TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

LALITPUR ENGINEERING COLLEGE
KHOLKA POKHARI, LALITPUR

A PROPOSAL OF MAJOR PROJECT

DefaceLab: DeepFake Detection using Deep Learning

SUBMITTED BY

ABHISHEK NEUPANE [LEC076BCT002]
RABINDRA ADHIKARI [LEC076BCT025]
SANJISH MAHARJAN [LEC076BCT032]
SUSHIL KAFLE [LEC076BCT045]

SUBMITTED TO

DEPARTMENT OF COMPUTER ENGINERING

2080 Bhadra 1

DefaceLab: DeepFake Detection using Deep Learning

Submitted by
ABHISHEK NEUPANE [LEC076BCT002]
RABINDRA ADHIKARI [LEC076BCT025]
SANJISH MAHARJAN [LEC076BCT032]
SUSHIL KAFLE [LEC076BCT045]

Project Coodinator
Er. Bishika Subedi

A project submitted in partial fulfillment of the requirements for the degree of
Bachelor of Computer Engineering

DEPARTMENT OF COMPUTER ENGINEERING
Lalitpur Engineering College
Tribhuvan University

2080-04-01

ABSTRACT

Deepfakes are realistic-looking fake media generated by deep-learning algorithms that iterate through large datasets until they have learned how to solve the given problem (i.e., swap faces or objects in video and digital content). The massive generation of such content and modification technologies is rapidly affecting the quality of public discourse and the safeguarding of human rights. Deepfakes are being widely used as a malicious source of misinformation in court that seek to sway a court's decision. Because digital evidence is critical to the outcome of many legal cases, detecting deepfake media is extremely important and in high demand in digital forensics. As such, it is important to identify and build a classifier that can accurately distinguish between authentic and disguised media, especially in facial-recognition systems as it can be used in identity protection too. This is what we tend to do. In this work, we will be using the quiet recent Vision Transformer Model, as it uses the attention based mechanism. Since, the Vision transformer has its own unique attributes and is very much reliable and provides a lot of significant merits in development compared to others, we plan to implement it for our work, for the detection of the AI generated images.

Keywords: deepfake detection; digital forensics; media forensics; deep learning; Visual Transformer; AI generated images

Contents

List of Figures

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
FPS	Frames Per Second
GAN	Generative Adversarial Network
ML	Machine Learning
RNN	Recurrent Neural Network

1 INTRODUCTION

Deepfake technology has revolutionized the world of digital media manipulation. By combining artificial intelligence and image/video processing, deepfakes have garnered widespread attention. Deepfakes involve the creation of realistic media portraying individuals in situations they never experienced or saying things they never said. As this technology becomes more sophisticated, concerns arise regarding its impact on politics, entertainment, and personal privacy. This report provides an overview of deepfakes, including their underlying processes, societal implications, ethical challenges and the way of detection. In navigating this landscape, it is crucial to find a balance between innovation and responsible use in our increasingly digitized society.

In the last few years, cybercrime, which accounts for a 67% increase in the incidents of security breaches, has been one of the most challenging problems that national security systems have had to deal with worldwide.[1] Deepfakes, at present time, are being widely used to swap faces or objects in video and digital content. This artificial intelligence-synthesized content can have a significant impact on the determination of legitimacy due to its wide variety of applications and formats that deepfakes present online (i.e., audio, image and video). Considering the quickness, ease of use, and impacts of social media, persuasive deepfakes can rapidly influence millions of people, destroy the lives of its victims and have a negative impact on society in general [1]. Deepfake technology has been driven by various motivations, including individual attacks, political manipulation, and the spread of false information. Its impact extends beyond personal attacks to manipulating satellite images and using stock images for identity protection. Cyber attackers continuously adapt their strategies, making it challenging to identify deepfake media and stay ahead of evolving threats.

The societal implications of deepfake technology are profound. Misinformation and disinformation fueled by deepfakes erode public trust, damage reputations, and violate privacy at personal and professional levels. Deepfakes can also disrupt democratic processes and contribute to societal polarization. Addressing the legal and ethical concerns surrounding deepfakes requires technological advancements, policy development, media literacy, and careful consideration of privacy rights and the manipulation of visual evidence. To tackle these implications, it is essential to advance deepfake detection methods, bolster cybersecurity measures, promote media literacy for individuals to discern manipulated content, and establish clear legal frameworks governing the responsible use of deepfake technology. By taking a comprehensive approach, we can effectively navigate the ethical challenges and societal impacts posed by deepfakes.

The deepfake technology holds importance in several areas. It offers creative expression and entertainment possibilities, enhances research and development in fields like computer vision, and aids forensic analysis in legal investigations. Deepfakes also emphasize the need for media literacy and critical thinking skills, promoting education and awareness. Ethical considerations and policy development are crucial in addressing the responsible use of deepfakes and protecting individuals' rights. Understanding the significance of deepfake technology enables us to navigate its implications effectively and harness its potential while mitigating potential harm.

We cannot dispute the influence deep fakes will have in the next years given all the benefits and cons that have been presented. Therefore, keeping an eye on deepfake content is crucial. This paper will provide an overview of the fundamental organizational structure of our project on how deepfake detections can be done.

