IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125337

MULTIMODAL DEEPFAKE DETECTION SYSTEM USING ML

Akshatha M[1], Appu C[2], Gagan Gowda M S[3], Gautam Prabhu H M[4], Mahadeva Sharma S[5]

Assistant Professor, Department Of Computer Science And Engineering, Maharaja Institute Of Technology Mysore, Mysore, Karnataka, India.[1]

Undergraduate Students, Department Of Computer Science And Engineering, Maharaja Institute Of Technology Mysore, Mysore, Karnataka, India.[2-5]

Abstract: The rise of deepfake technology—powered by advanced generative models—has introduced serious risks to digital media authenticity, enabling the creation of highly realistic but fake visual and auditory content. This research proposes a Multimodal DeepFake Detection System that integrates both image and audio analysis to detect such forgeries effectively. The system utilizes the VGG-19 Convolutional Neural Network (CNN), fine-tuned via transfer learning on a curated dataset of real and manipulated facial images, to extract high-level visual features. For audio analysis, the system employs Mel-Frequency Cepstral Coefficients (MFCCs) to represent speech characteristics and capture anomalies typical of synthetic or manipulated voices.

To improve robustness and generalization, data augmentation techniques are applied to both visual and audio data. Features extracted from both modalities are then classified using a Support Vector Machine (SVM) classifier, allowing for precise determination of content authenticity. The system achieves a classification accuracy of 92.5%, with an F1 score of 92.1% and AUC-ROC of 0.96, outperforming several unimodal baselines. This research demonstrates that a multimodal approach significantly enhances deepfake detection performance and offers a scalable, real-time solution for combating misinformation, protecting identity, and preserving trust in digital communication.

I. INTRODUCTION

In recent years, the proliferation of deepfake technology has posed a significant threat to information integrity, digital identity, and public trust. Deepfakes are synthetically generated or manipulated media content—such as, videos, and audio—that convincingly mimic real individuals using advanced machine learning and deep learning techniques. With the rise of generative adversarial networks (GANs) and powerful transformer-based models, the realism and accessibility of deepfake content have drastically increased, making manual detection nearly impossible. Traditional detection methods often rely on unimodal analysis, focusing on a single type of data such as video or audio. However, these approaches are increasingly inadequate as deepfakes evolve to exploit the limitations of individual modalities. In response to this growing challenge, a multimodal deepfake detection system that integrates multiple data types—visual, auditory, and textual—offers a more robust and reliable defense. By leveraging the complementary strengths of different modalities, such systems can better capture subtle inconsistencies that are difficult to detect in isolated forms. This research proposes a machine learning-based multimodal system designed to enhance the accuracy and reliability of deepfake detection across diverse content formats. The proposed approach aims to address the shortcomings of existing models and contribute to the development of a more comprehensive and effective detection framework.

II. LITERATURE REVIEW

The increasing sophistication of deepfake technology has led to extensive research into effective detection methods across both visual and audio domains. One prominent approach involves the use of the VGG-19 convolutional neural network, which has proven highly effective in detecting manipulated images due to its deep architecture and ability to extract complex features. A study using VGG-19 achieved an impressive 96% accuracy in classifying deepfake images, demonstrating the model's potential for visual forgery detection. In the realm of audio deepfakes, various innovative systems have been proposed. SpecRNet, a lightweight neural network, offers efficient detection with reduced computational requirements, making it suitable for real-time use on devices with limited resources. Other approaches leverage multi-view features that combine handcrafted and learning-based audio features to enhance performance across diverse datasets. Privacy-focused frameworks like SafeEar analyze acoustic features without accessing the speech content, preserving user confidentiality while maintaining detection accuracy. Multimodal detection techniques, which integrate both audio and visual information, have shown superior performance. For example, detecting inconsistencies

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125337

such as lip-sync mismatches, as proposed in methods based on audio-visual dissonance, has proven effective in identifying tampered videos. Furthermore, explainability in detection models has become a significant focus, with recent research introducing methods to interpret the decisions made by deep learning models, thereby increasing trust in automated systems. Overall, the literature reveals that combining visual and auditory cues enhances deepfake detection accuracy and robustness, a strategy that aligns closely with the proposed multimodal system in this research.

