International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

AI-Powered Deepfake Detection

Sahana Sunkad[1], Dr. Sridevi Malipatil[2]
[1]M. Tech, Department of CSE, RYM Engineering College-RYMEC, Ballari, VTU Belagavi, Karnataka, India
[2]Professor, Department of CSE, RYM Engineering College-RYMEC, Ballari, VTU Belagavi, Karnataka, India.
***Email ID**: sahanasunkad25@gmail.com[1], sridevi.siddu@gmail.com[2]*

Abstract
This research work presents a method that utilizes minimal training data and time to generate customized, photo- realistic talking head models. The technique employs few-shot learning, enabling the generation of satisfactory results from a single image, with improved fidelity using additional inputs. Unlike traditional warping-based approaches, the system synthesizes video frames directly using deep convolutional networks. The learning process is defined through adversarial training involving high-capacity generators and discriminators, which allows the system to quickly adapt to new identities through extensive meta-learning on large-scale video datasets. A person- specific parameter initialization further accelerates training and enhances performance. The proposed approach demonstrates the capability to produce lifelike talking head models of previously unseen individuals, including those depicted in portrait paintings. The paper discusses the ability of deepfake technology to produce artificial intelligence-generated digital content that looks real is examined closely. This research takes into account the wider societal ramifications as well as the complexities of AI algorithms in creating and detecting deepfakes. It highlights the critical requirement for advanced detection systems to stop exploitation and considers the continuous development of this powerful technology.

***Keywords:** Convolutional Neural Networks (CNNs), Multi-Task Cascaded Convolution Network (MTCNN), Librosa, and Random Forest Algorithms.*

1. Introduction
Recent years have seen notable advances in deep learning, a subset of machine learning techniques that combine representation learning with artificial neural networks. Deepfake technology, which automates the creation of synthetic video content, has raised concerns about election manipulation and cyberbullying. This study suggests an integrated system with a unique face forensics model and a convolutional approach for media manipulation detection in order to address these problems. In terms of construction, the system makes use of Convolutional Neural Networks (CNNs), which are often need to undergo substantial training on sizable datasets customized for each subject in order to produce convincing human head images. But in real-world applications, it becomes necessary to train customized talking head models using as little as one image as input. In order to address this difficulty, the system uses meta-learning methods on a sizable video dataset. It frames the learning procedure as adversarial training problems with high-capacity generators and discriminators, initializing parameters in a way that is unique to each individual to enable quick training despite the task's complexity. In the meantime, the detection side tackles the onslaught of altered media by utilizing deep learning and computer vision to create a hybrid face forensics framework that enhances detection performance by fusing traditional image forensic approaches with false face image forensic methods. This framework, embodied in a convolutional neural network architecture with two distinct feature extractors, aims to simultaneously extract content and trace features from facial images, aiming to mitigate the adverse societal impacts of manipulated media. This study examines the relationship between deep learning and digital forensics, with a focus on Deepfake technology and the production and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

identification of altered material. The goal of the paper is to determine whether Convolutional Neural Networks (CNNs) with limited training data and time restrictions can produce individualized, photo-realistic talking head models. In an effort to lessen the negative effects that these technologies have on society, it also explores the creation of a hybrid face forensics framework that uses convolutional methods to identify edited or altered media. The scope includes developing adversarial training strategies with high-capacity generators and discriminators, and investigating meta-learning techniques for few- and one-shot learning scenarios. In order to improve manipulation detection effectiveness, the research also looks into integrating traditional picture forensic methods with false face image forensic techniques. The study paper's overall goal is to advance knowledge and the creation of reliable techniques for manipulating media and identifying it, with ramifications for both digital integrity and the welfare of society.

