Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

Comparative Analysis of Deep fake Video and Audio Detection

Harish Chandra Prasad*[1], ***Dr. Arti Gautam Dinker***

[1,2]School of Information & Communication Technology Gautam Buddha University, Greater Noida, Uttar Pradesh, India.
***Corresponding Author**: harishfet77@gmail.com

Abstract

Deepfake ("deep learning + fake" = DF) refers to forged videos and audios generated using AI algorithms. While they can be a source of entertainment, they can also be harmful in various ways. Manipulating both audios and videos for harmful purposes has been a concerning issue for more than 10 years. The ability to detect these videos and audios through AI detectors is a motivating factor in achieving the best results for the project. This paper contains a comparative study of the existing research on deepfake issues, showcasing Accuracy, F1 Score, bar plots, and graphs. While exploration of deepfake videos has several approaches and datasets available, audio deepfakes have been relatively neglected. In this work, we propose the idea of joint deepfake video and audio detection using a hybrid deep learning model ensembling ResNet50 and EfficientNet B0. The dataset comprising real and synthetic voice recordings was selected from the Scene Fake repository on Kaggle. Key audio features, including Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, chroma, zero-crossing rate, and root mean square energy (RMSE), were extracted using the Librosa library. A Random Forest Classifier was trained to detect audio, while the DFDV dataset was utilized to extract facial frames from videos.

Keywords: Deepfake Detection; ResNet50; EfficientNet B0; MFCC; Random Forest Classifier; Kaggle

Introduction

In the past 10 or more years, the rapid increase in the production of deepfake videos and audios has risen by more than 80% due to the widespread use of smartphones, modern technology among today's youth, and the global availability of Wi-Fi and high-speed internet connections. Understanding the need for detecting these fraudulent videos and audios has therefore become a necessary practice in recent years. Convincing deepfakes are often purposefully created to spread false information. For instance, South Korea recently faced a deepfake "emergency" [1], which is just one example among many alarming cases reported in the news. Most of the currently available deepfake detectors focus on detecting either deepfake videos or deepfake audios using models such as ResNext CNN or LSTM [2]. The most common deep learning methods employed in this domain include Autoencoders (AEs) [3], Variational Autoencoders (VAEs) [3], and Generative Adversarial Networks (GANs) [4]. Furthermore, beyond fake images and videos, new technologies have enabled real-time human voice cloning [5, 6], a network-based speech synthesis technique capable of mimicking the voices of well-known speakers [5].

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

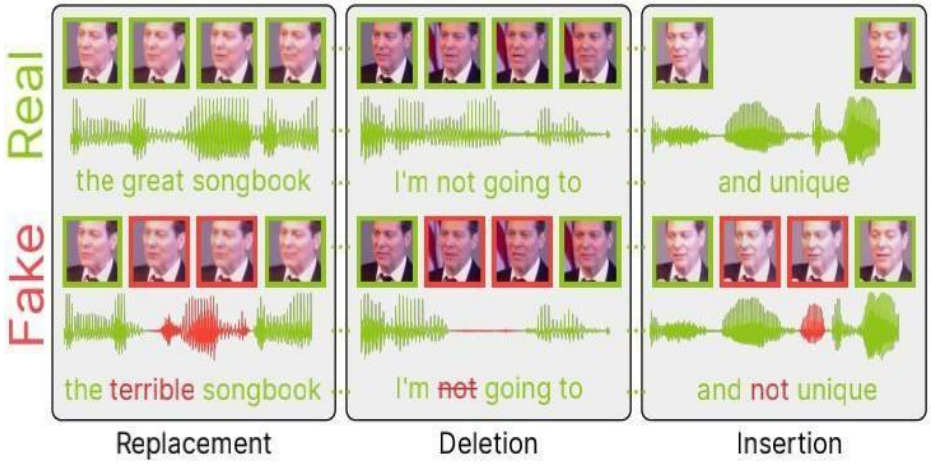

Fig. 1. Audio-Visual Manipulations in Deepfake Detection illustrate examples of word-level replacement, deletion, and insertion used to manipulate audio-visual content [16].