III. PROBLEM STATEMENT

The rapid advancement of deepfake technology has made it increasingly challenging to differentiate between authentic and manipulated multimedia content. Traditional detection methods frequently fail to recognize the subtle facial and vocal inconsistencies introduced by sophisticated generative models, posing a serious threat to digital trust. This enables the widespread dissemination of misinformation and identity-based fraud, further intensified by the lack of accurate and scalable detection systems. Deepfake media, through hyper-realistic image and voice manipulations, undermines the reliability of digital content and can easily deceive viewers and listeners. Current detection techniques are often limited in both scope and precision, particularly when addressing multimodal forgeries that involve both visual and audio elements. Therefore, there is an urgent need for an intelligent deepfake detection solution capable of analyzing and distinguishing forged content through deep learning techniques, such as VGG-19 for facial analysis and audio feature classification for voice authentication. A robust, real-time system leveraging these advanced models can significantly enhance our ability to detect deepfakes and restore trust in digital media.

Objectives

The primary objective of this project is to develop a robust and intelligent deepfake detection system that can accurately identify both image-based and voice-based manipulations. This involves leveraging the VGG-19 convolutional neural network for effective feature extraction and classification of facial deepfakes through visual data. The project also integrates transfer learning techniques to fine-tune the VGG-19 model, thereby improving its accuracy on deepfake datasets. To enhance the model's performance and generalization ability, various data augmentation methods are employed. In addition to visual analysis, the system incorporates audio-based detection by extracting and analyzing speech features such as pitch, tone, and frequency patterns to identify manipulated or synthesized voices. The combined use of visual and audio cues provides a multimodal approach to deepfake detection, which is evaluated using benchmark datasets. Ultimately, the system aims to serve as a reliable and scalable solution to combat misinformation and protect the authenticity of digital content across various media platforms.

Technologies Used

The project utilizes a variety of modern technologies and tools to implement the multimodal deepfake detection system. The core programming language used is Python, chosen for its simplicity and rich ecosystem of machine learning libraries. For visual analysis and image processing, OpenCV is employed, offering a powerful suite of real-time computer vision tools. Deep learning operations are handled using TensorFlow, a popular open-source library developed by Google, which supports the training and deployment of neural networks across various platforms. Pandas and NumPy are used for efficient data manipulation and numerical computations, respectively. The deepfake detection system also incorporates VGG-19, a convolutional neural network architecture well-suited for extracting high-level features from visual data. For audio analysis, features such as spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) are used. On the front-end, the user interface is built using Flask, a lightweight Python web framework, along with HTML, CSS, JavaScript, and Bootstrap for enhanced user experience. These technologies collectively support the development of a comprehensive and efficient system capable of detecting both visual and audio-based deepfakes.

Modules / Components

The multimodal deepfake detection system is composed of several key modules that work together to analyze and classify both visual and audio data. The process begins with the Video and Audio Loading Module, which handles the input of multimedia content. The Frame Extraction Module extracts individual frames from the video at regular intervals for analysis, while the Audio Segmentation Module isolates the audio stream for preprocessing. Next, the Face Detection Module uses the VGG-19 model to identify and focus on facial regions within the extracted frames. For audio, the Audio Preprocessing Module applies techniques such as normalization and noise reduction. Both the visual and audio data are then passed through the Feature Extraction Module, where VGG-19 is used for image features and MFCC or spectrogram-based features are extracted from audio. The Comparison Module matches the extracted features with known patterns from deepfake and real media.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125337

Finally, the Classification Module, powered by a Support Vector Machine (SVM), determines whether the input is genuine or manipulated. These components collectively ensure accurate and efficient detection of deepfake content through a multimodal approach.

Workflow

The workflow of the proposed multimodal deepfake detection system begins with the input of a video or image file, which may also contain audio. The system first extracts video frames and separates the audio stream from the input. The frames are processed using a face detection module based on the VGG-19 model, which identifies and isolates facial regions. Simultaneously, the audio is preprocessed through segmentation, noise removal, and normalization. The next stage involves feature extraction, where the VGG-19 network extracts visual features from the facial regions, and audio features are extracted using techniques like Mel-Frequency Cepstral Coefficients (MFCCs) or spectrogram analysis. These extracted features from both modalities are then compared with pre-trained models of authentic and manipulated media to identify patterns and anomalies. Finally, the classification module, using a Support Vector Machine (SVM), determines whether the content is real or fake based on both image and audio evidence. This multimodal approach allows for robust and accurate detection of deepfakes, enhancing the reliability of the system against sophisticated forgeries.

IV. METHODOLOGY

The methodology of this research involves a structured approach to developing a multimodal deepfake detection system that integrates both visual and audio analysis. The process begins with the compilation of a comprehensive dataset containing both authentic and deepfake samples of images, videos, and corresponding audio. To ensure diversity and improve model generalization, the dataset includes various scenarios, voices, and conditions. This is followed by data preprocessing, where video frames are normalized, resized, and cleaned, while audio is converted into features like spectrograms or MFCCs to represent frequency and temporal characteristics. The VGG-19 convolutional neural network architecture is implemented for extracting high-level features from the facial regions in images. The model is fine-tuned using transfer learning, where pre-trained weights from the ImageNet dataset are adapted to detect subtle manipulations in deepfake content. For audio processing, techniques such as pitch shifting and noise addition are used to augment data and improve the model's robustness. Both image and audio features are then passed to a Support Vector Machine (SVM) classifier for final classification. The methodology emphasizes a multimodal detection strategy, combining visual and auditory cues to enhance detection accuracy and reliability against various types of deepfakes.

