DeepTruth

Detecting DeepFakes using Deep Learning algorithms

By:

Hamza
27671
Omer Fayyaz Khan
28559
Muhammad Labib Amir
26217

Supervised by:
Dr. Jawaid Iqbal
(Assistant Professor/In-charge Program Cyber Security)

Faculty of Computing
Riphah International University, Islamabad
Fall 2024

A Dissertation Submitted To

Faculty of Computing,

Riphah International University, Islamabad

As a Partial Fulfillment of the Requirement for the Award of

the Degree of

Bachelors of Science in Cyber Security

Faculty of Computing
Riphah International University, Islamabad

Date: 12th November, 2024

Final Approval

This is to certify that we have read the report submitted by ***Hamza (27671), Omer Fayyaz Khan (28559), and Muhammad Labib Amir (26217)***, for the partial fulfillment of the requirements for the degree of the Bachelors of Science in Cyber Security (BSCY). It is our judgment that this report is of sufficient standard to warrant its acceptance by Riphah International University, Islamabad for the degree of Bachelors of Science in Cyber Security (BSCY).

Committee:

1 ______________________________

Dr. Jawaid Iqbal
(Supervisor)

2 ______________________________

Dr. Musharraf Ahmed
(Head of Department)

Declaration

We hereby declare that this document “DeepTruth - Detecting DeepFakes using Deep Learning algorithms” neither as a whole nor as a part has been copied out from any source. It is further declared that we have done this project with the accompanied report entirely on the basis of our personal efforts, under the proficient guidance of our teachers especially our supervisor **Dr. Jawaid Iqbal**. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from anywhere else, we shall stand by the consequences.

Hamza
27671

Omer Fayyaz Khan
28559

Muhammad Labib Amir
26217

Dedication

"DeepTruth - Detecting DeepFakes using Deep Learning Algorithms" is all about finding the truth and protecting authenticity in this digital age. In today's world, where DeepFake content is spreading like an epidemic on social media, it is becoming difficult to tell what is real and what is manipulated. The deceptive and misleading nature of DeepFakes can adversely affect individuals as well as they society as a whole.

We want to dedicate our work to all the amazing people who are actively developing cutting edge technology and methods to identify and minimize the impact of DeepFakes. They are the real heroes, fighting against digital deception and keeping the truth alive.

We also want to dedicate this project to all the instructors, researchers, and students who are fascinated by the perplexity of deep learning and how it can solve real problems. We hope that our project will widen the knowledge base in the field of DeepFake detection and inspire even more breakthroughs.

Acknowledgement

First of all, we are obliged to Allah Almighty the Merciful, the Beneficent and the source of all Knowledge, for granting us the courage and knowledge to complete this Project. Next, our families deserve our gratitude for their constant support and patience while we were working on this project. We sincerely appreciate Dr. Jawaid Iqbal, our supervisor, for his mentoring and insightful advice. Last but not least, we would also like to thank the entire teaching staff at the Faculty of Computing at Riphah International University for equipping us with the fundamental knowledge and abilities needed for this project.

Hamza

27671

Omer Fayyaz Khan

28559

Muhammad Labib Amir

29217

Abstract

DeepFakes are synthetic media forged by digital manipulation. Creating such content has become extremely convenient as the exponential growth in computing power has further strengthened deep learning algorithms. DeepFakes have given rise to serious matters including but not limited to interference in politics, extortion, and non-consensual pornography. We are offering detection of fake videos, generated using artificial intelligence, in this project. A two staged deep learning model is used in our system. Firstly, to draw out frame features of a video, ResNeXt convolutional neural network is used. These characteristics facilitate in identifying the footage as authentic or doctored by being passed into Long Short-Term Memory which is an improved recurrent neural network. This model sanctions higher accuracy in the classification of such videos. We made use of an extensive dataset, gathered from various origins, to test our methodology keeping in view real world scenarios and efficiency. Our cumulative dataset originated from Celeb-DF [1], DFDC [2], and Face Forensic++ [3]. To counter this menace, we intend to utilize artificial intelligence in our system. Our results point towards a higher probability of reduction in the proliferation rate of DeepFakes.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Chapter 1: Introduction

1.1 Introduction

The amalgamation of deep learning and fake, called DeepFakes, have turned out to be a disturbing invention. Such synthesized content obscures the thin line between reality and fantasy, by effectively manipulating human faces. Our research motivation and proposed solution's objective in addition to the opportunities and challenges related to DeepFakes, are the prime focus of this chapter. All this is followed by an outline for the entire report.

1.2 Opportunity & Stakeholders

Manipulation of videos has turned much easier due to the advancement in deep learning algorithms. Although potentially used for vile purposes, DeepFakes may also come in handy for recreation, such as for creating funny imitations. Notable stake, in the identification of DeepFakes, is held by various sectors.

1.2.1 Media & News Organizations

The opinion of the public can be manipulated by DeepFakes. Also, a loss of trust can occur towards known news sources.

1.2.2 Law Enforcement & Government

It is of vital importance to have efficacious DeepFake identification techniques for communal balance. Dissemination of infuriating and provoking material can be avoided this way.

1.2.3 Individuals

The use of DeepFake for blackmail, revenge porn, and harassment highlights the criticality for protection of individuals.

1.3 Motivations and Challenges

Due to the advancements in artificial intelligence and data manipulation, DeepFake material is now much more usual. Misleading people is now very convenient subject to the almost real but fake content. Politicians, the entertainment industry, and public can be widely affected by DeepFakes. This makes development of trustworthy DeepFake detection mechanisms inevitable.

1.3.1 Motivations for DeepFake Detection

Deepfakes may cause catastrophic events. A manipulated video of an actor lambasting his fans or a politician inciting public hatred for another country are excellent examples of this abuse. DeepFakes usually cause distrust, and controversies.

It is critical to develop artificial intelligence based DeepFake detection. Distribution of such content and it's after effects can be avoided by effective identification. Making use of the combination of ResNeXt and LSTM, our proposed solution aims at achieving this desired result.