1.1 Objectives

For the project focusing on the development of an integrated face forensics system to detect deepfake content effectively, the objectives are clearly outlined to address both the technical challenges and broader societal impacts. Here they are structured for clarity and focus:

- **Develop an Advanced Detection Framework:** Design and implement a hybrid face forensics model that integrates conventional image forensic techniques with cutting edge deep learning approaches, specifically convolutional neural networks (CNNs).
- **Implement Few-Shot and One-Shot Learning:** Incorporate meta-learning strategies to enable the system to effectively learn from a minimal number of images, including few-shot and one- shot learning capabilities. This is essential for adapting quickly to new and evolving deepfake methods with limited available data.
- **Optimize Adversarial Training Techniques:** Develop and refine adversarial training processes that involve high-capacity generators and discriminators. The objective is to enhance the model's ability to generalize from known to unknown manipulations, thereby improving its predictive accuracy against emerging deepfake technologies.
- **Build and Validate with Diverse Datasets:** Collect and compile a comprehensive and diverse set of data, including a custom DeepFake dataset alongside the public Face2Face dataset. Use these datasets to train and validate the effectiveness of the forensics framework under various conditions and across different types of deepfake manipulations.
- **Ensure Robustness and Scalability:** Ensure that the detection system is not only robust against a variety of deepfake attacks but also scalable and adaptable for widespread deployment across different platforms and technologies.

1.2 Literature Review

In recent years, the proliferation of deepfake technology has raised concerns about its potential misuse, prompting researchers to develop various detection methods. This literature survey explores different approaches proposed by researchers to detect deepfake content across different media types, including images, videos, and audio. By analyzing the methodologies, limitations, and potential advancements outlined in these studies, we aim to provide insights into the evolving landscape of deepfake detection and mitigation strategies.

- **Liu and Du [1]:** They delved into deepfake detection using Capsule Networks, which have shown promise in understanding hierarchical relationships in media. However, while their methodology offers insights into combating synthetic media manipulation, it lacks real time applicability and may not perform uniformly across various types of deepfakes.
- **Zang and Han [2]:** Their use of Recurrent Neural Networks (RNNs) for deepfake detection focused on analyzing temporal

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

dependencies within videos. While this approach addresses some challenges, such as the dynamic nature of deepfake manipulation, it may struggle with longer video sequences or subtle manipulations that could evade detection, posing limitations in real-world applications.

- **He et al. [3]:** They prioritized high-level semantic features for deepfake detection, aiming to accurately identify manipulated content. However, by relying solely on high-level features, their approach may overlook subtle manipulations or low-level artifacts, reducing its effectiveness in comprehensive detection across various manipulation techniques.
- **Guera and Abd-Almageed [4]:** Their utilization of Capsule Networks for deepfake detection highlighted challenges regarding scalability and interpretability, particularly in complex scenarios. While Capsule Networks offer unique capabilities, their practical applicability and effectiveness in detecting diverse deepfake content require further refinement.
- **Zhang et al. [5]:** By employing Convolutional Neural Networks (CNNs) for detecting manipulated facial content, they aimed to leverage the network's ability to learn complex patterns from images. However, their model's limited analysis of temporal dependencies in videos may hinder its ability to detect deepfakes relying on dynamic features rather than static facial characteristics.
- **Gül et al. [6]:** Their introduction of an Attention- Based Detection Network effectively focused on specific facial regions, a critical aspect of deepfake detection. However, attention-based models may struggle to capture broader features necessary for comprehensive detection across various manipulation techniques and media types.
- **Alnaim et al. [7]:** Their creation of a deepfake detection dataset focused on face masks addressed a pertinent need in the context of the infectious disease era. Nonetheless, attention-based models' limitations in capturing broader features could pose challenges in fully detecting manipulated content featuring individuals wearing face masks.
- **Hamza et al. [8]:** Leveraging Mel-Frequency Cepstral Coefficients (MFCC) for deepfake audio detection, they emphasized the importance of high-quality training data and advanced audio processing techniques. However, further exploration of advanced audio processing methods could enhance their approach's efficacy in detecting and mitigating deepfake audio content.
- **Waseem and Abu Bakar [9]:** Their comprehensive review of face and expression swap techniques laid a foundation for understanding these deepfake methods. To augment its utility, deeper exploration of specific detection methods alongside their respective strengths and weaknesses would provide valuable insights for effective countermeasures against sophisticated deepfakes.
- **Waqas et al. [10]:** Investigating the use of deepfake image synthesis for data augmentation, they highlighted ethical considerations and the need for validation of synthetic data. A deeper exploration of ethical frameworks and validation techniques could further enhance understanding and ensure responsible and impactful research practices in the field.
- **Waqas et al. [11]:** They explore using deepfake image synthesis as a method for data augmentation, aiming to supplement scarce annotated datasets. Their study demonstrates potential improvements in model training but also emphasizes the need for ethical frameworks and rigorous validation when employing synthetic images.
- **Patel et al. [12]:** This paper introduces an

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

improved Dense CNN (D-CNN) architecture for detecting deepfake images, trained on a large balanced dataset (140 k images). It achieves strong accuracy (~94 %) and shows solid generalizability across datasets. The work suggests further gains might come from incorporating inter-frame consistency or temporal features.

