DEEPFAKE: VIDEO FACE FORGERY DETECTION

[1] BITRAGUNTA VAMSI, [2] CHATARASUPALLI UDAY KIRAN, [3] CHIRUVELLA REDDY THARUN, [4] DEVARASETTY SAI ESWAR, [5] DR. D. PRASANNA. M.E., Ph.D, Associate professor
MAHENDRA ENGINEERING COLLEGE
[1] bitraguntavamsi@gmail.com, [2] udaykiranch70@gmail.com, [3] rt7684002@gmail.com, [4] Saieswar750@gmail.com

ABSTRACT

The rising incidents of video forgery in the digital space, stemming from breaches in information security, have led to a pressing need for monitoring visual and audio content to document such forgeries. The proliferation of counterfeit films heightens the potential for disorder and security threats. With the ease of posting, downloading, and sharing multimedia files online—encompassing audio, images, and videos—the increase in viruses contributes significantly to the prevalence of video forgeries. Recent technological advancements have facilitated mass media manipulation and made the production of false information more straightforward. The integrity of media is under severe threat due to the creation and wide distribution of deepfake content on social media, and identifying such content is believed to be challenging. A method for detecting deepfakes has been introduced to identify these video forgeries. To pinpoint deepfake videos, a Convolutional Neural Network (CNN) technique known as ResNet is utilized. The objective of this model is to enhance both the performance and accuracy of the detector in recognizing videos that have been modified using a specific method. The suggested approach extracts deep features through the ResNet CNN algorithm and then employs fundamental mathematical processes.

KEYWORDS: Video Dataset for Training, Uploading Videos and Preprocessing, Extracting Features with ResNet CNN Model, Classification through CNN, Detection of Forgery.

1. INTRODUCTION

The swift evolution of video editing applications has made it simpler for users to produce fabricated videos with the help of artificial intelligence algorithms, readily available video content, and intuitive tools. Manipulated videos, known as deepfakes, are becoming increasingly prevalent, leading to dangers such as the dissemination of false information, the evasion of facial recognition technology, and the creation of entertainment media. As the occurrence of fake videos rises, they contribute to societal discord and security concerns, underscoring the necessity for effective detection methods. Advances in deep learning have considerably improved the capability to identify fake videos, thanks to significant enhancements in processing power. A crucial technology involved in deepfakes is the Generative Adversarial Network (GAN), which creates fake videos by exchanging faces between individuals in a video. GANs utilize deep artificial neural networks (ANNs) that have been trained on target videos and facial images to precisely swap faces and expressions. By dividing the input image into individual frames, GANs produce highly convincing fake videos, which poses a formidable challenge for detection. While considerable research has concentrated on identifying copy-move forgeries in static images, the detection of inter-frame duplications in videos has not been as thoroughly explored, primarily due to challenges related to robustness and processing constraints. Present detection methods tend to be ineffective, computationally demanding, and lacking in accuracy. Furthermore, many current techniques are developed using limited datasets, which hinders the effectiveness of deep learning methods. These approaches also overlook the different frame rates found in videos, such as those encountered in films, rendering them unsuitable for real-time use.

2. LITERATURE REVIEW

Wu, Rongliang, et al. [1] introduced the Cascade EF-GAN, commonly referred to as the Cascade Expression Focal GAN. This model has been designed to minimize artifacts and blurring in generated facial images by incorporating local focuses, which enhance the preservation of identity-related features, particularly around the eyes, nose, and lips. Moreover, a new cascade transformation approach is proposed that segments significant facial expression changes into smaller steps. This technique effectively manages large expression transitions, allows for more realistic editing, and reduces overlapping artifacts. Comprehensive experiments conducted on two publicly available facial appearance datasets indicate that the Cascade EF-GAN outperforms current leading techniques in expression editing.

