International Journal of Research and Development in Engineering Science
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 7, Issue.2 (March – April), 2025 , Pages: 8 -12
Received: 18 July 2024 ; *Accepted:* 27 December 2024 ; *Published:* 12 March 2025
https://doi.org/10.63328/IJRDES-V7RI2P2
Original Paper : Cross-Sectional Research – Methodology Study Paper

A Robust Deep Learning approach for Discriminating between Authentic and AI-Generated Videos

Harshitha B [1], V. Deepa Priya [2], Nikitha B [3] , Subashini K [4], Magasakthi S [5]

[1-5] Department of Information Technology, Kamaraj College of Engineering and Technology, Virudhunagar, India;
harshithabhaskaran1@gmail.com , deepapriyakcet@gmail.com , nikithab2003@gmail.com , subashinikannan03@gmail.com , shenbasakthi15@gmail.com

** Corresponding Author: harshithabhaskaran1@gmail.com*

Abstract: The rapid advancement in processing power has empowered deep learning algorithms to produce remarkably convincing human-synthesized videos, commonly referred to as "deep fakes." This technological progression raises concerns about potential malicious applications, such as blackmail, manipulation through revenge porn, or the exploitation of political unrest using realistic face-swapping deepfakes. In response to these challenges, we propose a novel deep learning-based technique designed to reliably differentiate between genuine videos and those generated by artificial intelligence. Our approach introduces an innovative method for automatically detecting replacement and recreation deep fakes. Leveraging the capabilities of Artificial Intelligence (AI) to combat AI-driven threats, our system employs a two-step process. Initially, frame-level features are extracted using a Res-Net Convolutional Neural Network (CNN). Subsequently, these features serve as input for training a Deep Neural Network (DNN) based on a Recurrent Neural Network (RNN) architecture. To assess the effectiveness of our method, we conducted extensive evaluations on a substantial and diverse dataset. This dataset was meticulously curated by combining various sources, including Face-Forensic++, Deepfake Detection Challenge, and Celeb-DF, aiming to simulate real-time scenarios and enhance the model's performance on real-world data. Our results demonstrate the system's capability to discern alterations in videos, effectively distinguishing between deep fakes and authentic content. Furthermore, we showcase the simplicity and reliability of our approach, illustrating its competitive performance. This research contributes to the ongoing efforts in developing robust solutions for identifying and mitigating the risks associated with the proliferation of AI-generated deep fake content in various contexts.

Keywords: Res-Next CNN, RNN, Deep Neural Network, Computer Vision.

1. Introduction

A method called "deep fake" is used to simulate human images using neural network components such as auto encoders and GANs (Generative Adversarial Networks). These programs use a deep learning technique to superimpose target pictures onto source films, producing a realistic-looking deepfake video. It is impossible to distinguish the differences in these deepfake videos with the human eye since they are so realistic. In this study, we provide a novel deep learning-based technique that can reliably discriminate between real videos and artificial intelligence-generated phony ones. Our effective method of differentiating between pristine and deep fake videos is based on the limitations of the deep fake generating tools. Although the existing deep fake creation technologies may not be obvious to humans, they leave behind some distinct artifacts in the frames that trained neural networks may detect during the development process. We show that Res-Next Convolution Neural Networks can effectively capture the unique artefacts left in the Deep Fake videos produced by Deepfake generation tools. Res-Next Convolution Neural Networks are used by our system to retrieve frame-level characteristics. Then, using these attributes, a Deep Neural Network (DNN) based Recurrent Neural Network (RNN) is trained to determine whether or not the video has been altered, or if it is a deep fake or a true video.

We suggested testing our approach on a sizable collection of deepfake films gathered from several online video portals. We have attempted to improve the performance of the deep fake detection model using real-time data. We used a variety of accessible datasets to train our model in order to accomplish this. in order for our model to acquire

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 2, 2025 , https://doi.org/10.63328/IJRDES-V7RI2P2

the features from various types of photos. From the Face-Forensic++ [1], Deepfake detection challenge [2], and Celeb-DF [3] datasets, we sufficiently extracted videos. To get competitive outcomes in real-time scenarios, we assessed our model using a YouTube dataset, which contains a substantial amount of real-time data.

2. Related Work

Agreed that deep fakes cause a significant threat to everyone if used for harmful purposes such as phishing, scam, and identity theft, reducing the trustworthiness of the public data. A convolutional vision transformer is used in this project to detect the deep fake. This project adds a CNN module to the VIT architecture as CNN extracts the features (facial features in an image), and VIT takes those features as input to categorize them into a specific class. the datasets used are Face Forensics++Face swap, FaceForensics++,deepfakedetaction,FaceForensics++deepfake,FaceForensics+ face shifter, Face Forensics++ neural textures. the results are an accuracy of 91.5%, an AUC value of 0.91, and a loss value of 0.32, which indicates the difference between the predictions and the actual results. CNN and RNN had an accuracy of 92.6% (validation) and 91.88% (testing). CViT had 87.25 (validation) and 91.5(testing). The face recognition library had the best results compared to MTCNN and Blaze Face libraries as it had higher accuracy [2].

Discussed about AI-synthesized face-swapping videos, commonly known as Deep Fakes, are an emerging problem threatening the trustworthiness of online information. Developing and evaluating Deepfake detection algorithms calls for large-scale datasets is needed. The dataset used is the Celeb-DF dataset. Celeb-DF is generally the most challenging to the current detection methods, and their overall performance on Celeb-DF is lowest across all datasets, with an average AUC of 56.9 %. At the same time, FF-DF had the highest results across all methods, with an average AUC of 82.3 %. While the method that had the highest performance across all datasets was DSP-FWA which had an AUC of 87.4 %, and the lowest performance goes to the Head Pose method, with an AUC of 58.7 %. The celeb-DF dataset has proven that it still needs improvement [4].

Agreed that the huge use of digital images has been followed by a rise of methods to change image contents, using editing software like Photoshop. The digital image forensics research field is dedicated to detecting fake images to regulate the circulation of such fake content. Providing two possible network architectures (Meso 4 and MesoInception 4) to detect such forgeries efficiently with a low computational cost. The dataset used is the Deepfake dataset, the Face2Face dataset [5]. Both networks have reached close scores, around 90 %, considering each frame independently. A higher score is not expected as some images have facial extractions with a very low resolution. It is observed a decline of scores at the strong video compression level. The image aggregation significantly enhanced both detection rates. It even rose greater than 98 % with the MesoInception, network on the Deepfake dataset. Note that on the Face2Face dataset, the score is the same for both networks, but the misclassified videos are dissimilar.

Agreed that Deep faking greatly impacts our environment nowadays; it's created by putting faces using deep neural networks into original images/videos. Together with additional forms of misinformation shared through the digital social network, digital impersonations were created by deepfake, which have become a real problem with negative social influence. Accordingly, there is a serious need for successful methods to detect Deep Fakes. The dataset used is UADFV and a subset of the DARPA. The results are the SVM classifier reaches an AUROC of 0.89. This means that the difference between head poses evaluated from the central region and the whole face is a good feature for identifying Deep Fake generated photos.

DARPA GAN Challenge dataset, the AUROC of the SVM classifier is 0.843. The reason is that the synthesized faces in the DARPA GAN challenges are mostly blurry, leading to a struggle to predict facial landmark places accurately and, as a result, the head pose evaluation. They also calculated the performance using separate videos as the unit of analysis for the UADFV dataset. This is achieved by taking the average of the classification prediction on frames over separate videos [6]. Also, other related work in [7-29] has been proposed in recent years to address machine learning and its application in different fields.

3. Proposed Work

We have gathered the data from different available datasets like Face Forensic++(FF) , Deepfake detection challenge (DFDC) , and Celeb-DF and created our own new dataset. The pre-processing module focuses on refining videos by removing noise and detecting and cropping faces. Videos are split into frames, and a computational threshold of 150 frames. The dataset is then divided into training and testing sets with a 80-20 split for real and fake videos.

The model architecture involves a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), utilizing a pre-trained ResNet CNN for feature extraction and a Deep Neural Network (DNN) for classification.

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 2, 2025 , https://doi.org/10.63328/IJRDES-V7RI2P2

3.1. **Problem Statement:** Detecting deepfakes (fake videos made using AI) is tough, but important. People can misuse them for bad things like spreading false information or creating fake events. To stop this, we're using advanced computer programs that learn to recognize signs of fake videos.

3.2. **Objectives :** To create a model that will classify a video is deepfake or real. To provide a user-friendly system for used to upload the video and distinguish whether the video is real or fake.

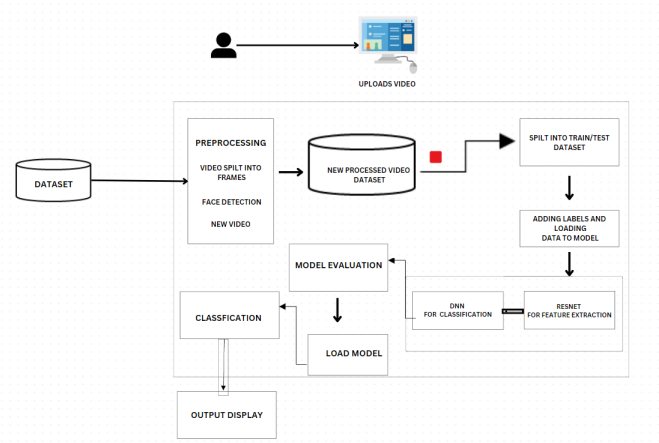

Figure.1 System Architecture

3.3. Algorithm Details

3.3.1. Preprocessing Details

Using glob, we imported all the videos in the directory in a python list. cv2.VideoCapture is used to read the videos and get the mean number of frames in each video. To maintain uniformity, based on mean a value 150 is selected as idea value for creating the new dataset. The video is split into frames and the frames are cropped on face location.

The face cropped frames are again written to new video using Video Writer. The new video is written at 30 frames per second and with the resolution of 112 x 112 pixels in the mp4 format. Instead of selecting the random videos, to make the proper use of DNN for temporal sequence analysis the first 150 frames are written to the new video.

3.3.2. Model Details

The model consists of following layers:

ResNet CNN: The pre-trained model of Residual Convolution Neural Network is used. The model's name is resnext50_32x4d () [6-9]. This model consists of 50 layers and 32 x 4 dimensions. Figure shows the detailed implementation of model.

stage	output	ResNeXt-50 (32×4d)
conv1	112×112	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 [1×1, 128 3×3, 128, C=32 1×1, 256] ×3
conv3	28×28	[1×1, 256 3×3, 256, C=32 1×1, 512] ×4
conv4	14×14	[1×1, 512 3×3, 512, C=32 1×1, 1024] ×6
conv5	7×7	[1×1, 1024 3×3, 1024, C=32 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax
# params.		25.0×10^6

Figure.2 Resnet Architecture

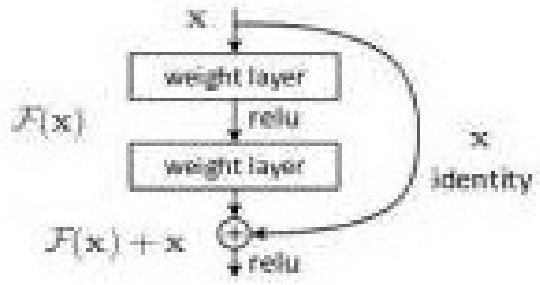

Figure.3 Resnet Working

3.3.3. Model Training Details

Train Test Split: The dataset is split into train and test dataset with a ratio of 80% train videos (4,200) and 20% (1,800) test videos. The train and test split are a balanced split i.e., 50% of the real and 50% of fake videos in each split.

Data Loader: It is used to load the videos and their labels with a batch size of 4.

Training: The training is done for 20 epochs with a learning rate of 1e-5 (0.00001), weight decay of 1e-3 (0.001) using the Adam optimizer.

Adam optimizer [21]: To enable the adaptive learning rate Adam optimizer with the model parameters is used.

Cross Entropy: To calculate the loss function Cross Entropy approach is used because we are training a classification problem.

Softmax Layer: A Softmax function is a type of squashing function. Squashing functions limit the output of the function into the range 0 to 1. This

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 2, 2025 , https://doi.org/10.63328/IJRDES-V7RI2P2

allows the output to be interpreted directly as a probability. Similarly, softmax functions are multi-class sigmoid, meaning they are used in determining probability of multiple classes at once. Since the outputs of a softmax function can be interpreted as a probability (i.e., They must sum to 1), a softmax layer is typically the final layer used in neural network functions. It is important to note that a softmax layer must have the same number of nodes as the output later.

In our case softmax layer has two output nodes i.e., REAL or FAKE, also Softmax layer provide us the confidence(probability) of prediction.

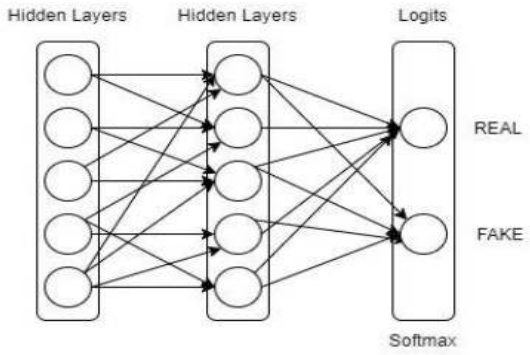

Figure.4 Softmax Layer

Confusion Matrix: A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. Confusion matrix is used to evaluate our model and calculate the accuracy.
Export Model: After the model is trained, we have exported the model. So that it can be used for prediction on real time data.

3.3.4. Model Prediction Details

The model is loaded in the application. The new video for prediction is pre-processed and passed to the loaded model for prediction. The trained model performs the prediction and return if the video is a real or fake along with the confidence of the prediction.

4. Results and Discussion

In this section, we present the outcomes of our simulation models and the analysis conducted using simulation software. Our simulation models were designed to evaluate the effectiveness of various deepfake detection techniques proposed in the literature. We employed state-of-the-art simulation software, such as TensorFlow, PyTorch, and OpenCV, to implement these models and conduct our experiments.

The simulation models were trained and tested using datasets consisting of both deepfake and real videos, sourced from reputable repositories such as Face-Forensic++, Deepfake Detection Challenge, Celeb-DF, and YouTube. We utilized a combination of frame-level features extraction techniques, including Res-Next Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to capture subtle artifacts indicative of deepfake manipulation. Furthermore, to facilitate a comprehensive comparison of the proposed techniques, we included high-quality visual representations of our results.

These images showcase the performance metrics, such as accuracy .

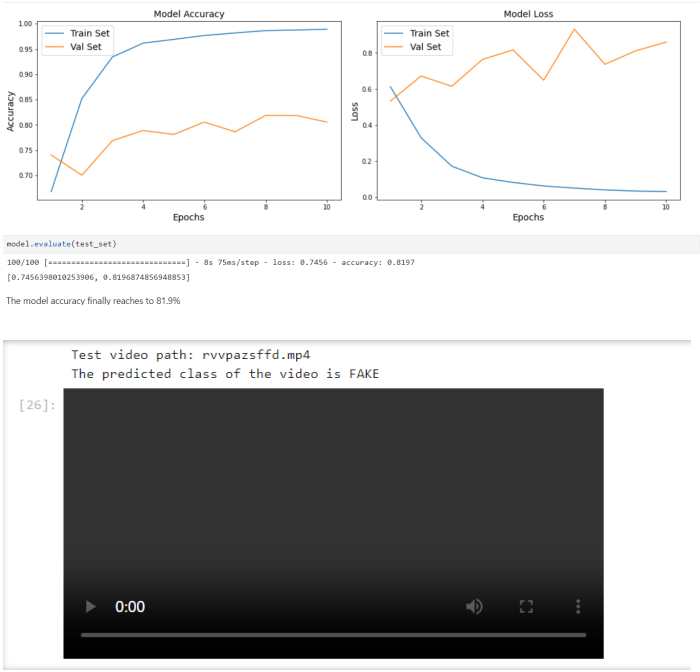

Figure.5 Result Analysis

5. Conclusion And Future Scope

In conclusion, our model only detects the video or image is fake or real. In future, we going to make our model detect audio too. this paper covers the detection and the generation of the deepfake, where the generation is used to help people know if the content they want is forged or not. The detection showed its best performance with the DNN and Resnet, and more accurate results are aimed to be established in the future for a more authentic, precise website. It is also intended to work with different datasets and generalize the dataset more with several augmentation techniques so that the system can detect any data inserted by the user.

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 2, 2025 , https://doi.org/10.63328/IJRDES-V7RI2P2

References

[1]. Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images” in arXiv:1901.08971.

[2]. Deepfake detection challenge dataset: https://www.kaggle.com/c/deepfake- detection challenge/data Accessed on 26 March 2020

[3]. Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu “Celeb-DF: A Large- scale Challenging Dataset for DeepFake Forensics” in arXiv:1909.12962

[4]. International Journal of Research Publication and Reviews, Vol 3, no 11, pp 540-544, November 2022

[5]. International Journal of Academic Engineering Research (IJAER), ISSN: 2643-9085 Vol. 6 Issue 3, March - 2022

[6]. Deep learning model for deep fake face recognition and detection Published online 2022 , International Research Journal of Modernization in Engineering Technology and Science Volume:03/Issue:03/March-2021

[7]. Deep Fake Face Detection Using Machine Learning and Various GAN Models, March 2022 DOI:10.46647//ijetms.2022.v06i02.001

[8]. Nirkin, Yuval, Lior Wolf, Yosi Keller, and Tal Hassner. "Deepfake detection based on discrepancies between faces and their co ntext." IEEE Transactions on Pattern Analysis and Machine Intelligence 44, no. 10 (2021): 6111-6121.

[9]. Zhao, Tianchen, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. "Learning self-consistency for deepfake detection." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 15023-15033. 2021.