1.3.2 Challenges in DeepFake Detection

Accurately identifying these forged videos is a tough problem due to the widespread availability of DeepFake creation tools.

DeepFake detection algorithms find it extremely arduous to compete with the ever-evolving tactics of content synthesizers. Adaptability and resiliency are necessary traits for such algorithms.

Knowledge of deep learning and programming languages is imperative for working with technologies such as ResNeXt and LSTM. Another challenge is the creation of a user-friendly graphical user interface.

Immense processing power is required to train deep learning models, primarily ones based on ResNeXt and LSTM networks. There is round-the-clock research on the topic of maintaining balance between the complexity and efficiency of a detection model.

A generalized approach to DeepFake detection methods is yet to be achieved. DeepFakes created using various methods may not be detected by a model trained on a specific type of DeepFake. Generalizability covering numerous DeepFake methods continues to remain a hurdle.

1.4 Significance of study

By addressing the urgent need for trustworthy DeepFake detection techniques, this study promotes digital trust and public safety. This study intends to lessen the negative effects of DeepFake material across several domains by creating a reliable detection system with ResNeXt and LSTM. The results have important ramifications for personal privacy, cybersecurity, and media authenticity. Improved DeepFake identification could help governments stop the spread of false information, help law enforcement protect people, and help journalists preserve their credibility.

All things considered, this study lays the groundwork for upcoming developments in detection technology, providing an essential instrument for defending both individual liberties and social stability in a world growing more digital.

1.5 Goals and Objectives

A reliable DeepFake detection tool is the prime goal of our work. Our project will:

1.5.1 Discern Authenticity

Precisely differentiate between DeepFakes and real content.

1.5.2 Combat Misinformation

Dissemination of deceptive and damaging material can be reduced by efficiently identifying manipulated content.

1.5.3 User-Friendly Interface

For ease of use, a user-friendly interface is necessary. This assists naive users with conveniently identifying forged content.

1.6 Scope of project

Our project’s desired end product is a solution to precisely identifying DeepFakes. This system will help discern between real and fake content making use of advanced artificial intelligence techniques. Public awareness regarding the evil of DeepFakes and an efficient identification system to mitigate its dissemination is another objective. Our approach is focused around preservation of trust and protecting people from falsified information and blackmail. Via these outcomes, our project aims at solving the DeepFake puzzle so as to provide a viable solution to fight back its negative consequences.

A two-stage deep learning mechanism is what we suggest for DeepFake detection. In the first stage, ResNeXt is utilized for extraction of frame features. These features assist with identifying inconsistencies that may point towards manipulation. Upon passing through LSTM, these features eventually help identify a video as real or fake.

1.7 Chapter Summary

In this preliminary chapter we studied the transformation and consequences of DeepFakes. By pointing out the effected parties such as media houses, politicians, and individuals, we have reiterated the criticality of effective DeepFake detection. Issues such as computing power and technical knowledge were studied for development of DeepTruth. Privacy of individuals, and authenticity of media highlights the significance our project. Preventing dissemination of forged content, assuring its authenticity, and a user friendly interface are the prime aims of our system.

Chapter 2:
Market Survey

Chapter 2: Market Survey

2.1 Introduction

A drastic increase in the use of DeepFakes has adversely affected trust and integrity of data. The two main reasons of manipulated content are damaging goodwill and regulating public responses. The main aim of this chapter is to examine the detection techniques. The pros and cons of current products are studied in detail.

2.2 Technologies & Products Overview

The existing detection systems and their methodology, accuracy, and application are the core elements of this section. By examining their pros and cons, we analyzed their effectiveness in real life scenarios.

Sensity detects face manipulations such as lip-sync, and face swapping using deep learning techniques, mostly convolutional neural networks (CNNs). To identify discrepancies in content, the technology also integrates pixel analysis. Sensity is appropriate for high-volume analysis in industries like media, and law enforcement since it allows integrations through APIs and provides enterprise-grade functionality.

DeepFake Detector uses CNNs trained on big datasets of authentic and altered media. It focusses on locating manipulation traces that are characteristic of DeepFakes. The accessibility of DeepFake Detector for individual and small-scale users is one of its advantages; nevertheless, its efficacy against really complex forgeries may be limited.

Using deep learning models, Deepware Scanner looks for patterns and anomalies in videos to identify DeepFakes. Because of its versatility and batch scanning possibilities, this tool is appropriate for organization that needs to verify big amounts of media.

Microsoft's technology examines image or video frames of altered material using deep learning algorithms trained on massive datasets such as FaceForensics++. By concentrating on minute variations in lighting and grayscale, it calculates the probability of manipulation. This technique is real-time and helps highlight possible misinformation.

2.3 Comparative Analysis

Products in the DeepFake detection realm are given in Table 2.1: Comparative Analysis:

Table 2.1: Comparative Analysis

Criteria	Sensity.ai	DeepFake Detector	Deepware Scanner	Microsoft Video Authenticator
Detection Accuracy	More than 95% accuracy in identifying face swaps, and lip-sync, using deep learning and pixel analysis	Reliable for simple manipulations, may struggle with complex DeepFakes	More than 90% accuracy, though results vary based on DeepFake complexity	Very high accuracy with real-time analysis, especially in media with subtle inconsistencies
Deployment Flexibility	Cloud-based, on-premise, and API integration for enterprise use	Primarily cloud-based, lacks on-premise or API options	Available as standalone and API, supports batch processing	Primarily cloud-based, with API access for select partners, designed for enterprise integration
Target Audience	Enterprise-focused, suitable for media, law enforcement, and financial institutions	Consumer-focused, ideal for individual users, journalists, and small businesses	Focused at both individual and enterprise users needing batch media verification	Enterprise-focused, such as news agencies and social media platforms needing real-time verification
Real-time Capabilities	Near-real-time detection, beneficial for live monitoring in high-risk sectors	Lacks real-time capability, best suited for post-upload verification	Supports near-real-time detection for some plans	Excels in real-time analysis, suited for live media verification
Usability	User-friendly interface, with API integration for easy deployment in enterprise setups	Simple, accessible platform, suitable for non-technical users	Straightforward interface with batch processing and API option	Easy to use, though enterprise setup may require some technical support
Cost	Custom pricing for enterprise	Cost-effective for individuals and small businesses	Tiered pricing, with basic and advanced plans	Pricing not publicly disclosed, customized based on enterprise needs