1.1 Background

At present context of time, the rapid advancements in mobile camera technology and the widespread use of social media platforms have made it easier than ever to create and share digital pictures. Deep learning has played a crucial role in developing technologies that were previously unimaginable. One notable example is modern generative models, which can produce highly realistic images, speech, music, and video. These models have been applied in various fields, such as enhancing accessibility through text-to-speech technology and generating training data for medical imaging.

There will always be drawbacks to any technological breakthrough. Since deepfakes are still relatively new and expanding quickly, their excessive use as a result of rising human interest has resulted in misuse of this technology. It is simple for widespread false information to proliferate among the populace when there is no controlling element and a weak mechanism in place to identify deep fakes.Since their initial emergence in late 2017, a variety of open-source deep fake generation techniques and tools have appeared, resulting in an increase in the amount of synthetic media clips. Others may be destructive to people and society, even though many are probably intended to be amusing. Due to the accessibility of editing tools and the strong demand for topic expertise, false digital contents have been growing in number and in realism up until recently.

Deep fakes are now widely disseminated on social media platforms, which encourages spamming and the spread of false information. Just picture a deep fake image of Donald Trump getting arrested which was trending on twitter or a deep fake of a well-known celebrity assaulting their supporters. These types of misinformation can brainwash the audience and are awful and endanger and mislead the general public.

Deep fake detection plays a crucial part in overcoming such a circumstance. Therefore, we provide a novel deep learning-based method that can successfully separate artificial intelligence-generated fake contents from authentic digital materials. In order to identify deep fakes and stop them from spreading across the internet, it is crucial to develop technology that can detect deepfakes.

1.2 Problem Statement

With the help of visual effects, convincing modifications of digital photographs and videos have been proven for many years. However, new developments in deep learning have dramatically increased the realism of fake content and made it more widely available. These purportedly artificial intelligence-generated works of media are also known as ”deepfakes”. It is easy to create deep fakes utilizing artificial intelligence techniques. However, it is extremely difficult to identify these Deep Fakes. Globally, it is found out that about 71% of total population using internet do not know what a deepfake is. Just under a third of global consumers of internet say they are aware of deepfakes [8]. In the past, there have been numerous instances of deep fakes being used to effectively incite political unrest, stage terrorist attacks, blackmail individuals, etc. In North America alone, the proportion of deepfakes more than doubled from 2022 to 2023. This proportion jumped from 0.2% to 2.6% in the US. It is up from 0.1% to 4.6% in Canada [7] and is rapidly growing. Therefore, it becomes crucial to identify these deep fakes and montior their spread through social media. Hence, with the growing curiosity we have taken a step forward in detecting the deep fakes using different transformer based models.

1.3 Objectives

- Our project can help in reduction in spread of false informations, that might mislead the people on the internet.
- Our project will distinguish and classify the video as deepfake or pristine.

1.4 Scope

At present time there are numerous tools available for creating false videos in the current deepfake technology landscape, but there are few trustworthy tools available for spotting them. The idea creation of a deepfake detection software to solve this discrepancy and stop the widespread dissemination of deepfakes is what our project is based upon. Users will be able to post images through our platform and segregate them as authentic or deepfake. This project can also be developed to include the development of a plugin for browsers that will automatically detect deep fakes. Notably, our idea can me implemented on different social sites as well as in various various governmental organizations. A synopsis of the program with the size of the input, bounds on the input, input validation, input dependency, the i/o state diagram, and the major inputs and outputs are explained in this report.

2 LITERATURE REVIEW

Since 2017, the Transformer has been very renowned as a new type of neural architecture which encodes the given input data into a powerful feature, with the help of attention mechanism. With the research article published in 2017, named ”Attention Is All You Need” [6], by taking it as a framework, numerious researches has been done recently upon the visions task as well [3]. This is where the Visual Transformer comes in.

While the Transformer Architecture has been a new standard for Natural Language Processing tasks, its application to computer vision remains limited. From many recent studies and researches done, the common points mentioned was that in vision, either the attention is applied in conjuction with convolutional networks, or used to replace certain components of convolutional network while keeping their overall structure in place [4].

The Transformer-liked architecture has been employed in the computer vision field in present context in three fundamental computer vision tasks, namely classification, detection and segmentation [2]. Our work lies within one of these fundamental tasks i.e the detection of artificially generated images of a human face.

The Visual Transformer has been gaining many contributions in recent years such that its capabilites has increased beyond the old traditional models. As a matter of fact, the Visual Transformer model has clear advantages over traditional models such as CNNs and RNNs in specific situations. It can process whole images in sections, which helps it understand the bigger picture. Also, it is good with working on different image sizes and tasks. Training these models on big datasets first and then adjusting it for specific tasks makes it really flexible compared to the traditional model [5]. One of its key characteristics, self-attention, not only helps us understand how the model works but also reminds us to pick the right model based on the job, data, and what we have.

In summary, with all these distinct attributes and merits the Vision Transformer imposes significantly compared to others at present time, it has led us to select it as the major model for our project development.

2.1 Existing Systems

2.1.1 Deepware

Deepware.ai is an innovative company at the forefront of deepfake detection technology. They specialize in developing advanced AI-driven solutions to combat the spread of manipulated media content. With a team of expert researchers and engineers, Deepware.ai leverages state-of-the-art machine learning algorithms and deep neural networks to accurately identify deepfakes. Their cutting-edge technology, combined with a user-friendly approach, empowers individuals and organizations to protect themselves from the potentially harmful consequences of deepfakes. Deepware.ai's commitment to continuous improvement and staying ahead of evolving deepfake techniques positions them as a trusted leader in the field, offering reliable and scalable solutions that contribute to a safer digital landscape.