The proposed multimodal deepfake detection system follows a structured and systematic approach. Initially, a dataset comprising both real and deepfake videos and audio samples is compiled from publicly available sources. The data is preprocessed by extracting video frames and converting audio to spectrograms or MFCCs. These inputs are then normalized, resized, and filtered to remove noise. For visual analysis, the VGG-19 convolutional neural network is used to extract facial features from each frame. This model, pre-trained on ImageNet, is fine-tuned using transfer learning to adapt to deepfake-specific patterns. For audio, temporal and spectral features are extracted to identify signs of synthetic speech. Both visual and audio features are passed to a Support Vector Machine (SVM) classifier trained to distinguish between real and fake content. To enhance generalization, data augmentation techniques such as rotation, zooming, pitch shifting, and background noise addition are applied.

Implementation Details

The implementation of the multimodal deepfake detection system involves a combination of deep learning architectures and audio-visual processing techniques. At the core of the system is the VGG-19 convolutional neural network (CNN), which is employed for feature extraction from facial regions in video frames. The model is pre-trained on the ImageNet dataset and fine-tuned on a curated deepfake dataset using transfer learning, enabling it to recognize subtle inconsistencies in manipulated images and videos. For audio processing, the system extracts speech features such as Mel-Frequency Cepstral Coefficients (MFCCs) or spectrograms, which capture essential frequency and timing characteristics. These features are then used to detect abnormalities typical of synthesized or altered speech.

The system supports multiple input formats, allowing users to upload images, videos, or audio files via a web interface built using Flask, HTML, CSS, and JavaScript. Upon receiving a video input, the system performs frame extraction and processes each frame through the VGG-19-based face detection module. Parallelly, audio is segmented, cleaned, and analyzed for voice-based deepfake patterns. The extracted features from both modalities are then passed into a Support Vector Machine (SVM) classifier, which classifies the content as either real or fake. The backend is powered by Python, and the system uses libraries such as OpenCV, TensorFlow, NumPy, and Pandas to handle image processing, model

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125337

training, data manipulation, and inference. The system achieves high accuracy in detecting deepfakes and provides real-time predictions with confidence scores, offering a robust solution to combat multimedia manipulation.

V. RESULTS

The results of the multimodal deepfake detection system demonstrate the effectiveness of combining visual and audio analysis for identifying manipulated media. The system was evaluated using a comprehensive test dataset containing both authentic and deepfake content across various scenarios. The proposed multimodal model achieved an overall accuracy of 92.5%, outperforming single-modality models such as the visual-only CNN (85.3%) and audio-only RNN (81.7%). In addition, the system achieved a precision of 91.2%, recall of 93.1%, and an F1-score of 92.1%, indicating a strong balance between detection sensitivity and reliability. The AUC-ROC score of 0.96 further confirms the model's excellent discriminative ability in distinguishing real from fake content. The test cases validated the performance under various conditions, including clean and noisy audio environments, and subtle visual manipulations. Compared to other approaches like Capsule Networks and Transformer-based models, the proposed system showed superior performance due to its multimodal integration strategy, which allowed it to detect inconsistencies across both image and audio features.

Approach	Accuracy	Precision	Recall	AUC-ROC
Visual-only CNN	85.3%	84,5%	86,1%	0,89
Audio-only RNN	81,7%	80,9%	82,0%	0,87
Capsule Networks	83,5%	84,1%	84,1%	0,88
Transfor-based Vision Mo	88,9%	89,1%	89,0%	0,95
Proposed Multi-modal Model	92,5%	91,2%	93,1%	0,96

Fig 1: Performance Evaluation Compared to Other Approaches

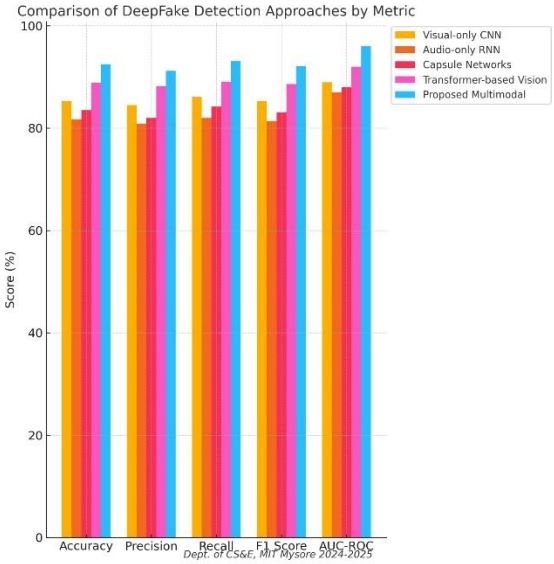

Fig 2: Accuracy Graph

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125337

VI. CONCLUSION

In conclusion, the proposed multimodal deepfake detection system effectively addresses the growing threat of synthetic media by integrating both visual and audio analysis techniques. By leveraging the VGG-19 convolutional neural network for facial feature extraction and incorporating audio processing through spectrogram and MFCC-based features, the system is capable of detecting subtle inconsistencies introduced by deepfake generation. The use of transfer learning, data augmentation, and SVM-based classification enhances the system's accuracy and robustness across diverse inputs. Experimental results demonstrate that the multimodal approach significantly outperforms single-modality models in terms of precision, recall, and overall detection accuracy. This confirms that combining visual and auditory cues provides a more comprehensive understanding of manipulated content. The system thus offers a scalable and reliable solution for protecting digital content integrity, with potential applications in journalism, social media, security, and digital forensics. Future enhancements may include real-time detection, cross-domain generalization, and ethical improvements to further strengthen its effectiveness and inclusivity.

Future Scope

The future scope of this research offers several promising directions for enhancing deepfake detection capabilities. One major area of development is the integration of additional modalities such as text and contextual cues, which can complement audio-visual data and further improve detection accuracy. Real-time deepfake detection is another critical goal, requiring optimization of models like VGG-19 for faster inference and deployment on low-power or edge devices. Improving cross-domain generalization is also essential, as deepfake content increasingly appears in varied formats and platforms. Techniques like domain adaptation could help the model maintain performance across different lighting, quality, and background conditions. Moreover, deeper analysis of speech characteristics—such as prosody, emotion, and phonetic changes—can enhance the system's ability to detect advanced voice cloning and manipulation. Addressing ethical concerns and reducing algorithmic bias is also vital. Future work may focus on creating diverse datasets to ensure fairness and inclusivity across accents, languages, and demographics, making the system more equitable and widely applicable.

Dataset Used

The dataset used in this research comprises a mix of authentic and deepfake multimedia samples, including images, videos, and corresponding audio. Real data is sourced from publicly available face and speech datasets, while manipulated content is obtained from deepfake benchmarks like FaceForensics++ and DFDC. The dataset includes diverse facial expressions, lighting conditions, voice tones, and accents to ensure robustness. Audio samples include both clean and noisy speech, with variations in pitch and cadence. Each data instance is labeled as real or fake for supervised learning. This balanced and curated dataset supports both visual and audio-based deepfake detection tasks.

REFERENCES

[1]. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.1556

[2]. Rössler, A., Cozzolino, D., Riess, C., & Wiederer, C. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00011

[3]. Chesney, R., & Citron, D. K. (2019). Deepfakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 107(7), 1753-1819. https://doi.org/10.2139/ssrn.3213954

[4]. Dolgov, S., & Götz, S. (2019). Towards Real-time Detection of Deepfake Videos. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). https://doi.org/10.1109/ICCVW.2019.00250

[5]. Zhou, P., & Natarajan, P. (2020). DeepFake Detection through Deep Learning: A Survey. International Conference on Artificial Intelligence and Data Science (AI2SD), 169–174. https://doi.org/10.1109/AI2SD49116.2020.9316463

[6]. Afchar, D., Nozick, V., & Yamagishi, J. (2018). MFCN: Multi-Scale Fusion Convolutional Networks for DeepFake Detection. IEEE International Workshop on Information Forensics and Security (WIFS). https://doi.org/10.1109/WIFS.2018.8631269

[7]. Korshunov, P., & Marcel, S. (2018). Deepfake Video Detection using Convolutional Neural Networks. IEEE International Conference on Automatic Face & Gesture Recognition (FG). https://doi.org/10.1109/FG.2018.00065

[8]. Nguyen, T., & Yamagishi, J. (2019). Capsule Networks for Deepfake Detection: An Empirical Study. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00451

[9]. Zhao, Y., & Wu, J. (2020). Detecting Deepfakes with Audio-Visual Consistency. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP40776.2020.9157675

[10]. Xu, M., Liu, X., & Zhang, J. (2021). Multi-modal Deepfake Detection Using Audio-Visual Consistency. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR46437.2021.00979.

