MINOR-2 PROJECT

REPORT

For

DeepFake Face Detection

Submitted By

SPECIALISATION	NAME	SAP ID
AIML	KUSHAGRA	500096089
AIML	TANNU KHETAN	500097428
AIML	ARCHIT JOSHI	500096754
AIML	ANANYA SINGH	500096298

Department of Informatics

School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007. Uttarakhand

Dr. Panduranga Raviteja

Project Mentor

TABLE OF CONTENTS

DECLARATION

We hereby declare that the project "DeepFake Face Detection" which is being submitted to the University of Petroleum and Energy Studies, Dehradun (Uttarakhand), as a partial fulfilment of the Bachelor of Technology in Computer Science degree, is an authentic record of our own work completed with the supervision of Dr. Panduranga Raviteja, Assistant Professor, Systematics Cluster, SoCS. The information presented in this project is the result of our own work and hasn't been previously submitted for consideration for any other degree.

CERTIFICATE

This certifies that Tannu Khetan, Kushagra, Archit Joshi and Ananya Singh have submitted the Minor Project titled **"DeepFake Face Detection"** in partial fulfilment of the requirements for the University of Petroleum and Energy Studies (Uttarakhand) to award a Bachelor of Technology in Computer Science Degree.

Dr. Panduranga Raviteja

Assistant Professor

Systematics Cluster, SOCS

ACKNOWLEDGEMENT

We would like to respectfully thank Dr. Panduranga Raviteja, assistant professor in the School of Computer Science Department's Systemics Cluster, for his important supervision, support, and assistance during our study. Without his generous support and direction, the project would never have come to be.

We would like to take this occasion to thank Dr. Panduranga Raviteja, Assistant Professor, Systemics Cluster, School of Computer Science, for his kind approval and encouragement. We would like to sincerely thank him for his guidance and counselling on occasion.

We would like to express our gratitude to each and every instructor in the School of Computer Science Department for their occasional guidance and assistance.

Date: 1-May-2024

1. Abstract:-

The goal of this project is to create a powerful deep fake face detection system that can recognise and analyze AI generated faces. The study intends to reliably identify small visual indicators suggestive of deep fake modifications by utilizing machine learning methods, such as convolutional neural networks (CNNs), Long Short Term Memory (LSTM) and facial recognition models, in conjunction with advanced picture analysis techniques. To distinguish between real and fake face photos, the approaches combine extensive feature extraction, pattern recognition, and classification. It is anticipated that the project's discoveries and understandings will greatly increase facial image manipulation detection methods, answering the escalating worries about deepfake technology abuse. The findings of this study have broad significance for a variety of fields where maintaining the authenticity and integrity of visual output is crucial, such as media, forensics, and cyber security.

Keywords: Convolutional Neural Network, Long Short Term Memory,Deepfake, Computer Vision

2. Introduction:-

Deep fake technology has become widely available because of the quick development of digital image modification tools. This has raised serious worries about the possible exploitation of modified facial photos for malevolent or misleading purposes. This research intends to address these issues by creating a reliable deep fake face detection system that correctly analyzes and recognises modified facial photos by utilizing machine learning algorithms and sophisticated image analysis techniques.

The goal of the project is to capture complex features and patterns in facial photos by utilizing cutting-edge image analysis methods like convolutional neural networks (CNNs) , Long Short Term Memory (LSTM) and facial recognition models. The technique seeks to improve the accuracy of deep fake detection by using a variety of datasets of real and altered face image training to distinguish between the two groups using features that have been learnt.

In order to properly discriminate between real and modified facial photos, the study also highlights the importance of feature extraction, pattern recognition, and classification approaches inside these sophisticated algorithms. The system's ability to recognise tiny visual clues suggestive of profound false alterations is made possible by the integration of different approaches, which ultimately aids in the precise identification of edited facial photographs.

The findings of this project have broad significance for a variety of fields where maintaining the authenticity and integrity of visual output is crucial, such as media, forensics, and cybersecurity. The project aims to ensure the integrity of visual content in an era marked by digital image manipulation by tackling the issues related to the exploitation of deep fake technology and developing trustworthy techniques for identifying modified facial photographs.

3. Literature Review:-

The Face Warping Artefacts study utilized a Convolutional Neural Network (CNN) model tailored to compare facial regions and their surroundings, targeting artefact identification. Two types of face artefacts were discerned in their research. They noted that existing deepfake algorithms generate images with limited resolution, necessitating further processing to align with original video faces. Temporal frame analysis wasn't integrated into their method.