Figure 1: Deepware

2.1.2 DuckDuckGoose

DuckDuckGoose offers an open-source browser extension that keeps tabs on all websites you visit and alert you once manipulated media is detected. Users should also appreciate the transparency of the DeepFake detector, as DuckDuckGoose provides detailed explanations for why a video was flagged to give you some insight on what to look for in a DeepFake. The team behind the tool has been dedicated to sharing their research findings and encouraging participants from the community to contribute to building a more reliable model with higher accuracy.

Figure 2: DuckDuckGoose

2.2 Proposed Systems

Our Deepfake Detection System is an online tool designed to help people identify and deal with deepfake content. It focuses on providing strong detection capabilities for deepfakes. Users can use the system to analyze and detect potential deepfakes by submitting images. The system uses advanced algorithms and machine learning techniques to accurately identify manipulated or fake media. This helps users recognize and address the risks that come with deepfakes, such as spreading false information, committing fraud, or violating privacy. The deepfake detection system aims to empower users by giving them the tools they need to protect themselves and others from the harmful effects of encountering deepfakes. With the help of cutting-edge algorithms, the system assists users in detecting and raising awareness about deepfakes, making the digital world a safer and more informed place.

3 FEASIBILITY STUDY

3.1 Economic feasibility

This is a low-budget project with no development costs. The total expenditure of the project is just computational power. The dataset and computational power required for the project are easily available. The computational power is easily provided by google collab. So, the project is economically feasible. The system will be simple to comprehend and use. As a result, there will be no need of trained personnel to use the system. This system will have the capacity to expand by adding more components.

3.2 Operational feasibility

The project is operationally feasible since after the completion of the project, it can be operated as intended by the user to solve the problems for what it has been developed.

3.3 Technical feasibility

The purpose of technical feasibility is to establish whether the project is possible in terms of software, hardware, manpower, and knowledge to complete. It will take into account determining resources in support of the suggested scheme. The system is platform independent because it is written in Python. Advanced machine learning libraries are available and the technology is cutting-edge. As a result, the system is technically possible.

4 METHODOLOGY

4.1 Software Development Life Cycle

Agile method of Software Development uses iterative approach. Agile method cycles among Planning, Requirement Analysis, Designing, Development and Testing stages. These cycle is called sprints. Each sprints are considered as a miniature project on itself. Using this method allowed us to update various parts of project at any point of project development. In this model an iterative approach was taken where working software was delivered after each iteration some new features is added to main system. It works in incremental and iterative approach. Agile model mainly focuses on customer collaborations, on individuals and iterations and welcomes changes at anytime in SDLC process. We prefer to use agile model in this system as it helps in developing realistic systems and promotes teamwork during software development. Also system is easy to manage and it can accommodate new changes at any stages of software development phase.

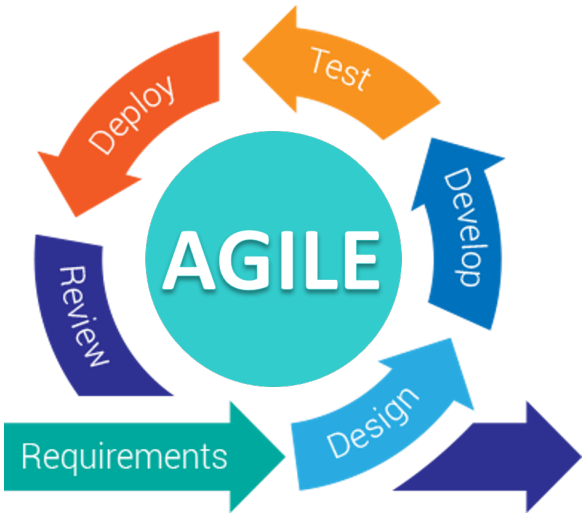

Figure 3: Agile Model

4.2 System Development Tools

Our static Deepfake detection System requires Python, Tensorflow, OpenCV, Machine Learning which are listed below:

1. Python
2. Pytorch
3. NumPy
4. OpenCV
5. Tensorflow

4.3 Functional Requirement

The functional requirements of the system are:

1. Detecting the Faces from Images and Videos.
2. Testing for realism of image.

4.4 Non Functional Requirement

These requirements are not needed by the system but are essential for the better performance of software. The points below focus on the non-functional requirement of the system.

- Reliability
- Usability
- Security
- Portability
- Speed and responsiveness
- Performance

4.5 Approach for Deepfake Detection

For deepfake detection, our approach combines the power of vision transformers with advanced techniques in computer vision. Vision transformers excel in capturing both global and local features within an image, making them suitable for identifying subtle inconsistencies introduced by deepfake manipulation.

Our approach involves the following steps:
Dataset Collection: We gather a diverse dataset comprising real and deepfake images. This dataset is crucial for training and evaluating our deepfake detection model.

Vision Transformer Architecture: We chose the Vision Transformer (ViT) architecture due to its ability to process image patches and learn relationships between them using self-attention mechanisms. ViT has shown impressive results in various computer vision tasks, and we adapt it for deepfake detection.

Preprocessing: We preprocess the dataset to extract image patches and resize them to a consistent input size. These patches retain essential information while reducing computational complexity. Additionally, we normalize pixel values to ensure consistent input for the model.