2.4 Problem Statement

The integrity of information is widely affected by the worldwide existence of DeepFake material in the cyber realm. Most of the existing methodology for Deepfake identification are severely constrained. Also, access to various technologies is limited to paid users only. Their resource-intensiveness frequently demands extensive computational power, which limits their applicability. Scalability concerns arise from the inability to cater for larger amounts of media, which in turn reduces its usefulness. Failing to provide real time detection is another critical deficiency in the existing systems. Lastly, the accuracy of such technologies is jeopardized by the disproportionate false positive and false negative results. In order to create a viable solution to DeepFakes, it is vitally important to address these weaknesses.

2.5 Chapter Summary

A thorough market analysis of the available DeepFake detection techniques is given in the Market Survey chapter. We discussed the importance of accurate detection methodologies to protect goodwill and stop the dissemination of forged content. Next, we performed an assessment of existing detection systems: Sensity.ai, DeepFake Detector, Deepware Scanner, and Microsoft Video Authenticator. Various key criteria were considered while assessing these products. Also, this study highlighted the distinctive advantages and disadvantages of each product.

Chapter 3:
Requirement and System Design

Chapter 3: Requirement and System Design

3.1 Introduction

A system's purpose, functions, and limitations are depicted effectively by requirement engineering. The needs of the users are assured by the resultant system design.

3.2 System Architecture

System architecture plays a vital role while formulating the system design. Figure 3.1: System Architecture displays how our model will proposedly work.

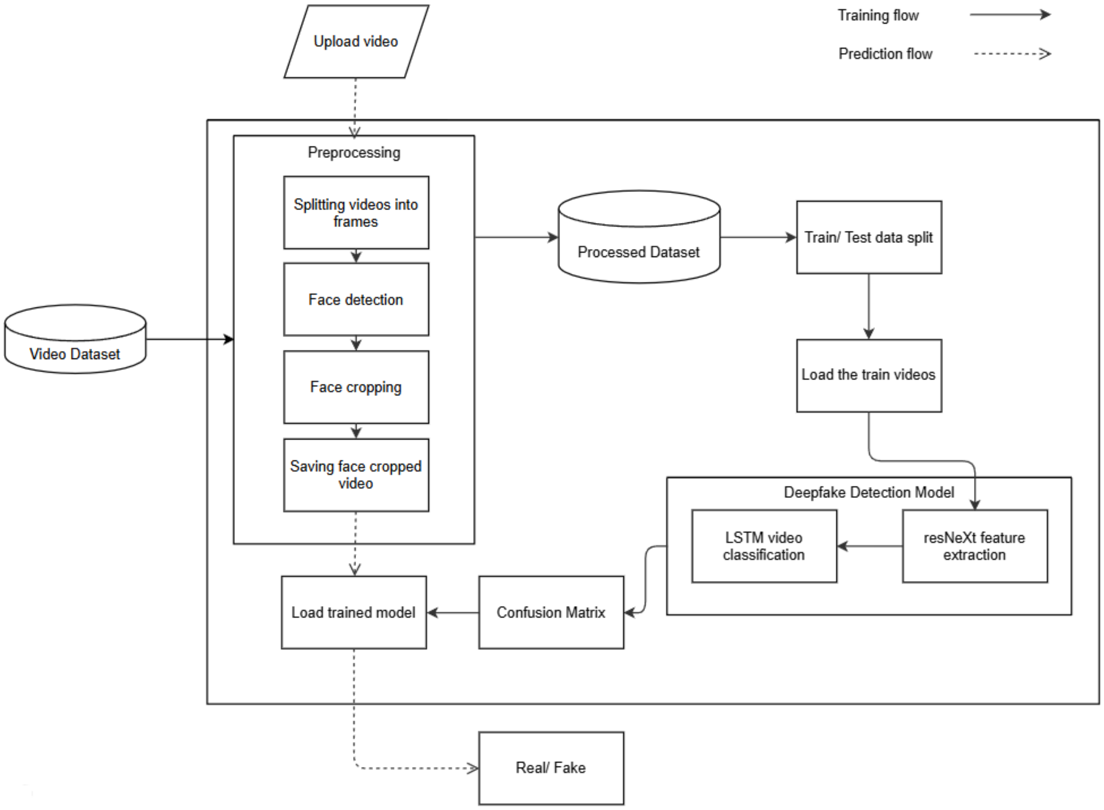


Figure 3.1: System Architecture

In this proposed model, to pull out characteristics at the level of frames of a video, the ResNeXt convolutional neural network is used. Extracting temporal relationships amongst frames, necessary for pointing out anomalies in the video, is made possible by putting those features through a LSTM-based recurrent neural network. Recognition of authenticity or manipulation of a video is enabled through the recurrent neural network.

3.3 Functional Requirements

Functional requirements of the user, our only stakeholder, are listed below:

3.3.1 User shall be able to upload videos up to a prescribed size and format

3.3.2 System shall be able to extract high level features from the uploaded video using ResNeXt CNN

3.3.3 System shall perform a dynamic temporal analysis on videos using LSTM RNN to capture dependencies and inconsistencies

3.3.4 System shall facilitate with the classification of videos as being authentic or forged subject to the features and patterns extracted

3.3.5 System shall return, to the user, the classification along with its confidence level and accuracy

3.4 Non-Functional Requirements

Non-functional requirements are listed below:

3.4.1 System shall ensure accurate identification of DeepFake content to achieve accuracy while minimizing false positives and false negatives

3.4.2 System shall ensure to withstand hostile attacks and changes in the input data while continuing to function consistently

3.4.3 System shall ensure efficient computation to provide real-time detection while minimizing processing time and resource usage