Additionally, other research has concentrated on detecting visual artifacts and fingerprints left by different generative frameworks [7–10], or on detecting local texture inconsistencies caused by face swapping [11, 12]. Another study leverages biometric information to identify distinct facial motion patterns in different individuals [13, 14]. More generally, human speech is characterized by a close relationship between uttered syllables and corresponding lip movements. When synchronization between these motions and audio breaks at any point, it may serve as an indicator of a deepfake. In Figures 1 and 2, the artifacts caused by face swapping or lip-sync manipulation are illustrated, showing that unnatural lip movements provide a useful signal for identifying audio-visual deepfakes [15].

Taking these challenges and prior work into consideration, we have developed a deepfake video and audio detector using advanced machine learning and deep learning techniques. The project is structured into two main components: detecting audio deepfakes and detecting video deepfakes.

For audio detection, a dataset containing both real and modified voice recordings was sourced from Kaggle's SceneFake repository. The audio files underwent preprocessing using the Librosa library, where critical features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, chroma, zero-crossing rate, and root mean square energy (RMSE) were extracted. An 80–20 random split was then applied to the dataset, and the training and testing subsets were used for implementing a Random Forest Classifier. This model was trained and evaluated using classification metrics and accuracy reports.

In parallel, the video detection component utilized the DFDV dataset to extract facial frames from videos using Haar Cascade frontal face detection. The dataset further underwent balancing checks and data augmentation to mitigate overfitting.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

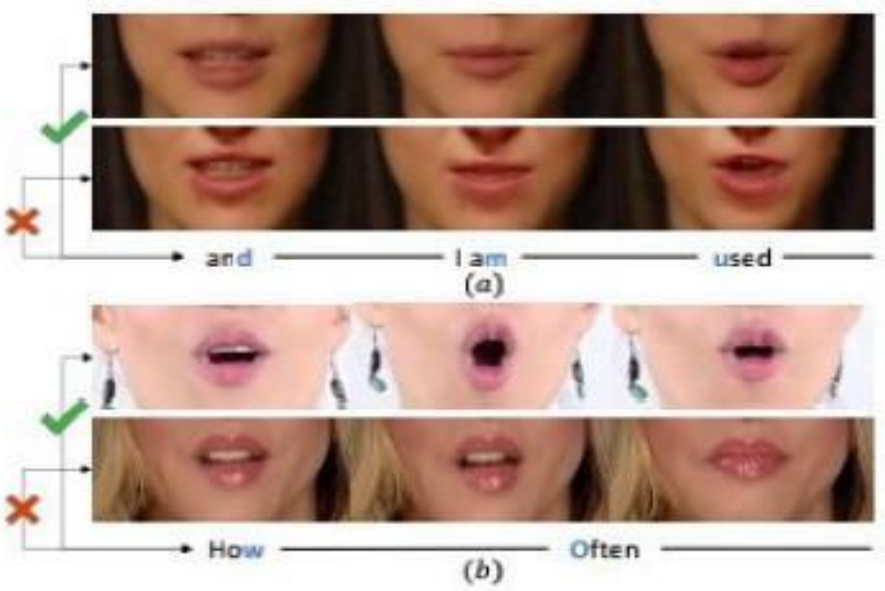

Fig. 2. Visual and Lip-Sync Inconsistencies in Deepfake Content.

Provides examples to show how changing the audio or video may throw off the synchronization patterns. The sentences are spoken by both videos, with the first row of video frames left unchanged and the second row faces wapped. There are inconsistencies between the words said and the lip movements in the altered films [15].

Subsequently, the dataset was divided into training, testing, and validation sets in a 75–12.5–12.5 ratio. A hybrid deep learning model was constructed by ensembling ResNet50 with EfficientNet B0, enabling effective feature extraction and classification. After training the model for 25 epochs, performance metrics were recorded, and the trained models were saved for deployment.

To facilitate user interaction with these models, a web application was developed using HTML, CSS, and JavaScript for the frontend, and Flask for the backend. The application provides users with the option to select either video or audio detection. Upon selection, users can upload their files, which are then processed on the server. For video uploads, individual frames are analyzed for face detection using the trained model, while audio files undergo feature extraction before being classified by the Random Forest model.

Related Work

In [2], Jadhav et al. recently proposed a deep learning-based framework for the detection of deepfake videos. In their paper, the authors used a two-stage model comprising Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). At the first stage, CNNs capture the spatial features of video frames, focusing on artifacts introduced during the generation process, such as inconsistencies in resolution and facial alignment. In the second stage, an RNN with Long Short-Term Memory (LSTM) units parses the sequence of consecutive frames to identify anomalies in motion and contextual coherence.