Training: During training, our vision transformer learns to differentiate between real and manipulated images. We use a binary cross-entropy loss function to optimize the model's weights. The self-attention mechanism in the ViT helps the model focus on relevant patches and capture intricate patterns indicative of deepfake manipulation.

Evaluation: We evaluate the model's performance using various metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into how effectively the model distinguishes between real and deepfake images.

Vision Transformer Architecture:
The Vision Transformer (ViT) architecture comprises the following components:

Patch Embedding: Input images are divided into non-overlapping patches. Each patch is linearly projected to obtain embeddings, which are then augmented with positional encodings to maintain spatial information.

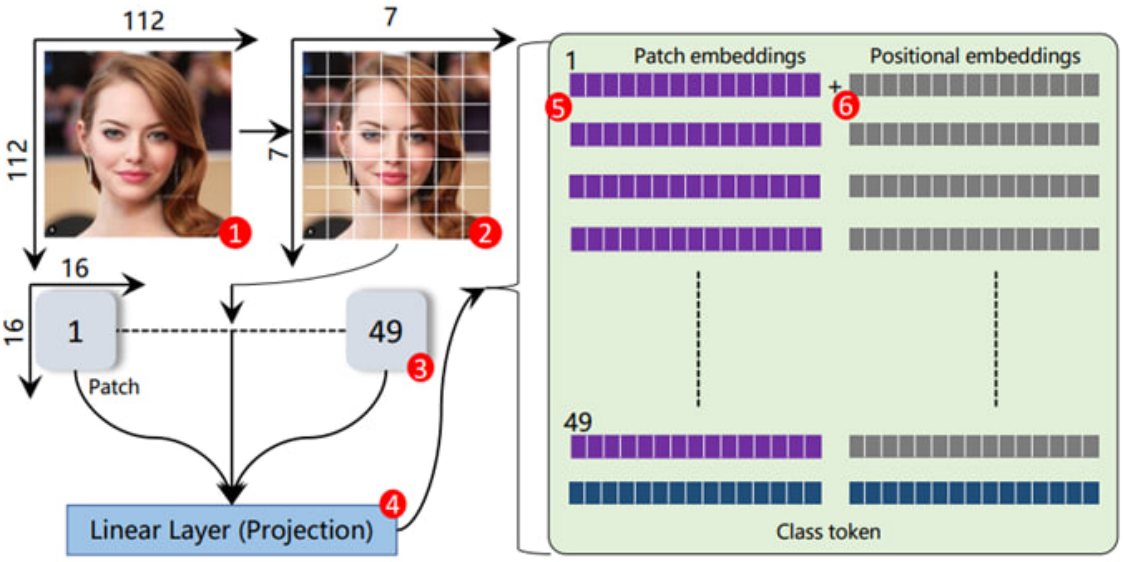

Figure 4: Patch Embedding

Transformer Encoder: The patch embeddings are fed into a stack of transformer encoder layers. Each layer consists of multi-head self-attention and feedforward neural networks. This enables the model to capture both local and global dependencies within the image.

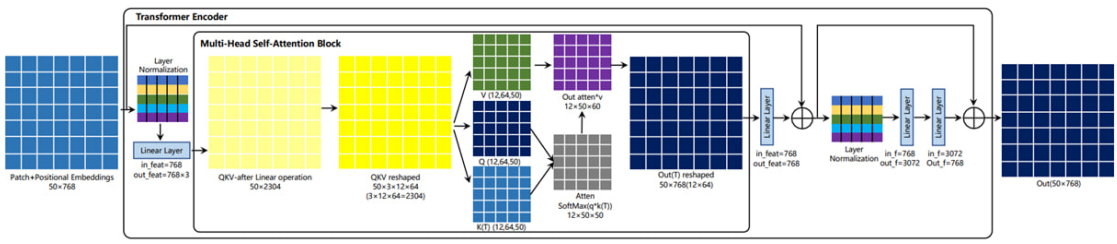

Figure 5: Transformer encoder

Classification Head: The final layer of the model serves as the classification head. It takes the transformed embeddings and predicts whether an image is real or manipulated.