A novel deepfake identification technique focused on eye blinking frequency, employing Long Short-Term Memory (LSTM) networks for temporal analysis of cropped blinking frames. However, contemporary deepfake algorithms are robust; merely assessing eye blinks may not suffice for detection. Recognition necessitates considering factors like teeth, wrinkles, and eyebrow positioning.

Capsule networks were employed for detecting manipulated images and videos, leveraging random noise during training, although not a recommended practice. While effective on their dataset, noise might hinder real-time performance, suggesting training on noise-free data.

Another method combined ImageNet pre-trained CNN models with LSTM for sequential frame processing, utilizing the dataset, which might be limited for real-time applications. Conversely, training on vast real-time data is proposed.

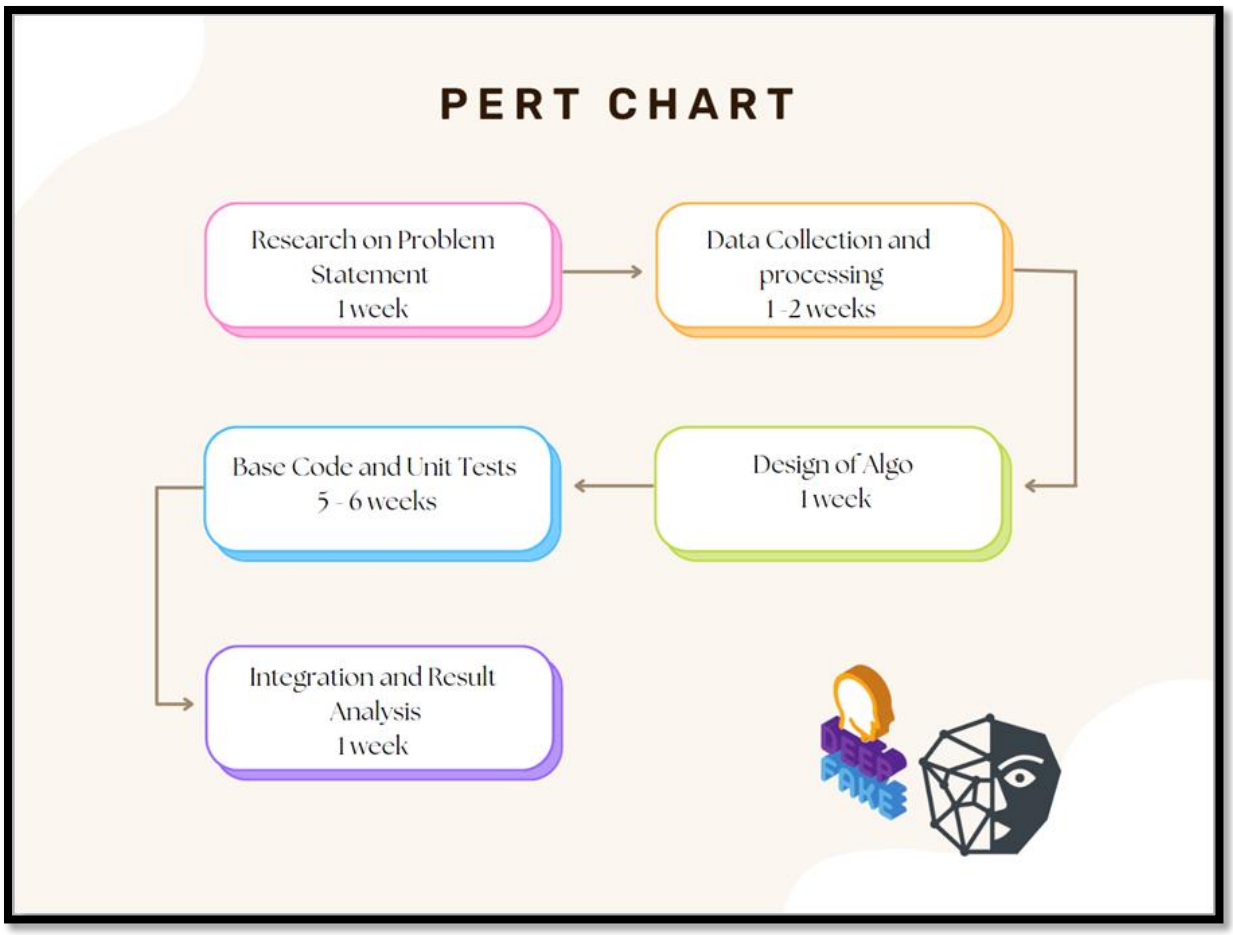
Extracting biological signals from both pristine and deepfake portrait videos, a CNN-LSTM ensemble was trained to assess temporal consistency and spatial coherence, with subsequent classification based on authenticity probability averages.

