DEEPFAKE VIDEO DETECTION SYSTEM

[1]NAGESH MALI, [2]ATHARVA TANKAR, [3]ANVIT GONDIKAR, [4]SURESH A

[1]PG Scholar Student, Vellore Institute of Technology, Tamil Nadu - 632014
[2]PG Scholar Student, Vellore Institute of Technology, Tamil Nadu - 632014
[3]PG Scholar Student, Vellore Institute of Technology, Tamil Nadu - 632014
[4]Assistant Professor, Vellore Institute of Technology, Tamil Nadu - 632014

Abstract:. The rapid advancement of deepfake methods in recent years has made facial video forgery capable of producing extremely convincing video content, posing serious security challenges. Additionally, detecting such fake videos has become much more difficult and imperative. Most detection methods currently employed view the problem as a straightforward binary classification task. Since the difference between forged and real faces is minimal, this article treats the subject as an independent fine-grained categorization problem. It has been proven that most face forgery methods available today leave similar artifacts in both the spatial and temporal dimensions, such as generative defects in the spatial domain and frame inconsistencies in the time domain. Moreover, the suggested spatial-temporal model is divided into two parts, each designed to obtain temporal and spatial forgery traces from a global viewpoint. The former half of the spatial domain is employed to capture the artifacts in successive frames, while the latter half captures artifacts within a single frame. A novel long-distance attention mechanism is employed in the construction of both parts. Finally, similar to previous fine-grained categorization approaches, the network is guided by attention maps to focus more on salient facial components.

Index terms: Attention mechanism; Face modification; Spatial artifacts; Temporal

I. INTRODUCTION

New AI, deep learning, and image processing technology have simplified the production of deepfake movies. A one-minute video that showcases the previous US president A video of Barack Obama uttering things he never said went viral in April 2018. Since they can convincingly mimic reality, distort the perception of viewers, and bring down the truth, deepfake films are hazardous. Such material might further spread as social networks expand, maybe making the issues surrounding conspiracy theories and disinformation worse. The faces of several popular celebrities, comedians, entertainers, and political figures were hacked and inserted into pornographic clips in one of the earliest examples of deepfakes. To achieve the Obama video highly realistic and believable, nearly 56 hours' worth of sample recordings were produced. There was not much alarm when early deepfake movies aimed at celebrities emerged. As opposed to traditional Hollywood-style fake videos, normally developed by hand from image manipulating programs such as Adobe Photoshop, they are a lot easier to produce. Face swapping is achieved in deepfake movies through using deep learning methodologies on large samples of video images; the higher the level of realism, the bigger the samples. Deepfake movies are, in the views of Stover and Floridi, a data catastrophe.

They also advocated advocating for the implementation of new technology and the responsible and ethical transmission of digital content on social networking sites. Development would need to be methods for detecting, halting, and fighting against deepfake digital content, including fake audio, video, photographs, paintings, and so forth. It will not be hard to accomplish this if there is a safe, sure, and trusted way of tracing the history of digital content. In order to ensure the authenticity and originality of digital content, users should be able to access accurate information regarding its actual point of origin and trace an item's history. Customers can prevent themselves from being cheated or convinced to accept counterfeit digital content by employing this method. There are means to verify that physical art is genuine at present. For example, you are issued a certificate of authenticity when you buy a piece of art. Also, this document from a reliable and established authority can be forged or discovered as unsigned. Also, it is more difficult to establish the provenance of artwork purchased in a secondary market. In a sense, getting the original provenance of the artwork takes the buyer quite a physical exertion and attestation. There
are no longer any recognized procedures for verifying the legitimacy of digital music, images, or films shared over the Internet. Such digital content cannot be submitted through COA. It is very hard to identify the real origin of a digital object posted in a credible and

legitimate manner. To measure digital content reliability, the average internet user relies on relevant blogs, comments, and reviews available on the internet. Proof of Authenticity (PoA) system is therefore acutely necessary for internet-delivered digital information in an attempt to pinpoint quality published sources as well as therefore combat deepfaked video, audio, as well as imagery content.

I.A MOTIVATION

The increasing occurrence of deepfake movies, audios, photos, and other types of digital content, which seriously jeopardize authenticity, truth, and confidence in the digital age, is what inspired this effort. This project was driven by the increasing prevalence of deepfake video, images, and other digital content, which severely undermines truth, authenticity, and trust in the age of the internet. Modern methods enabled the creation of fake digital content, which has the ability to cause misinformation and destabilize society. One of the key solutions for this issue is to Proof of Authenticity (PoA) for content, which supports trustworthy authentication of the content origin and history. It is challenging to build confidence in the authenticity of digital media since current solutions frequently lack reliable systems for tracking its provenance and full history.

I.B PROBLEM STATEMENT

Fake digital information, including deepfake audio, video, and image, has alarmingly increased the quick development of deep learning and artificial intelligence. Because such content can promote false narratives as true, distort reality, propagate misinformation, and undermine trust, this phenomena poses serious hazards to society. To solve this problem, a strong system for establishing and confirming the legitimacy of digital media is needed. Not with standing the fact that Proof of Authenticity (PoA) mechanisms have been advanced as a method for fighting malicious content, current solutions cannot effectively trace the history and origin of digital media.

II. SYSTEM ARCHITECTURE

The system architecture represents a deepfake detection framework that analyzes video inputs to verify their authenticity. The process starts with video input from the user, which undergoes preprocessing to enhance and standardize the data. Key features are then extracted and passed to a CNN-based classification model trained on a dataset. The system finally determines whether the video is real or fake and displays the result to the user. The use of CNN enhances the system's ability to identify subtle artifacts common in deepfake videos.

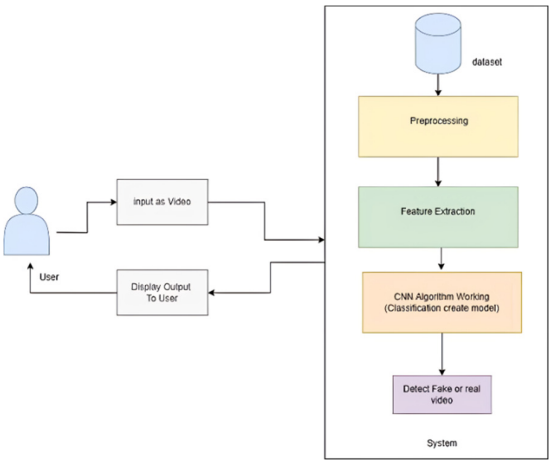

Figure 1: System Architecture

A set of video samples is employed at the beginning of the process, and it is passed through a controlled workflow to ensure accurate detection. To prepare the data for further analysis, the movies are initially subjected to preprocessing, which involves scaling, conversion of RGB images to grayscale, and frame extraction. Following preprocessing, the system extracts spatial and temporal features to differentiate genuine and fake videos through feature extraction. A Convolutional Neural Network (CNN), utilizing these extracted features to train a model capable of identifying deepfake patterns, classifies the films. The system verifies the processed video in the final step and determines whether it is genuine or not. Users interact with the system through submitting videos, which are then processed by this pipeline. The system gives the user the results of classification upon completion of the analysis. Through machine learning and deep learning approaches, this systematic methodology enhances the dependability and precision of deepfake detection and ensures an efficient and automatic verification process.

III. CONTENTS OF DATASET

The dataset used in this study was composed of frames extracted from deepfake and authentic video sources, divided into two sets: Frame Set 1 and Frame Set 2. The training set consisted of 1600 frames from Frame Set 1 and 1400 frames from Frame Set 2, while the testing set contained 1124 frames and 720 frames, respectively. The model demonstrated a high training accuracy of 97.54%, reflecting its ability to learn the distinguishing features between authentic and manipulated frames effectively. The notable disparity between the training and testing accuracies suggests potential overfitting, necessitating further improvements.

Dataset	Frames Set 1	Frames Set 2
Training Set	1600	1400
Testing Set	1124	720

Table.1 Frame Distribution

Dataset	Split Data	Accuracy
Training Set	80%	97.54%
Testing Set	20%	62.96%

Table.2 Dataset Accuracy

IV. DATASET IMAGES

A well-curated dataset is essential for achieving high accuracy in classification.

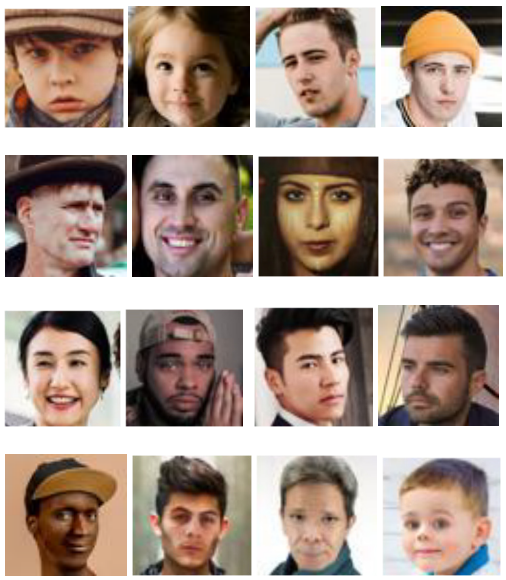

Figure 2: Sample of training and testing dataset

V. METHODOLOGY

This project seeks to create a Deepfake Video Detection system. Such systems are designed to identify and differentiate between authentic movies and those that have been altered using modern techniques. What follows in-depth and descriptive method that outlines the technologies used and how the system works.

V.A ALGORITHMS

CNN: A type of deep learning models that are specifically used for structured input, like images and videos, are referred to as convolutional neural networks, or CNNs. They perform very well on tasks such as segmentation, object detection, and image classification involving pattern and spatial information.

Convolution Layer: The core component of CNNs extracts spatial features such as edges, textures, and patterns by convolving filters, or kernels, with feature maps or input images. Each filter that operates on the input performs a convolution operation, which is an element-wise multiplication and addition.

Activation Function: After the process of convolution, nonlinear activation functions such as the Rectified Linear Unit (ReLU) are employed to introduce nonlinearity so that the network can recognize intricate patterns. The pooling layer reduces the feature maps to minimize their size and processing complexity. Max pooling, which captures the maximum value in each area, and average pooling, which captures the average of values, are some examples of popular pooling operations. Each neuron in the higher layer is connected with all the neurons in the lower layer through a fully connected layer takes the high-level features learned by the convolutional and pooling layers and uses them as a classifier.

V.B LIBRARIES

The main Python package for numerical computation is known as NumPy (Numerical Python). It provides tools and high-performance arrays for working with large datasets.

A library for computer vision that provides features to process images and videos is referred to as OpenCV (Open Source Computer Vision Library).

A high-level interface for constructing and training neural networks is referred to as Keras (High-level Neural Networks API). It simplifies model construction and sits on top of TensorFlow.

Pillow (Image Processing Library): Pillow is a Python image processing library which is a Pillow (PIL fork).

Matplotlib, also commonly referred to as the Data Visualization package, is a Python plotting library.

The Machine Learning and Deep Learning Library, or TensorFlow: TensorFlow is an open-source package utilized to construct and train machine learning

models. Applications that utilize deep learning extensively utilize it.

VI. RESULTS

The output of the model is going to be whether the video's frames are deepfake or a real.

Figure 3: Detection of fake video frame

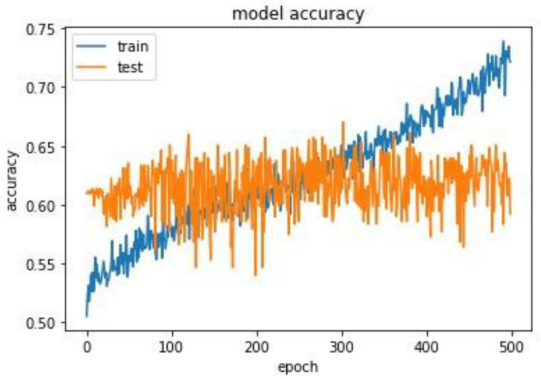

Figure 4: Model Accuracy graph

CONCLUSION

An application of deepfake video detection that has been designed using Python, TensorFlow, Keras, OpenCV, and NumPy is demonstrated here. The algorithm incorporates a Convolutional Neural Network (CNN) to correctly identify manipulated video frames. An ample dataset comprising genuine and deepfake films was obtained and preprocessed using OpenCV in a manner to ensure the effectiveness of the model. The CNN was trained on the processed frames, which allowed it to recognize subtle spatial irregularities and anomalies that are commonly exhibited by deepfake videos. Deepfakes in streaming videos can now be detected efficiently due to the successful incorporation of the proposed model into a real-time video analysis pipeline.The robustness of the system in identifying tampered video frames was proved with its high training accuracy of 97.54%. Future improvements may focus on increasing the model's accuracy through addressing audio-visual inconsistencies and incorporating hybrid structures such as Vision Transformers (ViTs). With the provision of an efficient and scalable means of ensuring the validity of digital content, this research aids the ongoing endeavors to curb the dissemination of misinformation.

