International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Face Forgery Detection Using Deep Learning

Shruthi T V[1], Rachana K R[2], Bhuvana S[3], Neha Appaji Y[4], Purnima R[5]

[1]Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student
Department of Artificial Intelligence and Data Science
East West Institute of Technology, Bengaluru

Abstract - The rising occurrence of forged content poses a threat to the authenticity of multimedia. This research suggests a simplified hybrid architecture as an effective method for detecting face forgeries. The framework includes CNN for dependable classification, EfficientNet for robust feature extraction, and MTCNN for precise face detection. MTCNN ensures that high-quality input is produced for feature extraction by accurately localizing facial regions. The CNN classifier utilizes the extracted features in order to distinguish between authentic and manipulated content, and EfficientNet, which has become famous for its good performance and computational efficiency, is able to capture face patterns at a subtle level. Transfer learning enhances the adaptability of the model toward new manipulation techniques because it pre-trains on a large-scale dataset before fine-tuning on data specifically related to deepfakes.

Key Words: Deep Learning, Deepfakes, Face Forgery, Multimedia forensics, CNN.

I. INTRODUCTION

Face Forgery technology has brought terrible challenges to the authenticity and reliability of digital content from social media to news and even in the courts. Forged content, which often alters facial features, expressions, or entire videos with fake identities, becomes a major threat to public trust in the media and sparked new demands for effective detection models. As forged content generation methods improve and continue to improve by leaps and bounds, they become more sophisticated, making earlier detection methods an ineffective tool for identifying subtlety manipulations hidden deep in high-resolution and realistic fakes.

Thus, in dealing with the challenges, the paper develops a face forgery detection framework that is resistant and relies on the power of Multi-Task Cascaded Convolutional Networks (MTCNN), EfficientNet, and Convolutional Neural Networks (CNN). MTCNN can detect faces precisely and analyze only relevant face regions, concentrating the model on areas of interest, which enhances accuracy. The next step involves efficient extraction, which captures the fine-grain with good efficiency. Finally, classification is done using CNN-assisted classifier, which will distinguish between original and fake videos.

A key element of this architecture is the application of transfer learning, where EfficientNet will first be pre-trained for a large-scale dataset, and it will be further fine-tuned using deepfake specific data. This will lead to enhanced generalization and fine-tuning by adapting to varying types of deepfake content.

II. LITERATURE REVIEW

The rapid speed in the development of artificial intelligence can create highly convincing fake media with very realistic manipulated facial images, through tools such as DeepFake, and videos generated via tools like Face2Face, posing significant challenges that bring together areas such as the authenticity of media information spread and digital forensics, which requires highly effective detection mechanisms.

With such an aim, the hybrid model CNN of detecting fake faces as a product of tools like DeepFake and Face2Face presented by Eunji Kim and Sungzoon Cho is proposed in paper [1]. The model makes use of both content feature extraction via ResNet-18 and trace feature extraction via multi-channel constrained convolution to detect traces that seem almost negligible for manipulation within facial images. This dual-extraction approach enhances the ability of the system to detect forgeries, with robust performance especially in compression scenarios. However, this model faces challenges such as high computational complexity, and reduced accuracy under extreme compression.

The paper [2] presents the MFM-LCFI framework for dealing with the challenges of detecting AI-generated fake faces. It addresses issues such as identifying subtle artifacts in low-quality images and improving generalization across different datasets. The framework integrates two innovative modules: the Multi-Feature Enhancement Module (MFEM), which captures subtle local discrepancies in spatial domains using shallow and global semantic features, and the Dual Frequency Decomposition Module (DFDM), which separates high- and low-frequency components while leveraging dual attention blocks to identify forgery clues in the frequency domain. This dual-branch design helps the model to analyze exhaustively spatial and frequency-based features, thereby significantly enhancing its ability to detect. The model has strong cross-dataset generalization, making it robust against unseen forgeries. However, the method comes with the challenge of efficiency and adaptability due to its reliance on computationally intensive dual-branch architecture and pre-trained models like EfficientNet-B4. This contribution helps to advance forgery detection. In this work, both spatial and frequency analysis are utilized so as to provide a better and robust solution for most deepfake detection challenges present nowadays.

