JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

UNMASKING THE DIGITAL ILLUSIONS THROUGH DEEPFAKES DETECTION SYSTEM

[1]**Ms. Pooja B. Aher**, [2]**Amena Inamdar**, [3]**Sameer Tupe**, [4]**Aman Bhondge**, [5]**Sweety Gangane**

[1,2,3,4,5]Computer Science and Engineering,
[1,2,3,4,5]Priyadarshini J. L. College of Engineering, Nagpur, India

Abstract : In this paper, a detection system for deepfakes in images, videos and audios has been developed. Throughout the year since deepfake has become a problem, there have been numerous attempts to develop detection systems to get a great accuracy, but still there has not been much progress. Whenever a deepfake detection system is created, a better deepfake processor comes into existence.

The main motive is to go through already existing Deepfakes detection system and also anything remotely related to detection of Fake images, videos, audios and one altogether and try to enhance the detection systems. There are multiple strategies that have been used[17] to detect Deepfakes but they do not provide much accuracy due to multiple reasons. Researches have been done to detect such videos or images through image segmentation, audio enhancement, continuous learning models, and what not.

Combining multiple features from multiple paper is the goal eventually to get the most accurate result, but even if there is a combination of multiple best features from various papers, it is not sure that all the features processing together will give the best efficiency. There have been researches before to make the best detection system through the use of multiple methods but they have not been much of a success. Therefore, there is a need to make pairs and dedicate a whole study to come to the realization of methods which will provide best accuracy for detection when combined together with methods for various aspects of detection. This paper analyses various papers made in the same direction, and tries to create what will be better for the future of deepfakes detection.

***IndexTerms* - *Deepfakes, Face tracking, Identity Leakage, Synthetic audio*.**

I. INTRODUCTION

The Deepfakes technology has risen in an era where the line between the reality and manipulation is becoming increasingly blurred. Media which is synthetic is being generated by multiple artificial intelligence in our day-to-day life, intelligences that are available to the disposal of mankind so easily that even a non-tech savy person can understand the whole editing system in seconds. Without realising the problems with this system, the internet continues to use this as a source of entertainment on multiple social medias, through the creation of fake videos of celebrities for fun. The long term effective of this not taken into consideration. These manipulations can have a deep reaching consequences which can be ranged from spreading disinformation and damaging reputation to undermining trust in media and institutions. In this paper, we will dive into the various ways in which a detection system has been created for this issue yet isn't working on its best potential.

In an era dominated by digital media, the rise of deepfake technology poses significant challenges to the authenticity and integrity of visual content. In response, this study presents a robust deepfake detection[5] system designed to discern between authentic and manipulated images and videos. Leveraging a diverse array of machine learning algorithms and feature-based methods, our model achieves high accuracy in detecting fake media. Through rigorous experimentation and evaluation on comprehensive datasets, we demonstrate the effectiveness of our approach in[15] accurately identifying[3] deepfakes. Our findings underscore the importance of employing a multi-modal approach to combat the proliferation of deceptive media and maintain trust in digital content.

With the advancement of machine learning and deep learning techniques, the creation of highly convincing fake videos and images, commonly known as deepfakes, has become increasingly accessible. These manipulated media can have serious implications, from spreading misinformation to damaging reputations and inciting unrest. Detecting deepfakes has thus emerged as a critical area of research, garnering significant attention from academia and industry alike. Here the ai m is to contribute to this field by proposing a novel deepfake detection system that leverages a combination of algorithms to improve accuracy. By synthesizing insights from existing research, we have identified a range of techniques that have shown promise in detecting deepfakes. Our approach involves evaluating these algorithms individually and in combination to identify the most effective strategies for deepfake detection.