The fraudulent Catcher system exhibits high accuracy in identifying fraudulent content, but formulating a differentiable loss function adhering to signal processing guidelines, especially without a discriminator, poses challenges in retaining biological signals.

4. Objective Justification:-

1. **Addressing Emerging Threats:** With the rapid advancement of AI and machine learning technologies, deep fake videos have emerged as a significant threat to society. These manipulated videos can deceive individuals, spread misinformation, and potentially cause harm to individuals, organizations, and societies at large. Therefore, developing effective tools to detect deep fake content, particularly in the realm of face manipulation, is imperative.
2. **Protecting Integrity and Trust:** The proliferation of deep fake videos undermines the integrity of visual media and erodes trust in the authenticity of online content. By focusing on deep fake face detection, the project aims to contribute to the preservation of trust in digital media and protect individuals and organizations from being misled by fraudulent content.
3. **Safeguarding Individuals' Reputations:** Deep fake videos can be used to fabricate compromising situations or false statements attributed to individuals, leading to reputational damage and potentially dire consequences. A robust deep fake face detection system can help mitigate the risks associated with such manipulations, thereby safeguarding individuals' reputations and privacy.
4. **Empowering Decision-Making:** In an era where visual content plays a crucial role in shaping public opinion and influencing decision-making processes, the ability to distinguish between authentic and manipulated videos is paramount. By providing accurate and reliable detection mechanisms, the project equips users with the tools necessary to make informed decisions based on authentic information.
5. **Contributing to Technological Advancement:** Research and development in deep fake detection not only serve immediate societal needs but also contribute to advancing the field of AI and machine learning. The project involves exploring innovative algorithms, methodologies, and techniques for detecting subtle manipulations in facial images, thus pushing the boundaries of technological capabilities in this domain.
6. **Fostering Ethical AI Practices:** As AI technologies become more pervasive, ensuring their ethical use is essential. By actively working on deep fake face detection, the project promotes responsible AI practices and encourages the development of ethical guidelines for the creation and dissemination of AI-generated content.
7. **Supporting Legal and Regulatory Frameworks:** The detection of deep fake content can aid law enforcement agencies, policymakers, and regulatory bodies in developing strategies to combat online misinformation and protect individuals' rights. By generating insights into the prevalence and nature of deep fake videos, the project contributes to the formulation of effective legal and regulatory frameworks to address this evolving challenge.

5. PERT Chart:-



6. MODEL -

CNN (Convolutional Neural Network):

- The CNN part of the model serves as a feature extractor. It takes in the input image and passes it through a series of convolutional layers, pooling layers, and activation functions.
- Each convolutional layer extracts features from the image by convolving filters (kernels) over the input image. These filters capture different patterns such as edges, textures, and shapes.
- The pooling layers downsample the feature maps, reducing their spatial dimensions while retaining important information.
- The activation functions introduce non-linearity to the network, allowing it to learn complex patterns.
- The output of the CNN is a set of high-level feature representations of the input image.

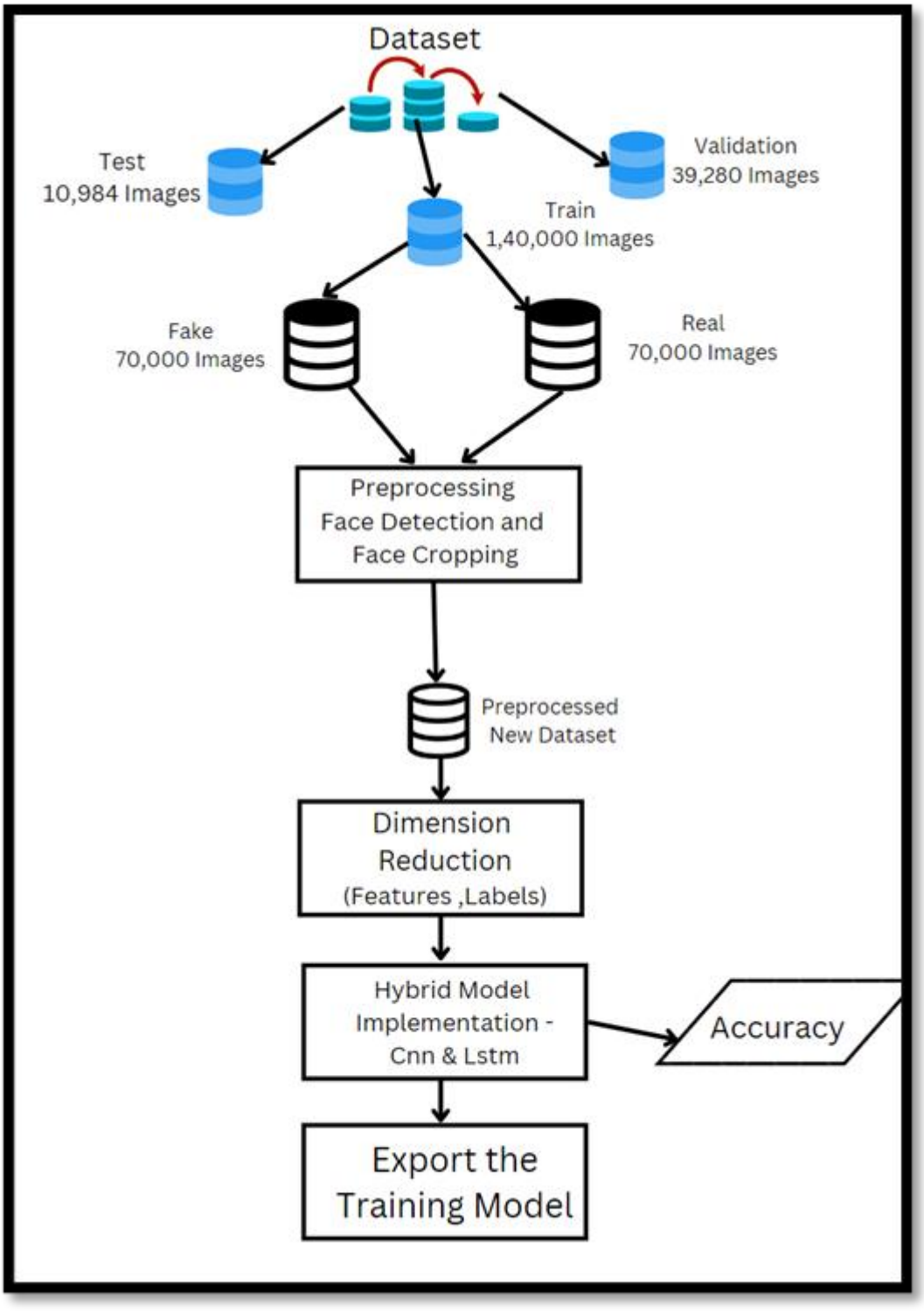
LSTM (Long Short-Term Memory):

- The LSTM part of the model acts as a sequence model, capable of capturing temporal dependencies in the data. In this case, it processes the sequence of features extracted by the CNN.
- The LSTM takes the sequence of features extracted by the CNN and learns to model the temporal dynamics.
- It consists of gates (input, forget, and output gates) and a memory cell, which allow it to selectively read, write, and reset information over time.
- By processing the sequential features, the LSTM learns to capture patterns that evolve over time, which is crucial for tasks involving sequential data like image/video analysis.
- The output of the LSTM is a representation of the temporal dynamics learned from the input sequence.

Hybrid Model:

- The hybrid model combines the strengths of both CNN and LSTM. The CNN extracts spatial features from individual frames/images, while the LSTM processes these features over time.
- By training the hybrid model on a dataset containing both fake and real images, it learns to distinguish between them based on the patterns present in the images.
- During training, the model adjusts its parameters (weights and biases) using optimization techniques like gradient descent and backpropagation to minimize the classification error (loss).
- Once trained, the model can predict whether an input image is fake or real based on the learned patterns and temporal dynamics.

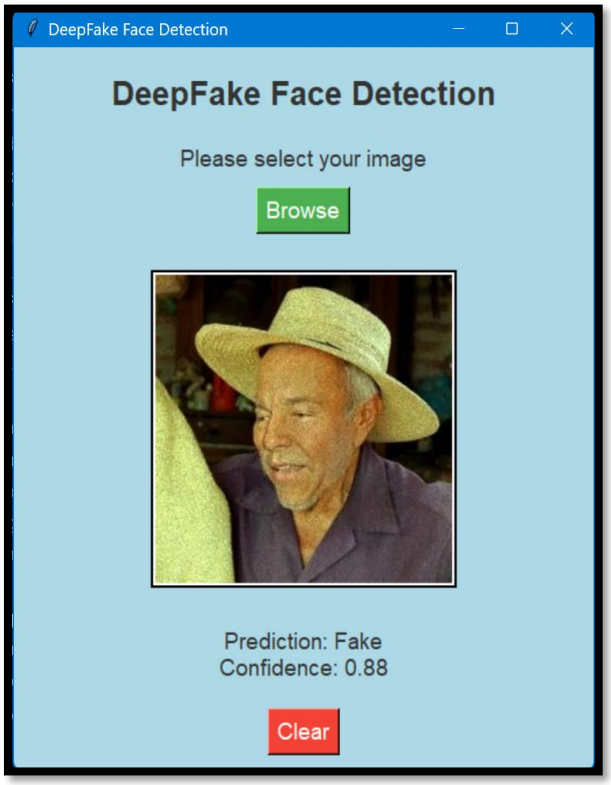
7. Training Workflow Diagram:-

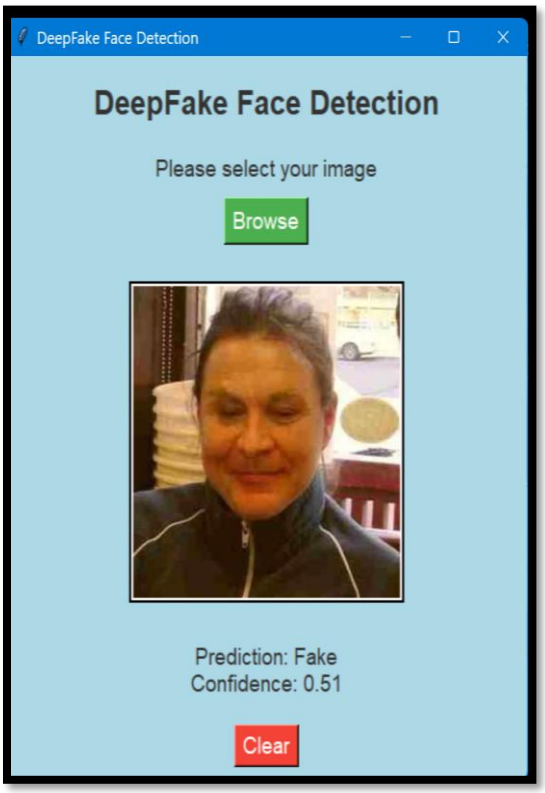


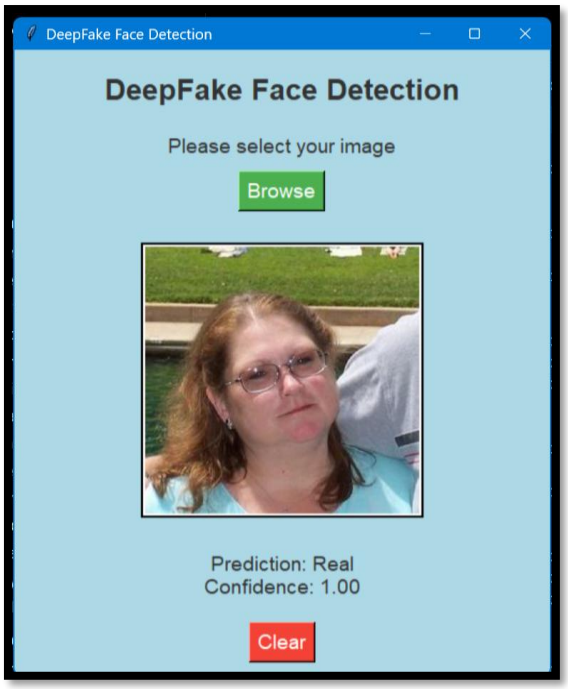
8. Result

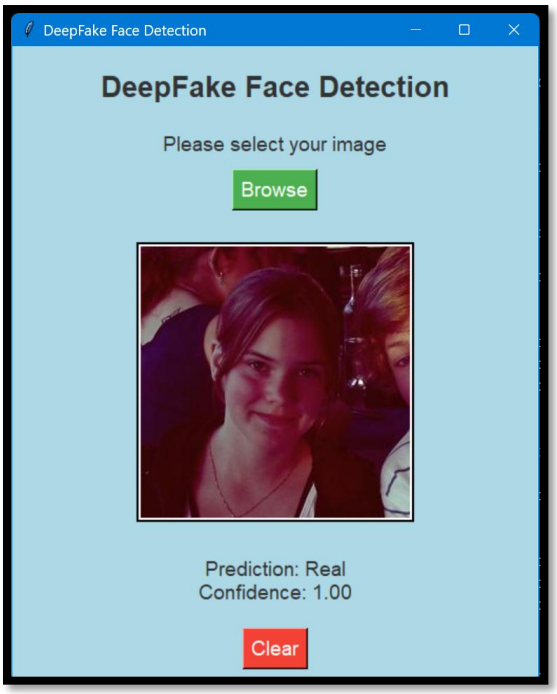
```
rch\hub\checkpoints\resnet18-f37072fd.pth
100%|                                                                                | 44.7M/44.7M [00:08<00:00, 5.31MB/s]
Epoch 1/10: 100%|                                                                  | 232/232 [10:35<00:00,  2.74s/batch]
Epoch 1/10, Loss: 0.4086819556418337
Epoch 2/10: 100%|                                                                  | 232/232 [10:29<00:00,  2.71s/batch]
Epoch 2/10, Loss: 0.27702097911055246
Epoch 3/10: 100%|                                                                  | 232/232 [10:59<00:00,  2.84s/batch]
Epoch 3/10, Loss: 0.22952460264258204
Epoch 4/10: 100%|                                                                  | 232/232 [09:52<00:00,  2.55s/batch]
Epoch 4/10, Loss: 0.19114851577760497
Epoch 5/10: 100%|                                                                  | 232/232 [11:37<00:00,  3.01s/batch]
Epoch 5/10, Loss: 0.18242112616063968
Epoch 6/10: 100%|                                                                  | 232/232 [08:59<00:00,  2.32s/batch]
Epoch 6/10, Loss: 0.15613171102437798
Epoch 7/10: 100%|                                                                  | 232/232 [09:01<00:00,  2.33s/batch]
Epoch 7/10, Loss: 0.12602984858478464
Epoch 8/10: 100%|                                                                  | 232/232 [09:06<00:00,  2.36s/batch]
Epoch 8/10, Loss: 0.11045788756806933
Epoch 9/10: 100%|                                                                  | 232/232 [08:58<00:00,  2.32s/batch]
Epoch 9/10, Loss: 0.10527063869322681
Epoch 10/10: 100%|                                                                 | 232/232 [10:22<00:00,  2.68s/batch]
Epoch 10/10, Loss: 0.1066550620307208
Accuracy: 0.8799351000540833
```

Accuracy of the model : 87.99%









9. Future Scope:-

- Explore video and multimodal detection.
- Address data scarcity and generalizability.
- Partner with social media platforms and raise public awareness.
- Investigate new architectures like Transformers and Explainable AI.

10. References:-

1. Wahidul Hasan Abir; Faria Rahman Khanam; Kazi Nabiul Alam; Myriam Hadjouni; Hela Elmannai, Sami Bourouis; Rajesh Dey; Mohammad Monirujjaman Khan (July 2022) :- Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods

2. Mj Alben Richards; E Kaaviya Varshini; N Diviya; P Prakash; P Kasthuri; A Sasithradevi (September 2023) :- Deep Fake Face Detection using Convolutional Neural Networks

3. Hasin Shahed Shad; Md. Mashfiq Rizvee; Nishat Tasnim Roza; S. M. Ahsanul Hoq; Mohammad Monirujjaman Khan; Arjun Singh; Atef Zaguia; Sami Bourouis (December 2021) :- Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network

4. Remya Revi K. ; Vidya K R; M. Wilscy (February 2021) :- Detection of Deepfake Images Created Using Generative Adversarial Networks: A Review

5. Xu Chang; Jian Wu; Tongfeng Yang; Guorui Feng (September 2020) :- DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network

6. Siddharth Bhamare; Shreeraj Bhamare (February 2024) :- Enhancing Deepfake Image Detection with Deep Convolutional Neural Networks

7. Suganthi ST, Mohamed Uvaze Ahamed Ayoobkhan, Krishna Kumar V, Nebojsa Bacanin, Venkatachalam K, Hubalovsky Stepan, and Trojovsky Pavel (22 February 2022) :- Deep learning model for deep fake face recognition and detection - PMC(nih.gov)