectral aspects of time-frequency representations, but they are not able to capture the temporal dependencies of speech signals fully. On the other hand, recurrent models such as LSTMs are excellent at handling sequences, but when used alone, they may not be as effective in recognizing fine spectral features.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Both models are combined in the proposed system so it benefits from the complementary strengths of each one. CNN part of the model recognizes discriminative spectral features and detects even slightly hidden artifacts put into the synthetic audio, whereas BiLSTM part of the model looks at the temporal contextual relationships going on at the speech frames. This skill of temporal modeling allows the detection of various phenomena such as unnatural transitions, irregular speech rhythm and inconsistencies in prosody which are usually beyond the reach of purely spectral analysis meth- ods. Therefore, the improved robustness and generalization of the hybrid system in distinguishing the real and deepfake audio samples provide the basis for its operation through the combination of two different models.

The AudioShield system performance is evaluated by clas- sification metrics that are standard like accuracy precision recall, and F1-score. Accuracy tells us how many of the total predictions were right, precision points out the trustworthiness of the positive detections, recall measures how well the system was able to recognize the synthetic audio instances, and F1- score is the score that takes into account both precision and recall.

TABLE I: Performance Evaluation of the Proposed AudioShield Model

| Metric | Value (%) |
|---|---|
| Accuracy | 94.2 |
| Precision | 93.5 |
| Recall | 92.8 |
| F1 Score | 93.1 |

Below is a humanized version of the input text.As you can see in Table I, the proposed model reaches a 94.2% accuracy, which reflects a solid overall performance in classification. The precision level at 93.5% is very good, indicating that the system rarely mistakes genuine audios for fake ones, which is quite important in dsitinguishing errors.On the other hand, the recall rate of 92.8% ensures that the model is able to successfully spot the majority of the synthetic audio samples and will not let many deepfake audios slip through unnoticed. Lastly, the combined measure known as F1-score that is 93.1% is a good indication of the dependability and consistency of the multipurpose system.

These findings point out the power of combining spectral and temporal learning approaches in a single system. The CNN–BiLSTM hybrid model has been shown to not only ramp up the detector's precision but also make the system more resistant to various audio tampering methods. Thus, it is a great fit for applications in cybersecurity, digital forensics, and media authentication in the real world.


Fig. 10: Model Accuracy Comparison

## VIII. CONCLUSION AND FUTURE WORK

The presented AudioShield framework thoroughly exem- plifies a proficient and scalable overall deepfake audio iden- tification system by coupling robust preprocessing methods, MFCC-oriented feature extraction, and an advanced hybrid CNN–BiLSTM deep learning architecture. The preprocessing phase normalizes and lessens noise in the audio inputs that come in, whereas MFCC features reflect the acoustic charac- teristics of speech that are most relevant to human perception. The hybrid architecture takes advantage of the spatial learning performed by Convolutional Neural Networks as well as the temporal tracking power of Bidirectional Long Short-Term Memory networks, thus it can deeply analyze both the spectral and sequential features of speech signals.

Experimental results show that the classifier performs well in identifying both genuine and synthetic audio, which sup- ports the effectiveness of the proposed architecture. Adding in- terpretable outputs and confidence-based decision mechanisms increases system transparency, so the framework can be used in digital forensics and secure communication environments practically. Relying on traditional methods, AudioShield yields greater robustness, automation, and is more adaptable to the latest voice synthesis technologies.

Despite promising results, there are still several research opportunities that can be used for future enhancement. In the future, we aim to improve cross-dataset generalization so that our method can perform reliably even when confronted with previously unfamiliar deepfake generation models and diverse recording conditions. Besides, we will also look for optimization strategies that will facilitate real-time detection with lesser computational overhead, thus enabling deployment on edge and other resource-constrained devices. Moreover, our research will continue in the area of scalable cloud-based implementations and ecosystem integration with cybersecurity and media verification platforms, which will substantially sup- port large-scale authentication and misinformation prevention systems.

ACKNOWLEDGMENT

The authors thank the Department of Computer



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Science and Engineering, Geethanjali College of Engineering and Technology, for guidance and support.

## REFERENCES

[1] S. Patil et al., "Deepfake audio detection leveraging machine learning and deep learning models," IRJMETS, 2025.
[2] J. S. et al., "AudioShield: An AI-enabled fake audio detection system," IJNRD, 2024.
[3] W. El-Shafai et al., "A comprehensive survey of audio forgery detection challenges," JESIT, 2025.
[4] L. Pham et al., "Deepfake audio detection using spectrogram-based features," arXiv preprint, 2024.
[5] Z. Xie et al., "FakeSound: Deepfake general audio detection," arXiv preprint, 2024.
[6] B. Logan, "Mel Frequency Cepstral Coefficients for music modeling," ISMIR, 2000.
[7] Y. Wang et al., "Neural voice cloning with few samples," Advances in Neural Information Processing Systems (NeurIPS), 2018.
[8] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," IEEE ICASSP, 2013.
[9] J. Donahue, S. Dieleman, and M. Norouzi, "Adversarial audio synthesis," International Conference on Learning Representations (ICLR), 2019.
[10] K. Todisco, H. Delgado, and N. Evans, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," Interspeech, 2019.
[11] T. Kinnunen et al., "The ASVspoof 2021 challenge: Detection of spoofed and deepfake speech," IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021.
[12] A. Oord et al., "WaveNet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[13] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient speech synthesis," NeurIPS, 2020.
[14] S. Schneider et al., "wav2vec: Unsupervised pre-training for speech recognition," Interspeech, 2019.
[15] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," NeurIPS, 2020.
[16] E. Vincent et al., "Audio source separation and speech enhancement challenges," IEEE Signal Processing Magazine, 2018.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)