This paper [3] gives an excellent overview of the development methods, challenges, and the datasets in deepfakes, especially focusing on swaps of faces and expressions. This paper discusses the evolutionary ways of manipulation methods which progressed from traditional autoencoder-based approaches to more modern, using GANs, CNNs, etc. Deepfake manipulation

techniques can be categorized into face generation, facial attribute modification, face swapping, and expression swapping which provides a structured overview of the field. The paper also points out the societal impacts of deepfakes, including threats to biometric security, risks of identity theft, and the spread of misinformation. It provides in-depth comparative analysis of various data sets used in training and evaluating deepfake detection models and conducts a review of state-of-the-art detection techniques including artifact-based, biometric-based, which may improve the detection accuracy as well as the robustness. Even though the current study exhaustively covers exploration of the domain, several challenges remain unaddressed, such as the possibility of low resolution or compression of videos posted on the social media. Moreover, both deepfake generation and detection are computationally intensive, and thus this restricts accessibility to small applications. This work sheds light on the current landscape of deepfake technology and detection. The critical challenges are discussed and directions for future research are provided.

This paper [4] proposes a CNN-based framework for deepfake detection by employing Diverse Gabor Filters to improve feature extraction. The proposed approach involves a uniform Gabor function that can offer linear, elliptical, or circular filters, which are then put into CNNs by backpropagation to allow for adaptive learning. The framework introduces Dual Scale Large Receptive Field Network (DSLRFN), a compact architecture created for deepfake detection; the model is using a self-attention mechanism together with a reduced parameter structure while maintaining high accuracy levels. Evaluation on four benchmark datasets demonstrates that the approach achieves competitive detection performance with a significantly reduced model size, cutting parameters by 64.9% without compromising accuracy. The framework offers notable advantages, including a compact architecture for efficient deployment, mathematically defined filters that improve interpretability, and robust performance across multiple datasets.

III. METHODOLOGY

The proposed deepfake detection system utilizes a hybrid approach by integrating MTCNN for face detection, EfficientNet for feature extraction, and CNN for classification. Below is a detailed explanation of each stage of the process:

1. Video Upload and Preprocessing:

• User Interaction: The process begins when a user uploads a video for analysis. The system is designed to handle videos in various formats such as MP4, AVI, and MOV.

• Video Frame Extraction: Once the video is uploaded, the system splits the video into individual frames. Frame extraction allows the system to analyze each video frame independently, detecting faces and other features that can help differentiate real content from forged content.

• Preprocessing: During preprocessing, the video frames are resized and normalized to ensure they meet the input size requirements of the face detection and feature extraction models. This step ensures consistent input quality for further processing.

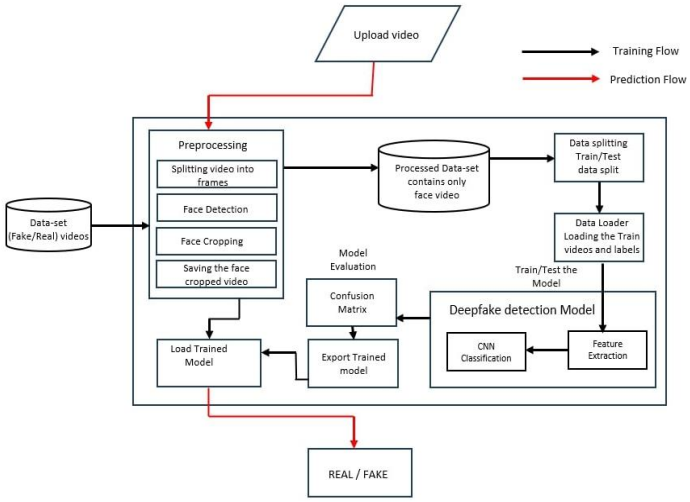

Figure 1: System Architecture

2. Face Detection with MTCNN:

• MTCNN (Multi-task Cascaded Convolutional Networks) is employed to detect faces in each frame of the video. MTCNN is a robust face detection model that uses a cascade of three neural networks (P-Net, R-Net, and O-Net) to detect faces and their landmarks (e.g., eyes, nose, and mouth) in varying poses and lighting conditions.