3.4.4 System shall bear an intuitive user interface to improve usability and accessibility

3.5 Hardware and Software Requirements

DeepTruth has the following hardware as well as software requirements:

3.5.1 Hardware Requirements

CPU: A modern multi-core CPU (Intel Core i7)

GPU: CUDA enabled GPU (NVIDIA GeForce GTX/ RTX)

RAM: 16GB or above

Storage: 128GB SSD

3.5.2 Software Requirements

Operating System: Windows 10

Programming Language: Python 3.11

Framework: PyTorch 2.3.0, Django 5.0.6

Libraries: OpenCV, Face-recognition

Client-side browser: Chrome, Firefox, or Edge

3.6 Threat Scenarios

Threat scenarios related to video upload systems, are possible weaknesses or attacks that might jeopardize the integrity, security, or operation of the system. The main threats are, unsanctioned entry, integrity of data and its tampering. We have listed important threat scenarios in the following text:

3.6.1 Malicious File Upload

An attacker can attempt to upload a tampered or invalid file to obstruct routine functions. This is possible due to the exploitation of system deficiencies such as buffer overflow. System crashes or unpredictable behavior may be the resultant. Data corruption, which puts the integrity of data at risk, may be the outcome of such attacks.

3.6.2 File size in Excess of 100 MB

A threat actor might take edge of the system, by loading overly large files that have dense resolutions or exceptionally high rate of frames. The results of such an attack may include DoS, rapid decline in system capabilities, or depletion of resources. Eventually, the user experience will also be adversely affected bringing a bad name to the said product.

3.6.3 No Face Detection

The detection system may approve fake content, if a malicious user uploads a video without a detectable face. Post classification of such content by the system, the process integrity will be severely affected.

3.7 Threat Modeling Techniques

Threat modelling is the act of identifying and considering potential threats that might impact a system's security, learning an attacker's perspective and incorporating safeguards. We have mapped the risks using the STRIDE model and it is depicted below as applied to our particular model viewpoint of an attacker, and implementing preventative measures is known as threat modelling. We have adopted the STRIDE model to detect and lessen risks for our model which is displayed in the Table 3.1: STRIDE Model below:

Table 3.1: STRIDE Model

STRIDE | Threats	Malicious File Upload	File Size in Excess of 100 MB	No Face Detection
Spoofing			√
Tampering	√		√
Repudiation	√		
Info disclosure			
Denial of service	√	√	
Elevation of privilege	√		

In each case, the analysis carried out according to the STRIDE model reveals several threats together with their corresponding vulnerabilities. For example, in the Malicious File Upload situation there are Risks: Tampering, Repudiation, Denial of Service (DoS), and Elevation of Privileges, which can result in system or data corruption or an unlawful breach. Compared to this, the scenario when the file size is exceeding 100 MB is belong to DoS wherein possibilities of system crashes due to depletion of resources can be seen. Finally, the presented No Face Detection scenario looks at the risk of spoofing and tampering to claim that with no faces to recognize, illegal content can easily go unnoticed in an uploaded video. By all counts, this makes the STRIDE analysis a highly accurate approach to identify and minimize risks while ensuring that a system is designed and implemented to be both safer and more resilient.

3.8 Threat Resistance Model

In another model named Threat Resistance Model, an outlook to the measure that how the system would put up with or resist numerous kinds of security threats is defined with a possible aim to ensure the reliability of the system in case of attacks. Examples of activity groups within this paradigm are the preventive, detective, and corrective activities describes how the system would endure or resist different kinds of security threats. Layers of defense, such as preventive, detective, and corrective actions, are commonly included in this paradigm.

3.8.1 Preventive Measures

You need to validate the video so as to ensure that only genuine video files are accepted and any other files or non- video files are rejected. More advanced, control the size of the video file to prevent DoS attack that is attributed to huge file size. Further, fix face recognition to eliminate or promote videos with faces and avoid those that do not contain any faces.

3.8.2 Detective Measures

It can be seen that patterns marked by heuristic approaches and deep learning models include altered videos and corresponding fraud. Moreover, all file upload operations must be recorded for purposes of detecting any shifts in behavior and having accountability in the event of an attack.

3.8.3 Corrective Measures

When there is such case of malicious or incorrect content, display the appropriate error messages including – Only video files allowed and face not detected. Also, need to constantly observe the uploaded files to control all the illicit actions.

3.9 Chapter Summary

The specifications, architecture, and security protocols for DeepTruth are outlined in this chapter. It explains the system architecture, which makes use of LSTM to identify temporal discrepancies across video frames and ResNeXt CNN for feature extraction. User video uploading and classification are examples of functional needs; accuracy, security, and usability are examples of non-functional objectives. Essential software and hardware requirements are noted, including a powerful CPU, GPU, and PyTorch. The STRIDE model is used to analyze potential security issues, such as malicious file uploads and large files. To improve system resilience and preserve data integrity, preventive, detective, and corrective actions are suggested.

Chapter 4:
Proposed Solution

Chapter 4: Proposed Solution

4.1 Introduction

Rapid advancement in deep learning technologies has revolutionized DeepFakes. These videos pose significant challenges, particularly in the realms of security, privacy, and misinformation. In order to make certain how authentic the video content is and to alleviate its negative effects on people and the society, detection of DeepFakes is of utmost importance.

Our project incorporates the strengths of ResNeXt CNNs and LSTM based RNNs, to counter these challenges. Capturing frame level characteristics utilizing ResNeXt convolutional neural networks, is where our model starts with. Extracting temporal relationships amongst frames, necessary for pointing out anomalies in the video, is made possible by putting those characteristics through a LSTM-based recurrent neural networks. An extensive dataset was created by amalgamating numerous renowned datasets such as Celeb-DF, DFDC, and Face Forensics++.

A summary of every step included in our project including but not limited to gathering data, preprocessing it, training the model, and classifying the input, is covered in the following sections. Advancement of DeepFake detection methods and value addition to digital forensics are sole targets of this approach.

