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 8, August 2025

DOI: 10.17148/IARJSET.2025.12835

Deepfake Creation and Detection of Multimedia Data Using Machine Learning

Varun Kumar G[1], Dr. Harish G[2], Dr. Smitha shekar B[3]

Student, Computer Science and Engineering M.Tech, Dr Ambedkar Institution of Technology, Bengaluru, India[1]

Associate Prof, Computer Science and Engineering M.Tech, Dr Ambedkar Institution of Technology, Bengaluru, India[2]

Professor, Dept. Of Computer Science & Engineering, Dr. Ambedkar Institute of Technology, Bengaluru, Karnataka, India[3]

Abstract: This paper presents a multi-modal deepfake detection system capable of analyzing images, videos, and audio for signs of manipulation. The project addresses the limitations of single-modality detection systems by combining various machine learning and deep learning techniques to identify complex forgeries. The proposed system utilizes CNN and ResNet for detecting spatial inconsistencies in images, an LSTM on frame sequences for identifying temporal anomalies in videos, and a combination of Librosa and a Random Forest classifier for detecting synthetic audio patterns. The system aims to be resilient against adversarial attacks and provide accurate, real-time results through a user-friendly, Flask-based web interface. The anticipated outcomes include superior detection accuracy, balanced precision and recall, and enhanced generalization across diverse datasets.

Keywords: Deepfake Detection, Multi-modal Detection, Machine Learning, Deep Learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM).

I. INTRODUCTION

The rapid advancement of deepfake technology, which leverages deep learning models to create highly realistic synthetic media, poses a significant and growing threat to society. These manipulated digital assets, including images, videos, and audio, are often indistinguishable from genuine content to the human eye and ear. The proliferation of deepfakes has severe implications, from the spread of disinformation and political propaganda to the potential for identity theft, harassment, and fraud. As the accessibility of these sophisticated generation tools increases, the need for equally powerful and robust detection systems has become a critical area of research and development.

A major challenge in the field of deepfake detection lies in the rapid pace at which generation techniques are evolving, often outpacing the capabilities of existing detection methods. Many current systems are limited to a single modality, focusing exclusively on either images, videos, or audio. This specialized approach makes them vulnerable to more complex forgeries and results in poor generalization across diverse deepfake methods. Furthermore, these single-modality systems often struggle with high false positive or negative rates and typically lack an integrated, user-friendly platform for real-time analysis, creating a gap between the generation and detection capabilities.

This research paper proposes a solution to these challenges by presenting a comprehensive, multi-modal deepfake detection system. Our objective is to develop a platform that can simultaneously analyze images, video, and audio for signs of manipulation, combining machine learning and deep learning techniques to identify subtle, multi-faceted anomalies. The system is designed to be resilient against adversarial attacks and aims to provide superior accuracy and balanced precision and recall. Ultimately, this work seeks to contribute to the ongoing effort to combat misinformation, protect individual privacy, and provide a reliable tool for verifying the authenticity of digital media.

II. EXISTING SYSTEM

Simple Swap (SimSwap): A program that swaps a person's face from one image onto another without changing their facial expressions or how they are looking. It uses a special part of its code to put the new face on the target and another part to make sure the original facial features are kept.

Identity-aware Dynamic Network (IDN): A small, fast program designed for face swapping on mobile phones. It can change its settings on the fly to swap faces, even for people it hasn't seen before.

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 8, August 2025

DOI: 10.17148/IARJSET.2025.12835

MRI-GAN: A system that uses a type of AI called a Generative Adversarial Network (GAN) to find small differences in images to tell if a video is fake.
DeepVision: This method detects deepfakes by checking for unusual patterns in how often a person blinks, which helps it find things that pixel-based programs might miss.

Dual Attention Forgery Detection Network (DAFDN): A network that uses special attention mechanisms to find peculiar marks left behind by image warping.
Hybrid Face Forensics Framework: A system that uses a mix of different detection methods to more accurately find manipulated content in videos, even when the videos are highly compressed.
Web-based platforms: Some platforms allow users to upload a video to easily check if it's real or fake. These systems often use a combination of different AI models, like ResNeXt and LSTM, to analyze the video frames.
Recognition Pipeline: A two-part process that uses a CNN to find features in individual frames and a Recurrent Neural Network (RNN) to spot strange, inconsistent frames that are a sign of face-swapping.

III. PROPOSED SYSTEM

The proposed system is a multi-modal deepfake detection platform that can analyze images, videos, and audio for signs of manipulation. It is designed to overcome the limitations of single-modality detection systems by combining various machine learning and deep learning techniques.

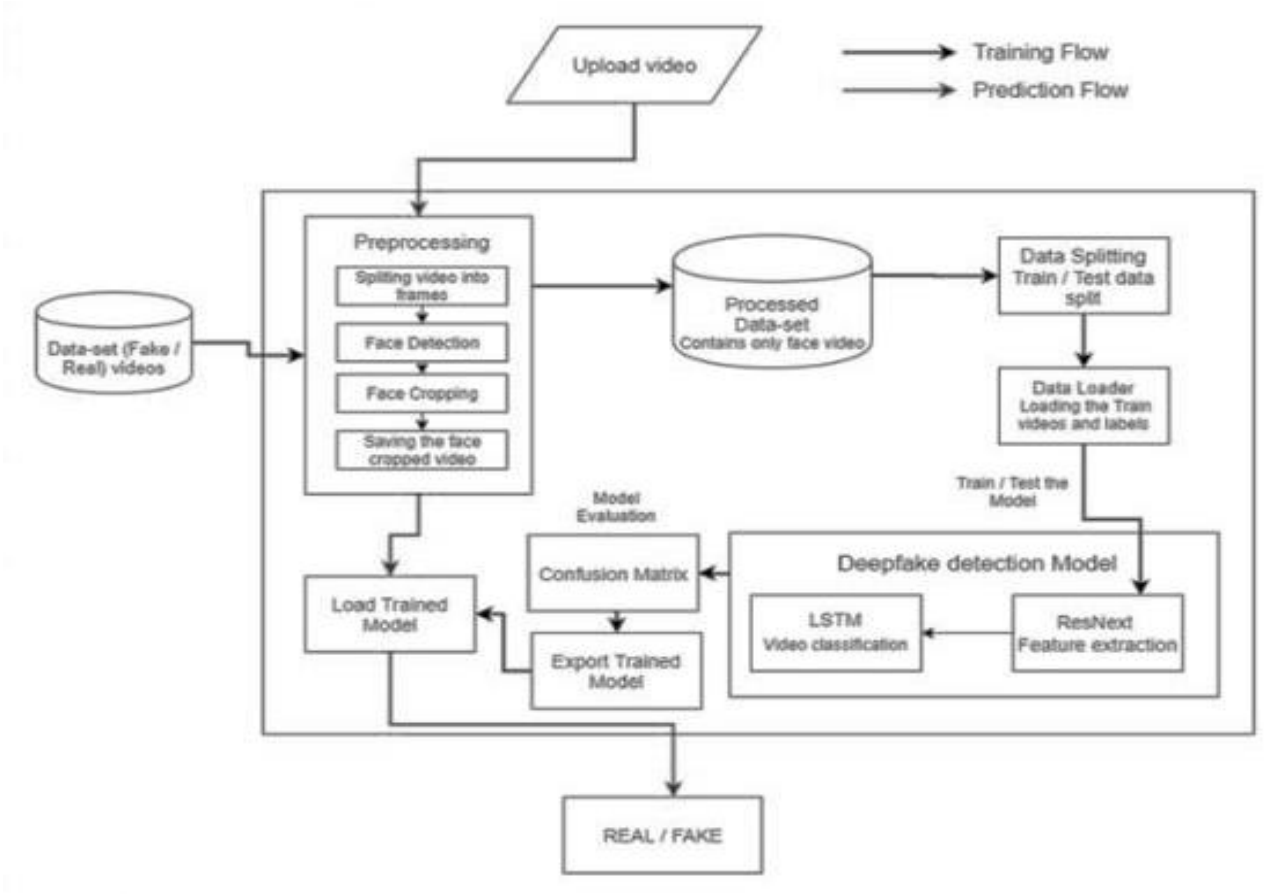

Fig.1 Proposed System Architecture

The system's technical components include:

- For Images: It uses a combination of Convolutional Neural Networks (CNN) and ResNet models to detect spatial inconsistencies and visual artifacts.
- For Videos: It employs an LSTM model to analyze temporal relationships and identify anomalies across a sequence of frames, such as unnatural facial movements or jitter.

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 8, August 2025

DOI: 10.17148/IARJSET.2025.12835

For Audio: It uses the Librosa library for audio signal processing and feature extraction, which are then used by a Random Forest classifier to detect synthetic audio patterns.

The system is hosted on a Flask-based web server with a responsive user interface built using HTML, CSS, and JavaScript, which allows for real-time detection. The workflow involves the user uploading or recording a media file, which is then processed by the appropriate detection model, and the results are displayed on the web interface with a confidence score. The final output for images includes a heatmap of manipulated regions, for videos a timeline of suspicious segments, and for audio a classification result with a confidence score. The system is expected to achieve high detection accuracy, balanced precision and recall, and be resilient to adversarial attacks.
Sources

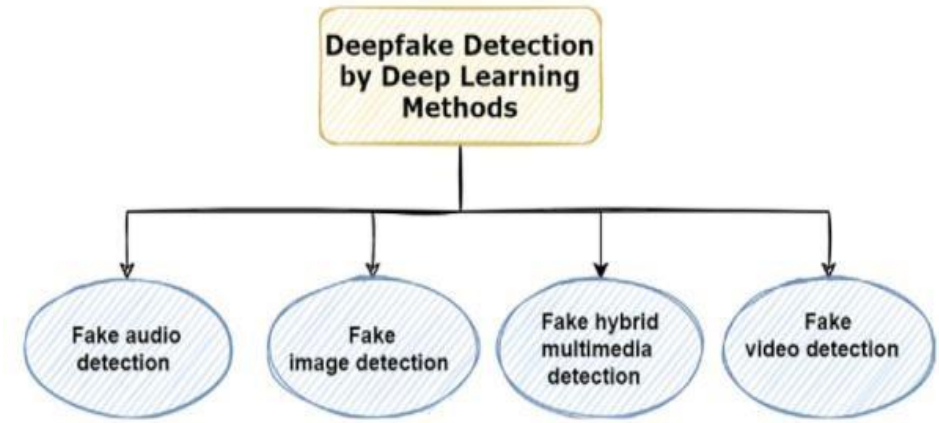

Fig.2 Proposed flow Architecture

IV. RELATED WORK

The related work in deepfake detection can be broadly categorized into several areas, with different studies focusing on specific aspects of the problem. One significant area of research has been the development of efficient face-swapping frameworks. For instance, **Simple Swap (SimSwap)** is an notable example, designed to achieve high-fidelity face swapping while preserving key attributes like facial expressions and gaze direction. It does this by using a specific module for identity injection and a unique loss function that ensures the original facial features of the target are maintained. Similarly, the **Identity-aware Dynamic Network (IDN)** was created as a lightweight network for real-time, subjectagnostic face swapping, making it practical for use on mobile devices due to its efficiency and small number of parameters.

Another major focus has been on using generative models and advanced AI architectures to detect synthesized media. The **MRI-GAN** framework, for example, utilizes Generative Adversarial Networks (GANs) to identify deepfakes by pinpointing perceptual differences in the videos. This approach achieved a respectable accuracy on the DeepFake Detection Challenge Dataset. Building on this, the **Multi-modal Multi-scale Transformer (M2TR)** takes a more sophisticated approach by using transformer models to capture subtle manipulation artifacts at various spatial scales. This system also incorporates frequency domain analysis, which allows it to detect forgeries that are not obvious in standard RGB color channels. The researchers behind M2TR also contributed a new, high-quality dataset, SR-DF, to help advance the field.

Beyond purely visual analysis, some research has explored training strategies and human-centric methods to improve detection. The **CYBORG** training strategy is a prime example, where a new component is added to the loss function that encourages the deep learning model to focus on human salient regions, or areas that a human would naturally pay attention to. By blending data-driven optimization with human-derived "coaching," this method aims to make the AI more effective at spotting manipulations. Another study by Jihyeon Kang et al. also took a practical approach by focusing on three common traces of deepfake creation: residual noise, warping artifacts, and blur effects, and using a specialized network to detect these pixel-level irregularities.
Temporal analysis and the detection of behavioral inconsistencies have also been a fruitful area of research. Methods like **DeepVision** moved beyond pixel-based analysis to verify anomalies by measuring the period and duration of eye blinks, which are often a sign of unnatural manipulation in deepfakes. Other recognition pipelines have combined different types

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 8, August 2025

DOI: 10.17148/IARJSET.2025.12835

of networks to capture these temporal irregularities. For example, one two-step approach uses a **Convolutional Neural Network (CNN)** to extract features from individual video frames, and then an **RNN** to capture the erratic, inconsistent frames that often result from the face-swapping process.

Finally, some of the most recent and promising work has focused on creating more robust and practical detection systems. The **Dual Attention Forgery Detection Network (DAFDN)** is a network designed to extract traces left by image warping using specialized attention modules. A web-based platform by A. Jadhav et al. also demonstrates a practical application, using a combination of **ResNeXt and LSTM** models to classify videos as real or fake, with the entire process being available through a user-friendly interface. These platforms represent a move towards making deepfake detection more accessible and integrated into real-world applications.

V. FUTURE WORK

Future work in deepfake detection should focus on several key areas to keep pace with the rapidly evolving nature of deepfake technology. A primary area for future research is the development of real-time detection capabilities. The current proposed system, while robust, can be further optimized to enable analysis as media is being created or streamed, which is essential for applications like social media monitoring and live video verification. The focus should be on building more efficient algorithms and leveraging powerful hardware to reduce latency without sacrificing accuracy. Furthermore, ongoing research is needed to ensure that detection models can effectively identify advanced deepfakes that may use new and emerging generation methods, as deep learning technologies continue to advance. This includes creating and training models on diverse, high-quality datasets that represent a wide array of manipulation techniques and video qualities to improve generalization.

Another crucial area is enhancing the robustness and resilience of detection systems against adversarial attacks. As detection methods become more sophisticated, malicious actors are likely to develop deepfakes specifically designed to fool these algorithms. Future work must, therefore, concentrate on creating frameworks that can maintain accurate and reliable performance even when faced with deliberate attempts to deceive the model. This could involve exploring new techniques that not only identify the signs of manipulation but also understand the specific ways in which a deepfake might be trying to evade detection. The goal is to build a system that is not only accurate but also trustworthy in highstakes scenarios.

In addition to technical improvements, future efforts should also address the broader societal and ethical implications of deepfake technology. A key area for research is the establishment of international standards for the responsible creation and detection of deepfakes. This would provide clear guidelines for developers, researchers, and policymakers, promoting the ethical use of AI and synthetic media. The research should also consider how to make deepfake detection more accessible to the general public, perhaps by integrating tools into consumer applications to enhance media literacy and public awareness. This would empower individuals to verify the authenticity of content they encounter, helping to combat misinformation and disinformation campaigns at a grassroots level

.

Finally, the development of comprehensive and interpretable systems is vital for future work. The proposed system already includes confidence scores and visualizations like heatmaps and timelines to highlight suspicious areas. Future work can build on this by creating even more detailed and understandable reports for human analysts, particularly in legal and forensic contexts where video and photo evidence are used. The improvement of AI and machine learning technologies means that the opinions of human experts may no longer be sufficient to verify manipulated digital content, making it essential for detection systems to provide clear, reliable, and trustworthy outputs that can be presented as evidence.

VI. CONCLUSION

Deepfakes are a significant threat to digital trust and societal stability, as their increasing realism and accessibility can lead to misinformation, social unrest, and privacy violations. To counter this, the proposed multi-modal deepfake detection system aims to be a more comprehensive solution than existing single-modality tools. By combining advanced AI techniques like ResNet, LSTM, and Random Forest classifiers, the system analyzes images, videos, and audio to provide more accurate and reliable verification. The document highlights the necessity of such automated systems, as even human experts can be deceived by sophisticated forgeries. Ultimately, the project's goal is to offer a practical, userfriendly tool to combat the malicious use of synthetic media, while also laying the groundwork for future advancements in real-time detection and ethical standards for deepfake technology.

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 8, August 2025

DOI: 10.17148/IARJSET.2025.12835

REFERENCES

[1]. Renwang Chen, Xuanhong Chen, Bingbing Ni, Yanhao Gen et al., "Simple Swap (SimSwap)," an efficient framework for generalized and high-fidelity face swapping.

[2]. Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu et al., "a lightweight Identity-aware Dynamic Network (IDN) for subject-agnostic face swapping by dynamically adjusting the model parameters".

[3]. Pratikkumar Prajapati, Dr. Chris Pollett, "a novel framework for using Generative Adversarial Network (GAN)based models, called MRI-GAN, that utilizes perceptual differences in images to detect synthesized videos".

[4]. Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen et al., "a Multi-modal Multi-scale Transformer (M2TR), which operates on patches of different sizes to detect local inconsistencies in images at different spatial levels".

[5]. Aidan Boyd, Patrick Tinsley, Kevin Bowyer, Adam Czajka, "a training strategy to ConveY Brain Oversight to Raise Generalization (CYBORG)".

[6]. Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens et al., "demonstrated that, with careful pre- and post-processing and data augmentation, a standard image classifier trained on only one specific CNN generator (ProGAN) is able to generalize surprisingly well to unseen architectures".

[7]. Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, "a novel and improved deep-CNN (D-CNN) architecture for deepfake detection with reasonable accuracy and high generalizability".

[8]. Jihyeon Kang, Sang-Keun Ji, Sangyeong Lee, Daehee Jang et al., "a technique for detecting various types of deepfake images using three common traces generated by deepfakes: residual noise, warping artifacts, and blur effects".

[9]. Luca Guarnera, Oliver Guidice, Sebastiano Battiato et al., "a new approach aimed to extract a Deepfake fingerprint from images is proposed".

[10]. Yi-Xiang Luo, Jiann-Liang Chen, "a Dual Attention Forgery Detection Network (DAFDN), which embeds a spatial reduction attention block (SRAB) and a forgery feature attention module (FFAM) to the backbone network".

[11]. Tackhyun Jung, Sangwon Kim, Keecheon Kim, "a method called DeepVision which is implemented as a measure to verify an anomaly based on the period, repeated number, and elapsed eye blink time".

[12]. Eunji Kim, Sangzoon Cho, "a hybrid face forensics framework based on a convolutional neural network combining the forensics approaches to enhance the manipulation detection performance".