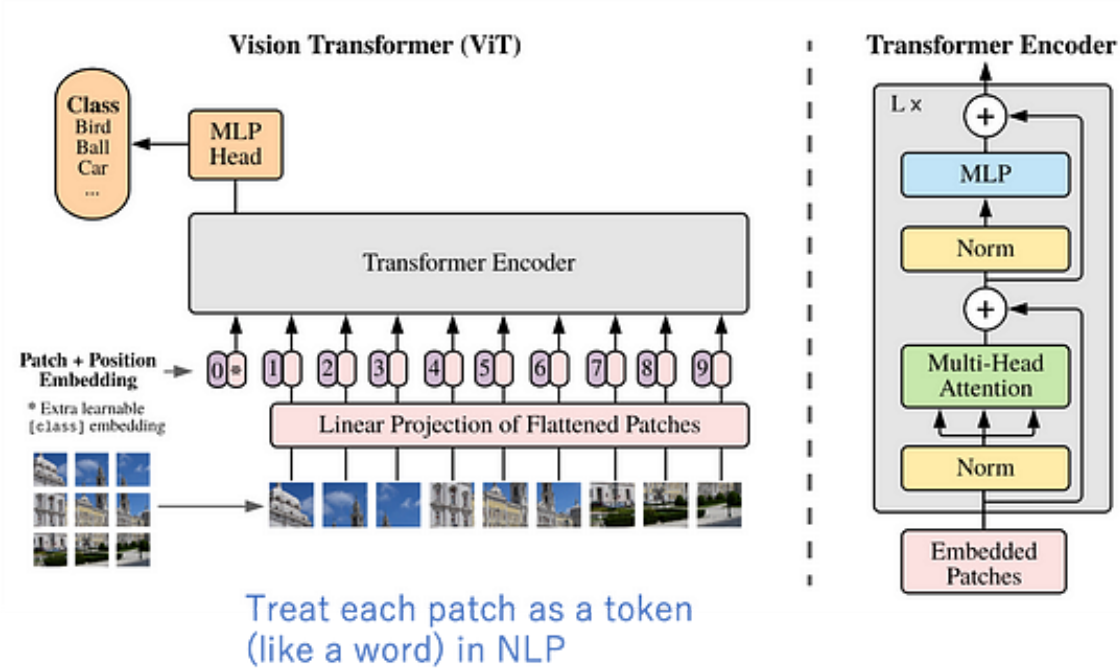

Figure 6: Vision Transformer Architecture

Steps:

Dataset Overview:
Our dataset includes a diverse collection of real and deepfake images sourced from various public datasets and proprietary sources. It covers a wide range of scenarios, lighting conditions, and facial expressions to ensure the model's robustness. The dataset was meticulously curated to represent real-world variations and challenges that deepfake detection might encounter.

Data Collection Process:
Our dataset acquisition process involved the following steps:

1. **Public Datasets:** We sourced a significant portion of our dataset from publicly available deepfake and real image datasets. These datasets were chosen for their variety and relevance to our detection task.

2. **Proprietary Sources:** To further enhance the diversity of our dataset, we collaborated with proprietary sources that provided deepfake and real images. These sources helped us ensure a comprehensive representation of different contexts and manipulation techniques.

3. **Video Conversion:** To include videos in our dataset, we first converted them into individual frames (images) to facilitate compatibility with the vision transformer architecture. This step involved extracting frames at a consistent frame rate from each video, resulting in a sequence of images for each video.

4. **Frame Selection:** To avoid redundancy and maintain dataset balance, we carefully selected frames from videos to represent various stages of manipulation, expressions, poses, and lighting conditions.

5. **Annotation and Labeling:** Each image was labeled as either ”real” or ”deepfake.” Annotations were done manually to ensure accurate labeling for training and evaluation.

Preprocessing Techniques:
Before training the model, we preprocess the dataset as follows:

1. **Resize:** All images, whether sourced from videos or other datasets, were resized to a consistent resolution, such as 224x224 pixels. This resizing ensured a uniform input size for the vision transformer.

2. **Augmentation:** To increase the dataset's diversity and improve the model's ability to generalize, we applied various data augmentation techniques. These techniques included random rotations, horizontal flips, brightness adjustments, and minor deformations.

3. **Normalization:** Pixel values of the images were normalized to a specific range to ensure consistent input for the model during training and inference.

By meticulously preprocessing the dataset, we enhanced the model's capacity to learn relevant features and intricate patterns required for accurate deepfake detection.
This comprehensive approach leverages the capabilities of vision transformers while customizing them to address the specific challenges associated with deepfake detection. By integrating diverse data sources and applying rigorous preprocessing steps, we developed a robust and efficient deepfake detection model that demonstrates impressive performance across various scenarios.

5 BLOCK DIAGRAMS

5.1 System Architecture

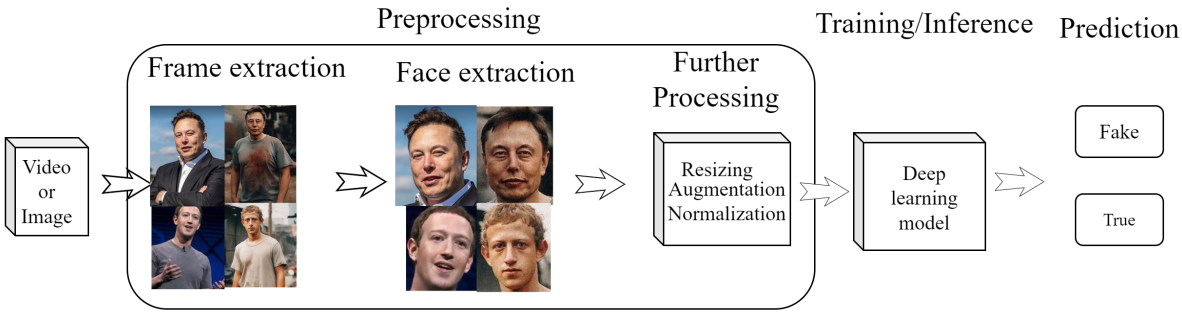

Figure 7: System Architecture

The system architecture of our project involves multiple steps. Initially, frames are extracted from the input video or obtained directly from an image. These frames then undergo face extraction, where faces are identified and cropped using face detection algorithms. The extracted faces are resized to a standardized size and undergo normalization to ensure consistent pixel values. The preprocessed face images are then fed into a deep learning model for classification. The model analyzes the features and patterns in the images to determine whether they are real or fake. Finally, the system produces the output, indicating the authenticity of the input video or image as either real or fake.