Li, Yuezun, et al. [2] presented Celeb-DF3, an innovative, challenging, and extensive dataset for the development and evaluation of DeepFake recognition algorithms. The Celeb-DF dataset comprises nearly 2 million frames, equivalent to 5,639 DeepFake videos. It includes excerpts from publicly available YouTube videos featuring 59 celebrities of varying ages, genders, and ethnicities that serve as the basis for the original source videos. The DeepFake films were generated using an advanced DeepFake separation technique. The algorithm utilized to produce the DeepFake videos in Celeb-DF plays a significant role in improving visual quality. Modifications to the fundamental DeepFake creation algorithm were made to address specific visual artifacts found in previous datasets.

Shen et al. [3] revealed that the latent space of GANs can be interpreted to enable semantic editing of faces. This study explores the inner workings of Generative Adversarial Networks (GANs), specifically examining how the latent space can be utilized for precise control over facial features in generated images. The fundamental innovation is the revelation that the high-dimensional latent space within a GAN is not merely chaotic data, but rather contains organized regions corresponding to meaningful semantic attributes. By pinpointing specific directions in the latent space, they achieved accurate and realistic modifications of facial characteristics. This implies that by skilfully navigating along a particular direction, one can systematically change specific facial features, such as age, gender, emotion, or even the inclusion of items like eyeglasses. The crux of this method is the ability to disentangle the latent representation, ensuring that alterations to one attribute do not unintentionally impact others. The implications of this research highlight the potential for highly controllable and intuitive facial editing, paving the way for creative uses in image manipulation, animation, and virtual reality. This work marks a shift from basic, global image transformations to more precise and semantically significant modifications.

Nirkin et al. [4] introduced FSGANv2, a refined approach to subject-agnostic face swapping and re-enactment. The tasks of face swapping and re-enactment present significant challenges in the field of computer vision, as they demand the seamless combination of facial features from a source face onto a target face while ensuring realism and coherence. This paper builds upon the foundational FSGAN framework, which sought to tackle these challenges in a way that is not dependent on specific individuals. The advancements in FSGANv2 have led to more realistic and stable outcomes in face swapping and re-enactment, addressing some of the limitations found in the initial FSGAN. These advancements likely involve enhancements to the blending algorithms, which play a vital role in the smooth integration of source and target faces. Furthermore, the improved method probably integrates more robust strategies for managing variations in lighting, pose, and facial expressions—common factors that can introduce artifacts in face swapping. The importance of FSGANv2 resides in its capability to generate high-quality results in face swapping and re-enactment with minimal

need for manual adjustments, which makes it a significant asset for various uses, such as in visual effects, content creation, and virtual interaction.

Nguyen et al. [5] offered a review of deep learning methods used for the creation and detection of deepfakes, highlighting key trends, challenges, and prospective future paths. The rise of deepfakes—AI-generated synthetic media that can convincingly imitate real individuals—raises substantial ethical and societal concerns. This survey paper provides an in-depth look into the swiftly changing realm of deepfake technology, exploring both the methodologies employed to fabricate these forgeries and the strategies devised for their detection. The survey underscores the growing sophistication in deepfake generation and the urgent need for effective detection methods. It likely investigates various deep learning models utilized in deepfake production, including GANs and auto encoders, while also analyzing techniques aimed at enhancing their realism, such as advanced training approaches, improved loss functions, and the integration of contextual details. Regarding detection, the survey probably discusses a variety of methods, including those that examine facial characteristics, inconsistencies in eye blinking, and artifacts found in audio and video signals. The paper likely points out the continuous "arms race" between creators of deepfakes and those developing detection systems, where advancements in one domain are promptly met with countermeasures in the other. It concludes by pinpointing significant challenges, like the necessity for more resilient and adaptable detection methods that can address novel and unseen deepfake techniques, and suggests potential future research directions in this essential field.

