DOI: 10.22413/ijastems/2026/v12/i6/3

Deep Learning Based Framework for DeepFake Image Detection Using YOLO Models

V. V. N. Bhaskar, S. Zeba Nikhat, P. M. Sruthi, K. Sandhya, R. Vyshnavi
Department of Computer Science & Engineering
Aditya College of Engineering, Madanapalle, India
Email: bhaskarvvn@gmail.com

Abstract— In the digital era, the rapid advancement of AI has led to the creation of highly realistic fake images known as deepfakes. This paper presents a deep learning-based framework for deepfake image detection using YOLO (You Only Look Once) models, designed with a strong focus on societal protection and public awareness. The system leverages the YOLO algorithm to efficiently detect facial regions, as most deepfake manipulations are concentrated on human faces. Based on these features, the system classifies the image as either real or deepfake. The proposed system improves detection accuracy and processing speed, making it suitable for real-time and practical applications such as social media verification and digital forensics. The image is then processed using a preprocessing module After preprocessing, the YOLO-based face detection module identifies and extracts facial regions from the image. These detected face regions are then passed to the classification module, where deep learning techniques analyze various visual features to determine whether the image is real or manipulated. Finally, the evaluation and result visualization module displays the detection outcome along with a confidence score, providing a clear and understandable output to the user.

I.INTRODUCTION

In the digital age, the rapid growth of AI has enabled the creation of deepfake images that appear highly realistic. Although these technologies have useful applications in entertainment and media production, they also pose serious threats to online authenticity, cybersecurity, and digital trust.

This Paper titled "Deep Learning–Based Framework for Deepfake Image Detection Using YOLO (You Only Look Once) Models", aims to develop an intelligent and reliable detection system. The proposed framework integrates advanced deep learning techniques to analyze and identify manipulated facial content.

The system leverages powerful object detection models such as YOLO, along with other detection approaches, to accurately extract spatial features from images. The core idea of the paper is to detect and localize facial regions using YOLOv11 models and then analyze these regions for inconsistencies or manipulation artifacts. By combining spatial feature extraction with deep learning classification methods, the system improves detection accuracy and robustness.

The backend implementation is carried out using Python for model training and processing in a local development environment using VS Code. Server-side operations and database management are handled using PHP and MySQL through XAMPP (localhost). The frontend interface is developed to provide real-time visualization and user interaction. The system displays detection results along with confidence scores, making it user-friendly, interactive, and efficient. Overall, this paper aims to enhance media verification systems by providing a scalable, accurate, and real-time deepfake detection solution.

II. LITERATURE SURVEY

Nelson et al. (2025) proposed a study titled "Deepfake Detection in Manipulated Images/Audio/Videos: A Three-Stage Multi-Modal Deep Learning Framework." The research focuses on detecting deepfake content across images, audio, and videos using advanced deep learning techniques. The authors introduced a three-stage framework that combines spatial and temporal feature extraction models.

Somnath Banerjee et al.(2025) presented a research paper titled *"Neural Architecture Search Based Deepfake Detection Model using YOLO,"* published in the International Journal of Advanced Research in Science, Communication and Technology. The study focuses on developing an advanced deep learning framework for detecting deepfake videos by combining the powerful object detection capability of YOLOv8 with an automated model optimization technique known as Neural Architecture Search The main objective of the research was to improve the accuracy and efficiency of deepfake detection systems by automatically designing an optimal neural network architecture for classification.

Mehmet Karakose et al.(2024) proposed a study titled *"A YOLOv3-Based Method for Detecting Deepfake Manipulated Facial Images,"* published in the Turkish Journal of Science and Technology. The research focuses on applying the YOLOv3 architecture to detect manipulated facial images generated through deepfake techniques. The authors aimed to develop an efficient deep learning–based system capable of identifying altered facial regions in images by utilizing the object detection capability of the YOLO framework. In the proposed method, facial images were first collected and organized into labelled datasets containing both authentic and deepfake samples.

Yatin Patel et al.(2024) presented a research paper titled *"Deepfake Image Detection Using Machine Learning and Deep Learning,"* published in the Educational

DOI: 10.22413/ijastems/2026/v12/i6/3

Administration: Theory and Practice. The study focuses on exploring the effectiveness of machine learning and deep learning techniques for identifying manipulated deepfake images. With the rapid growth of deepfake technology and the increasing use of synthetic media, the authors highlight the need for reliable automated systems capable of detecting manipulated visual content. The research proposes a deep learning–based framework that utilizes CNNs to distinguish between real and fake images. The methodology begins with collecting a dataset containing both authentic and manipulated images. These images undergo preprocessing steps such as resizing, normalization, and dataset splitting into training and testing sets. Proper data preparation ensures that the model can effectively learn discriminative visual patterns associated with deepfake manipulations.

III. PROPOSED METHOD

The proposed system aims to develop an advanced and efficient deep learning–based framework for detecting deepfake images using YOLO models. Unlike traditional methods, the proposed approach integrates face detection and classification into a single unified architecture, improving both speed and accuracy.

In the proposed system, the input image is first pre-processed through resizing, normalization, and enhancement techniques to improve image quality.
The image is then passed to the YOLO model, which performs real-time face detection by identifying facial regions using bounding boxes.

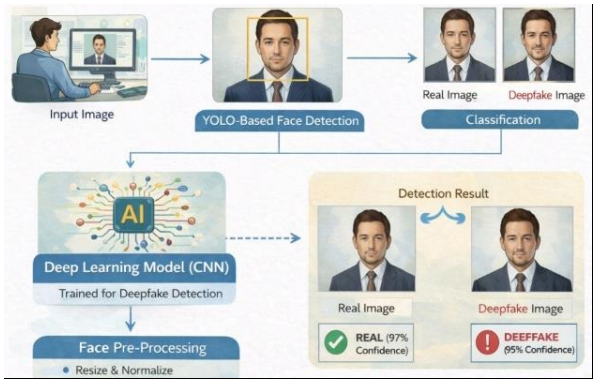

Figure 1: System Architecture

YOLO works as a single-stage object detection model, making it faster and more efficient compared to multi-stage detection systems.

These features are extracted and processed to generate structured data that can be used by machine learning models for emotion classification.The classification model then predicts whether the detected face is **real or fake**, along with a confidence score indicating the reliability of the prediction. The system is trained on a large dataset of real and manipulated images to ensure robustness and high accuracy. The system is designed to accurately identify whether a facial image is real or manipulated. The process begins with an input image provided by the user. This image is passed to a YOLO-based face detection module, which detects and localizes the face region using bounding boxes. YOLO ensures fast and precise face detection in real time.

Once the face is detected, it undergoes face pre-processing, which includes resizing and normalization. This step standardizes the image and improves the performance of the deep learning model. The preprocessed face image is then fed into a Convolutional Neural Network CNN trained specifically for deepfake detection. The CNN extracts important spatial features and identifies subtle artifacts or inconsistencies commonly found in manipulated images.

Based on the extracted features, the system performs classification, labeling the image as either Real or Deepfake. Finally, the detection result is displayed along with a confidence score (e.g., 97% Real or 95% Deepfake), indicating the model's certainty in its prediction.

The process begins with an input image, which is provided to the system for analysis. First, YOLO-based face detection is applied to identify and locate the face region in the image.

After detecting the face, face preprocessing is performed, which includes resizing and normalizing the image to make it suitable for model processing. The processed facial image is then passed to a deep learning model based on CNN that has been trained specifically for deepfake detection. The CNN analyzes facial features and patterns to distinguish between authentic and manipulated images. Based on the extracted features, the system performs classification to determine whether the image is a real image or a deepfake image

Finally, the detection result is displayed with confidence scores indicating the probability of the image being real or fake. This methodology improves detection accuracy by combining fast face detection using YOLO with powerful feature learning provided by CNN models.

IV. SIMULATION RESULTS

The framework integrates detection and classification modules to process input images and identify deepfakes. Simulation results demonstrate improved accuracy and faster processing compared to traditional methods. The environment supports performance analysis under different conditions, ensuring system reliability. network: organizational scaling, technology changes and application deployment.

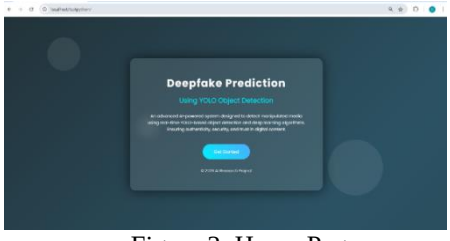

Figure 2: Home Page

DOI: 10.22413/ijastems/2026/v12/i6/3

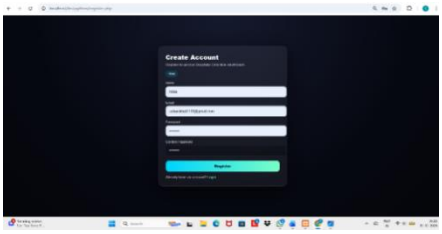

Figure 3: Login Page

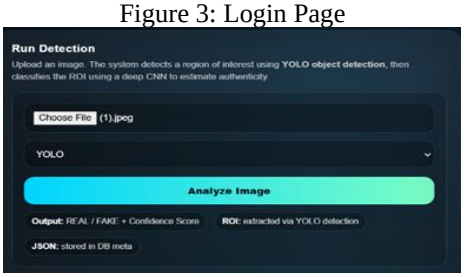

Figure 4: Detection Page

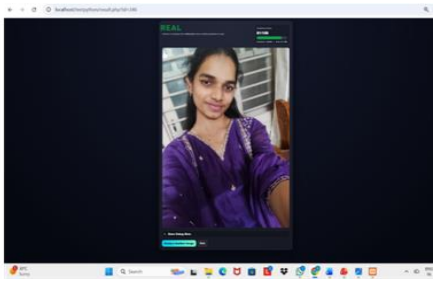

Figure 5: Detection of Real Image

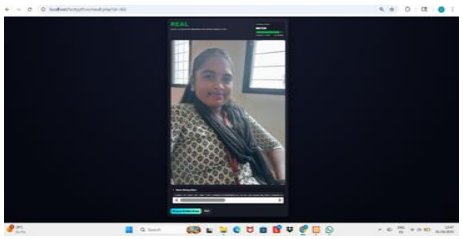

Figure 6: Detection of Real Image

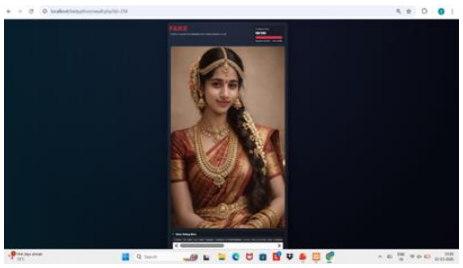

Figure 7: Detection of Fake Image

Figure 8: Detection of Fake Image

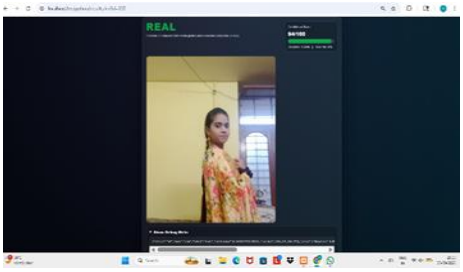

Figure 9: Detection of Real Image

V. CONCLUSION

The deep learning-based framework for deepfake image detection using YOLO models was successfully developed to address the growing misuse of AI in generating manipulated digital images. The proposed system effectively detected and analyzed facial regions using a YOLO-based architecture, followed by deep learning–based classification to determine image authenticity. The model analyzed the input images and classified them as real or fake, generating results along with corresponding confidence scores.

The system was implemented and validated for real-time performance, demonstrating high detection accuracy and faster processing compared to conventional approaches. It eliminated the need for manual verification by enabling automated detection, thereby improving efficiency and reducing human effort. The inclusion of confidence scores enhanced interpretability, making the system accessible to non-technical users. The proposed system is highly beneficial for public use in applications such as social media platforms, news verification systems, digital forensics, cybersecurity, banking security systems, legal investigation platforms, and identity authentication processes to help prevent misinformation and digital manipulation.

Overall, this paper work contributes toward building a safer digital ecosystem by empowering users to identify manipulated content and helping to reduce the problems caused by deepfake technology.

VI. REFERENCES

[1] Leema Nelson, Harshita Batra and Radha P “Deepfake Detection in Manipulated Images/ Audio/ Videos: A Three-Stage Multi-Modal Deep Learning Framework” published in Inteligencia Artificial Iberoamerican Journal of Artificial Intelligence, ISSN No:1988-3064 Volume 28, Issue 76, pp.20-39, December 2025.

[2] Somnath Banerjee et al. “Neural Architecture Search Based Deepfake Detection Model using YOLO,” published in the International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), ISSN No:2581-9429, Volume 5, Issue 1,pp.375-383, January 2025.

[3] Mehmet Karakose, Hasan Yetis, Mert Cecen, “A YOLOv3-Based Method for Detecting Deepfake Manipulated Facial Images,” published in the Turkish Journal of Science and Technology, Volume 19, Issue 2, pp. 315-324, February 2024.

[4] Yatin Patel Mayank Jain “Deepfake Image Detection Using Machine Learning and *Deep Learning” published in Educational* Administration: Theory and Practice, ISSN No: 2148-2403 Volume 30, Issue 5,pp.14987-14993May 2024

[5] Jordan J. Bird and Ahmad Lotfi ,"CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images" published in ResearchGate.net and IEEE Access, ISSN No: 2169-3536, Volume 12 ,pp.15642 - 15650, January 2024.

[6] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen Yann ," Deep Learning for Deepfakes Creation and Detection: A Survey", published in Computer Vision and Image Understanding, ISSN No:2331-8422, August 2022.