5.2 Use Case Diagram

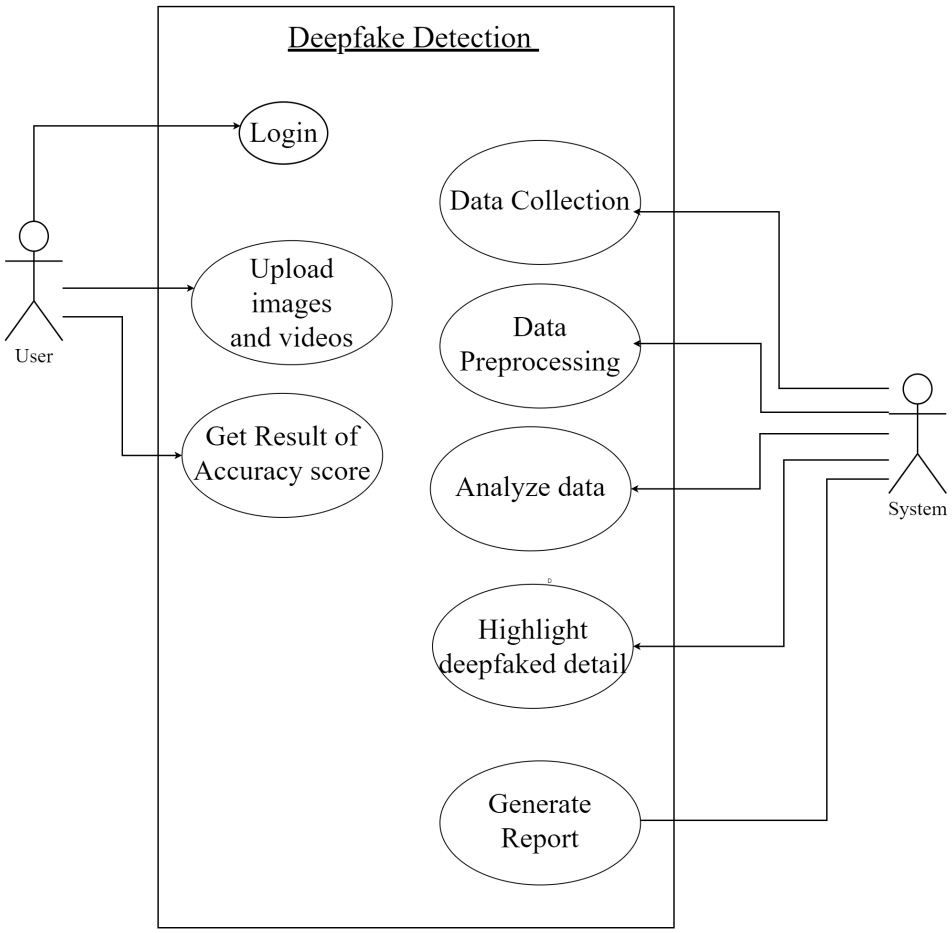

Figure 8: Use Case Diagram

The use case diagram for our system illustrates various interactions and roles of the system's users. The primary actors involved are the "User" and the "System." The User interacts with the system by initiating the deepfake detection process, either by uploading a video or an image. The User can also access the system to view the detection results. On the other hand, the System is responsible for managing the system, including user authentication, system configuration, and monitoring the overall functionality. The use case diagram shows the main use cases, such as "Upload Media," "Detect Deepfake," and "View Results," which represent the key functionalities of the system.

5.3 Sequence Diagram

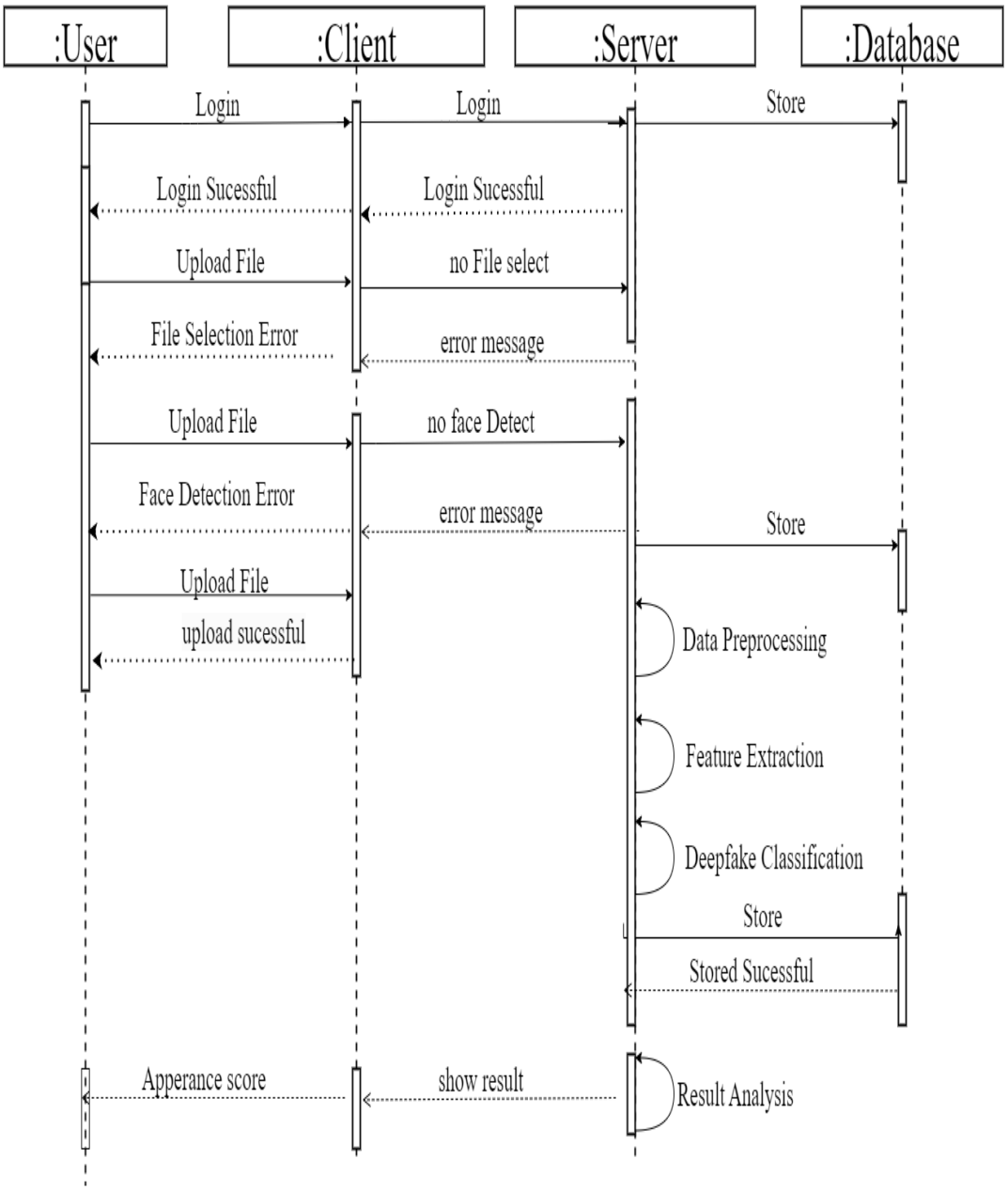

Figure 9: Sequence Diagram

The sequence diagram shows that the user initiates the process by accessing the system and providing their login credentials. The Server checks the login credentials and verifies the user's identity. Once authenticated, the user proceeds to upload a file containing the video or image to be analyzed for deepfakes.Then face detection algorithms is used to detect and extract faces from the uploaded media. This collected data undergoes further processing, including resizing and normalization, to prepare it for deep learning modeling. Finally, the processed data is fed into the deep learning model, which analyzes the features and patterns to classify the media as either real or fake.

5.4 Dataflow Diagram

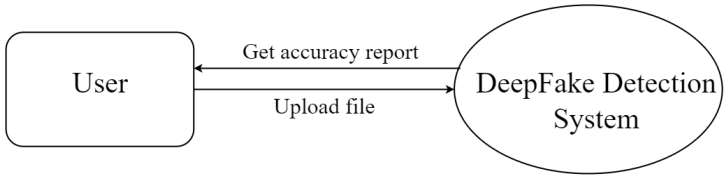

Figure 10: Level 0 DFD

DFD level – 0 indicates the basic flow of data in the system.

- User: User input to the system is uploading video.
- System: In system it shows all the details of the Video and output shows the fake video or not.
  and output flow

Hence, the data flow diagram indicates the visualization of system with its input feed to the system by User.

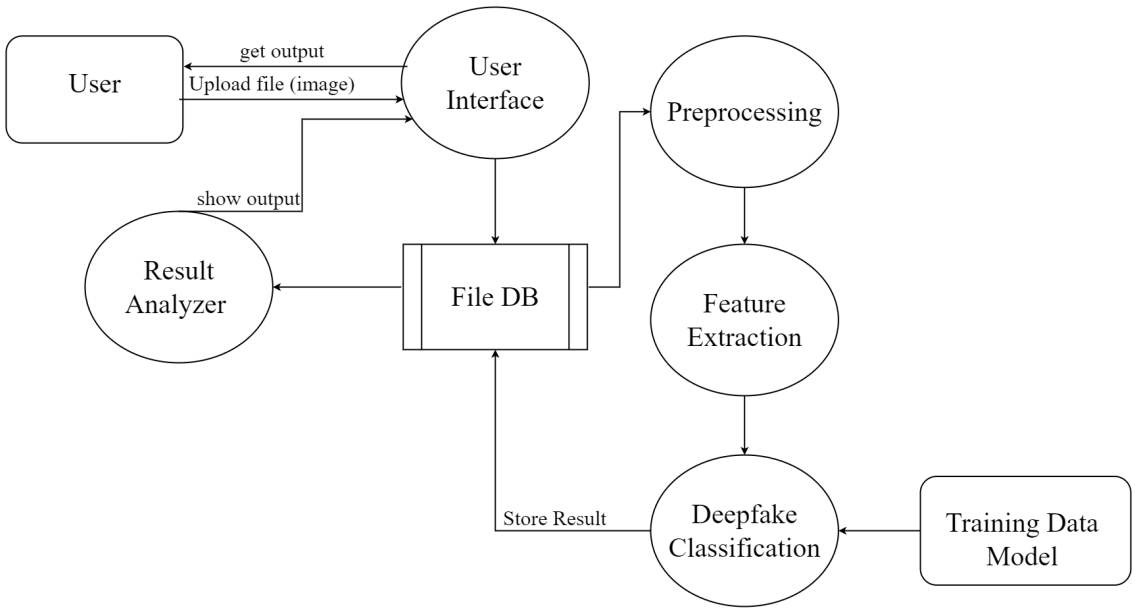

Figure 11: Level 1 DFD

DFD Level – 1 gives more in and out information of the system.Where system gives detailed information of the procedure taking place.

5.5 Activity Diagram

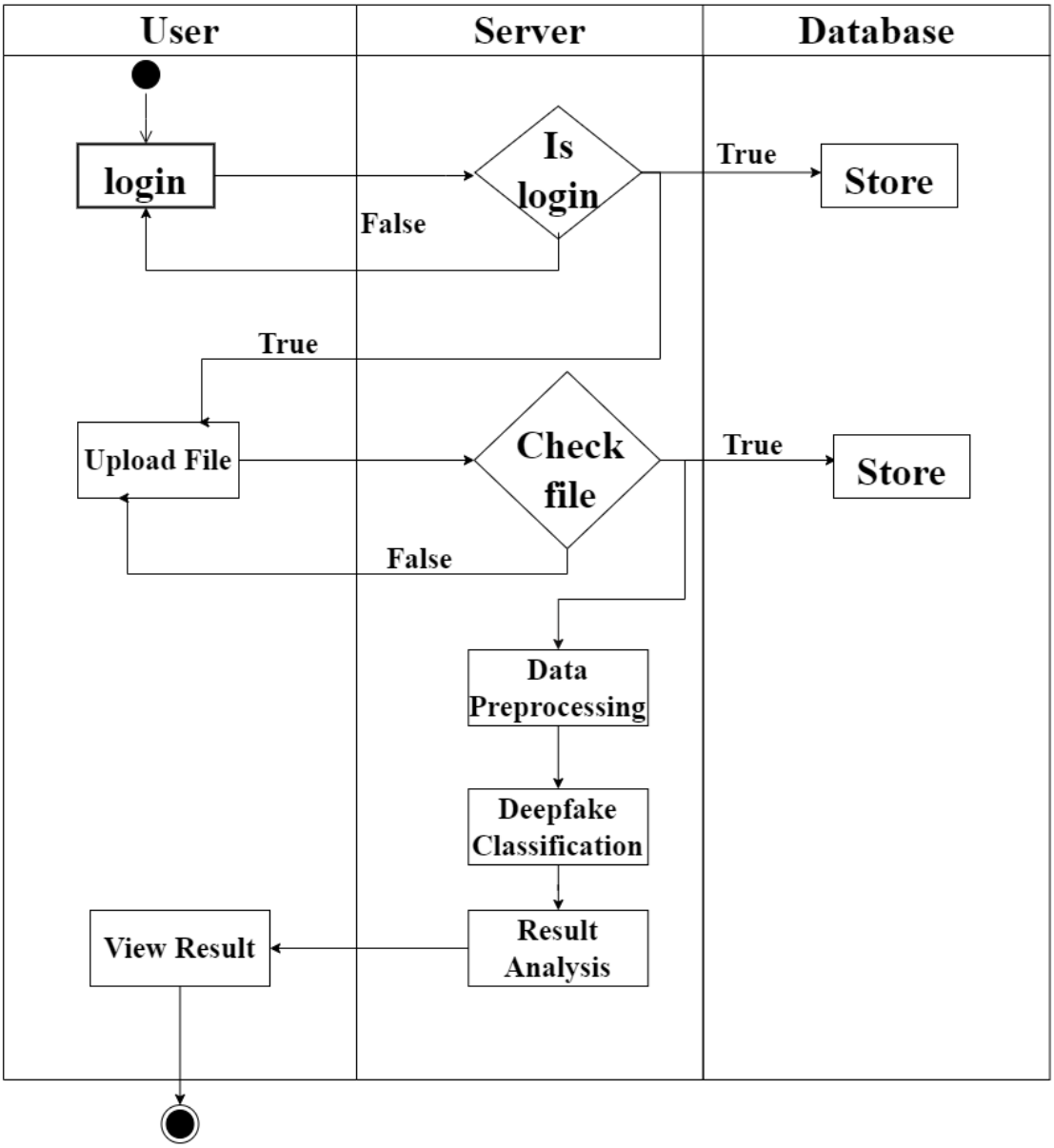

Figure 12: Activity Diagram

The activity diagram shows that the user initiates the process by accessing the system and providing their login credentials. The Server checks the login credentials and verifies the user's identity. Once authenticated, the user proceeds to upload a file containing the video or image to be analyzed for deepfakes.

6 EXPECTED OUTCOMES

- User-friendly interface for easy upload and clear result presentation.
- Accurate identification of manipulated media content.
- Robust performance against different deepfake techniques and adversarial attacks.

References

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images"

[2] Y. Liu et al., "A Survey of Visual Transformers," in IEEE Transactions on Neural Networks and Learning Systems.

[3] Han, K., Xiao, A., Wu, E., Guo, J., XU, C., and Wang, Y. (2021). "Transformer in Transformer".

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations"

[5] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, Zhiqiang He. (2021). "A Survey of Visual Transformers".

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I. (2017). "Attention is all you need. In Advances in neural information processing systems (Vol. 30)".

[7] Douglas Blakey, "Forced verification and AI/deepfake cases multiply at alarming rates: Sumsub" (2023). https://www.electronicpaymentsinternational.com/news/forced-verification-and-ai-deepfake-caeses-sumsub/

[8] iproov, "How To Protect Against Deepfakes – Statistics and Solutions" (2022). https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection