

# Comparative Analysis of Deep Learning Models for Detecting Deepfake Audio Using MobileNet and Explainable AI

S Harish Kumar<sup>1</sup>, Ashok Dasari<sup>2</sup>, S Divya Sree<sup>3</sup>, K Chandu<sup>4</sup> and R Arun Kumar<sup>5</sup>  
{[21691A2842@mits.ac.in](mailto:21691A2842@mits.ac.in)<sup>1</sup>, [ashokd@mits.ac.in](mailto:ashokd@mits.ac.in)<sup>2</sup>, [21691A2828@mits.ac.in](mailto:21691A2828@mits.ac.in)<sup>3</sup>, [21691A2816@mits.ac.in](mailto:21691A2816@mits.ac.in)<sup>4</sup>,  
[21691A2804@mits.ac.in](mailto:21691A2804@mits.ac.in)<sup>5</sup>}

UG Student, Department of CST, Madanapalle Institute of Technology & Science, Andhra Pradesh, India<sup>1, 3, 4, 5</sup>

Assistant Professor, Department of CST, Madanapalle Institute of Technology & Science, Andhra Pradesh, India<sup>2</sup>

**Abstract.** Deepfake audio, which has the creation of highly realistic synthetic voices, has become a major cybersecurity concern. It can be used for misinformation, fraud, and unauthorized access, making accurate detection crucial. This paper presents a hybrid deep learning approach that improves both the accuracy and interpretability of deepfake audio detection. In this study the model integrates CNNs, RNNs, and transformers to extract and analyze features from audio files effectively. To ensure transparency in decision-making, we use XAI techniques such as SHAP, LIME, and Grad-CAM to highlight the key factors influencing predictions. Our experimental results demonstrate high detection accuracy, resilience against adversarial attacks, and improved trustworthiness of model decisions. This research contributes to strengthening cybersecurity defenses by making deepfake detection both reliable and interpretable.

**Keywords:** Deepfake audio, cybersecurity, deep learning, Explainable AI, hybrid model, adversarial robustness, transparency.

## 1 Introduction

Recently, deep learning has enabled the creation of highly realistic synthetic voices, commonly known as deep-fake audio. These artificially generated audio clips can closely mimic human speech, which makes it difficult to distinguish between genuine and fake recordings. While this technology has potential benefits in areas like entertainment and voice assistance, it also poses serious cybersecurity threats. Deepfake audio can be misused for identity theft, financial fraud, misinformation campaigns, and unauthorized access to secure systems, leading to significant consequences [1]. Traditional ML models often fail to predict deepfake audio effectively because they rely on handcrafted features and are limited in handling complex patterns. To address this, DL models such as VGG16, ResNet, and custom CNNs can be used for more accurate detection. However, a major challenge with deep learning models is their black-box nature, especially in cybersecurity applications where explainability is crucial [2]. This research focuses on integrating hybrid deep learning models with Explainable AI (XAI) techniques to develop a robust deepfake audio detection system. GANs are employed to synthesize audio content, which helps in training models to recognize deepfake characteristics. To enhance interpretability, XAI techniques such as LIME, Grad-CAM, and SHAP are incorporated. These methods provide insights into how the model

identifies deepfake audio, making the detection process more transparent and trustworthy [3]. By combining advanced deep learning architectures with explainability tools, this research focuses on the accuracy, reliability, and transparency of deepfake audio detection. The proposed approach ensures that cybersecurity professionals and organizations can confidently use AI-based detection systems to safeguard against the growing threat of deepfake audio. The structure of the paper as follows. Section II reviews existing research in deepfake audio detection, discussing various approaches, their advantages, and their limitations. Section III describes the methodology, including the hybrid deep learning models used, dataset details, preprocessing steps, and the implementation of Explainable AI techniques. Section IV presents the experimental setup, evaluation metrics, and the performance analysis of the proposed model. Section V discusses the findings, comparing the proposed approach with existing models and emphasizing the impact of XAI in improving model transparency. Finally, Section VI has the conclusion of the paper and summarizes the key outcomes and gives the insights for the future research directions for improving the deepfake audio detection in cybersecurity

## **2 Literature**

### **2.1 Previous Work on Deepfake Audio Detection**

The detection of deep-fake audio has gained attention, leading to the development of various methodologies aimed at distinguishing synthetic voices from true ones.

In 2024, researchers introduced a cross-domain audio deepfake detection dataset comprising over 300 hours of speech data. This dataset was meant to optimize the generalization skills of detection models across several domains, addressing the limitations of previous datasets that were often outdated and domain-specific. The study emphasized the importance of diverse and up-to-date data in training robust deepfake audio detectors [4]. Another research in the year 2024 posited the Sonic Sleuth model, An AI model specifically designed to detect audio deepfakes. Leveraging state-of-the-art DL methodologies, such as original CNN design, the achieved result amounted to 98.27% accuracy and an error rate of 0.016 in equivalent on a large real and synthetic sound dataset. The work demonstrated the importance of training with percet simulated ardualization, environmental noise and acoustive conditions(able to use inputs in a wide variety of acoustic conditions, such as changing classes in the data allowing for generalisation across a wide range of scenarios for inputs across different types of audio data [5]. In [6], the M2S-ADD model, which attempts to detect audio authenticity cues when converting from mono to stereo. This method exploits the essential distinctions between mono and stereo audio files to address the problem of deepfake detection using audio, providing a new angle on audio deepfake detection. In [7] of 2022, the Mel-frequency cepstral coefficients (MFCC) features are used with machine learning classifiers to detect deepfake audio based on. This methodology targeted at extracting voice features and using classification tools, allowed a proof of concept of the possibility to rely on traditional audio-based features and modern machine-learning algorithms to deal with the problem of synthetic content detection. In [8] a comprehensive survey was released in 2023 that covered a lot of methods for generating and detecting audio deepfakes. It covered different approaches used in generating deepfake audio, and their corresponding detection approaches, providing interesting perspectives on the problems and progress in the area. The ADD 2023 competition in [9] was proposed to promote the detection of deepfake audio rather than binary classification. This challenge mimicked actual use-cases and interest by concentrating on

objectives like detecting rescaled intervals in partially generated audio and traced the source of the filler noises. The project work was targeted at developing and improving detection and analysis methods critical for the construction of convincing and reliable evidence in applications such as audio forensics or journalism. In [10], a study of 2023, proposed the In 2023, a study suggested the HM-Conformer model, a system for detecting audio deepfakes using conformers. This model uses hierarchical Token pooling and aggregation at several classification levels are used to increase detection performance. The approach improved the accuracy in identify spoofing attacks from text-to-speech or voice transformation systems. In 2024 [11] the researchers introduced the CLAD (Contrastive Learning-based Audio Deepfake Detector) model to improve resistance to manipulation. This method used contrastive learning to reduce variations caused by manipulations, thereby improving identifying robustness. The study highlighted the vulnerabilities of existing detectors to simple manipulations and demonstrated that CLAD significantly reduced false acceptance rates across various tests. [12] focused on the development of the Codecfake dataset, an open-source large-scale dataset designed for detecting ALM-based deepfake audio. The researchers proposed a countermeasure strategy to achieve universal detection of deepfake audio and address domain ascent bias issues. Their method showed promising results in the detection an ALM based sound and also in generalization across different test conditions. [13] presented a deep learning model for the detection of deepfake audios. The methodology entailed converting input audio data to multiple spectrograms, and testing on diverse classification models such as CNN, RNN, and transfer learning from the computer vision models. The target domain: Imbalanced training procedures are not effective Visualization As detailed in Table 1, the systems achieved very competitive EER in the standard benchmark data, which suggests the possibility of the selected spectrograms and deep learning methods for boosting the performance of the audio deepfake detection.

## 2.2 Related work on XAI

Incorporating Explainable AI (XAI) into deepfake audio detection systems has become a research direction to improve the transparency and reliability of such models. [14] proposed a novel approaches for transformer-based audio deepfake detectors. Researchers created methods for integrating outcomes into these models to assure their explainability, a crucial factor in AI-based detection solutions.

## 3 Dataset

We use the 'Fake or Real' dataset, provided by the researchers at York University, to train our deepfake audio classifiers. Our dataset includes both real and deepfake audios, which enables our model to correctly distinguish true positive samples and fake ones. Audio was converted to frequency spectral data audiosong27. Spectrograms provide rich features, and hence are more appropriate to be used with deep learning models [15].

### 3.1 Feature selection

The features in this study we extracted by the deep learning model layers as follows:

- 1) *Convolutional Layers:* During training, MobileNet's convolutional layers automatically determine which spectrogram segments are crucial for sorting. While deeper

layers of the network capture complicated patterns and structures in the spectrogram, early levels may concentrate on basic elements like edges and textures [16].

2) *Global Average Pooling*: The GAP following the convolutional Layers lowers the spatial dimensions of feature maps into a single value per channel. This operation effectively "selects" the required features by averaging the values over each feature map, which reduces the feature count while keeping important data [17].

3) *Dense Layers*: The final dense layers in MobileNet further process the extracted features and refine the feature selection for the final classification task. These layers assign weights to different features and output a prediction based on the learned importance of each feature [18].

### 3.2 Feature Extraction

The features in this study were extracted as below:

*MobileNet's Convolutional Layers*: MobileNet uses depthwise separable convolutions as the primary feature extraction technique. The conventional convolution method is divided into two easier [19] steps by depthwise separable convolutions:

- **Depthwise Convolution**: every input channel is convolved with its own filter, meaning that each feature map is processed independently.
- **Pointwise Convolution**: A  $1 \times 1$  convolution is used to include the results of the depthwise convolutions.

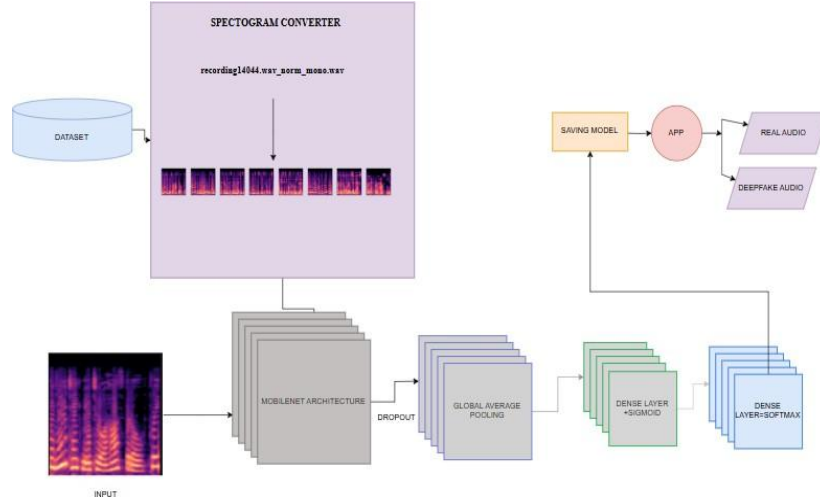
The features extracted by the MobileNet model during the deepfake audio detection task can be categorized into low, mid, and high level features. These features are derived from the audio spectrograms and processed through various layers of the MobileNet architecture. The Table 1 demonstrates these features as they are extracted at different stages of the model.

## 4 Proposed Model

In this study we explored various pre-trained deep learning model and evaluated to determine the best-performing model for classifying deepfake audio. Given the importance of Feature extraction and classification within the framework of detecting synthetic audio, we initially experimented with three prominent models: MobileNet, InceptionV3, Custom CNN, and VGG16. Based on their shown efficiency in picture classification tasks and their adaptability to audio data in the form of spectrograms, these models were chosen.

Based on the effectiveness of each model the Mobilenet outperformed over the other models with an of accuracy of 91% making it the best-performing model for this deepfake audio detection task. Its ability to efficiently process audio spectrograms, combined with its lightweight architecture, allowed it to deliver high performance without compromising computational efficiency. Fig 1 illustrates an end-to-end pipeline for deepfake audio detection using MobileNet. The process begins with the spectrogram conversion of raw audio files into visual representations. These spectrograms are then fed into the MobileNet architecture, which acts as a feature extractor. After feature extraction, Dropout and Global Average Pooling layers help regularize the model and reduce spatial dimensions. The extracted

features pass through Dense layers with sigmoid and softmax activations to differentiate the audio as either real or fake. The trained model is then saved and deployed in an app for real-time classification of incoming audio samples. This pipeline effectively uses MobileNet's lightweight structure to deliver efficient, accurate audio classification.



**Fig. 1.** Proposed model.

#### 4.1 Mathematical Representation

The base of MobileNet is depthwise separable convolutions, which reduces the complexity and parameters when compared to conventional convolutions. A mathematical representation of the MobileNet architecture used in this work is as follows:

##### 4.1.1 Depthwise Separable Convolution:

Depth wise separable convolutions are a key feature of the MobileNet architecture. They consist of two procedures include depthwise and pointwise convolution.

Depthwise Convolution: Each input channel is combined with its own filter:

$$Y_{\text{depthwise}} = W_{\text{depthwise}} * X \quad (1)$$

Where:

- $Y_{\text{depthwise}}$  is the output of the depthwise convolution,
- $W_{\text{depthwise}}$  is the depthwise filter,
- $X$  is the input feature map.

Pointwise Convolution: A  $1 \times 1$  convolution is used to aggregate the results of the depthwise convolution:

$$Y_{\text{pointwise}} = W_{\text{pointwise}} * Y_{\text{depthwise}} + b \quad (2)$$

Where:

- $Y_{\text{pointwise}}$  is the output of the pointwise convolution,
- $W_{\text{pointwise}}$  is the  $1 \times 1$  filter,

- $b$  is the bias term.

**Table 1.** Feature Levels in the MobileNet Model for Audio Classification.

Level	Features
Low-Level	Edges (horizontal, vertical, diagonal), Textures (periodicity, noise), Frequency Patterns (specific spectral distributions)
Mid-Level	Harmonics (peaks at regular intervals), Rhythm (amplitude and frequency variations), Formants (important for phoneme distinction in speech)
High-Level	Speech Patterns (complex interactions), Audio Artifacts (subtle features in deepfake audio)

*Final Output:* The final output of the depthwise separable convolution is:  
 $Y_{\text{output}} = Y_{\text{pointwise}}$

(3)

#### 4.1.2 ReLU Activation

After each convolution operation, the ReLU activation function is applied:

$$f(x) = \max(0, x)$$

(4)

- $f(x)$  is the result after using ReLU,
- $x$  is the input to the activation function.

#### 4.1.3 Global Average Pooling (GAP)

Global Average Pooling (GAP) decreases the spatial dimensions of the feature maps to a single value per channel. It is calculated as the average of all values in a given feature map:

$$GAP(Y) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Y_{i,j}$$

(5)

Where:

- $Y$  is the input feature map,
- $H$  and  $W$  are the height and width of the feature map,
- $Y_{i,j}$  is the value at position  $(i, j)$  in the feature map.

This results in a 1D vector representing the extracted features.

#### 4.1.4 Fully Connected Dense Layer

The output from the GAP is sent through entirely connected (dense) layers. The equation for a dense layer is:

$$Z = W_{\text{dense}} \cdot X + b$$

(6)

Where:

- $Z$  is the dense layer output,
- $W_{\text{dense}}$  is the dense layer's weight matrix,
- $X$  is the input to the dense layer (from GAP),
- $b$  is the bias term.

The output of the dense layer is passed through an activation function, such as ReLU or sigmoid.

#### 4.1.5 Output Layer for Binary Classification

For binary classification (e.g., real vs. fake audio), a sigmoid activation function is applied:

$$P(\text{real}) = \sigma(Z) = \frac{1}{1+e^{-z}} \quad (7)$$

Where:

- $\sigma(Z)$  is the sigmoid function,
- $Z$  is the output from the final dense layer.

The sigmoid function produces a value ranges from 0 and 1, a value closer to 1 indicating "real" audio, and a value near to 0 indicating "fake" audio.

#### 4.1.6 Loss Function

The model is trained using the binary cross-entropy loss, which measures the difference between true labels and predicted probabilities:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (8)$$

- $N$  represents number of samples,
- $y_i$  represents the true label for the  $i$ -th sample (0 or 1),
- $p_i$  is the predicted probability for the  $i$ -th sample (from the sigmoid output).

This loss function helps minimize the difference between predictions and actual labels during training.

#### 4.2 XAI integration

In this study we integrated MobileNet with Explainable AI (XAI) techniques that enhanced the transparency and decision making of the model, which is important in the Deepfake audio detection. While MobileNet is excellent at extracting the relevant features from the spectrograms, the black-box nature of deep learning models often makes the predictions unclear, integrating XAI makes it more enhanced [20].

1) *LIME*: LIME was applied to the spectrograms fed into MobileNet. For a given audio input, LIME identifies the regions of the spectrogram that are most important in the classification. The key advantage of LIME is that it provides local explanations, which help explain why the model generated a specific prediction for a given audio sample.

2) *Grad-CAM*: For a given prediction, Grad-CAM generates a heatmap to highlight the important areas of the spectrogram, indicating where the model "focuses" when making a choice. This heatmap is then superimposed on the original spectrogram, allowing us to visually inspect which frequencies and time frames MobileNet is attending to during classification.

### 5 Comparative Result Analysis

1) *InceptionV3*: In the study to compare we used the InceptionV3 pre-trained model, was also explored for the audio classification task. It was fine-tuned with custom

layers to adapt to the audio data represented as spectrograms. While InceptionV3 achieved good performance, its accuracy was slightly lower compared to MobileNet, with a recorded accuracy of around 87%. Despite its slightly lower performance, InceptionV3 was still able to capture relevant features for audio classification, but MobileNet outperformed it in terms of overall accuracy and computational efficiency [21].

2) *VGG16*: We tested VGG16 model, a deep convolutional network known for its simple yet powerful architecture in the audio classification task. The model was trained using the binary cross-entropy loss function and Adam optimizer at a learning rate of  $1e-5$ . During training, the model attained 85% validation accuracy. The results show that VGG16 can extract significant characteristics from audio spectrograms and classify them accordingly [22].

3) *Custom CNN Model*: We created the Custom CNN model with convolutional layer, pooling and dense layers, specifically designed for the audio classification task. It used the Adam optimizer and binary cross-entropy loss function. The model achieved validation accuracy of 90%, outperforming the VGG16 model. The custom model showed better generalization to the audio data, suggesting that a more tailored architecture for the task at hand can provide improved performance in classifying real versus fake audio.

4) *MobileNet*: MobileNet model outperformed the models with an accuracy of 91.5%, showing its effectiveness in distinguishing between real and synthetic audio signals. Table 2 shows the achieved metrics of the models.

**Table 2.** Model Performance Comparison.

Model	Accuracy (%)	Precision	Recall	F1 score	Epoch
Inception Net	90.17	0.505	0.529	0.539	5
VGG 16	93.37	0.508	0.421	0.472	5
Mobile Net	91.57	0.507	0.551	0.507	5
Custom CNN	88.33	0.506	0.503	0.504	20

5) *Grad-CAM Visualizations*: The Grad-CAM visualizations provide a heatmap that highlights the regions in the spectrogram that had a role in the model's decision. The red and yellow colors indicate high importance, and the blue and green colors represent less influence.

*Inception Net*: Inception Net achieved an accuracy of 90.17%, with a precision of 0.505, recall of 0.529, and an F1 score of 0.539. This indicates that while Inception Net performed well, its precision and recall were not optimal.

*VGG16*: VGG16 showed a higher accuracy of 93.37%, but with a precision of 0.508 and recall of 0.421, resulting in a relatively low F1 score of 0.472. VGG16 classified more samples as real, resulting in a lower recall.

*MobileNet*: MobileNet, a lightweight model, showed an accuracy of 91.57%, a precision of 0.507, recall of 0.551, and gave F1 score of 0.507. Despite slightly lower accuracy compared



to VGG16, MobileNet demonstrated a better balance, made it more effective for real-time applications where computational efficiency is also critical.

*Custom CNN:*The Custom CNN achieved an accuracy of 88.33% after 20 epochs of training, with recall and precision both around 0.506 and an F1 score of 0.504. Although the accuracy was lower than the other models, the custom architecture still performed reasonably well with a good balance of precision and recall.

### 5.1 Interpretation of Visualizations

The table 3 provides visualizations of how the SHAP, Grad-CAM, and LIME methods explain the classsifiacion ability of different models in the deepfake audio detection task. Each technique highlights the regions in the spectrogram that the models consider most important when classifying audio as either real or fake.





*SHAP Visualizations:* The SHAP plots show the contribution of each feature (pixel) in the spectrogram to the model’s output. The green and red regions indicate positive and negative contributions, respectively. In the visualizations:

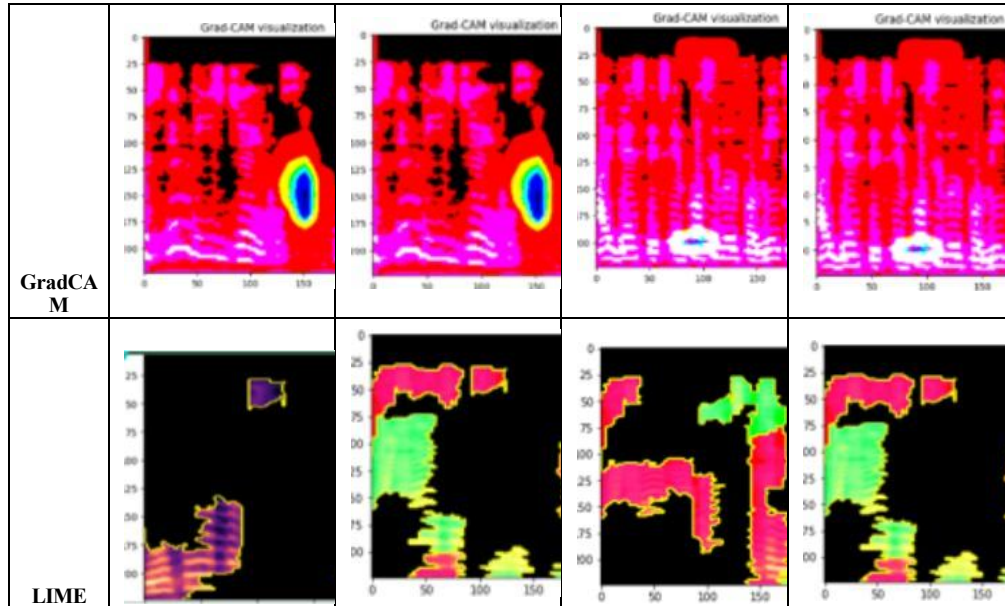
- **VGG16** and **MobileNet** show a more spread-out focus across the spectrogram, with some areas of high importance (green) corresponding to certain frequencies and time intervals.
- **InceptionV3** appears to be more uniform, suggesting it might be focusing on a broader set of features rather than specific time-frequency regions.

*Grad-CAM Visualizations:* The Grad-CAM visualizations provide a heatmap that highlights the regions in the spectrogram that had a role in the model’s decision. The red and yellow colors indicate high importance, and the blue and green colors represent less influence.

- **MobileNet** exhibits a focused area of attention, with clear highlights around specific frequency bands, suggesting it pays more attention to distinct features within the audio signal.
- **Custom CNN** and **VGG16** appear to have a broader focus, indicating that they might consider more general features across the spectrogram.
- **InceptionV3** shows varied focus, indicating that it may be relying on a range of features in the spectrogram.

**Table 3.** XAI Visualization Methods for Different Models.

Models	Custom CNN	VGG16	MobileNet	InceptionV3
SHAP				



*LIME Visualizations:* The LIME heatmaps provide further insights into the model’s reasoning by showing the areas of the spectrogram that influenced the decision. The yellow and green regions indicate high importance.

- Similar to Grad-CAM, **MobileNet** highlights specific regions, with concentrated focus on certain frequencies that are most relevant to the classification task.
- **VGG16** and **Custom CNN** also show significant focus but with less specificity compared to **MobileNet**.

*Interpretation of Results:*

- **MobileNet** shows the most focused attention in its visualizations, suggesting that the model is more discriminative and effectively identifies specific features in the spectrograms that distinguish real audio from fake audio. **VGG16** and **Custom CNN** demonstrate more generalized focus, suggesting they might be using broader features for classification.
- **InceptionV3** displays multiple areas of focus, indicating that it is drawing from various features, which might make it more versatile but less precise compared to **MobileNet** in terms of focusing on the key aspects of the spectrogram.

## 6 Conclusion

This study proposes a deepfake audio detection model with MobileNet for classifying audio samples into positive and negative samples. The lightweight architecture of MobileNet with depthwise separable convolutions efficiently learns useful features from audio spectrograms, such as low-level, mid-level, and high-level patterns. Although other network architectures like VGG16 were slightly more accurate, MobileNet had a higher recall and better

computationally, and hence suitable for real-time applications.

In order to improve transparency of the model, the XAI techniques (LIME, SHAP and Grad-CAM) were incorporated into the model to interpret the portions of the spectrogram that contributed to predictions. This has the advantage of achieving both accuracy and interpretability, both of which should be important for practical application.

The system evidences effective deepfake detection, and provides a solid baseline for further enhancements by data extension, use of more sophisticated architectures, and more effective XAI methodologies.

## 7 Future Work

Improving focus on the dataset, both in improving generalization across different audio types and deepfake techniques. Using alternative architectures or hybrid models, especially including transfer learning, may improve the system's results. Optimizing Explainable AI methods to allow the model to provide more detailed reasons for decision-making may improve its interpretability. Real-time deployment testing in real-world environments such as live phone calls or streams would provide more information about how reliable and efficient the system is. Integration of multimodal approaches, including combining audio and visual data to detect deepfake video, might further increase its accuracy. These are areas in which the system could be made more reliable, scalable, and applicable for real-world purposes.

## References

- [1] Qais, A. Rastogi, A. Saxena, A. Rana and D. Sinha, "Deepfake Audio Detection with Neural Networks Using Audio Features," 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSPP), Hyderabad, India, 2022, pp. 1-6, doi: 10.1109/ICICCSPP53532.2022.9862519
- [2] Govindu, Aditi, et al., "Deepfake audio detection and justification with Explainable Artificial Intelligence (XAI)," 2023.
- [3] Qais, Abu, et al., "Deepfake audio detection with neural networks using audio features," 2022 international conference on intelligent controller and computing for smart power (ICICCSPP). IEEE, 2022.
- [4] Y. Li, et al., "Cross-domain audio deepfake detection: Dataset and analysis," arXiv preprint arXiv:2404.04904, 2024.
- [5] Pandey, Ashutosh, and DeLiang Wang. "On cross-corpus generalization of deep learning based speech enhancement." *IEEE/ACM transactions on audio, speech, and language processing* 28 (2020): 2489-2499.
- [6] J. Frank, and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," arXiv preprint arXiv:2111.02813, 2021.
- [7] "Deepfake Audio Detection via MFCC Features Using Machine Learning," *IEEE Access*, vol. 134018–134028, 2022, doi: 10.1109/access.2022.3231480.
- [8] Y. Patel et al., "Deepfake Generation and Detection: Case Study and Challenges," in *IEEE Access*, vol. 11, pp. 143296-143323, 2023, doi: 10.1109/ACCESS.2023.3342107.
- [9] J. Yi, et al., "ADD 2023: Towards Audio Deepfake Detection and Analysis in the Wild," arXiv preprint arXiv:2408.04967, 2024.
- [10] H. Shin, et al., "Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods," *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

IEEE, 2024.

- [11] H. Wu, et al., "Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning," arXiv preprint arXiv:2404.15854, 2024.
- [12] Y. Xie, et al., "The codecfake dataset and countermeasures for the universal detection of deepfake audio," IEEE Transactions on Audio, Speech and Language Processing, 2025.
- [13] L. Pham, et al., "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2). IEEE, 2024.
- [14] G. Channing, et al., "Toward Robust Real-World Audio Deepfake Detection: Closing the Explainability Gap," arXiv preprint arXiv:2410.07436, 2024.
- [15] R. Reimao and V. Tzerpos, "FoR: A Dataset for Synthetic Speech Detection," 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpED), 2019, pp. 1–8, doi: 10.1109/SPED.2019.8906599.
- [16] G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.
- [17] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.
- [18] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.
- [19] G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.0486, 2017.
- [20] Hullur, A., Gurav, A., Govindu, A., & Godse, P. (2023). Deepfake Audio Detection and Justification with Explainable Artificial Intelligence (XAI). Research Square.
- [21] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2818-2826).
- [22] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR).