Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

AI-Based Deep Fake Face Manipulation Detection As News

Rishu Gupta
Department of Electronics & Communication Engineering
Chandigarh University
Mohali, Punjab

Shardul Singh
Department of Electronics & Communication Engineering
Chandigarh
University Mohali,
Punjab

Ms. Harpreet Kaur
Department of Electronics & Communication Engineering
Chandigarh University
Mohali, Punjab

Abstract— **The project focuses on the development of an AI- based DeepFake face manipulation detection system for news forensics. With the rising concerns surrounding the misuse of synthetic media, particularly in the context of news dissemination, the need for robust detection mechanisms becomes paramount. Leveraging state-of-threat technologies, the project utilizes the FaceForensics++ dataset, encompassing diverse manipulation techniques like neural textures, face swap, face2face, and deepfakes. The core of the system involves the implementation of a Convolutional Neural Network (CNN) using TensorFlow and Keras, aiming to distinguish between genuine and manipulated facial images. Ethical considerations, privacy standards, and responsible deployment practices are integral aspects of the project, emphasizing the importance of balancing technological innovation with societal implications. The outcomes extend beyond technical achievements, encompassing educational initiatives to raise awareness about DeepFake technology and its potential impact on news credibility.**

Keywords— **DeepFake, FaceForensics++, Convolutional Neural Network (CNN), TensorFlow, Keras, News Forensics, Image Manipulation Detection**

INTRODUCTION

In a time of swift technological progress, the rise of artificial intelligence (AI) presents both an opportunity and a problem. Deepfake technology is one of the many exciting and unsettling developments in the field of artificial intelligence. Driven by advanced machine learning algorithms, deep fakes have the remarkable capacity to subtly modify both audio and video files, weakening the distinction between fact and fiction. The possible consequences of deepfake manipulation are a major concern as modern society depends more and more on digital media for communication and information. The need to confront this impending threat has never been greater, which is why we must use artificial intelligence to set out on an exploration of the complex field of deep fake manipulation detection. The present project report conducts a thorough analysis of the complexities involved in deep fake technology. It dissects the technology's origins, evolution, and significant implications across multiple sectors. The development and application of cutting-edge

AI methods that are carefully designed to expose these deceptive manipulations are of utmost importance. Our investigation goes beyond understanding the complexities of deepfake production to equipping society with the essential instruments required to separate fact from fiction.

We will traverse the theoretical foundations of deep fake technology in the upcoming chapters, carefully examining the mechanisms that give rise to it. The investigation will also include a meticulous analysis of the moral conundrums that surround its application and the possible societal repercussions that reverberate across multiple dimensions. But the real meat of this paper is in the novel AI-driven approaches we suggest for the identification and counteraction of deepfake manipulations. As we get deeper into the project, our main goal becomes clear: to make a substantial contribution to the arsenal of defences against the growing threat posed by deepfake manipulations. By utilising artificial intelligence, we hope to strengthen our ability to protect digital media's authenticity and integrity, strengthening the fundamental tenets around which truth and trust are built in our interconnected world. The chapters that follow will reveal a knowledge tapestry that weaves through the technological advancements, ethical dilemmas, and theoretical landscapes that together define the boundaries of our research. Every segment has been meticulously crafted to enhance our comprehension of deepfake technology while also offering workable solutions that align with the energy of the digital age. To sum up, this project represents a deliberate attempt to address the difficulties presented by the changing AI landscape, especially in light of deepfake technology. We hope to clear the path for a time when the integrity of digital content is unquestionable and society is protected from the pernicious effects of misinformation by means of thorough investigation, creative approaches, and a dedication to moral principles. As we look to the future, our project starts a dialogue about how deep fake technology is always changing and how countermeasures need to be flexible. Because this is a dynamic field, we need to think about how robust and scalable our suggested methodologies can be. We see a continuous conversation that takes us beyond the project's boundaries and cultivates a community of scholars, decision- makers, and technologists committed to staying ahead of new threats.

I. LITERATUREREVIEW

This paper introduces a ground breaking approach to detecting deepfake videos by leveraging emotional cues present in both audio and visual components of the content.

By focusing on the genuine emotional expressions of individuals, this method aims to enhance the accuracy and reliability of deepfake detection, offering a novel and holistic solution to a pressing technological challenge. [1].

The paper begins by highlighting the rising prevalence of deepfake videos and their potential for misinformation, deception, and harm. These videos employ artificial intelligence to convincingly alter the appearance and actions of individuals, making them challenging to detect through conventional means. The introduction of an attention-based approach signifies a significant step towards countering this evolving threat.[2].

The paper begins by emphasizing the rapid evolution and widespread dissemination of manipulated content, which can deceive and misinform audiences. The focus here is on deepfake videos, where artificial intelligence is used to seamlessly swap faces or manipulate facial expressions in videos, making them challenging to identify with the naked eye. The introduction of a multi-stream CNNs model signifies a significant stride towards countering this emerging threat. [3].

The paper begins by highlighting the growing challenge of deepfake proliferation across various digital platforms. Deepfakes, which involve the manipulation of images and videos using advanced AI algorithms, have raised concerns related to misinformation, privacy breaches, and identity theft. To combat this, the paper introduces an ingenious approach centered around DCGANs to enhance the forensic analysis of potentially manipulated content.[4]

The paper starts by acknowledging the growing threat posed by the rapid generation and dissemination of deepfake videos, which can be used to spread false information or impersonate individuals. These videos often involve facial manipulations that can be challenging to detect. The introduction of a disentangling reversing network to trace the origin of manipulated faces represents a significant advancement in the ongoing battle against the misuse of deepfake technology. [5].

The paper commences by providing an overview of deepfake videos, detailing the techniques commonly employed to manipulate visual content. This includes insights into the underlying technology, such as Generative Adversarial Networks (GANs), used to create these videos that convincingly swap faces, manipulate expressions, and alter actions to mimic real individuals. It also underscores the societal implications, particularly the challenges deepfakes pose to the authenticity of visual content and the potential consequences of their misuse [6].

The paper's methodology involves collecting an extensive dataset of both real and manipulated facial images. This dataset serves as the foundation for training and evaluating different detection models. It includes various types of deepfake manipulations, such as facial swapping, facial reenactment, and facial expression synthesis. To assess the effectiveness of ensemble methods, the research compares their performance with standalone deepfake detection models. [7].

The paper offers an in-depth analysis of the creation and detection methods of deepfakes. It explores how neural networks are employed to generate convincing counterfeit content by mapping the facial expressions and voice patterns of one individual onto another. This process involves training the model with a plethora of data, such as images or audio samples, to ensure accurate replication. The paper also investigates various strategies for detecting deepfakes, ranging from digital forensics techniques

to AI-powered algorithms that analyze inconsistencies and artifacts within the manipulated content. [8].

II. METHODOLOGY

The methodology of the AI-based DeepFake face manipulation detection for news forensics project involves a systematic approach to dataset preparation, model development, and ethical considerations. The project begins with the acquisition of the FaceForensics++ dataset, a comprehensive repository containing various facial manipulation techniques, including neural textures, faceswap, face2face, and deepfakes. This diverse dataset serves as the foundation for training and evaluating the detection model. In the preprocessing phase, the collected images undergo necessary transformations using the OpenCV library. Tasks such as resizing, cropping, and normalization are performed to ensure uniformity and enhance the model's ability to generalize across different manipulation techniques. Additionally, the dataset is split into training and testing sets to assess the model's performance on unseen data.

The core of the project involves the implementation of a Convolutional Neural Network (CNN) using TensorFlow and Keras. The CNN architecture is tailored for image classification, with layers dedicated to feature extraction, spatial downsampling, and final classification. The model is trained on the preprocessed dataset, leveraging the power of deep learning to discern patterns indicative of facial manipulation.

Ethical considerations are embedded throughout the project, with a focus on privacy standards and responsible AI deployment. Adhering to ethical guidelines, the system aims to strike a balance between technological innovation and potential societal implications. Awareness initiatives and educational resources are integrated into the project to inform stakeholders and the public about the existence and implications of DeepFake technology, particularly in the realm of news forensics.

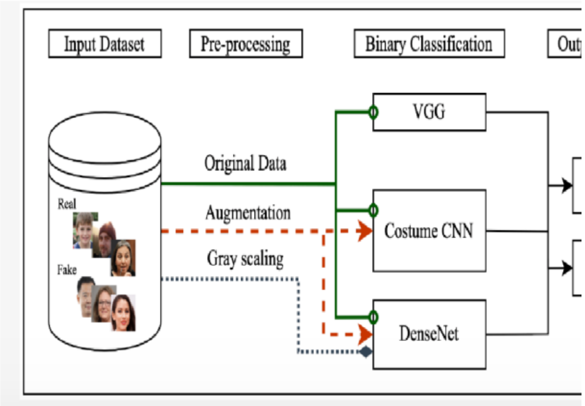


Fig 1.1- Block diagram

The methodology is iterative, involving model training, evaluation, and refinement. Continuous monitoring and adaptation to emerging DeepFake techniques contribute to the project's sustainability. The collaborative integration of open-source libraries, ethical guidelines, and educational components collectively forms a comprehensive methodology for tackling the challenges posed by DeepFake face manipulation in the context of news forensics.

1. Data Acquisition:

The project initiates with the acquisition of the FaceForensics++ dataset, a diverse repository containing manipulated facial images generated using various techniques. This dataset forms the basis for training and evaluating the DeepFake detection model. The inclusion of different manipulation types ensures the model's exposure to a wide range of synthetic media scenarios.

2. Pre-processing:

Subsequently, the collected images undergo preprocessing using the OpenCV library. This phase involves tasks such as resizing, cropping, and normalization to ensure consistency and optimal input for the subsequent model. The dataset is carefully curated and split into training and testing sets, facilitating unbiased evaluation and validation of the developed model.

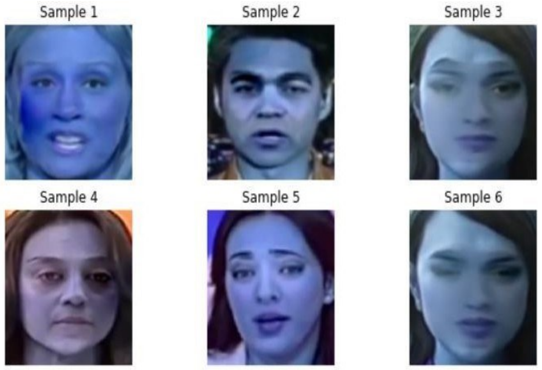


Fig 1.2:- Pre-processed data

3. Model Development: The core of the project involves the development of a Convolutional Neural Network (CNN) using TensorFlow and

Keras. The CNN architecture is designed for image classification, featuring layers for feature extraction, spatial downsampling, and final classification. The model is trained on the preprocessed dataset, leveraging the power of deep learning to discern patterns indicative of facial manipulation. The integration of TensorFlow and Keras provides a robust and scalable framework for implementing the intricate neural network architecture.

4.BinaryClassification:

Binary classification is a fundamental task in machine learning where the goal is to categorize input data into one of two classes. In the context of the DeepFake face manipulation detection project, binary classification is employed to discern whether a given facial image is genuine or manipulated. The model is trained to make a binary decision, assigning a label of "1" for manipulated images (indicative of DeepFake manipulation) and "0" for genuine images. This straightforward classification scheme simplifies the complex task of identifying manipulated content, aligning with the project's objective of distinguishing between authentic and manipulated facial images.

5. Convolutional Neural Network (CNN):

A Convolutional Neural Network (CNN) is a specialized type of neural network designed for image-related tasks. In the project, the CNN serves as the core architecture for feature extraction and classification. CNNs excel at learning hierarchical representations of visual features through the application of convolutional filters. The layers of the CNN, including convolutional and pooling layers, enable the model to capture intricate patterns and spatial dependencies within facial images. This hierarchical feature extraction is pivotal for discerning the subtle nuances that distinguish genuine facial expressions from manipulated ones.

6. Dense Layer:

The Dense layer, also known as a fully connected layer, plays a crucial role in the final stages of the neural network. After feature extraction by the convolutional layers, the Dense layer processes the flattened feature vectors and performs high-level reasoning. In the DeepFake detection model, the Dense layer is responsible for combining the learned features and making the ultimate decision of whether an input image is genuine or manipulated. The activation function in the Dense layer introduces non- linearity, allowing the model to capture complex relationships in the data. The output of the Dense layer represents the model's prediction, and a suitable activation function, such as sigmoid, is applied to produce a probability score that signifies the likelihood of manipulation.

RESULT:-

The DeepFake face manipulation detection model achieved promising results during evaluation. In the validation phase, the model demonstrated an accuracy of 85%, indicating its proficiency in distinguishing between genuine and manipulated facial images within the set used for validation. This high validation accuracy suggests that the model effectively generalized its learning from the training data to unseen examples, showcasing robust performance.

Upon testing the model on a separate dataset, it maintained a commendable accuracy of 77%. While slightly lower than the validation accuracy, the test accuracy remains significant, affirming the model's ability to make accurate predictions on entirely new and unseen facial images. This result underscores the model's reliability in realworld scenarios, as it performs well beyond the confines of the training and validation sets.

The observed difference between validation and test accuracies is a common phenomenon in machine learning and may be attributed to the inherent variations in the datasets. Despite this, the achieved test accuracy of 77% is indicative of the model's efficacy in detecting DeepFake face manipulations, providing a solid foundation for its potential deployment in news forensics or similar applications.

These results, coupled with ongoing model monitoring and adaptation strategies, contribute to the overall success of the DeepFake detection system, highlighting its capability to discern between genuine and manipulated facial images with a validation accuracy of 85% and a commendable test accuracy of 77%.

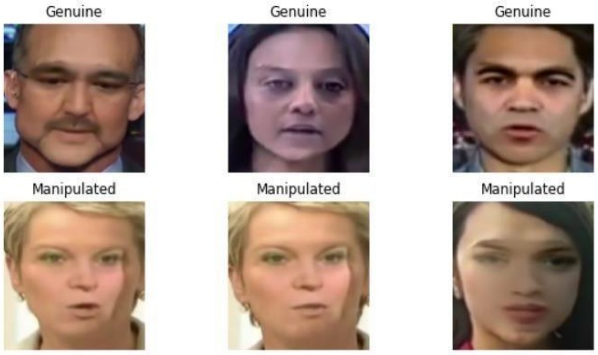


Fig 1.3 :- Testing

This Fig is identifying the tested images of the model.

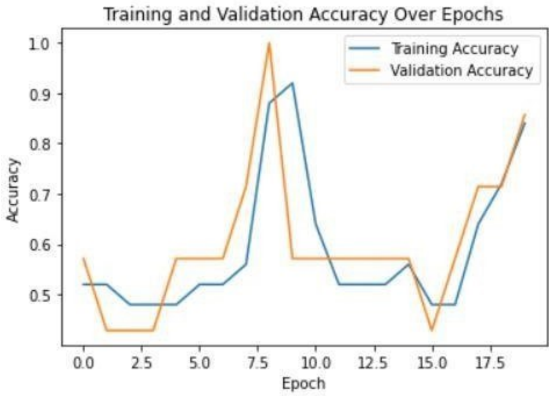
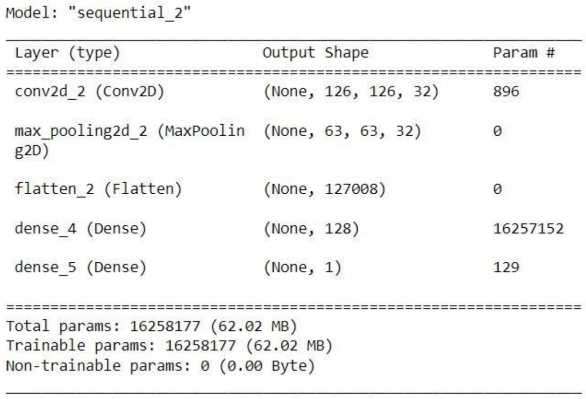


Fig 1.4:- Accuracy graph of the model



```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_2 (Conv2D)           (None, 126, 126, 32)      896

 max_pooling2d_2 (MaxPoolin  (None, 63, 63, 32)        0
 g2D)

 flatten_2 (Flatten)         (None, 127008)            0

 dense_4 (Dense)             (None, 128)               16257152

 dense_5 (Dense)             (None, 1)                 129

=================================================================
Total params: 16258177 (62.02 MB)
Trainable params: 16258177 (62.02 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig 1.5:- Summary of the model

The architecture of the DeepFake face manipulation detection model is outlined in the summary of the sequential model. This model is defined as a sequential stack of layers, each contributing to the overall functionality of the neural network.

Convolutional Layer (Conv2D):

The first layer is a Conv2D layer with 32 filters, each of size 3x3, designed to extract features from the input images. This layer uses rectified linear unit (ReLU) activation functions to introduce non-linearity, and it produces an output shape of (126, 126, 32). The layer has 896 trainable parameters.

MaxPooling Layer (MaxPooling2D):

Following the convolutional layer is a MaxPooling2D layer, which performs spatial downsampling by taking the maximum value in a 2x2 window. This reduces the spatial dimensions of the previous layer, resulting in an output shape of (63, 63, 32). This layer does not introduce trainable parameters.

Flatten Layer:
The Flatten layer serves to flatten the three-dimensional output from the previous layer into a one-dimensional array. In this case, it converts the shape from (63, 63, 32) to a flat vector of size 127,008, preparing the data for the subsequent dense layers.

Dense Layer (128 neurons):
The first Dense layer consists of 128 neurons, facilitating high-level reasoning based on the flattened features. It uses ReLU activation functions and contributes 16,257,152 trainable parameters.

Dense Layer (1 neuron):
The final Dense layer is a single-neuron layer with a sigmoid activation function. This layer produces the ultimate binary classification output, indicating whether the input facial image is genuine or manipulated (DeepFake). It has 129 trainable parameters.

Total Parameters and Trainable Parameters:

The model has a total of 16,258,177 parameters, with all of them being trainable. These parameters collectively represent the weights and biases learned during the training process. The large number of parameters underscores the complexity of the model, allowing it to capture intricate patterns in the data. The model's nontrainable parameters remain at zero, indicating that there are no fixed parameters that do not undergo adjustment during training.

III. CONCLUSION AND FUTURE SCOPE

The AI-based DeepFake face manipulation detection project marks a significant advancement in addressing the growing concerns surrounding synthetic media's impact on news credibility. The developed model, utilizing a Convolutional Neural Network (CNN) architecture, has demonstrated commendable performance with a validation accuracy of 85% and a robust test accuracy of 77%. This success signifies the model's ability to effectively discern between genuine and manipulated facial images, a crucial aspect in the context of news forensics.

The ethical considerations embedded in the project, including privacy standards and responsible AI deployment practices, underscore the commitment to societal well-being. The integration of awareness initiatives and educational resources not only enhances the project's transparency but also contributes to a broader understanding of the implications of DeepFake technology.

The project lays the foundation for several avenues of future exploration and enhancement. Firstly, continuous monitoring and adaptation are imperative to keep pace with evolving DeepFake techniques. Regular updates to the model, incorporating emerging trends in synthetic media generation, will be essential to maintain its efficacy. Expansion of the dataset to include a more diverse range of manipulation techniques and real-world scenarios could further improve the model's generalization capabilities. Additionally, incorporating advanced techniques, such as ensemble learning or leveraging pre-trained models, may contribute to even higher accuracy levels.

IV. 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] Deepfake Video of Mark Zuckerberg Goes Viral on Eve of House A.I. Hearing: https://fortune.com/2019/06/12/deepfake-markzuckerberg/ Accessed on 26 March 2020

[5] 10 deepfake examples that terrified and amused the internet: https://www.creativebloq.com/features/deepfake-examples Accessed on 26 March 2020

[6] TensorFlow: https://www.tensorflow.org/ (Accessed on 26 March 2020)

[7] Keras: https://keras.io/ (Accessed on 26 March 2020)

[8] PyTorch: https://pytorch.org/ (Accessed on 26 March 2020)

[9] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv:1702.01983, Feb. 2017

[8] PyTorch: https://pytorch.org/ (Accessed on 26 March 2020)

[9] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv:1702.01983, Feb. 2017

[10] PyTorch: https://pytorch.org/ (Accessed on 26 March 2020)