Their approach exploits a weakness in the current state-of-the-art GAN-based deepfake generation tools, specifically fixed-size face synthesis and affine warping, which result in detectable artifacts. The framework was trained on a massive dataset of real and manipulated videos, achieving competitive accuracy in distinguishing authentic videos from deepfakes, while maintaining a relatively simple yet effective architecture for this challenging task.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

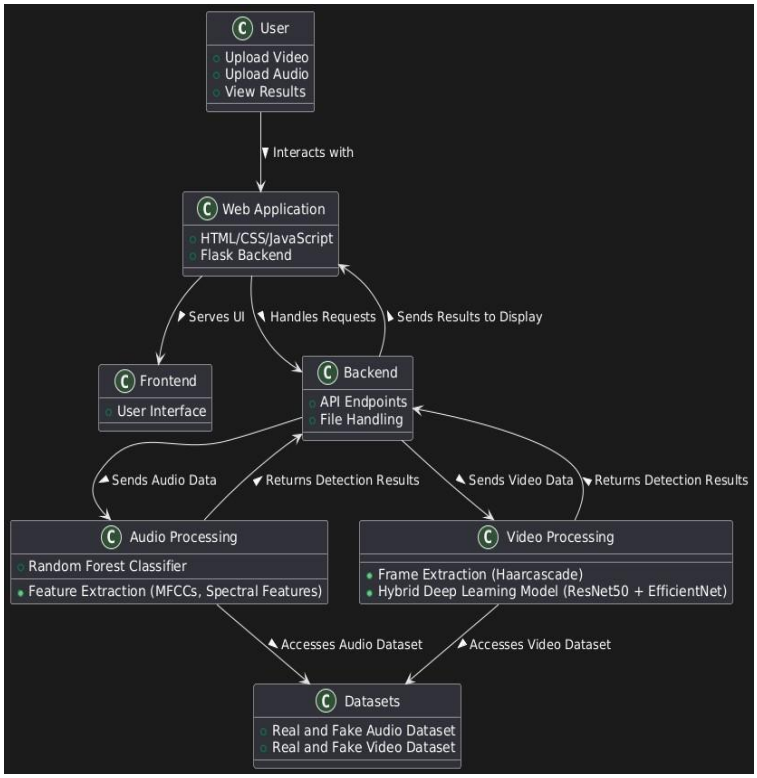

Fig. 3. Deepfake Detection System: User Flow and Scenario Analysis. A user-case diagram of the project showing step-by-step performance and user feasibility.

In [8], Sheng-Yu Wang et al. proposed a universal forensic detector capable of distinguishing real images from CNN-generated ones, regardless of the underlying architecture or dataset. This exposed a systematic weakness—often described as "fingerprints"—in CNN-generated images, which allowed the authors to build a classifier trained on one generative model that generalized well across others. To achieve this, they aggregated a broad dataset termed ForenSynths, comprising images synthesized with 11 different models, including ProGAN, StyleGAN, BigGAN, CycleGAN, and Deepfakes. The binary classifier was initially trained on ProGAN-generated images and real images to recognize these CNN-related artifacts. Interestingly, the use of blurring and JPEG compression as augmentation techniques significantly improved the model's generalization ability.

To further test robustness, the model was evaluated on images generated by unseen architectures such as StyleGAN2, and the results showed that it could successfully identify them. This highlights the importance of dataset diversity and augmentation in enhancing generalization. The findings revealed that images generated by existing CNN architectures still carry noticeable artifacts, making them detectable. However, the authors cautioned that future GAN developments may eliminate such artifacts, thereby making detection significantly more challenging.

In [15], Zhou et al. suggested a joint audio-visual deepfake detection framework that integrates cues from both audio and video modalities. Since many deepfake videos exhibit inconsistencies between the visual content and the associated audio, the framework employs a deep neural network to process concurrent data streams. This allows it to capture temporal patterns and synchronization mismatches that frequently arise

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

when audio and video are generated independently. The approach relies on feature extraction techniques that account for subtle artifacts in both modalities, which are characteristic of manipulation.

Testing against large-scale datasets demonstrated that such multimodal methods outperform single-modality approaches by effectively exploiting cross-modal inconsistencies. This joint detection framework represents a significant advancement in combating deepfakes, particularly in addressing the temporal inconsistencies introduced during their generation.

3. Dataset and Methods

This section discusses the datasets and models used in the development of our deepfake detector, as well as for the comparative analysis with existing deepfake detection methods. Two primary datasets were employed: the SceneFake repository on Kaggle for audio deepfake detection, and the Deepfake Detection Challenge (DFDC) dataset on Kaggle for video deepfake detection.

Dataset Description

The datasets used for this project include the SceneFake repository on Kaggle, which was utilized for deepfake detection in audio and for training the model to produce accurate results, and the Deepfake Detection Challenge (DFDC) dataset on Kaggle, which was employed for deepfake detection in videos.

- **SceneFake Repository on Kaggle:**This dataset was compiled from the acoustic scene dataset of the DCASE 2022 competition and the Logical Access (LA) dataset from ASVspoof 2019. The LA dataset contains real and fake voices, as well as synthetic speech, across its three subsets: training, development, and test. The acoustic scene dataset consists of 64 hours of audio clips, each 10 seconds long, recorded across 10 different acoustic scenes [17]. URL: https://www.kaggle.com/datasets/mohammed- abdeldayem/scenefake
- **Deepfake Detection Challenge (DFDC) on Kaggle:** This dataset consists of natural, authentic videos with and without deepfakes, designed to closely resemble the format and structure of the Training and Public Validation/Test sets. It provides a large-scale benchmark for deepfake video detection models, and the challenge is evaluated using log loss as the scoring metric [18].

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \qquad (1)$$

Where:

- nnn = the number of videos being predicted
- y^i\hat{y}_iy^i = the predicted probability of the video being **FAKE**
- yi=1y_i = 1yi=1 if the video is **FAKE**, 000 if **REAL**
- log⁡f0(⋅)\log(\cdot)log(⋅) = the natural logarithm (base eee)

A smaller log loss is better. When logarithm is used, very strong penalties are imposed for being both highly confident and wrong. Worst case, your error score will grow to infinity when you predict something to be true when it isn't. This is prevented by limiting forecasts by a very small value away from extremes. URL: https://www.kaggle.com/c/deepfake-detection-challenge/overview

Models and Methods Used

The models and methods used in extensive researches are ResNext CNN for feature extraction and LSTM for sequence processing in videos [2]. The model that we have used is ResNet 50 and EfficientNet B0. Deepfake Audios were detected via MFCC Features using Machine Learning.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

RESNEXT CNN FOR FEATURE EXTRACTION: It is a residual block, named as ResNext Block, in the CNN architecture of ResNext, which is also similar to an Inception module where summation of set of transformations is done with a technique called "split-transform-merge" or branched routes inside a single module. Along with the depth and width, another dimension, known as the cardinality C (number of transformations), plays a significant role while comparing the ResNext Block with that of a Residual Block. A set of compounded transformations can be written as:

$$F(x) = \sum_{i=1}^{\Sigma} T_i(x)$$

Where: Ti(x)=T_i(x) =Ti(x)= arbitrary function. Analogous to a simple neuron, TiT_iTi should project xxx into an (optionally low-dimensional) embedding and then transform it.

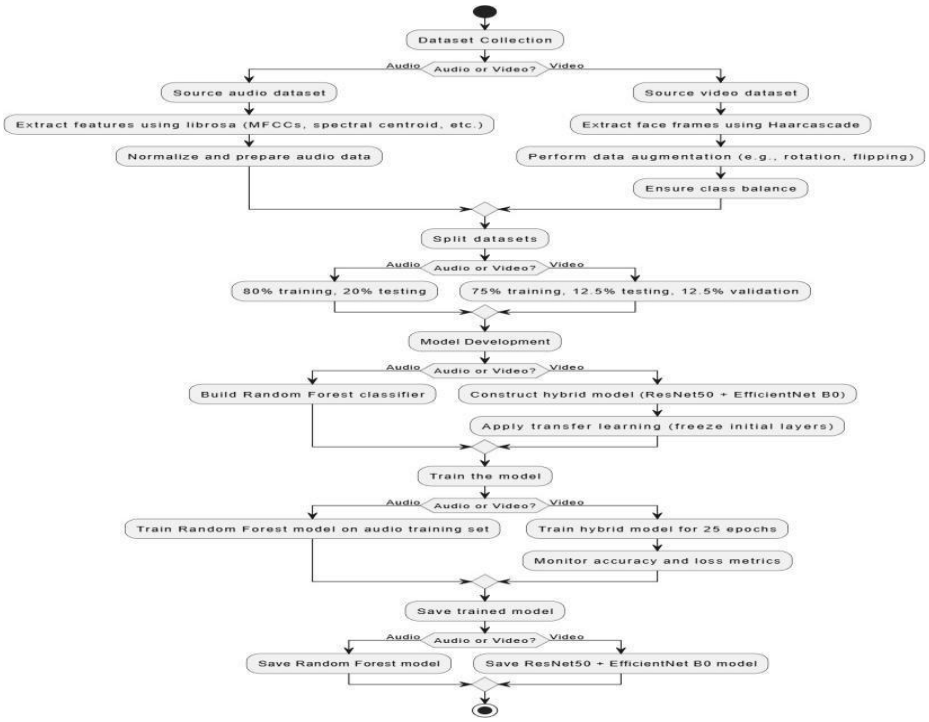

Fig. 4. Training workflow showing data processing, feature extraction, and model training steps for audio and video detection.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

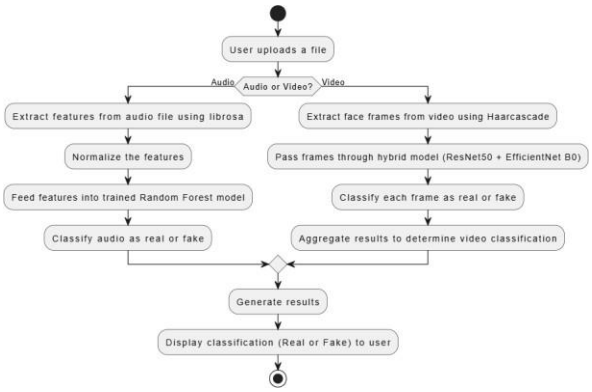

Fig. 5. Testing workflow showing how uploaded user content is processed and classified as real or fake.

- **LSTM FOR SEQUENCE PROCESSING:** Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that can store and recall information over time. LSTMs are designed to handle sequential data, such as video streams, sentences, and time series signals.
- **RESNET50:** ResNet-50 is one of the models in the family of ResNet developed to solve problems in training deep neural networks. It was developed by researchers in Microsoft Research Asia and has found widespread popularity because of its depth and high efficacy in tasks related to image classification.
- **EFFICIENTNETB0:** EfficientNetB0 is a CNN architecture that is part of the EfficientNet family. It is the base model and the smallest version in the EfficientNet family, serving as a foundation for larger models with progressively higher capacities.
- **MFCC**: MFCC (Mel-Frequency Cepstral Coefficients) is a widely used method for feature extraction in audio signal processing. It converts audio signals to a representation that best illustrates their most important features for detection or classification.

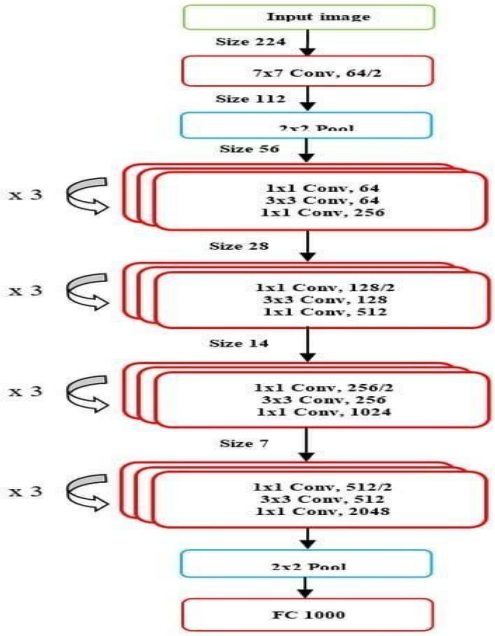

Fig. 6. Architecture of ResNet50 used in the model for deepfake video detection.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

Comparative Analysis and Result

This section focuses on the comparison of the earlier present models and the model that we have proposed for the deepfake detection for both videos and audios.

- **Earlier model (deepfake detection for video):** ResNext and LSTM. It divides the video into frames and takes out its features using ResNext CNN along with RNN using LSTM. Such an approach could be used to detect visible artifacts between fake and real videos introduced by GANs [2]. This model requires a large amount of labelled data for effective training and has high computational cost.
- **Proposed model (deepfake detection for video):** Our model combines ResNet50 and EfficientNet B0, both pre-trained on ImageNet, as feature extractors. It uses these models in a frozen state, leveraging their robust feature extraction capabilities. This significantly reduces training time and is more efficient for smaller datasets.
- **Earlier model (deepfake detection for audio):** Previous approaches employ complex preprocessing including MFCC extraction and dimensionality reduction using PCA. They often require significant computational resources for deep-learning models like VGG-16, making them less practical for real-time applications [19].
- **Proposed model (deepfake detection for audio):** Our model utilizes straightforward preprocessing with features like MFCC, spectral centroid, and RMSE. The Random Forest classifier is computationally efficient and easier to train compared to deep learning models. Our implementation achieves 92% accuracy while maintaining simplicity and interpretability.

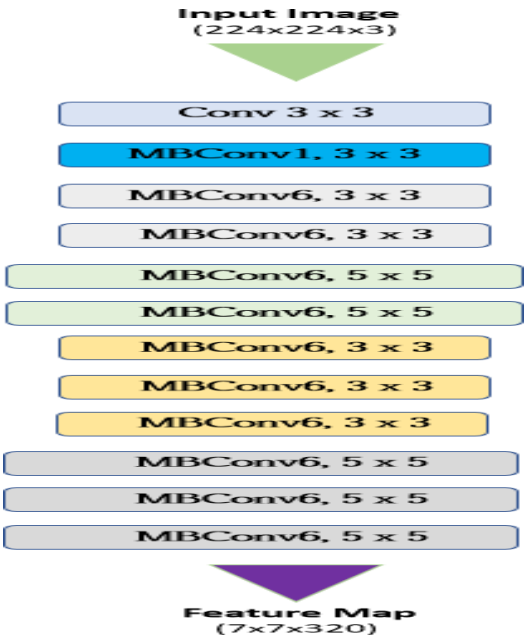

Fig. 7. Architecture of EfficientNetB0 used in our hybrid model for deepfake video detection.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

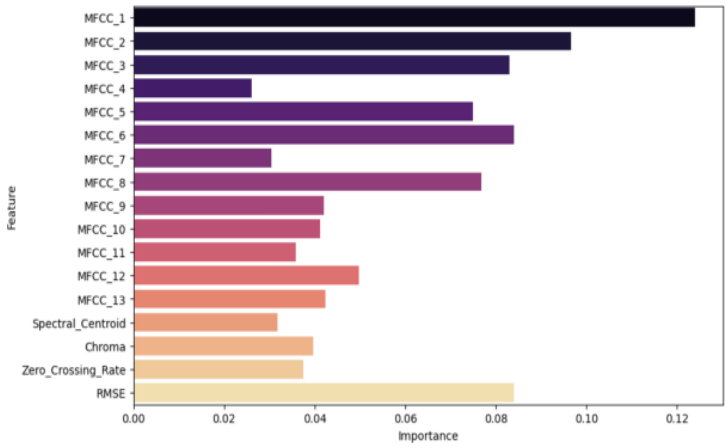

Fig. 8. Feature importance from Random Forest Classifier, with MFCCs (particularly MFCC_1, MFCC_2, and MFCC_3) being the most influential.

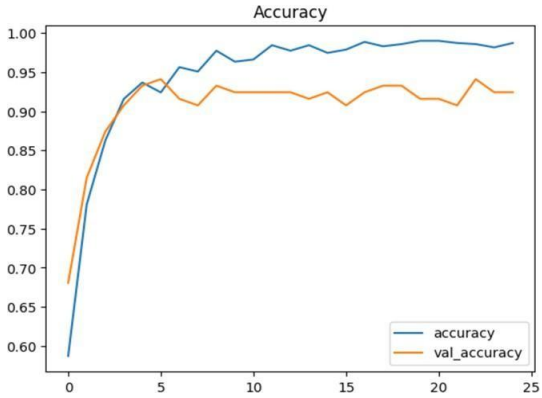

Fig. 9. Model accuracy over 25 epochs showing training (blue) and validation (orange) performance.

