

**Department of Computer Science and Engineering**

**DEEP LEARNING LAB – CSE 3183**

***A mini project report on***

**Celebrity Faces Deep Fake Detection**

**Group members:**

**Name: Boya Yashasvi Naidu**

**Register no: 210962070**

**Section: A**

**Roll no: 15**

**Name: M Sai Sandeep**

**Register no: 210962074**

**Section: A**

**Roll no: 17**

**Department of Computer Science and Engineering**

**Manipal Institute of Technology, Manipal.**

**April 2024**

**DEEP FAKE DETECTION**

**Abstract:**

In this paper, we combined multiple datasets to create a novel dataset enriched with diverse content, facilitating a comprehensive analysis of deepfake characteristics. Our approach integrates spatial and temporal information extracted from videos to discern between authentic and deepfake content. To achieve this, we employed a deep learning architecture consisting of ResNeXt for robust feature extraction and an LSTM network for sequence processing of temporal data. The ResNeXt model effectively captures spatial patterns within frames, while the LSTM model processes temporal dynamics across video sequences. Our method achieved a notable accuracy of 79.315% in distinguishing deepfake videos from authentic ones.

**Introduction:**

The rapid evolution of technology has brought about a revolution in the manipulation of digital media, a journey that extends from the early days of lithography to the contemporary era dominated by sophisticated deepfake technology. This transformative process, fueled by the accessibility of powerful tools such as PyTorch and a widespread technical knowledge base, has democratized the creation of manipulated media, transcending the realm of skilled professionals and extending its reach to a broader audience. Convolutional autoencoders, in particular, have played a pivotal role in lowering barriers, empowering individuals to engage in the creation of hyper-realistic manipulated content through widely-used applications like FaceApp and FakeApp.

However, the ascendancy of deepfake technology has not been without its challenges. The democratization that allows for benign applications also opens the door to potential misuse, prompting growing concerns about the technology's malicious implications. While instances of constructive applications exist, a significant portion of deepfake tools has regrettably been exploited for harmful purposes, including the creation of fake celebrity pornography and other detrimental content. The unsettling realism of deepfake videos has intensified concerns regarding their involvement in generating illegal material, spreading fake news, and perpetrating political hoaxes, ultimately contributing to societal tensions and a pervasive erosion of trust in digital media.

In response to these escalating challenges, our research paper addresses the imperative need for robust deepfake detection. We recognize the dual-use nature of this evolving technology and seek to contribute to the preservation of digital integrity by presenting a novel solution that leverages advanced neural network architectures. As guardians of truth in the age of manipulated media, we aim to explore innovative approaches that can safeguard against the malicious use of deepfake technology and restore trust in the authenticity of digital content.

**Literature Review:**

Various approaches have been proposed to tackle this challenge, leveraging techniques ranging from analysing patterns of eye blinking to exploiting discrepancies in head poses and employing advanced neural network architectures. For instance, DeepVision utilizes a combination of Fast-HyperFace and EAR algorithms to track changes in eye blinks, achieving an accuracy rate of 87.5%. Another approach focuses on analysing the temporal structure of video sequences using optical flow fields, demonstrating promising performance in distinguishing between deepfake and original videos.

The presented study focuses on the development of DeepVision, an algorithm designed to detect Deepfakes by analyzing patterns of eye blinking. Deepfakes, generated using GANs, pose significant challenges for integrity verification due to their increasing sophistication. Deep Vision utilizes various factors such as gender, age, activity, and time to track changes in eye blinks, which are involuntary and influenced by cognitive and behavioral factors. The algorithm employs a combination of Fast-HyperFace and EAR algorithms for face and eye detection, respectively, and compares measured blinking data with a pre-configured database to determine authenticity. Experimental results demonstrate DeepVision's effectiveness in identifying Deepfakes, achieving an accuracy rate of 87.5%. However, limitations exist, particularly regarding mental illnesses affecting blinking patterns. Despite these challenges, DeepVision represents a promising approach to enhance integrity verification and combat the spread of Deepfakes.[1]

The paper introduces a novel approach for detecting deepfake videos by analyzing the temporal structure of video sequences using optical flow fields. Unlike existing methods that focus on frame-based analysis, this approach aims to identify discrepancies in motion across frames between naturally generated videos and synthetically created deepfake videos. Optical flow fields, computed using the PWC-Net model, are utilized to capture apparent motion between consecutive frames. These flow fields are then inputted into a semi-trainable CNN, such as VGG16 or ResNet50, for classification into fake or original videos. Transfer learning is employed due to limited dataset size, where only a portion of the network is fine-tuned on the deepfake dataset while the rest is trained on pre-existing features. Preliminary experimental results on the FaceForensics++ dataset demonstrate promising performance in distinguishing between deepfake and original videos. The study suggests future research directions, including evaluating the reliability of optical flow fields across different datasets and neural network architectures, as well as exploring potential synergies with existing frame-based detection methodologies to enhance overall detection accuracy. [2]

This paper introduces a novel method for detecting Deep Fakes by exploiting discrepancies in head poses estimated from facial landmarks. Deep Fakes are created by replacing parts of original images with synthesized faces, leading to inconsistencies in landmark locations. The proposed approach compares head poses estimated from central face regions and entire faces, leveraging the mismatch in landmark locations. Using support vector machine (SVM) classifiers trained on these differences, the method effectively distinguishes between real and Deep Fake images/videos. Experimental results demonstrate the effectiveness of the approach on various datasets, achieving high classification accuracy. The study highlights the potential of utilizing 3D head pose estimation as a reliable cue for detecting AI-generated fake images/videos, contributing to the ongoing efforts to combat misinformation and digital impersonations. [3]

The paper proposes a two-stage analysis methodology for detecting deepfake videos. In the first stage, a Convolutional Neural Network (CNN) is utilized to extract features at the frame level. These features capture visual inconsistencies introduced by face-swapping processes. In the second stage, a temporally-aware Recurrent Neural Network (RNN), specifically a convolutional LSTM structure, is employed to analyze temporal inconsistencies between frames. The integration of CNN and LSTM enables end-to-end learning, allowing the system to effectively distinguish between manipulated and unaltered videos. Experimental results demonstrate the efficacy of the approach, achieving a high detection accuracy, particularly with short video sequences. The methodology offers a promising solution for automated deepfake detection, addressing the growing concerns surrounding the proliferation of manipulated media. [4]

"An Efficient Deep Video Model for Deepfake Detection" tackles this challenge with a novel sequential-parallel network architecture, aiming to strike a balance between effectiveness and computational efficiency. Inspired by the human visual system, the model analyses both spatial and temporal features. It dissects individual frames for static inconsistencies like unnatural facial expressions or scene manipulations, similar to how we scrutinize details in a photograph. Additionally, it examines motion patterns across frames, looking for subtle inconsistencies in blinking or head movements, analogous to analysing footprints for signs of tampering. To achieve efficiency without sacrificing accuracy, the model harnesses knowledge distillation. Imagine transferring the expertise of a seasoned detective to a rookie – knowledge distillation works similarly, transferring knowledge from a complex, resource-intensive model to a smaller, faster one. This enables real-time deepfake detection while minimizing computational strain. While initial results are promising, the authors acknowledge the ever-evolving nature of deepfakes. Continuous evaluation and refinement alongside exploration of complementary techniques like attention mechanisms and ensemble methods are crucial for ensuring the model's long-term effectiveness in this ongoing battle against deception. [5]

Building upon prior success with Convolutional Neural Networks (CNNs) for deepfake detection, "Deepfake Detection Using XceptionNet" proposes a novel framework. This framework leverages the pre-trained XceptionNet architecture for efficient feature extraction from images or videos. The extracted features are then fed into a classifier to distinguish between real and deepfake media, achieving, according to the paper, high accuracy while maintaining computational efficiency. This approach takes advantage of XceptionNet's depthwise separable convolutions, known for their ability to balance accuracy and efficiency. However, the paper acknowledges potential areas for further exploration. These include addressing potential biases and limitations in the training dataset, ensuring the model's generalizability across diverse deepfake creation techniques, incorporating explainability mechanisms to build trust and transparency, and exploring the potential of multi-modal fusion (e.g., combining image, audio, and text analysis) for even more robust detection. By addressing these aspects, researchers can build upon the promising foundation laid by this work and continue to advance the fight against deepfakes. [6]

The evolution of deepfake technology has necessitated continuous advancements in detection methodologies to combat malicious applications. Transfer learning, a technique widely adopted in the field of computer vision, has proven effective in adapting pre-trained models for specific tasks. Additionally, the integration of the YOLO (You Only Look Once) architecture, renowned for its real-time object detection capabilities, has become pivotal in the quest for robust deepfake detection. Previous research has explored the efficacy of transfer learning in enhancing model performance for deepfake detection, leveraging pre-existing knowledge from diverse datasets. Furthermore, the hybridization of transfer learning with the YOLO V7 architecture, tailored specifically for human cropping, represents a novel approach to address the challenges posed by the nuanced nature of manipulated content. This literature review examines the landscape of deepfake detection algorithms, focusing on the synergistic utilization of transfer learning and YOLO V7, shedding light on the strengths, limitations, and contributions of this hybridized approach in fortifying the defence against emerging threats in the deepfake landscape. [7]

As the prevalence of deepfake content continues to rise, the imperative to develop effective detection techniques has led to a proliferation of research endeavours. This literature review explores the landscape of deepfake detection methodologies, encompassing traditional forensic analysis, machine learning-based approaches, and the fusion of advanced neural network architectures. Researchers have extensively investigated the vulnerabilities and limitations of each technique, addressing challenges such as the rapid evolution of deepfake generation methods and the dual-use nature of detection models. Studies have compared the efficacy of traditional methods, including image and video analysis, with state-of-the-art machine learning algorithms, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs). Additionally, the review delves into the nuanced intricacies of ensemble methods that combine multiple detection techniques to enhance overall robustness. Through a comprehensive examination of existing literature, this review aims to provide insights into the strengths, weaknesses, and potential synergies of various deepfake detection approaches, contributing to a holistic understanding of the evolving landscape and guiding future research directions in the pursuit of more effective deepfake detection systems. [8]

The rapid evolution and widespread adoption of deepfake technology have underscored the critical need for robust detection methodologies. This comprehensive review has highlighted various approaches, ranging from analysing physiological cues like eye blinking patterns to leveraging advanced neural network architectures and optical flow fields for temporal analysis. While each method offers unique advantages and insights, they collectively contribute to the ongoing efforts to combat the proliferation of manipulated media

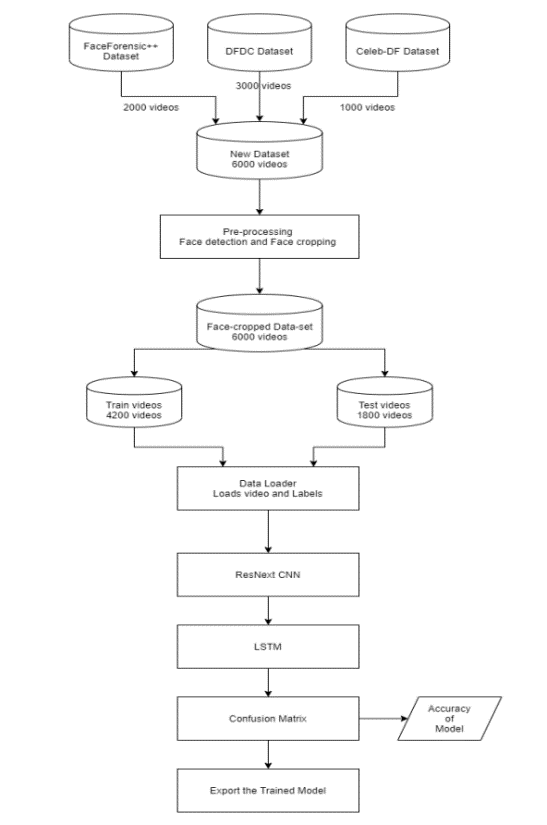
**Research Gap & Objective:**

One of the paper[4] used CNN and RNN in their model. The proposed model uses the LSTM in place of RNN as the video is a sequence of images and we need to keep track of all the past images for the proper detection of deep fake.

The objective of this research is to develop a robust and efficient deepfake detection model leveraging both spatial and temporal information within video frames. This is achieved by combining feature extraction from a pre-trained ResNext CNN model with sequential analysis via LSTM. The aim is to accurately classify videos as either deepfake or authentic by identifying imperceptible traces and distinct artifacts left by the deepfake generation process. The proposed methodology is evaluated on a diverse dataset comprising real and deepfake videos, with the goal of achieving high accuracy while minimizing false positives and false negatives. Additionally, the research aims to compare the performance of the proposed model against baseline approaches and assess its robustness under various real-world scenarios.

**Methodology:**

*Creating Deepfake Videos:*

**** To effectively detect deepfake videos, it is crucial to understand their creation process. Many tools, such as GANs and autoencoders, utilize a source image and target video to generate deepfakes. These tools break down the video into frames, identify faces within each frame, and replace the source face with the target face. The modified frames are then merged using various pre-trained models that enhance video quality by eliminating residual artifacts left by the deepfake generation process. Our approach to detecting deepfakes follows a similar method. Deepfakes produced using pretrained neural networks appear highly realistic, making them difficult to distinguish with the naked eye. However, these tools often leave subtle traces or artifacts in the videos that may go unnoticed. The goal of our research is to identify these imperceptible traces and distinct artifacts to classify videos accurately as either deepfake or authentic.

*Data Acquisition and Pre-processing:*

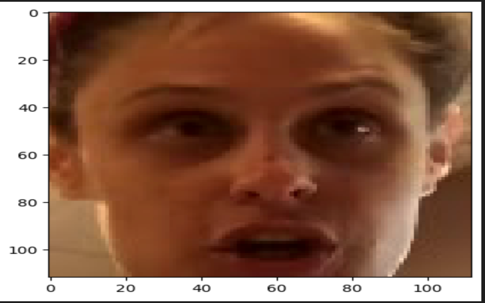


Fig 1:Sample test image

To enhance the model's efficiency for real-time prediction, we collected data from various existing datasets including FaceForensic++ (FF), the Deepfake Detection Challenge (DFDC), and Celeb-DF. To ensure a diverse and balanced dataset for accurate and real-time detection across different video types, we combined these datasets to create a new dataset.

To mitigate training bias, we maintained an equal distribution of 50% real and 50% fake videos in the dataset. Specifically, we processed the DFDC dataset by removing audio-altered videos using a Python script. From this preprocessed dataset, we selected 1500 real and 1500 fake videos. Additionally, we included 1000 real and 1000 fake videos from the FaceForensic++ dataset, along with 500 real and 500 fake videos from the Celeb-DF dataset.

In total, our dataset comprises 3000 real videos and 3000 fake videos, amounting to 6000 videos overall. This balanced distribution ensures that the model is trained on a diverse set of real and fake video samples, which is essential for robust and reliable real-time detection.

In this preprocessing step, the videos are processed to remove unwanted noise and focus only on the necessary content, which in this case is the face. The process begins by splitting each video into individual frames. For each frame, facial detection techniques are applied to identify and crop out the face. These cropped frames are then used to reconstruct a new video containing only the detected faces.

Model Architecture:

Our model employs a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), specifically Long Short-Term Memory (LSTM), for video classification tasks such as identifying deepfake versus pristine videos.

**ResNext Feature Extraction**: To extract frame-level features, we utilize a pre-trained ResNext CNN model. ResNext is a variant of the Residual CNN architecture optimized for deep neural networks. For our experiments, we leverage the `resnext50\_32x4d` model, which consists of 50 layers and dimensions of 32 x 4.

**Fine-tuning with Additional Layers**: Instead of building the CNN from scratch, we fine-tune the pre-trained ResNext model by adding additional layers tailored to our classification task. These layers are adjusted to ensure proper convergence during gradient descent, including selecting an appropriate learning rate.

**LSTM for Sequence Processing:** The 2048-dimensional feature vectors extracted by the ResNext model serve as input to the LSTM network. We use a single-layer LSTM with 2048 latent dimensions and 2048 hidden units, incorporating a dropout probability of 0.4 to prevent overfitting. The LSTM processes frames sequentially, allowing for temporal analysis by comparing frames at different time steps within the video.

**Activation Functions and Output Layer:** The model includes a Leaky ReLU activation function for non-linearity. A linear layer with 2048 input features and 2 output features (for binary classification) enables the model to learn correlations between input and output. An adaptive average pooling layer with an output size of 1 is used to reshape the data for downstream processing.

**Batch Training and SoftMax Layer:** Training is performed in batches of size 4 to optimize efficiency. A SoftMax layer is employed to compute class probabilities during prediction, providing a confidence score for the model's output.

By combining feature extraction from ResNext with sequential analysis via LSTM, our model is tailored to effectively process video data and make accurate classifications between deepfake and pristine videos, leveraging both spatial and temporal information within the video frames.

*Hyperparameter Tuning:*

Hyperparameter tuning is crucial for maximizing model accuracy. After extensive iterations on the model, we identified the optimal hyperparameters for our dataset. We utilized the Adam optimizer, which adapts the learning rate based on the model parameters. The learning rate was fine-tuned to 1e-5 (0.00001) to facilitate convergence towards a better global minimum during gradient descent. Additionally, a weight decay of 1e-3 was employed.

Given that this is a classification problem, we utilized the cross-entropy loss function. To leverage our available computational resources effectively, we adopted batch training with a batch size of 4. Through experimentation, we determined that a batch size of 4 is ideal for training in our development environment, balancing training efficiency and resource utilization.

**Results and Conclusion:**

The image you sent shows the results and conclusion of a study on a video classification model. The model leverages a combination of feature extraction from a pre-trained ResNext CNN model and sequential analysis using LSTM. The model was evaluated using a dataset consisting of 3,000 real videos and an equal number of deepfake videos. The model successfully achieved an overall accuracy rate of 79.3% in accurately categorizing videos as either deepfake or authentic based on the extracted features and sequential analysis. This performance demonstrates the effectiveness of integrating deep learning techniques, specifically CNNs and LSTMs, for robust video classification tasks, particularly in distinguishing between real and manipulated videos.

A graph showing the performance of training and validation accuracy

Description automatically generated

Fig 1: Accuracy vs Epoch

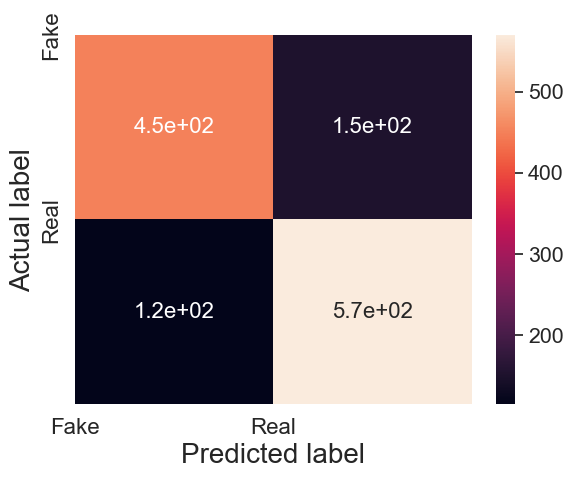


Fig 2: Confusion Matrix

**Precision** is the ratio of correctly classified real images (true positives) to the total number of images classified as real (true positives + false positives). In this case, the model is precise about 75% of the time it predicts a real image.

**Recall** is the ratio of correctly classified real images (true positives) to the total number of actual real images (true positives + false negatives). Here, the model correctly identifies nearly 80% of the real images.

**Future Work:**

1. **Hybrid Architectures:**

Explore hybrid architectures that combine the strengths of convolutional neural networks (CNNs) like ResNeXt and recurrent neural networks (RNNs) like LSTM with autoencoders. By integrating autoencoders into the feature extraction process, the model can learn both discriminative and reconstructive features, potentially improving detection accuracy while maintaining computational efficiency.

1. **Spatiotemporal Attention Mechanisms:**

Develop spatiotemporal attention mechanisms to enhance the model's ability to identify subtle inconsistencies in both spatial and temporal dimensions. Attention mechanisms can selectively focus on relevant frames or regions within the video sequence, while autoencoders can detect anomalies introduced by deepfake manipulations.

1. **Graph Neural Networks (GNNs):**

Investigate the applicability of graph neural networks (GNNs) for deepfake detection, particularly in scenarios where the relationships between different frames or regions within a video can be represented as a graph. GNNs, in combination with autoencoders, can capture complex dependencies and interactions, offering potential improvements in detecting manipulations across video sequences.

**References:**

[1]. Xin Yang, Seoungwon, Jinsub Yoon, and Hyoungseob Kim. ” DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, pp. 1207 1212, September 2019.

[2]. Irene Amerini, Roberta Marzario, Domenico Vitulano, Silvia Ruggeri, Matteo Barni Hyoungseob Kim. ” Deepfake Video Detection through Optical Flow Based CNN: 2019 IEEE International Conference on Computer Vision Workshop (ICCVW)

3420-3423

[3]. Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8261-8265). IEEE.

[4]. Ali Abdulzahra Mohsin Albazony, Mohammed Hasan Jafaar, Abeer Mohammed Salih. DeepFake Videos Detection by Using Recurrent Neural Network (RNN). 2023 IEEE International Conference on Electrical, Computer and Communication Engineering (ICECCE).

[5]. Sun, R., Zhao, Z., Shen, L., Zeng, Z., Li, Y., Veeravalli, B., & Yang, X. (2023). An efficient deep video model for deepfake detection. In Proceedings of the International Conference on Artificial Intelligence Applications and Innovations (ICAIAI) (pp. 1242-1254). Springer, Singapore.

[6]. Preetha Theresa Joy, Ashok V (2023).Deepfake Detection Using XceptionNet**.** 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE).

[7]. Nawaf Waqas; Sairul Izwan Safie; Kushsairy A. Kadir; Sheroz Khan; Abdul Ali Khan (2023). Transfer-Learning and YOLO V7 Hybridised for Human Cropping for Deepfake Detection Algorithms. 2023 IEEE 9th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA).

[8]. Sonia Salman; Jawwad Ahmed Shamsi (2023). Comparison of Deepfakes Detection Techniques2023.3rd International Conference on Artificial Intelligence (ICAI).

Dataset: @inproceedings{Celeb\_DF\_cvpr20,   author = {Yuezun Li and Xin Yang and Pu Sun and Honggang Qi and Siwei Lyu},   title = {Celeb-DF: A Large-scale Challenging dataset for DeepFake Forensics},   booktitle= {IEEE Conference on Computer Vision and Patten Recognition (CVPR)},   year = {2020}}