3. EXISTING METHODOLOGIES

One method based on meta-learning is the meta-deepfake detection (MDD) algorithm. That learns effective face representations with a meta-optimization goal on both the objective and artificial source fields. The source domain is moved to the target area by the MDD. The gradient from the Meta-instruct and meta-test is integrated using meta-optimization to improve the generality of the model. The MDD does not need special model adjustments to handle unseen domains. During training, the source domains were separated into meta-train domains (Ttrains) and meta-test domains (Ttests) in order to accomplish domain generalisation. Both genuine and fictional face pairs are included in this data, and the patterns are unique to this area. These pairs facilitate the collection and comparison of data between authentic and fraudulent photos. As a result, it also enhances inter-class separability, a distinct distribution of sample attributes that enhances model quality and encourages distinction during training. During optimisation, the network may be able to identify more distinctive features with less effort. In actuality, characteristics acquired by supervised learning are less likely to generalise when exposed to secret manipulation methods.

4. PROPOSED METHODOLOGIES

Current deepfake detection systems encounter considerable challenges, especially in coping with unfamiliar manipulation techniques and in generalizing across various domains. Conventional supervised learning methods depend heavily on labeled datasets, making it difficult for models to adjust to previously unseen fake generation methods. Furthermore, these systems frequently find it hard to distinguish between authentic and altered faces due to overlapping feature distributions and insufficient inter-class separability. This hinders their performance when faced with deepfakes generated using sophisticated or innovative techniques. Additionally, traditional models require regular updates or retraining to adapt to new domains, which can be both computationally intensive and time-consuming. To enhance the recognition of AI-generated fake videos, the proposed Deepfake Manipulated Video Identification system incorporates advanced meta-learning algorithms along with a hybrid Convolutional Neural Network (CNN). The

system initiates with a training phase in which a dataset of real and manipulated videos is analyzed by breaking down the videos into individual frames. Each frame undergoes face detection and feature extraction through ResNet CNN, effectively capturing the spatial discrepancies between authentic and fake faces. Moreover, an analysis of temporal consistency is performed to spot irregularities across sequential frames. The training dataset is separated into meta-train (Ttrains) and meta-test (Ttests) domains, facilitating domain generalization and aiding the model in adapting to unforeseen manipulation techniques. To enhance learning efficiency, the system employs linear bottlenecks and inverted residual blocks to minimize spatial information loss while improving memory utilization, ensuring that the model can generalize effectively across various domains. During the testing phase, the system processes videos uploaded by users by converting them into frames, implementing ResNet-based face detection, and extracting both spatial and temporal features. These features are classified using the trained model, which compares them to previously learned patterns to ascertain whether the video is genuine or manipulated. If a video is identified as manipulated, it is labeled as a deepfake and incorporated into the training dataset as part of a dynamic feedback loop. This approach enables the system to continually learn from emerging deepfake techniques, ensuring its adaptability and long-term effectiveness. By merging hybrid CNNs for feature extraction, meta-learning for domain generalization, and a feedback loop for ongoing improvement, the system provides an efficient and resilient framework for detecting deepfake videos while preserving high accuracy and adaptability to new threats.

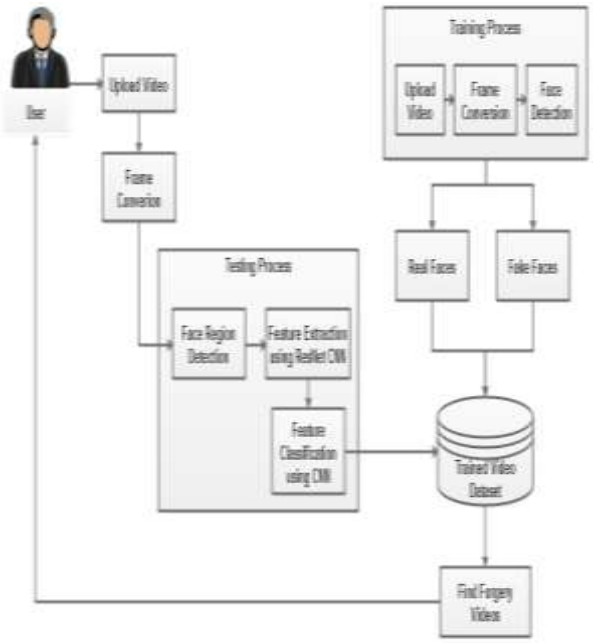

Fig.1 Proposed Architecture

5. IMPLEMENTATION

Data Preprocessing:

Videos are divided into separate frames, which are then resized and normalized to ensure uniformity. This stage is crucial for properly preparing the data for subsequent analysis and enhances the efficiency of the system.

Face Detection and Feature Extraction:

Each frame is subjected to face detection to focus on the face, followed by the application of ResNet to extract features. This procedure emphasizes subtle variations between authentic and altered faces, which is essential for making a distinction.

Loss Function:

The loss function plays a vital role in steering the optimization process during training, ensuring an equitable performance balance between genuine and fabricated videos while minimizing classification errors and learning robust features.

Frame Extraction:

Videos are dissected into individual frames, facilitating an in-depth examination of each one. This phase aids in spotting minor irregularities that may suggest alterations.

Face Detection and Feature Extraction (Repeated):
Each frame experiences face detection paired with feature extraction via ResNet, which captures intricate details necessary for differentiating genuine faces from those that have been manipulated.

Temporal Feature Analysis:
The system also investigates inconsistencies across successive frames. These temporal discrepancies are often found in deepfake videos and assist in detecting manipulations.

Optimization Techniques:
To enhance memory usage, avert the loss of information, and accelerate learning, linear bottlenecks and inverted residual blocks are incorporated into the CNN architecture.

Video Processing:
During testing, user-uploaded videos are divided into frames that undergo the same face detection and ResNet-based feature extraction procedures as those used during training.

Feature Classification:
The features extracted from each frame are evaluated against patterns learned previously. This analysis allows the system to determine whether the video is genuine or manipulated based on both spatial and temporal characteristics.

6. METHODOLOGY

A Convolutional Neural Network (CNN) integrated with Residual Networks (ResNet) merges the strengths of CNNs for image recognition with ResNet's advantages in dealing with the vanishing gradient issue in deep networks.
The procedural steps include:

- Data Preparation
- Frame Extraction
- Face Detection and Feature Extraction
- Temporal Feature Analysis
- Define the CNN Architecture
- Optimization Techniques
- Loss Function
- Feature Classification

7. EXPERIMENTAL RESULTS

The deepfake datasets are sourced from the KAGGLE website, which contains both genuine and modified images. The edited images feature faces that are generated in various ways and the framework is developed using Python environments. Each image consists of a 256 by 256 jpeg depicting either a real or altered human face. System performance can be measured through Precision, Recall, F1 Score, and Accuracy.

7.1 Precision: High precision indicates that when the model identifies a video or image as a deepfake, it is usually correct. Conversely, low precision suggests the model may mistakenly classify genuine videos as deepfakes, resulting in wrongful accusations.

7.2 Recall: Also referred to as Sensitivity or True Positive Rate, high recall suggests that the model effectively detects the majority of deepfakes, albeit at the risk of misclassifying some real videos as deepfakes. If recall is low, the model may overlook many deepfakes, failing to recognize them.

7.3 F1 Score: The F1 score is beneficial in situations that require a balance between precision and recall. For deepfake detection, this balance is critical since failing to identify a deepfake (low recall) and incorrectly tagging genuine content (low precision) can both carry severe repercussions.

7.4 Accuracy: A high accuracy rate indicates that the model predominantly makes correct predictions regarding both deepfake and real content.

DISCUSSION OF PROPOSED WORK

By integrating advanced techniques like SE-ResNet and CBAM, which effectively identify altered facial regions, the proposed innovation enhances Deepfake detection through an upgraded ResNet framework.

Moreover, multi-branch architectures are created to simultaneously handle inconsistencies in the spatial (RGB) and frequency domains. ResNet transformer is employed to ensure temporal coherence, thereby enabling a temporal loss function that maintains consistency from frame to frame. Additionally, lightweight ResNet models designed for adversarial training and knowledge distillation improve real-time detection and resilience against malicious attacks. These advancements focus on enhancing the precision, effectiveness, and interpretability of Deepfake video detection.

Table 1. Accuracy Measurement

Algorithm	Accuracy (%)
Meta Deepfake Detection (MDD)	80
ResNet CNN	93

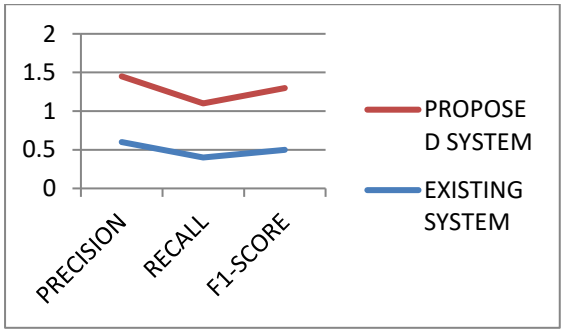

Fig.2 Performance Output

8. CONCLUSION

The proposed system for detecting deepfakes effectively tackles the problem of recognizing manipulated videos by utilizing sophisticated deep learning approaches. By employing a convolutional neural network (CNN) alongside ResNet for feature extraction, the system captures both the spatial and temporal aspects of video sequences. The addition of a sequence descriptor guarantees a strong representation of the features extracted, which are then classified by a detection network comprised of fully connected layers to identify videos as either authentic or deepfake. This approach not only achieves precise detection but also enhances the system's ability to adapt to new deepfake methods. Through the integration of state-of-the-art architectures and analysis of temporal elements, the project presents a dependable and efficient solution for ensuring the authenticity of digital media in a time marked by the rise of deepfake manipulation.

REFERENCES

[1] Wu, Rongliang, Gongjie Zhang, Shijian Lu, and Tao Chen."Cascade ef-gan: Progressive facial expression editing with local focuses." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5021-5030. 2020.

[2] Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, and SiweiLyu. "Celeb-df: A large-scale challenging dataset for deepfake forensics." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3207-3216. 2020.

[3] Shen, Yujun, JinjinGu, Xiaoou Tang, and Bolei Zhou. "Interpreting the latent space of gans for semantic face editing."In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9243-9252. 2020.

[4] Nirkin, Yuval, Yosi Keller, and Tal Hassner. "FSGANv2: Improved subject agnostic face swapping and reenactment." IEEE Transactions on Pattern Analysis and Machine Intelligence 45, no. 1 (2022): 560-575.

[5] Nguyen, ThanhThi, Quoc Viet Hung Nguyen, Dung Tien Nguyen, DucThanh Nguyen, Thien Huynh-The, SaeidNahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. "Deep learning for deepfakes creation and detection: A survey." Computer Vision and Image Understanding 223 (2022): 103525.

[6] Rana, MdShohel, et al. "Deepfake detection: A systematic literature review." IEEE access 10 (2022): 25494-25513.

[7] Guarnera, Luca, et al. "The face deepfake detection challenge." Journal of Imaging 8.10 (2022): 263.

[8] Almutairi, Zaynab, and HebahElgibreen. "A review of modern audio deepfake detection methods:challenges and future directions." Algorithms 15.5 (2022): 155.

[9] Taeb, Maryam, and Hongmei Chi. "Comparison of deepfake detection techniques through deep learning." Journal of Cybersecurity and Privacy 2.1 (2022): 89-106.

[10] Patel, Yogesh, et al. "Deepfake generation and detection: Case study and challenges." IEEE Access (2023).