1.1. Video Deepfakes

Video deepfakes typically refer to the use of artificial intelligence, particularly deep learning[14] algorithms to manipulate or alter videos, typically by replicating the faces of the individuals and attaching them to other bodies. Deepfakes in videos were easier to detect earlier as the motion of the person in the video or the face expressions or the smoothness was troubling at some point during this edited video, however now days, deepfakes are using refining algorithms which make them look like nothing before. The smoothness, the accuracy, the precision and what not have been enhanced during this. This technology enabled the creation of extremely realistic videos where individuals look like they are saying or doing something that they have not done actually. Detecting such deepfakes mainly involve a combination of manual as well as automated methods. These are:

i. Visual Inspection
ii. Analysis of Facial features
iii. Machine Learning Algorithms
iv. Source Authentications
v. Blockchain and Watermarking
vi. Forensic Analysis

However, till date, there has been ongoing research on video deepfakes and so far, various techniques and approaches have been developed to address these challenges. Looking in the background, one comes across FaceForensics, which is a benchmark dataset and framework currently that is using advance research for deepfakes detection. It consists of a huge collection of videos with manipulated facial content. Another such approach is using the capsule networks, which basically are a type of neural networks that capture hierarchical relationships between facial features. Temporal Analysis is also another one of such techniques are the ones that examine expressions and facial movement consistency. It analyses how facial features change frame by frame. Studies have also shown and investigated the use of Gaze analysis in which, eye movements and gaze patterns have been identified to further identify anomalies that may suggest the presence of deepfakes.

Apart from this, GANs are mostly used to generate deepfake videos so a countermeasure analysis has been done for the same where the aim is to identify the characteristics, patterns and artifacts associated with GAN image generation. Eventually some research focuses on Audio Visual analysis too by analyzing auditory cues in videos with reference to the video movement. As the technologies are evolving, the research continues to approach a new perspective each time and to enhance the already existing approaches.

1.2. Image Deepfakes

Images depicting individuals in situation or contexts they were not actually present in by creating a realistic manipulated image, come under the category of Image deepfakes. It is somewhat similar to video deepfakes but here the images are not moving is it makes it both easier and tougher to detect the fakeness index in such situation. Detecting such images involve some basics techniques such as follows-

i. Reverse Image search
ii. Examination of Facial Features
iii. Metadata Analysis
iv. Error Level Analysis
v. Forensic Software

Previously done research on the same includes Deep Learning Models such as CNN to detect the manipulation and identify patterns and also detect anomalies that are indicative of the manipulation. GANs countermeasures a part of image as well as video deepfakes both. An approach of error level analysis is also taken into consideration where the analysis of compression of artifacts in images to detect regions that may have been digitally altered are looked into. Manipulation Detection Benchmark data set was also introduced to facilitate research on image manipulation detection. It contains of a huge collection of images with different manipulation types including alterations which aims to aid in the evaluation and development of detection algorithms.

Another such technique that is being used is feature based analysis that focus on the extraction of characteristics and features. This may include texture patterns, inconsistencies in lighting and also pixel values of statistical properties. By advancing detection capabilities, research aims to mitigate the potential risks associated with the proliferation of fake images.

If comparing video deepfakes with image deepfakes, it is often considered that detecting image deepfakes is easier. This is due to majorly three reasons listed below-

- Images are static, that means they do not move, because there is no movement, it makes it easier to analyze the individual frames and identify the inconsistencies.
- Reverse image search and metadata analysis are techniques that are widely available as well as well established which makes it easier to provide effective means of detecting images that have been manipulated.
- Sometimes algorithms to detect deepfakes leave behind a trail of telltail artifacts which are easy to detect through careful analysis if there is not motion.