REFERENCES

[1] H.R. Hasan, K. Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts," *IEEE Access*, vol. 7, pp. 41596–41606, 2019.

[2] G. Pang, B. Zhang, Z. Teng, Z. Qi, J. Fan, "MRE-Net: Multi-rate Excitation Network for Deepfake Video Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[3] S. Saif, S.S. Ali, S. Kausar, A. Jameel, "Generalized Deepfake Video Detection through Time-distribution and Metric Learning," *IEEE IT Professional*, 2022..

[4] Q. Yin, W. Lu, B. Li, J. Huang, "Dynamic Difference Learning with Spatio-temporal Correlation for Deepfake Video Detection," *IEEE Transactions on Information Forensics and Security*, 2023.

[5] W. Lu, L. Liu, B. Zhang, J. Luo, X. Zhao, Y. Zhou, J. Huang, "Detection of Deepfake Videos Using Long-distance Attention," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[6] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I.E. Davidson, T.F. Mazibuko, "An Improved Dense CNN Architecture for Deepfake Image Detection," *IEEE Access*, 2023.

[7] M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 97213–97230, 2022.

[8] M.K. Johnson, H. Farid, "Exposing Digital Forgeries in Complex Lighting Environments," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 450–461, 2007.

[9] J. Siegelman, "Analytical Survey of Deepfake Video Detection and Emotional Insights" *Issues in Information Systems*, vol. 25, no. 2, pp. 96–112, 2024.

[10] T. Koike, M. Ito, H. Yamasaki, "A Lightweight Temporal Attention Network for Real-time Deepfake Detection on Mobile Devices" *Sensors*, vol. 24, no. 3, 2024.

[11] N.M. Alnaim, Z.M. Almutairi, M.S. Alsuwat, H.H. Alalawi, A. Alshobaili, F.S. Alenezi, "DFFMD: A Deepfake Face Mask Dataset for the Infectious Disease Era," IEEE Access, vol. 11, pp. 8837–8846, 2023.

[12] A. Hamza, A.R. Javed, F. Iqbal, N. Kryvinska, A.S. Almadhor, Z. Jalil, R. Borghol, "Deepfake Audio Detection via MFCC Features Using Machine Learning," IEEE Access, vol. 10, pp. 55614–55624, 2022.

[13] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I.E. Davidson, T.F. Mazibuko, "An Improved Dense CNN Architecture for Deepfake Image Detection," *IEEE Access*, 2023.

[14] S.Ramachandran, A.V.Nadimpalli, A. Rattani, "Experimental Evaluation of Deepfake Detection Using Deep Face Recognition," *arXiv preprint* arXiv:2110.01640, 2021.

[15] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, K.A. Lee, "ASVspoof 2021: Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, pp. 199–214, 2023.

[16] C.C. Hsu, S.N. Chen, M.H. Wu, Y.F. Wang, C.M. Lee, Y.S. Chou, "GRACE: Graph-Regularized Attentive Convolutional Entanglement for Deepfake Detection," *arXiv preprint* arXiv:2406.19941, 2024..

[17] P. Liu, Q. Tao, J.T. Zhou, "From Single-modal to Multi-modal Deepfake Detection: A Survey," *arXiv preprint* arXiv:2406.06965, 2024.

[18] H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, O. Etzioni, "Tug-of-War Between Deepfake Generation and Detection," *arXiv preprint* arXiv:2407.06174, 2024.

[19] S.A. Khan, D.T. Dang-Nguyen, "Deepfake Detection: A Comparative Analysis," *arXiv preprint* arXiv:2308.03471, 2023.

[20] S. Ramachandran, A.V. Nadimpalli, A. Rattani, "Experimental Evaluation of Deepfake Detection Using Deep Face Recognition," *arXiv preprint* arXiv:2110.01640, 2021.

★★★