• Handling Challenges: One of the main challenges in face forgery detection is the quality and manipulation of faces in the video. MTCNN is highly efficient in detecting faces even in challenging scenarios like occlusion, varying angles, or low resolution, which makes it ideal for handling deepfake videos.

• Output of MTCNN: The output of MTCNN is a set of bounding boxes around detected faces, along with key facial landmarks. These faces are extracted and passed on for further analysis.

3. Feature Extraction with EfficientNet:

• After the faces are detected by MTCNN, EfficientNet is used for feature extraction. EfficientNet is a state-of-the-art Convolutional Neural Network (CNN) architecture that achieves high accuracy with fewer parameters. This makes it particularly well-suited for processing large amounts of video data while minimizing computational costs.

• Feature Extraction: EfficientNet processes the detected faces to extract relevant features such as texture, shape, and movement patterns. These features represent the facial attributes that can help differentiate between real faces and faces generated by deepfake techniques.

• Why EfficientNet: EfficientNet's ability to scale the network efficiently (with fewer parameters) while achieving high performance is critical for processing large-scale video datasets.

It balances accuracy and computational efficiency, making it a powerful tool for deepfake detection.

4. Classification with Convolutional Neural Networks (CNN):

• After the feature extraction step, a Convolutional Neural Network (CNN) classifier is used to classify the extracted features as "real" or "fake." The CNN is trained on a large dataset of real and deepfake faces, learning the subtle differences in facial expressions, textures, and artifacts that are commonly found in deepfakes.

• Training the CNN: The CNN classifier is trained using labeled data (real and fake images). It learns patterns such as unusual facial movements, inconsistencies in lighting or shadows, and artifacts (e.g., blurring, mismatched skin tones) that often appear in face forged videos.

• CNN Model Structure: The CNN model consists of several convolutional layers followed by pooling layers, which reduce the dimensionality of the feature maps while preserving important spatial features. Fully connected layers at the end of the network output a classification result: "real" or "fake."

• Detection Process: The CNN model processes the features extracted by EfficientNet and determines whether the video frame contains a real or fake face based on its learned patterns.

5. Face Forgery Detection Decision:

• The output of the CNN classifier is a binary classification result:

- Real: If the classifier determines that the features match those of a genuine, unaltered face.
- Fake: If the classifier identifies inconsistencies or artifacts typically associated with face forgery.

• Thresholding: A confidence threshold can be applied, where if the CNN outputs a probability higher than a certain value (e.g., 90%), the result is considered definitive. Otherwise, the system might flag the video for further review or analysis.

6. Display Detection Result to User:

• Once the system has classified the video, the final result (real or fake) is displayed to the user. The user is informed of the detection outcome along with a confidence score, which reflects the model's certainty about the classification.

• Visual Feedback: For deeper user engagement, the system may highlight specific frames or facial features where discrepancies were detected. This helps the user understand the reasoning behind the decision.

7. System Evaluation and Continuous Improvement:

• Evaluation Metrics: The system's performance is evaluated using common metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics help assess how well the model distinguishes between real and fake videos, especially in terms of minimizing false positives and false negatives.

• Continuous Learning: As the system encounters new deepfake videos, it can be periodically retrained with fresh datasets to improve its accuracy. The system can also adapt to emerging deepfake techniques through fine-tuning.

• User Feedback: If the system misclassifies a video (e.g., marking a real video as fake or vice versa), user feedback can be used to further refine the model and reduce misclassification in future predictions.