1.3. Audio Deepfakes

Audio deepfakes refers to the generation of or manipulation of recordings to make it sound like someone is saying something that they did not actually say. Basic techniques in detection of such audio fakes are as follows-

i. Spectral Analysis
ii. Voice Biometrics
iii. Linguistic Analysis
iv. Acoustic Analysis
v. Contextual Analysis

Previously, in Spectral Analysis, researchers have explored techniques such as Fourier analysis and spectrogram analysis to identify anomalies and inconsistencies in the frequency of audio signals. These helps detect unnatural shifts in pitch or spectral artifacts. Voice biometrics and speaker verification algorithms have also been developed to authenticate individuals based on their voice characteristics. In linguistics and semantics, the examination of content and structure of the spoken languages done through audio recordings. The linguistic inconsistencies, unusual patterns and semantic errors need to be investigated through this. Acoustic analysis is another such way on which research is done in which the focus is on extracting and analysing acoustic features from audio recordings such as reverberation and spatial characteristics. This aims to develop a robust method to detect deepfaked audio.

If one compares the detection process of image, video and audio deepfakes, it comes to a conclusion that it is both easier as well as harder. It is harder to detect audio fakes on a level because audio deepfakes may leave behind fewer artifacts as compared to visual deepfakes making it potentially harder to detect through traditional forensic analysis. On the other hand, the[17] human auditory perception is not that sensitive to the subtle changes as compared to visual lookout, which makes it easier potentially to create an audio fakes that are convincing.

1.4. Issues with Deepfakes

Initially, the use of deepfakes for entertainment purposes such as editing photos of our friends to depict funny scenarios or editing amusing videos of actors was generally acceptable. However, due to the advancements of deepfakes and them being increasingly realistic, the potential harm has grown significantly. Using deepfakes to generate images of deceased grandparents or parents may have been considered permissible in the past but the growing clearness and realistic portrayal of such images have raised a concern and has potential negative impacts. These negative impacts include Misinformation and Fake News, Privacy Violation, Identity Theft and Fraud, Erosion of Trust, Social Engineering Attacks, Political Manipulation,[4] Fake Pornographic Content development and many more such things.

II. METHODOLOGY

For developing a deepfakes detection system, one can use traditional machine learning methods like Support Vector Machine or Random Forest and so on, and we can also use big shot alternatives like Recurrent Neural Networks (RNN) or Convolutional Neural Network (CNN)[11] or Generative Adversarial Networks (GAN)[18] . In this paper, the work and development has been done on the basis of usage of Convolutional Neural Network (CNN).

RNNs are suitable for sequential analysis of data, and hence are not a good choice for image-based tasks where spatial relationships are crucial. Similarly, GANs are used to generate deepfakes and this might bring up the fact that one can reverse and counter evaluate this and detect those deepfakes but the point to keep in mind here is that whenever we will develop a deepfake detection with the help of GAN reversal, a better deepfake developer will be formed looking at the issues and anomalies of the detection system and looking at the points where the detection became easier. This, instead of making our task easier , might just help in making a better and enhanced deepfake generator. On the same page, Traditional machine learning algorithms like SVM or random forest[12] are mostly used for feature extraction and not complex pattern detection. Ultimately, CNN approach becomes the best approach to detect deepfakes due to the mentioned reasons.

CNNs are chosen for their ability to extract intricate patterns within images. They have been successful in various computer vision tasks, making them ideal for differentiating between real and AI-generated images. The hierarchical architecture of CNNs allows them to learn and represent features at different levels of abstraction, enabling them to detect even the most convincing deepfake content.

2.1. Implementation Process:

Dataset preprocessing involves resizing, normalizing, and augmenting images to improve their quality and diversity. Design and training of a CNN model with appropriate architecture, including[7] convolutional layers for feature extraction and fully connected layers for classification. Training the network to differentiate between genuine and manipulated images by minimizing a specific loss function[2] . Implementing transfer learning techniques to leverage pre-trained models and enhance the system's performance.

2.2. System Framework:

2.2.1. Data Collection and Preprocessing:

Create a varied dataset comprising both authentic and deepfake images. Then, preprocess the dataset by resizing, normalizing, and augmenting the images to enhance quality and diversity.

2.2.2. CNN Model Architecture Design:

A Convolutional Neural Network (CNN) was built with an appropriate architecture for effective feature extraction and classification. Convolutional layers were utilized to extract intricate patterns within the images, while fully connected layers were incorporated to facilitate the classification process.

2.2.3. Training Process:

The CNN model was trained on the preprocessed dataset, with a focus on distinguishing between authentic and manipulated images. A specific loss function was employed to minimize the disparities between predicted and actual labels[1] , enhancing the model's ability to accurately classify the images.

2.2.4. Transfer Learning Integration:

Transfer learning techniques were implemented to leverage pre-trained models and enhance the system's performance. The pre-trained models were fine-tuned to adapt to the nuances of deepfake detection, improving their capability to discern manipulated content.

2.3. Evaluation and Testing:

The performance of the CNN model was evaluated using various metrics such as accuracy, precision, recall, and F1-score[3] . Additionally, the system's robustness and reliability were tested by assessing its capability to detect progressively sophisticated deepfake content.

2.4. Deployment and Real-Time Implementation:

The system was prepared for deployment in real-world scenarios, with a focus on ensuring seamless integration into existing platforms or as a standalone solution for deepfake detection. Additionally, the system was monitored and updated periodically to adapt to evolving deepfake generation techniques.

2.5. Continuous Improvement and Adaptation:

The system was continuously updated with new data and adapted to emerging deepfake trends and technologies. Additional layers were incorporated or the architecture was fine-tuned as necessary to enhance the model's performance and accuracy.

2.6. Flow Diagram:

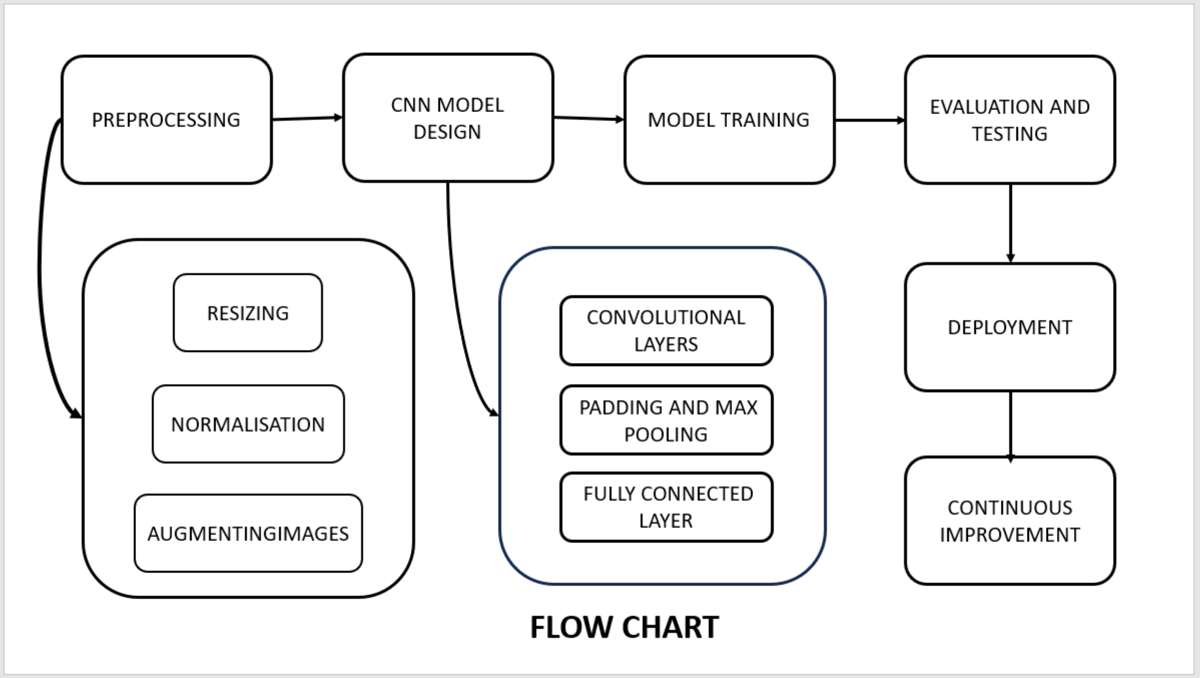
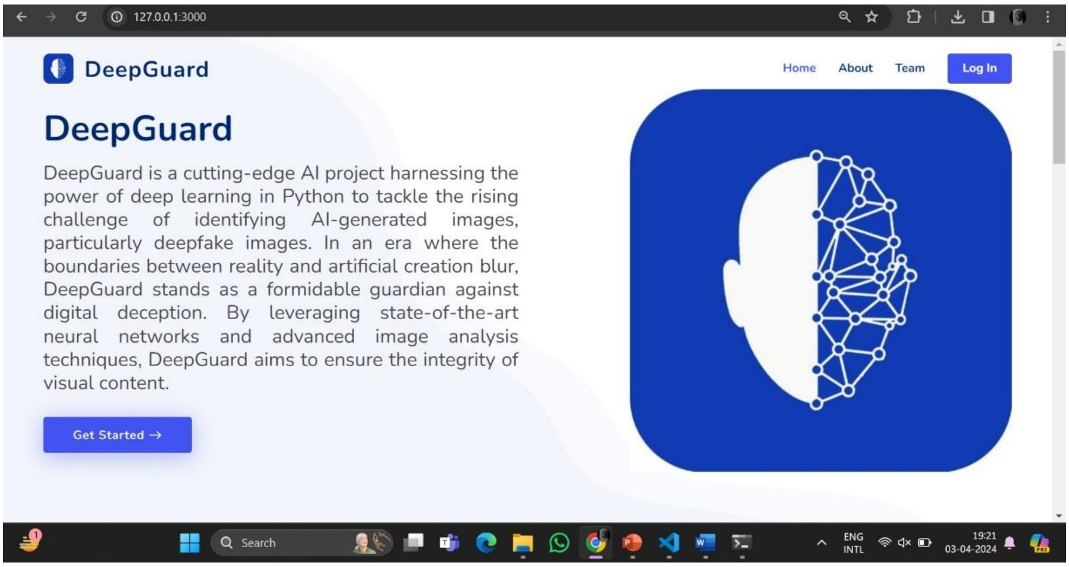


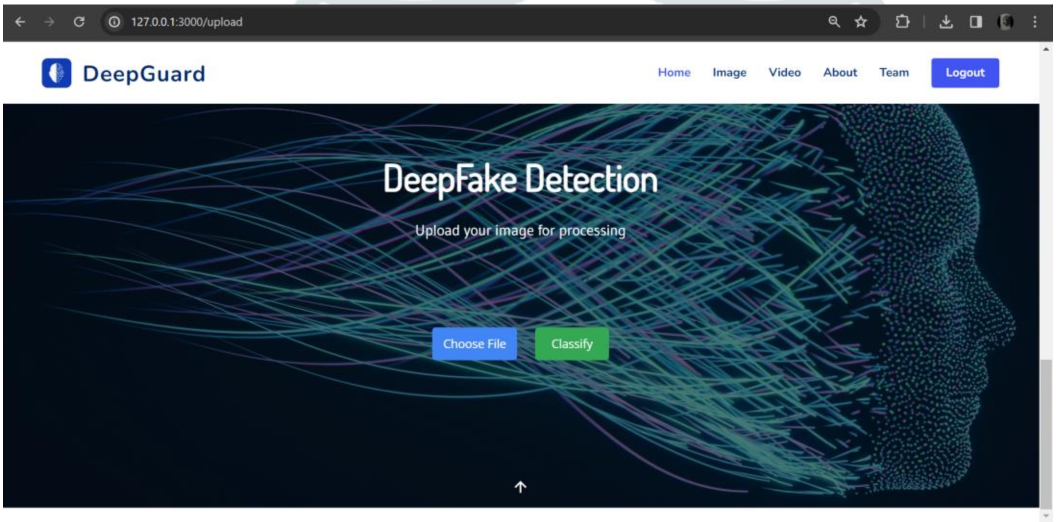
Fig: Flow of the project



DeepGuard

DeepGuard is a cutting-edge AI project harnessing the power of deep learning in Python to tackle the rising challenge of identifying AI-generated images, particularly deepfake images. In an era where the boundaries between reality and artificial creation blur, DeepGuard stands as a formidable guardian against digital deception. By leveraging state-of-the-art neural networks and advanced image analysis techniques, DeepGuard aims to ensure the integrity of visual content.

Fig: Landing page of the website (Locally hosted in the screenshot)



DeepFake Detection

Upload your image for processing

Fig: Uploading Page for Image/Video

III. RESULTS AND DISCUSSION

The paper describes the improvement of an in-intensity research application incorporating pix, video, and audio. It uses a mixture of system getting to know and function-based strategies to resolve the common hassle of reaching deep. The mission involves several steps, together with dataset preprocessing, CNN architecture layout, version schooling, integration of switc h studying, and continuous development. Challenges and advances in deepfake detection are mentioned, and vital barriers to authenticity and integrity in virtual media are highlighted Deepfake technology requires numerous intensity detection techniques in various media, which include visible inspection, faces trait evaluation, spectral analysis, voice biometrics, speech analysis. By deep throw creation techniques is to increase search capabilities to preserve pace with it.

The adaptive approach uses convolutional neural networks (CNNs) due to their efficiency in extracting complex patterns from images The device is rigorously tested on all data sets to verify its effectiveness Transfer-learning strategies are used use to leverage pre-learned models and consequently improve the performance of the recognition tool.

System evaluation and testing are important add-ons to compare the overall performance of CNN models using metrics such as accuracy, precision, don't forget, F1-rating etc. The robustness and reliability of the system are often tested on sophisticated lights, considering real-time implementation in order to ensure seamless integration and successful collision.

Continuity and efficiency occupy a central position, with new records often being used to redesign and introduce new layers designed to meet deeper functionality and emerging technologies, changing design as needed and providing the definition accuracy and efficiency quickly over the years.

Accuracy:

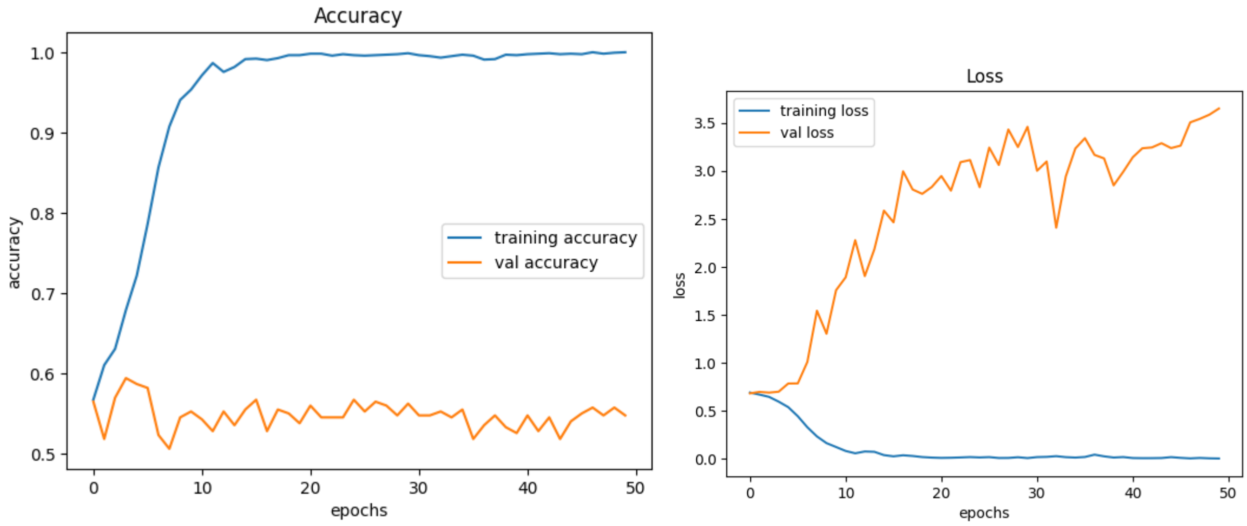
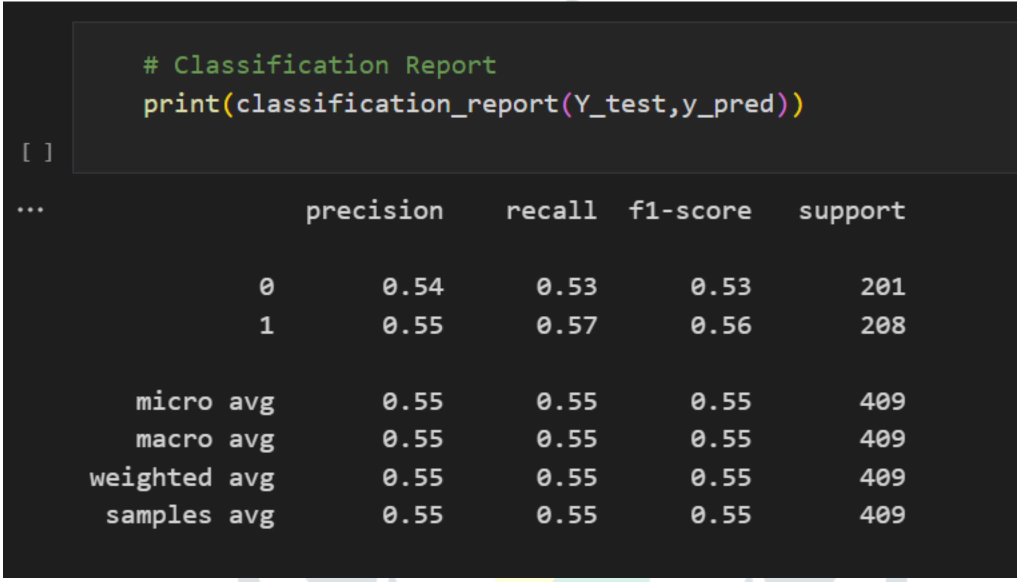


Fig: Accuracy of Training and Loss



```
# Classification Report
print(classification_report(Y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.54      0.53      0.53       201
           1       0.55      0.57      0.56       208

   micro avg       0.55      0.55      0.55       409
   macro avg       0.55      0.55      0.55       409
weighted avg       0.55      0.55      0.55       409
 samples avg       0.55      0.55      0.55       409
```

Fig: Classification Report

REFERENCES

[1] S. Siddiq, K. Kaur and R. Dhir, "Automatic Detection of Cloudy and Non-Cloudy SAR Images Using Convolutional Neural Networks," *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, 2023, pp. 350-355, doi: 10.1109/ICSCCC58608.2023.10176417.

[2] Rimsha Rafque, RahmaGantassi, RashidAmin1, Jaroslav Frnda, Aida Mustapha, Asma HassanAlshehri, "Deep fake detection and classifcation using error-level analysis and deep learning", In2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) 2020 Oct 13 (pp. 1-9). IEEE.

[3] "Front Matter," 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 1-609, doi: 10.1109/UBMK59864.2023.10286684

[4] Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, Ruibo Fu, "An Initial Investigation for Detecting Vocoder Fingerprints of Fake Audio", In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia 2022 Oct 14 (pp. 61-68).

[5] Liu J , Zhu K , Lu W , Luo X , Zhao X . A lightweight 3D convolutional neural network for deepfake detection. *Int J Intell Syst*. 2021; 36: 4990-5004. https://doi.org/10.1002/int.22499

[6] J. Bagate, R. Jadhav, N. Jain, G. Gaikwad and A. Arimpur, "Galaxy Shape Classification Using Machine Learning," 2023 7th International Conference on Computer Applications in Electrical Engineering-Recent Advances (CERA), Roorkee, India, 2023, pp. 1-6, doi: 10.1109/CERA59325.2023.10455372.

[7] Manuel Gil-Martín *, Marcos Sánchez-Hernández and Rubén San-Segundo, Human Activity Recognition Based on Deep Learning Techniques, Speech Technology Group, Information Processing and Telecomunications Center, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid (UPM), 28040 Madrid, Spain.

[8] S. Baek, H. H. Yoo, J. H. Ju, P. Sriboriboon, P. Singh, J. Niu, J.-H. Park, C. Shin, Y. Kim, S. Lee, Ferroelectric Field-Effect-Transistor Integrated with Ferroelectrics Heterostructure. *Adv. Sci*. 2022, 9, 2200566. https://doi.org/10.1002/advs.202200566

[9] A. Tabbakh and S. S. Barpanda, "A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer “TLMViT” for Plant Disease Classification," in IEEE Access, vol. 11, pp. 45377-45392, 2023, doi: 10.1109/ACCESS.2023.3273317.

[10] Talaat, F.M., ZainEldin, H. An improved fire detection approach based on YOLO-v8 for smart cities. *Neural Comput & Applic* **35**, 20939–20954 (2023). https://doi.org/10.1007/s00521-023-08809-1

[11] Lopez R, Gayoso A, Yosef N. Enhancing scientific discoveries in molecular biology with deep generative models. Mol Syst Biol. 2020 Sep;16(9):e9198. doi: 10.15252/msb.20199198. PMID: 32975352; PMCID: PMC7517326.

[12] M. Uğurlu and İ. A. Doğru, "A Survey on Deep Learning Based Intrusion Detection System," 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 223-228, doi: 10.1109/UBMK.2019.8907206. keywords: {Deep Learning;Intrusion Detection System;Cybersecurity},

[13] Iqbal S, Ghani Khan MU, Saba T, Mehmood Z, Javaid N, Rehman A, Abbasi R. Deep learning model integrating features and novel classifiers fusion for brain tumor segmentation. Microsc Res Tech. 2019 Aug;82(8):1302-1315. doi: 10.1002/jemt.23281. Epub 2019 Apr 29. PMID: 31032544.

[14] C. Chanachan, P. Thanesmaneerat, T. Mahasukon, J. Polvichai and S. Dumkor, "Car Driver's Behaviors Detections Using Ensemble Model," 2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), Jeju, Korea, Republic of, 2023, pp. 1-6, doi: 10.1109/ITC-CSCC58803.2023.10212694. keywords: {Deep learning;System performance;Fatigue;Data models;Behavioral sciences;Automobiles;Integrated circuit modeling;Vehicles;Accidents;Logistics;Driver's Behaviors Detections;Mediapipe;YOLO;Image Processing;Deep Learning;LSTM},

[15] Alabsi BA, Anbar M, Rihan SDA. CNN-CNN: Dual Convolutional Neural Network Approach for Feature Selection and Attack Detection on Internet of Things Networks. Sensors (Basel). 2023 Jul 19;23(14):6507. doi: 10.3390/s23146507. PMID: 37514801; PMCID: PMC10384372.

[16] Rajagopal H, Mokhtar N, Tengku Mohmed Noor Izam TF, Wan Ahmad WK. No-reference quality assessment for image-based assessment of economically important tropical woods. PLoS One. 2020 May 19;15(5):e0233320. doi: 10.1371/journal.pone.0233320. PMID: 32428043; PMCID: PMC7236984.

[17] Han, Xian-Feng & Laga, Hamid & Bennamoun, Mohammed. (2019). Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 1-1. 10.1109/TPAMI.2019.2954885.

[18] Riya Nimje1 , Shreya Paliwal2 , Jahanvi Saraf3 , Prof. Sachin Chavan4 , “ Investigation of Different Deep Neural Networks for the Diagnostics of Brain Tumor, Skin Cancer and Breast Cancer” ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022