IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14623

A Survey on Detection of Deep Fake Images Using CNN Model

Mr. Kumar K[1], Satya Karthik R[2], Sandeep Kumar Jena[3] , Shamanth S Joshi[4], Sudhanva H Rao[5]

Associate Professor, Dept of Computer Science, K S Institute of Technology, Bengaluru, Karnataka[1]

Dept of Computer Science, K S Institute of Technology, Bengaluru, Karnataka[2-5]

Abstract: The increasing prevalence of AI-generated deepfake images has become a significant concern in the context of misinformation and digital security. Deepfake technology, driven by generative adversarial networks (GANs) and sophisticated AI algorithms, enables highly realistic image alterations, making it challenging to differentiate genuine visuals from manipulated ones. This study introduces a deepfake image detection system utilizing convolutional neural networks (CNNs) for classification. By employing deep learning techniques, the system evaluates the authenticity of images and identifies alterations with high precision. Through training on diverse datasets, the model aims to bolster media integrity and strengthen digital security. The findings underscore the importance of reliable deepfake detection in minimizing the risks of manipulated content, offering valuable applications in fields such as journalism, social media verification, and digital forensics.

Keywords: Deepfake detection, image authenticity, machine learning, CNN, digital security, media verification, digital forensics

I. INTRODUCTION

The rapid progress in artificial intelligence (AI) has significantly impacted various sectors, driving innovation while introducing new risks. Among the most pressing concerns is the emergence of deepfake technology, which uses generative adversarial networks (GANs) and other advanced AI methods to produce highly realistic fake images and videos. Although deepfakes offer promising uses in fields like entertainment, visual effects, and assistive technologies, they also present serious challenges related to digital security, misinformation, and identity misuse. Malicious actors have exploited deepfakes to disseminate false narratives, conduct political manipulation, perpetrate cyber fraud, and carry out identity impersonation. The growing sophistication of AI-generated media has raised critical concerns in areas such as journalism, forensic analysis, and law enforcement, where conventional detection tools often fall short. As deepfake creation tools become more widely accessible, the demand for reliable detection systems continues to rise.

This research proposes the development of a deepfake image detection framework based on convolutional neural networks (CNNs) to accurately classify and flag manipulated visuals. By integrating deep learning approaches with advanced image examination, the system aims to improve the trustworthiness of digital content verification. The study explores state-of-the-art detection strategies, implements a robust classification algorithm to differentiate authentic images from synthetic ones, and assesses the model's performance using publicly available datasets and evaluation metrics. This work contributes to ongoing efforts to protect digital information and uphold the credibility of visual media in the face of evolving AI threats.

II. LITERATURE SURVEY

1) Detection of Deep Fake Images Using Convolutional Neural Networks

Deepfake technology has evolved rapidly, making it increasingly difficult to distinguish between authentic images and those that have been artificially manipulated. These synthetic visuals, created using advanced AI techniques such as generative adversarial networks (GANs), can replicate facial features and expressions with a high degree of realism. This poses serious challenges in areas like digital security, journalism, and social media, where the authenticity of visual content is crucial.

2) An Improved Dense CNN Architecture for Deepfake Image Detection

The detection of deepfake images has emerged as a pressing issue in today's digital landscape, particularly as generative adversarial networks (GANs) continue to advance, producing synthetic visuals with remarkable realism. This research investigates an enhanced Dense Convolutional Neural Network (CNN) architecture tailored to improve the model's

DOI: 10.17148/IJARCCE.2025.14623

ability to extract complex features and resist adversarial manipulations. The proposed framework utilizes deep, hierarchical feature representations to effectively differentiate between authentic and manipulated images, focusing on the subtle visual discrepancies often introduced by GAN-generated content.

3) Detection of Deep Fakes in Face Images Based on Machine Learning

The swift progression of deepfake technology has led to growing concerns about the credibility of digital facial imagery. This research introduces a machine learning-based method for detecting deepfake face images by identifying critical features that distinguish real faces from AI-generated ones. The approach applies multiple machine learning algorithms to examine altered facial structures, thereby enhancing the precision and resilience of detection. Performance is assessed using benchmark datasets, confirming the method's capability in accurately recognizing manipulated visuals. These results advance the field of deepfake detection and provide valuable guidance for developing trustworthy solutions to counter face-based digital forgeries.

III. OBJECTIVES

The primary objective of this study is to develop a CNN-based deepfake image detection system capable of distinguishing real images from AI-generated forgeries. The specific goals include:

1) Implementing a machine learning-based classification model to improve detection accuracy: The system will integrate state-of-the-art deep learning architectures such as Convolutional Neural Networks (CNNs) and transformer-based models to classify images based on authenticity, ensuring high detection precision.

2) Evaluating the performance of the proposed system using benchmark datasets: To ensure reliability, the model will be tested on widely used datasets such as the DeepFake Detection Challenge Dataset and Celeb-DF. Evaluation metrics like accuracy, precision, recall and F1-score will be used to assess model effectiveness.

3) Enhancing the adaptability of the system to evolving deepfake generation methods: Since deepfake technology continuously evolves, the detection system must be capable of adapting to new forgery techniques. This will be achieved by incorporating continual model updates based on new datasets.

4) Contributing to digital security and misinformation prevention by providing a robust detection mechanism: The system aims to act as a safeguard against malicious deepfake usage in various domains, including journalism, social media verification, and law enforcement, ensuring the integrity of digital media.

IV. METHODOLOGY

Developing a robust deepfake detection system requires a systematic approach that integrates various AI techniques, computational models, and data-driven strategies. This section outlines the system's design, architecture, and implementation strategy. The methodology is structured to ensure accuracy, efficiency, and scalability while detecting AI-generated image manipulations

A) System Design and Architecture

The system is designed as a modular pipeline that includes data preprocessing, feature extraction, model inference, and result interpretation. This approach ensures flexibility, allowing new AI models and techniques to be integrated over time. The system is structured into multiple layers to streamline image processing and classification tasks efficiently.

B) Technology Stack Selection:

The technology stack is chosen to maximize efficiency and adaptability. The system uses deep learning frameworks optimized for image analysis, along with high-performance libraries for handling large-scale datasets. Additionally, the use of cloud-based or on-premise AI computing resources ensures optimal model training and inference performance.

C) Architecture Design:

The detection model follows a multi-layered architecture. It starts with a feature extraction layer, utilizing convolutional neural networks (CNNs) to capture intricate details within images. A classification layer follows, employing techniques to determine image authenticity. The system is structured to accommodate real-time processing, ensuring that deepfake detection can be applied in dynamic environments.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14623

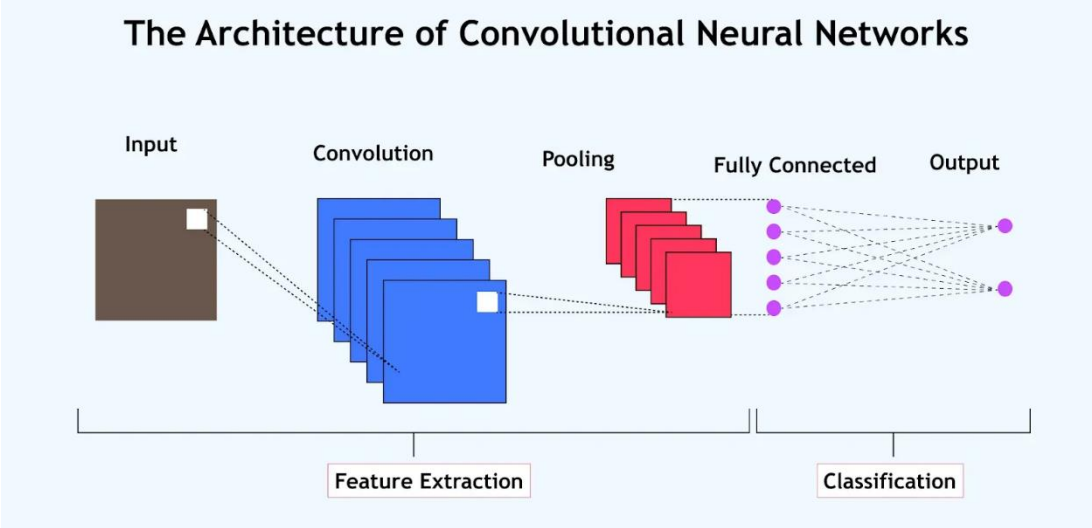

Figure 1.1 CNN Structure for Feature Extraction and Classification

D) Dataset Selection:

Publicly available deepfake datasets are utilized for training and testing the AI model. These datasets include a diverse range of images with varying levels of manipulation, ensuring that the model learns to identify both subtle and overt modifications.

E) Model Selection & Training:

Various CNN-based classification models are tested and compared to identify the most effective one for deepfake detection. The training process involves fine-tuning hyperparameters, employing data augmentation techniques, and ensuring that the model generalizes well across different datasets.

F) Evaluation Metrics:

The model's performance is assessed using standard evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into how well the system distinguishes between real and manipulated images, guiding further improvements in the model's architecture. Accuracy reflects the overall correctness of the model, while precision indicates how many of the images identified as deepfakes are actually false. Recall measures the model's ability to detect all instances of manipulated content, and the F1-score provides a balanced evaluation by combining precision and recall.

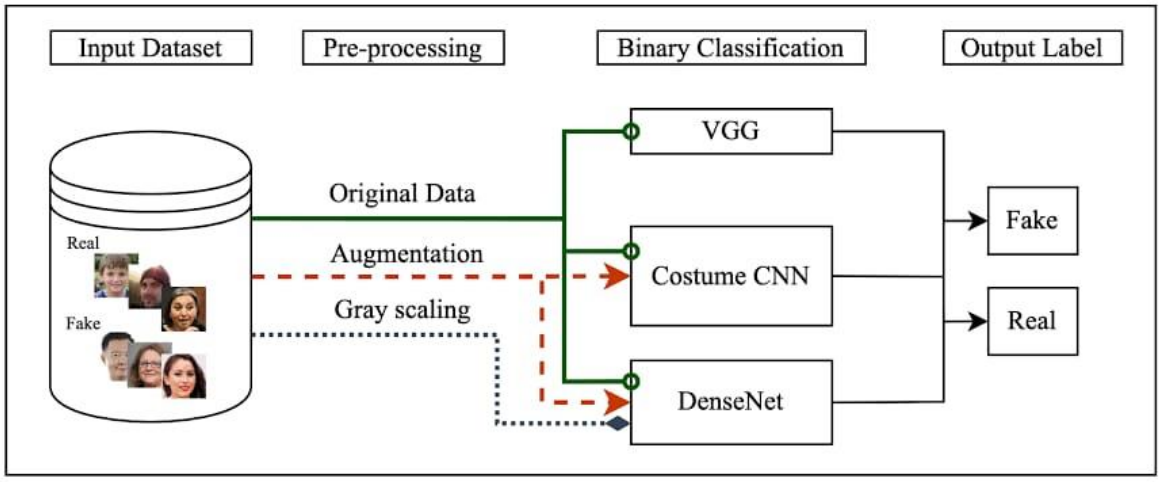

Fig 1.2 Model Training

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14623

V. APPLICATION REQUIREMENTS

To ensure the effective implementation of the deepfake detection system, the following application requirements have been established:

A. Hardware

1) Standard PC or Server

a) Processor: A multicore processor with a high clock speed suitable for handling AI workloads. AMD Ryzen 7 or Intel Core i7 with good multi core performance.
b) RAM: Minimum of 16GB RAM to handle feature extraction and other computation load, if larger dataset is used 32GB is preferable.
c) Storage: An SSD of 1TB is preferable for faster data access and model storage.
d) NVIDIA RTX 4050 GPU (or higher) for optimized AI model training and inference

B. Software

1) Development Tools

Integrated Development Environment (IDE): which supports both Python and JavaScript for the development of a CNN model and a Webapp

2) Web Framework

a) Backend: Flask server to process the image and handle requests
b) Frontend: React.js with Vite, axios, and Tailwind CSS

C) Machine Learning Framework

TensorFlow or OpenCV : Deep learning frameworks for training and fine-tuning the model

D) Version Control: Git or GitHub for efficient collaboration and controlling different versions of the webapp

E) Data:

1) Relevant datasets for:

Training of the CNN model for the detection of Deepfake Images, testing, and efficiency calculation of the implemented model.

VI. CONCLUSION

The system presented in this study successfully detects deepfake images by utilizing CNN-based classification models, thereby contributing to improved digital security. The findings confirm that deep learning methods can achieve high precision in identifying altered images, helping to mitigate threats related to misinformation and fraudulent content. Future enhancements will focus on refining the model to boost detection accuracy, incorporating more diverse datasets to enhance generalizability, and enabling real-time analysis for video-based applications. Adapting the system to keep pace with evolving deepfake generation techniques will further reinforce its effectiveness in ensuring media authenticity and combating digital deception

VII. ACKNOWLEDGEMENT

We wish to extend our sincere appreciation to **Mr. Kumar K** for the invaluable and constructive input provided throughout the planning and development of this project. We are truly grateful for his generous dedication of time. Additionally, we'd like to express our thanks to the esteemed professors of KSIT for their unwavering support and encouragement.

REFERENCES

[1] G. Aggarwal, A. K. Srivastava, K. Jhajharia, N. V. Sharma, and G. Singh, "Detection of Deep Fake Images Using Convolutional Neural Networks," *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, 2023

[2] Y. Patel *et al*., "An Improved Dense CNN Architecture for Deepfake Image Detection," in *IEEE Access*, vol. 11, pp. 22081-22095, 2023

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14623

[3] Sabah Abdul kareem, H., & Mahdi Altaei, M. S. (2023). Detection of Deep Fakes in Face Images Based on Machine Learning. *Al-Salam Journal for Engineering and Technology*

[4] TensorFlow: Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." arXiv preprint arXiv:1603.04467.

[5] OpenCV: Bradski, G. (2000). "The OpenCV Library." Dr. Dobb's Journal of Software Tools.

[6] DFDC (Deepfake Detection Challenge): Facebook AI. (2020). "Deepfake Detection Challenge Dataset." arXiv preprint arXiv:2006.07397.