IV. RESULTS:

Figure 2: HOME PAGE

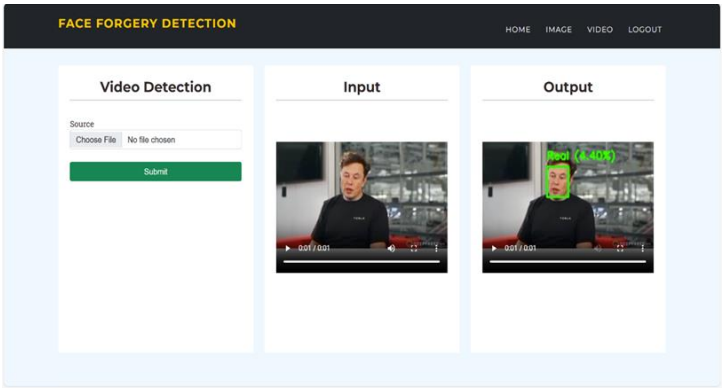

Figure 3: DETECTING VIDEO AS REAL

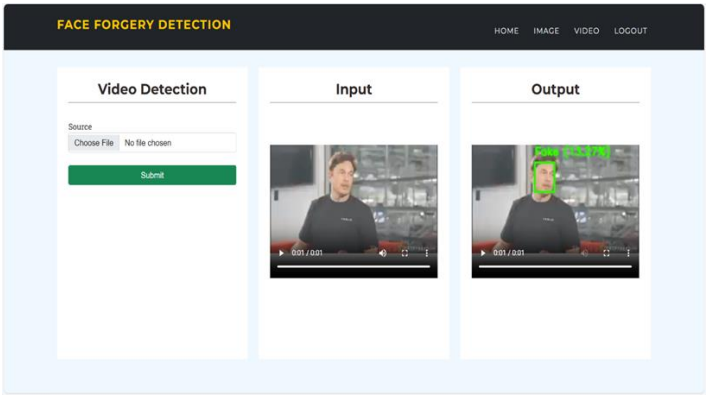

Figure 4: DETECTING VIDEO AS FAKE

V. CONCLUSION

In this paper, a strong face forgery detection system is proposed to defeat the challenges of manipulated facial media, such as deepfakes. The proposed technique relies on a CNN-based model trained on publicly available datasets that detects forged content by determining the subtle inconsistencies in patterns and facial features. This results in a high percentage of detection accuracy at around 96%, which makes this system effective against diverse types of face forgery techniques.

This work is a foundational step towards mitigating the misuse of face forgery technologies in social media, journalism, and security domains. Although the system performs well on benchmark datasets, further work is required to enhance its robustness in detecting forgeries generated by emerging techniques and in real-world, noisy environments. Future research directions include multimodal approaches that use audio-visual signals, real-time detection mechanisms, and lightweight models for deployment on edge devices.

By advancing the state-of-the-art in face forgery detection, this research contributes to the broader effort of ensuring the authenticity and integrity of digital media in an increasingly AI-driven world.

REFERENCES

[1] E Kim, S Cho: Exposing fake faces through deep neural networks combining content and trace feature extractors, IEEE Access (Volume: 9), 2021.
[2] Shuai Liu, Qian Jiang, Xin Jin, Zhenli He, Wei Zhou, Shaowen Yao: Multiple Feature Mining Based on Local Correlation and Frequency Information for Face Forgery Detection, 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI).
[3] Saima Waseem, Syed Abdul Rahman Syed Abu Bakar, Bilal Ashfaq Ahmed, Zaid Omar, Taiseer Abdalla Elfadil Eisa, Mhassen Elnour Elneel Dalam: DeepFake on Face and Expression Swap: A Review, IEEE Access (Volume: 11), 2023
[4] Ahmed H. Khalifa, Nawal A. Zaher, Abdallah S. Abdallah, Mohamed Waleed Fakhr: Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition, IEEE Access (Volume: 10), 2022.