- **Tran et al. [13]:** Presenting a Meta Deepfake Detection (MDD) approach, they employ meta- learning to capture transferable knowledge across domains. The model can directly detect deepfakes in unseen domains without retraining, using novel loss functions (Pair-Attention and Center Alignment) to enhance generalization.
- **Kang et al. [14]:** They propose a steganalysis-based detection technique targeting residual noise, warping artifacts, and blur—common across various deepfake types. Using a pixel-level steganalysis network plus facial landmark analysis, the method improves unified detection performance across multiple forgery types.
- **Malik et al. [15]:** This survey offers a comprehensive overview of deepfake detection approaches for human faces in images and video. It categorizes methodologies by detection type (image vs. video), highlights performance differences, and identifies research gaps—particularly the need for more robust generalization and real-world applicability.

Through the analysis of research papers conducted in the literature survey, the solution to addressing the challenges associated with deepfake proliferation and its potential societal impacts lies in the development of a comprehensive deepfake detection system. This system, named DeepGuard, encompasses a multi- faceted approach to combatting the spread of manipulated media content. Central to DeepGuard is the implementation of advanced machine learning and deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based models, tailored specifically for deepfake detection tasks. Additionally, the integration of meta-learning and generalization techniques aims to enhance the system's robustness and adaptability across diverse deepfake scenarios. Alongside these technical advancements, DeepGuard also prioritizes ethical considerations and societal implications, striving to uphold principles of truthfulness, privacy, and integrity in digital media. By leveraging state-of-the- art technologies and ethical frameworks, DeepGuard seeks to empower users with the tools and knowledge needed to navigate the increasingly complex landscape of digital content with confidence and trust.

2. Method

The proposed system integrates a generator network designed to transform input facial landmarks into output frames through a series of convolutional layers. This generator network employs adaptive instance normalization to modify embedding vectors, thereby enhancing the realism of the synthesized images. During meta-learning, the system predicts the adaptive parameters of the generator by processing sets of frames from the same video through an embedder, averaging the resulting embeddings, and utilizing the aggregated information. Additionally, the system evaluates the performance of the generator by comparing the generated images with the actual data, thus ensuring the fidelity of the synthesized output. This approach enables the creation of highly realistic facial images based on input landmarks, facilitating applications such as personalized talking head models and manipulation detection systems. Overall, the proposed system leverages advanced deep learning techniques to address challenges in computer vision and digital forensics, showcasing its potential for enhancing both synthetic media generation and manipulation detection capabilities. This architectural design offers a complete defense against altered media by combining elements for both producing and identifying deepfake content. It makes realistic deepfake photos and movies by using sophisticated manipulation techniques along with

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

facial landmark detection. It simultaneously identifies and flags questionable content by utilizing CNNs, MTCNN, and audio analysis. With this method, deepfake content can be produced for research purposes and effectively detected to stop its spread (Figure 1).

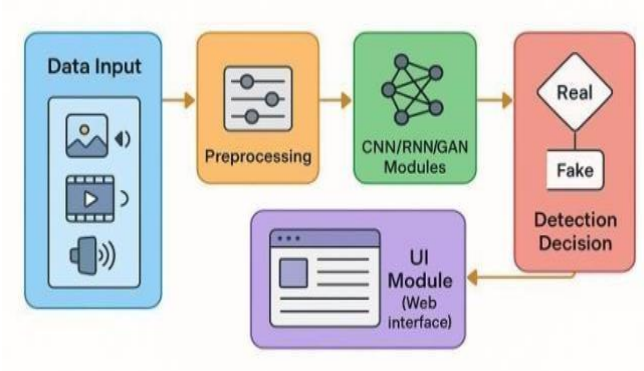

Figure 1 A Deepfakes Creation & Detection System Architecture

Several essential elements are included in the system architecture for both the development and detection of deepfakes:

2.1 Deepfake Creation

- **Facial Landmark Detection:** This component responsible for detecting and extracting facial landmarks from input photos or video frames is called Facial Landmark Detection. These checkpoints are important points of reference for the upcoming steps in the construction of the deepfake.
- **Deep Fake Image Creation:** This component creates deepfake images by utilizing extra approaches together with the extracted facial landmarks. These photos are edited to change the person's actions, facial expressions, or other traits from the original photos.
- **Deep Fake Video Creation:** Utilizing video manipulation techniques, this component builds on the foundation of deepfake images to produce deepfake films. These approaches to smoothly include deepfake images into the original video footage could be frame interpolation, morphing, or other techniques.

2.2 Deepfake Detection

- **CNN-Based Deepfake Detection**: This method looks for patterns in photos and videos that suggest deepfake modification by using Convolutional Neural Networks (CNN). To differentiate between real and known deepfake content, these CNN models are trained on a dataset.
- **MTCNN-Based Deepfake Detection: This** part handles face alignment and detection by utilizing Multitask Cascaded Convolutional Networks (MTCNN). It assists in the identification of deepfake content by recognizing faces and their alignment within picture or video frames.
- **Deep Fake Audio Detection:** This method looks for indications of audio tampering by analyzing audio information using techniques like feature extraction using librosa. The audio is then classified as either real or altered using a Random Forest classifier, which aids in the detection of deepfake audio information. When combined, these elements provide a complete solution that addresses audio tampering, image and video alteration, and deepfake content detection.

2.3 Workflow

The two primary procedures of creating and detecting deepfakes are depicted in this workflow diagram.

Creating Deepfakes

- The user takes a picture or a video, which is subsequently analyzed to determine important facial features using facial landmark detection.
- The deepfake picture/video production procedure uses the landmark data as input to produce a deepfake image or movie.
- The user is then given access to the deepfake image or video that was created.

Deepfake Detection

- The user submits a picture or a video that they believe to be a deepfake. Convolutional

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

neural networks (CNN) and Multitask Cascaded Convolutional Networks (MTCNN) are used to process the input data in order to detect deepfakes.

- The detection model generates a detection result that indicates whether or not the input image/video contains a deepfake, and it then returns the result to the user.

Audio Detection

- The user supplies an audio recording, which the Librosa library processes in order to analyze audio data.
- The Random Forest technique is used to do feature extraction on the audio material. The user receives a detection result from the Random Forest-based detection model that indicates whether the audio recording contains deepfake content. Using facial landmark detection, CNN, MTCNN, Librosa, and Random Forest algorithms, this data flow diagram offers a high-level overview of the data flow and interactions between various components involved in the deepfake development and detection processes.

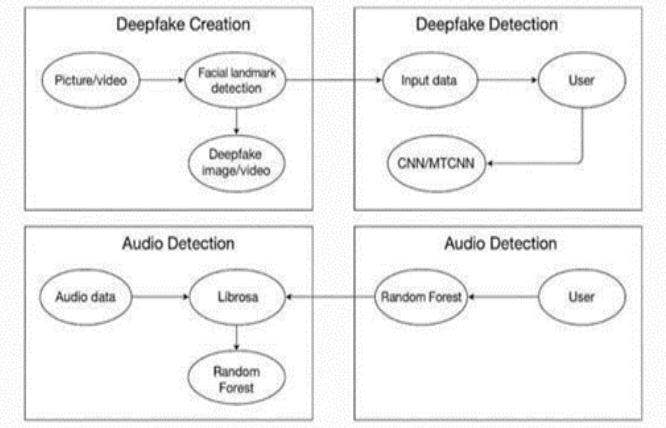

Figure 2 Workflow for Deepfakes Creation & Detection

The generator network converts input facial landmarks into output frames through a number of convolutional layers, as shown in fig 2. The embedding vectors are modified by the generator network via adaptive instance normalization. In order to predict the generator's adaptive parameters during meta learning, we run sets of frames from the same video through the embedder, average the resultant embeddings, and utilize the results. Next, we run a separate frame's landmarks through the generator and compare the output image to the actual data.

3. Results and Discussion

3.1 Results

Model Performance Accuracy

The model achieved a training accuracy of 96%. This indicates that the model correctly classified 96% of the examples in the training dataset. However, it's crucial to evaluate the model's performance on unseen data (validation or testing set) to avoid overfitting in deepfake detection. The validation accuracy is also 96%. While this is promising, a more robust evaluation for deepfakes should involve a separate testing set to assess generalizability (Figure 3).

Figure 3 Training and Validation Accuracy

Loss

Training Loss: The training loss generally decreases over epochs, indicating that the model is learning from the training data.

Validation Loss: The validation loss also decreases, suggesting that the model is not overfitting to the training data. However, a more comprehensive evaluation using a separate testing set is recommended for deepfake detection tasks (Figure 4).

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

Figure 4 Training and Validation Loss

3.2 Discussion

Deep Fake Creation is where users can make deep fakes from uploaded images or videos, and Deep Fake Detection is where they can check if uploaded video or audio content is real or fake (Figure 5). Image deepfake detection feature determines whether the uploaded image is real or fake. Audio deepfake detection feature, which determines whether the uploaded audio is real or fake. Video deepfake detection feature, which determines whether the uploaded Video is real or fake.

Conclusion

In conclusion, a series of convolutional layers converts input facial landmarks into output frames, with the generator network playing a critical part in this process. Through adaptive instance normalization, embedding vectors are modified throughout its operation. By running sets of frames from the same video through the embedder, averaging the resultant embeddings, and utilizing the results, sets of frames from the same video can be processed via the generator to assist predictive modeling of the adaptive parameters during meta-learning. In addition, the generator processes each frame landmark independently, and the resulting images are contrasted with the original data. This research demonstrates how well the generator network generates realistic facial pictures based on input landmarks, highlighting its potential uses in manipulation detection systems and personalized talking head models. These developments highlight deep learning's development and show how it might be used to solve problems in digital forensics and computer vision.

Acknowledgements

We would like to thank the faculty members of the Department of Computer Science and Engineering, Rao Bahadur Y. Mahabaleshwarappa Engineering College, Bellary, for their valuable support and infrastructure. We also acknowledge the contributions of various open-source libraries and tools, such as Python, Scikit-learn, which made the implementation and evaluation of machine learning models possible.

References

[1]. Y. Liu and S. Du, "DeepFake Detection Based on Capsule Network," in IEEE Access, 2021.
[2]. S. Zang and W. Han, "Deepfake detection using recurrent neural network," in Multimedia Tools and Applications, 2021.
[3]. Y. He et al., "Deepfake Detection Based on High-Level Information," in IEEE Transactions on Information Forensics and Security, 2021.
[4]. D. Guera and W. Abd-Almageed, "Deepfake Detection Using Capsule Networks," in IEEE/CVF International Conference on Computer Vision Workshop, 2021.
[5]. X. Zhang et al., "Deepfake Detection Using Convolutional Neural Networks," in IEEE Access, 2021.
[6]. A. Gül et al., "Deepfake Detection via Attention Based Detection Network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2021.
[7]. N. M. Alnaim et al., "A Deepfake Face Mask Dataset for Infectious Disease Era with Deepfake Detection Algorithms," 2023.
[8]. A. Hamza et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning," in Proceedings of the IEEE/CVF Conference on Computer

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3526-3533
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0518

Vision and Pattern Recognition Workshops, 2023.

[9]. S. Waseem and S. A. R. Abu Bakar, "DeepFake on Face and Expression Swap: A Review," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023.

[10]. N. Waqas et al., "DEEPFAKE Image Synthesis for Data Augmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022.

[11]. N. Waqas et al., "DEEPFAKE Image Synthesis for Data Augmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022.

[12]. Y. Patel et al., "An Improved Dense CNN Architecture for Deepfake Image Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023.

[13]. V.-N. Tran et al., "Generalization of Forgery Detection with Meta Deepfake Detection Model," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022.

[14]. J. Kang et al., "Detection Enhancement for Various Deepfake Types Based on Residual Noise and Manipulation Traces," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023.

[15]. A. Malik et al., "DeepFake Detection for Human Face Images and Videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023.