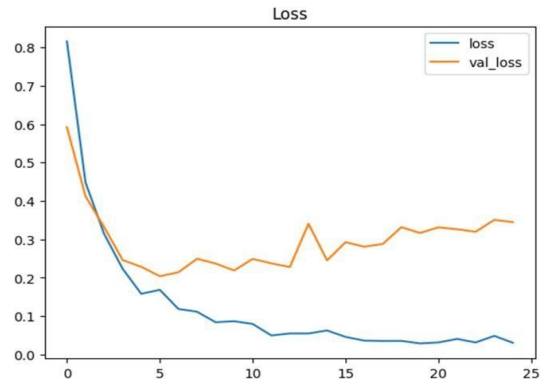

Fig. 10. Model loss over 25 epochs showing training (blue) and validation (orange) trend.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

Conclusion and Scope of Future Work

In this paper, we addressed the increasing issue of deepfake videos and audios by proposing a novel approach for detecting deepfakes in both modalities. We provided comparative research on existing models and methods, analyzing various datasets to achieve high accuracy with our proposed models. Our dual-component system includes a user-interactive website that facilitates seamless upload and analysis of videos and audios.

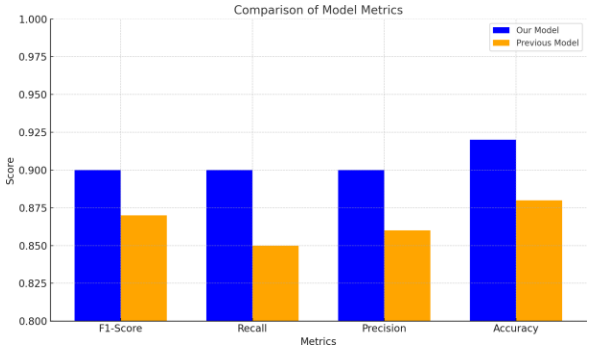

Fig. 11. Performance comparison between Previous Model (orange) and Our Model (blue) across key metrics.

Our solution creates a reliable system for identifying deepfake content by combining advanced machine learning and deep learning algorithms. The ensemble approach using ResNet50 and EfficientNetB0 for video detection, combined with Random Forest and MFCC features for audio detection, achieves 92% accuracy. This implementation balances performance with practical considerations like computational cost, processing time, and scalability. Current deepfake detection systems still face challenges with generalization, real-time processing, and user accessibility. Our approach makes progress in these areas through efficient architectures and an intuitive user interface. Future work will focus on improving cross-domain generalization and reducing computational requirements while maintaining high detection accuracy.

6. Acknowledgement

We express our gratitude to our mentor for her guidance and support during our research. Her knowledge and direction led us through this project. We also thank our parents and friends who supported us throughout and kept us motivated.

References

[1] Detecting Deepfakes: Can You Trust What You See? YouTube, 2024. [Online]. Available: https://www.youtube.com/watch?v=wJYY0ngBwT0. Accessed: Nov. 25, 2024.

[2] A. Jadhav, A. Patange, J. Patel, H. Patil, and M. Mahajan, "Deepfake Video Detection Using Neural Networks," IJSRD - International Journal for Scientific Research and Development, vol. 8, no. 1, pp. 1016–1022, 2020.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning Internal Representations by Error Propagation. Technical Report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

Received: 2025-08-28
Accepted: 2025-08-31
Published Online: 2025-09-20
DOI: 10.70454/JRIST.2025.10102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926

[5] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," arXiv preprint arXiv:1806.04558, 2019.

[6] A. Smith, "Deepfakes are the most dangerous crime of the future, researchers," The Independent, 2020. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/news/deepfakes-dangerous-crime-artificial-intelligence-a9655821.html. Accessed: May 31, 2021.

[7] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proceedings of the International Conference on Computer Vision (ICCV), 2019.

[8] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot...for now," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[9] N. Yu, T. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in Proceedings of the International Conference on Computer Vision (ICCV), 2019.

[10] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[11] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[12] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[13] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[14] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Protecting world leaders against deep fakes," in Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 2020.

[15] H. B. Zhuo, Y. L. Zhang, and N. Wang, "Generative Adversarial Networks in Biometric Presentation Attack Detection: A Comprehensive Survey," IEEE Access, vol. 8, pp. 19536–19555, Feb. 2020.

[16] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset," in Proceedings of the 32nd ACM International Conference on Multimedia, 2024.

[17] "SceneFake," Kaggle, Apr. 20, 2024. Available: https://www.kaggle.com/datasets/mohammedabdeldayem/scenefake.

[18] "Deepfake Detection Challenge," Kaggle. Available: https://www.kaggle.com/c/deepfake-detection-challenge/overview.

[19] A. Hamza, A. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake Audio Detection via MFCC Features Using Machine Learning," IEEE Access, vol. 10, Dec. 2022.