4.2 Proposed Model

Our proposed DeepFake video detection solution shall use a combination of advanced deep learning techniques, namely Long Short-Term Memory RNN and a pre-trained ResNeXt CNN. ResNeXt is Residual CNN network, used for feature extraction at the frame level. LSTM on the other hand is used to sequentially process the frames so that the temporal analysis of the video is made possible. With this combination, we shall be able to extract important information specific to each frame of a video while also successfully analyzing temporal patterns in the frames. Our model shall be thoroughly trained on a variety of datasets, including DeeperForensics-1.0, DFDC, Celeb-DF, and Face Forensics++, so that it can handle real-time problems efficiently.

4.3 Data Collection

An extensive dataset was created by amalgamating numerous renowned datasets such as Celeb-DF, DFDC, and Face Forensics++. An equal number of authentic and fake videos are used to ensure non-partisan training. During the development phase a new dataset was created consisting only of face cropped footages by preprocessing an existing dataset.

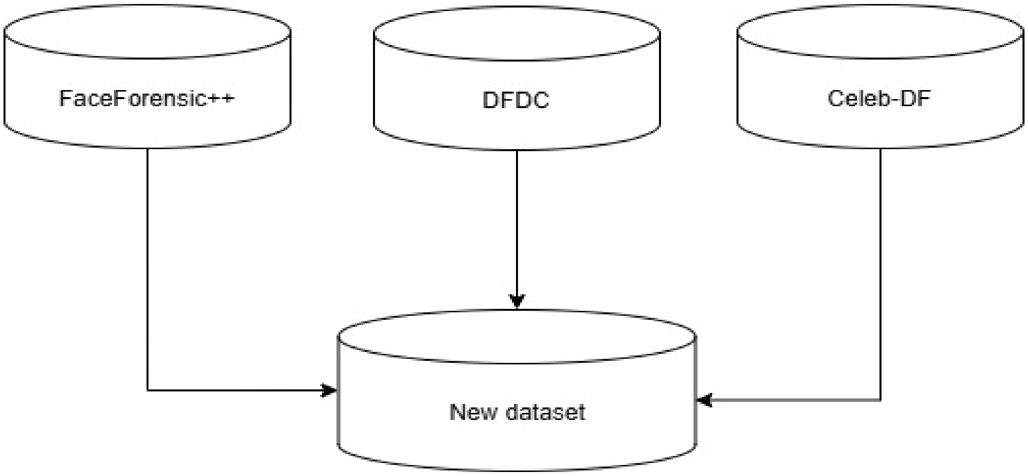


Figure 4.1: Data Collection

4.4 Data Pre-Processing

In order to exclusively focus on the detected faces, the phase of preprocessing includes splitting the videos into frames, recognizing faces in those frames, and then accurately cropping them. The mean of frames per footage is calculated to sustain uniformity in the dataset. Excluding frames that lacked a face, a new dataset was cultivated which contained frames of cropped faces based on the mean frame count to maintain uniformity.

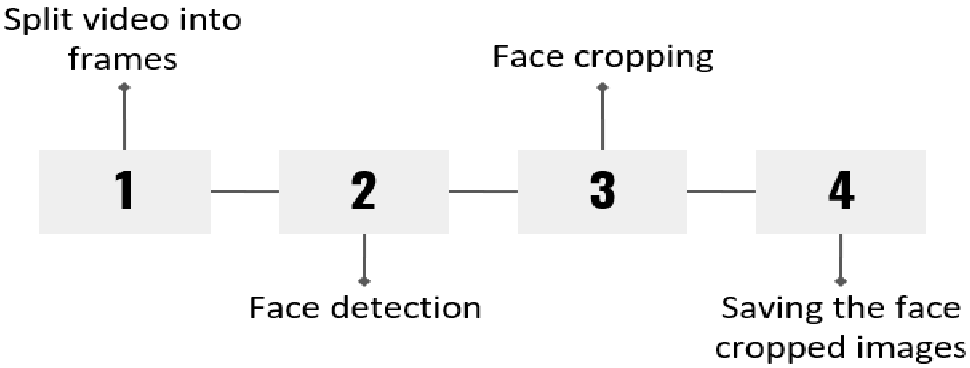


Figure 4.2: Data Pre-Processing

4.5 Tools and Techniques

Our DeepFake detection system leverages a powerful set of tools and frameworks, including but not limited to the following:

4.5.1 PyTorch Framework

This is an open source framework developed by Facebook artificial intelligence research and development for efficiency by providing user friendly configuration and procedural computing for research as well as manufacturing. It is suitable for developing such models since deep learning models can be created for different purposes and due to the increased use of deep learning networking. Moreover, PyTorch can be utilized in GPU acceleration especially when passing large data sets or complex structures such as ResNeXt and LSTM. Actually, owing to the compatibility with CUDA, PyTorch significantly improves the GPU acceleration for such models especially those including big datasets or complex structures of ResNeXt and LSTM.

4.5.2 ResNeXt CNN

ResNeXt is a refined version of ResNet which is an improved CNN model. Now, owing to a “cardinality” dimension, which has been added to the convolutional layers, and which groups them, performances are enhanced, but without leading to an exponential increase in the number of calculations to be made. ResNeXt is used to pre-extract features from each frame of the video where every frame’s primary features contain spatial data that can be used to identify DeepFake modifications. This makes it suitable for realizing fine detailed features from pictures or frames especially in video analysis and processing.

4.5.3 LSTM RNN

For example, Recurrent Neural Network (RNN), long short-term memory (LSTM) that enables for effective streaming data analysis and recognition of temporal dependency within long sequence of values. In light of the fact that using a time sensitive approach it can detect temporal inconsistency involved in the data, LSTMs are well placed to frame sequence analysis within DeepFake detection. It allows funding the LSTM to find small changes which may mean a change of content if comparing them across a sequence of frames.

4.5.4 OpenCV

It is an open source computer vision library that in developed for the purpose of undertaking computer vision operations in real time. This has made it suitable to be applied in motion tracking, object detection and face recognition of the sort of applications dealing with imagery and videos. With OpenCV the foundation preprocessing layers were established with these functions to support the face detection and frames from a video stream.

4.5.5 Django

Django is a high-level web framework developed in Python which strives to promote sensible and efficient construction of Web applications. It gives them a chance for using an easily understandable and conceptually simple online frontend of the DeepFake detection model, which accepts movies and returns feedback regarding their authenticity. Django is efficient in the backend operation such as handling of user requests and video uploads, response handling among others. Also, the framework is elastic and secure as shown below.

4.5.6 MTCNN

Multi-Task Cascaded Convolutional Network then makes use of machine learning to detect faces in the imported pictures or frames from a video, as the latter was introduced in the PyTorch framework. It is most commonly utilized in and also capable of performing in face recognition tasks as well as face alignment tasks. This methodology is efficient and can be applied with big data sets and tract real time applications.

4.6 Evaluation Metrics

For evaluating the effectiveness of the model the confusion matrix and all the related parameters are very important, especially when the task is carrying a high level of sensitivity as in case with cybersecurity. Each metric is explained below along with how it relates to our project:

4.6.1 Confusion Matrix

A confusion matrix is used to measure the performance of a model to classify, in order to compare the predicted labels to the actual ones. Thus, False Positives (FP), when real videos are classified as DeepFakes; True Positives (TP), where the model correctly identifies positives; True Negatives (TN), where it correctly identifies negative samples;

and False Negatives (FN), when DeepFakes are not detected. This matrix is effective for evaluating the resilience and suggestion of improvement because it gives an overview about the performance of the model.

4.6.2 Accuracy

In this type, accuracy is the number of true positive and additional true negative cases in relation to the total number of predicted cases which include true positives, false positives, true negatives and false negatives cases. Because it indicates the proportion of the correct predictions made by the model, it also measures its general effectiveness. A high level of accuracy when detecting DeepFake, also suggests that here we are dealing with a reliable model capable of detecting DeepFakes with very little probability of errors, making it possible for it to be used practically.

4.6.3 Precision

Precision is the ratio of TP to all the instances, which were predicted as positive hence is given by TP / (TP + FP). It measures the capacity of the model in putting a tag on specific true positives while minimizing on the false positives. Prime example, high precision is rather essential in the areas such as cybersecurity, because it allows ensuring that real content, in this case, objects being an actual DeepFake is rarely ranked as such. This not only improves the credibility of the model but also makes it less complicated for use since needless fake alarms are reduced.

4.6.4 Recall

The generalization of TP to all actual positives is referred to as recall, sensitivity and can be defined as TP/(TP + FN). They show how accurately the model can predict positive cases. This leads to high recall being required in DeepFake identification in order to ensure that the model capture many of the real DeepFakes and with lower risk of ignoring possible threats. In this case, a large value highlights the applicability of the model to security-sensitive applications because, with its help, it became possible to distinguish DeepFake materials.

4.6.5 F1 Score

The measure which is achieved by harmonic mean of both precision and recall is called F1 score. It is calculated as $F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$. It balances both measures, it is particularly helpful in such cases where the class distribution is not equal. In regard to DeepFake detection, a higher F1 score means that the model can get a good balance between distinguishing the modified information correctly and avoiding as many both false positives and false negatives as possible while also identifying DeepFakes.

4.6.6 Specificity

Specificity or the true negative rate is equal to the percentage of TN among all actual negatives, which is ‘TN + FP’. It evaluates how good the model is at identifying negatives and in return minimizes the number of false positives. Further, High specificity within DeepFake identification means that actual data is seldom flagged as DeepFake enhancing user satisfaction and reliability by reducing false positives and unnecessary content reporting.

4.7 Chapter Summary

In order to detect manipulated videos, this chapter introduces our DeepFake detection approach, which combines two cutting-edge deep learning models: For temporal analysis, LSTM RNN is adopted, while for frame-level feature extraction, ResNeXt CNN is used. To ensure fairness and reliability of training all necessary data collecting and preprocessing procedures are considered in the chapter including creating a unified dataset from all sources. There are some fundamental tools and techniques as well including the following; model training, video processing, and interface development are enabled by tools like PyTorch, OpenCV, Django, etc. The design principles of the evaluation criteria stress high detection accuracy and dependability in security-concerned applications.

Chapter 5:
Implementation and Testing

Chapter 5: Implementation and Testing

5.1 Introduction

This chapter also describe the introduction of DeepTruth, environment setting, functions, and main algorithms. Moreover, it also goes through the integration of its components and how it uses various protocols for communication as well as the kind of logs it can have. This is accompanied by different DeepTruth test scenarios and their outcomes and a brief overview about those.

5.2 Security Properties Testing

Since the project has not gone live, this has been limited, and security testing for the deployed environment have not been done. Yet, to prevent DeepFake detection models from becoming misused, the system should handle sensitive media content securely. Two examples of the security measures are the encryption of the data used in communication and the storage processes, together with limiting access to the uploaded videos.

5.3 System Setup

This part also outlines the ways of preparing the environment to detect deepfakes, which creates and implements deep learning models.

5.3.1 Environment Configuration

The deepfake detection system is established using Python 3.0, and PyTorch is used for model deployment and training. The backend is developed using Django web framework to design an efficient and user friendly interface to upload videos to the system for classification.

5.3.2 Key Functions and Algorithms

The main activities and mechanisms utilized in DeepTruth application are ResNeXt CNN for the extraction of the frame-level features. Temporal sequences are analyzed then using LSTM RNN to identify deepfakes. The input videos are first preprocessed in a way that the frames are extracted from videos, face is detected from the frames, and such frames containing face is cropped to a uniform resolution so that the data fed to the network is standardized.

5.4 System Integration

This section describes how elements of the DeepFake detection system connect to permit efficient operation from the video input to its classification. The system co-ordinates aspects of database management, standard computer communication protocols and neural network probabilistic models for safe and accurate authentic processing of uploaded content.

5.4.1 Components Integration

During feature extraction and sequential processing, this detection system comprises elements such as LSTM and ResNeXt, amongst others. Following that, feed the videos into ResNeXt; then, LSTM will receive features extracted from the videos to classify. elements like LSTM and ResNeXt. After videos are fed into ResNeXt, LSTM is given the extracted features to classify.

5.4.2 Communication Protocols/APIs

Django has provided setup for the REST API framework allowing for the system's front and backend communication. The API also handles model inference, video submission and result retrieval.

5.4.3 Database/Log Management Setup

To monitor usage or look for areas to improve, user interaction, model analytics, and system data are stored in an encrypted database.

5.5 Test Cases

The full functionality is covered by the test cases which are in Table 5.1: Test Cases; identification with no errors, real time processing and uploading of videos. To incline the technology to be more temperate, cases like high file sizes or unauthorized formats have been tested.

Table 5.1: Test Cases

Sr. no.	Case Description	Expected outcome	Actual outcome	Status
1	Upload video of enlisted formats	Real/Fake	Real/Fake	Pass
2	Upload corrupt or invalid video files	Only video files allowed	Only video files allowed	Pass
3	Upload video size above 100 MB	Maximum file size 100 MB	Maximum file size 100 MB	Pass
4	Upload video with no faces	No faces detected.	No faces detected.	Pass
5	Press upload without video selection	Please select a file	Please select a file	Pass
6	Upload video with insufficient frames	Video has less than required frames	Video has less than required frames	Pass

The cover page of DeepTruth in Figure 5.1: Homepage has a simple but appealing design focused at user friendliness.

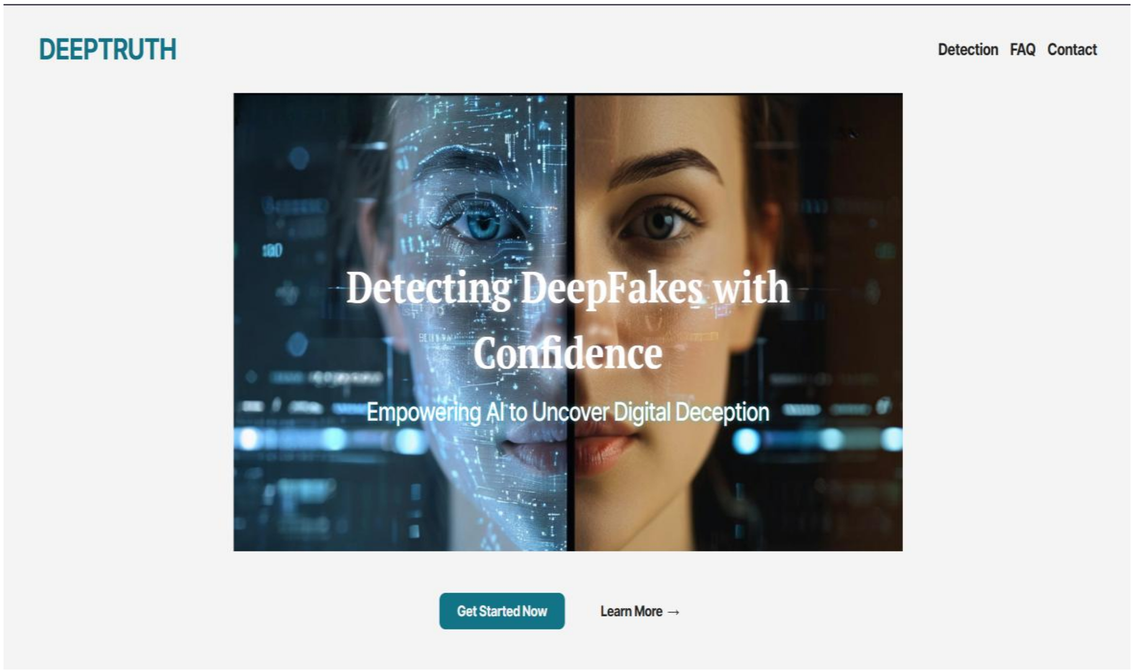
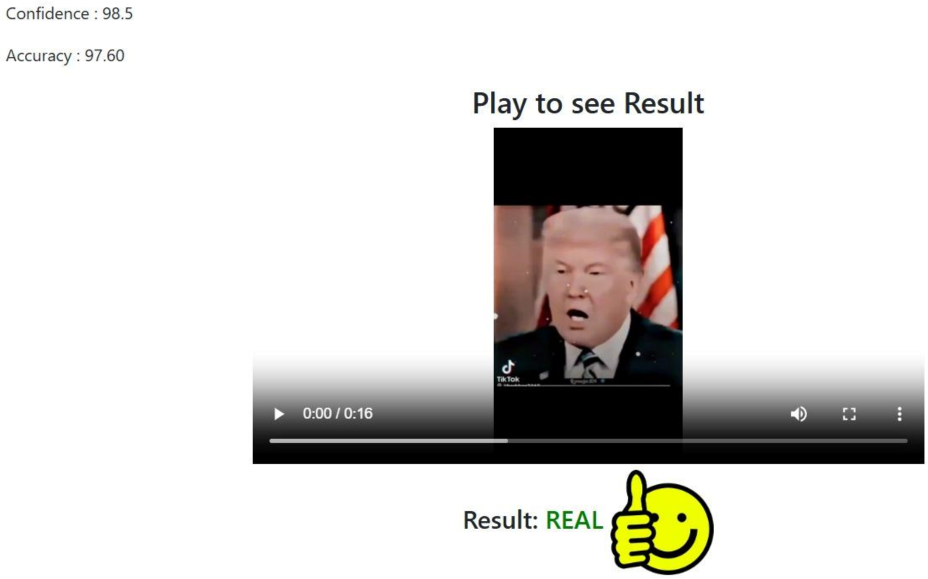

DEEPTRUTH
Detection FAQ Contact
Detecting DeepFakes with Confidence
Empowering AI to Uncover Digital Deception
Get Started Now
Learn More →

Figure 5.1: Homepage

Once a user clicks the ‘Detection’ tab as shown in Figure 5.2: Test Case 1, a file upload option appears with a slider for switching between 40, 60, and 100 frames.



Confidence : 98.5

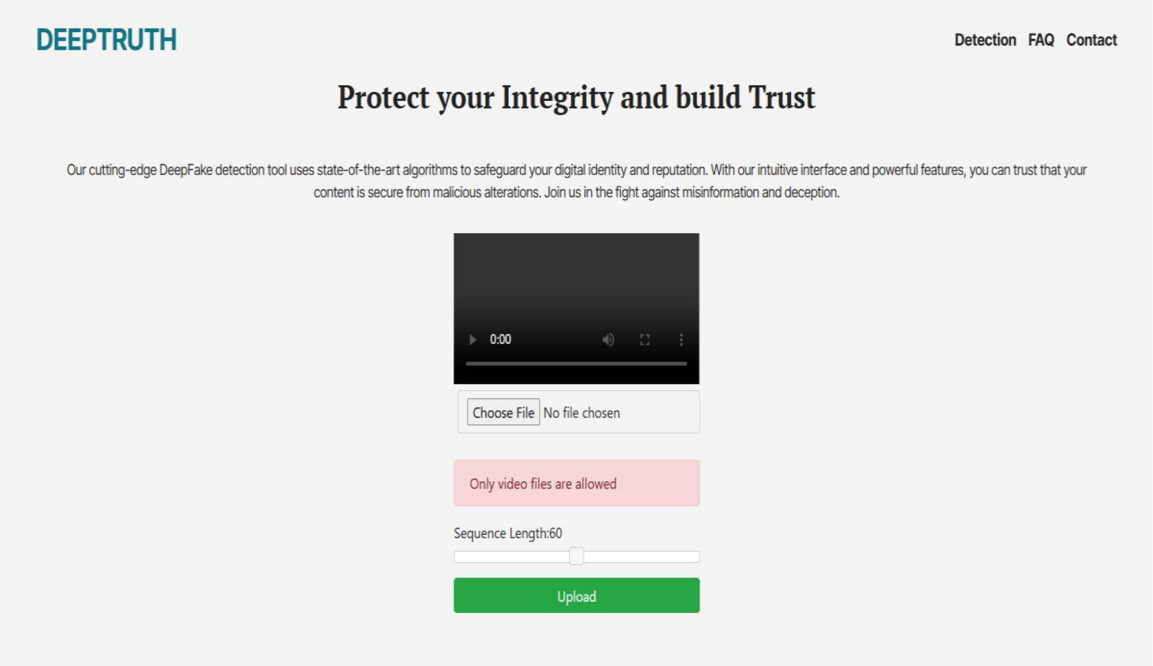
Accuracy : 97.60

Play to see Result

Result: REAL

Figure 5.2: Test Case 1

As depicted in Figure 5.3: Test Case 2, once user attempts to upload a non-video file an error message is notified instantly.



DEEPTRUTH

Protect your Integrity and build Trust

Our cutting-edge DeepFake detection tool uses state-of-the-art algorithms to safeguard your digital identity and reputation. With our intuitive interface and powerful features, you can trust that your content is secure from malicious alterations. Join us in the fight against misinformation and deception.

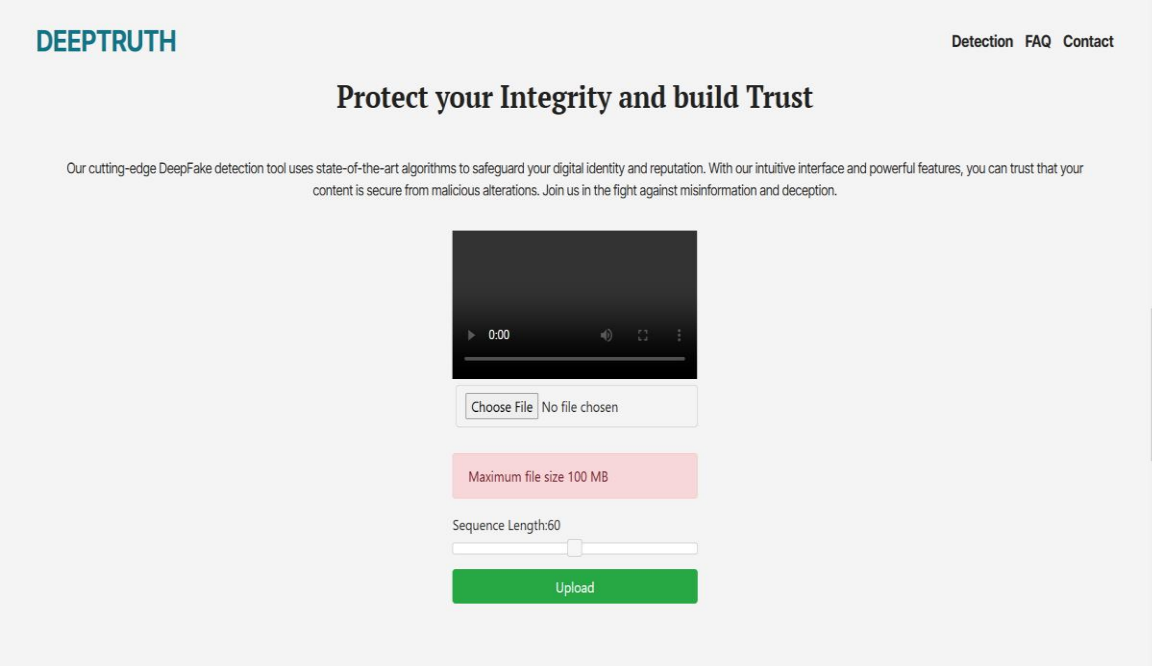
Choose File No file chosen

Only video files are allowed

Sequence Length:60

Figure 5.3: Test Case 2

As shown in Figure 5.4: Test Case 3, once a user attempts to upload a video file larger than 100 MB an error message is notified instantly.



Protect your Integrity and build Trust

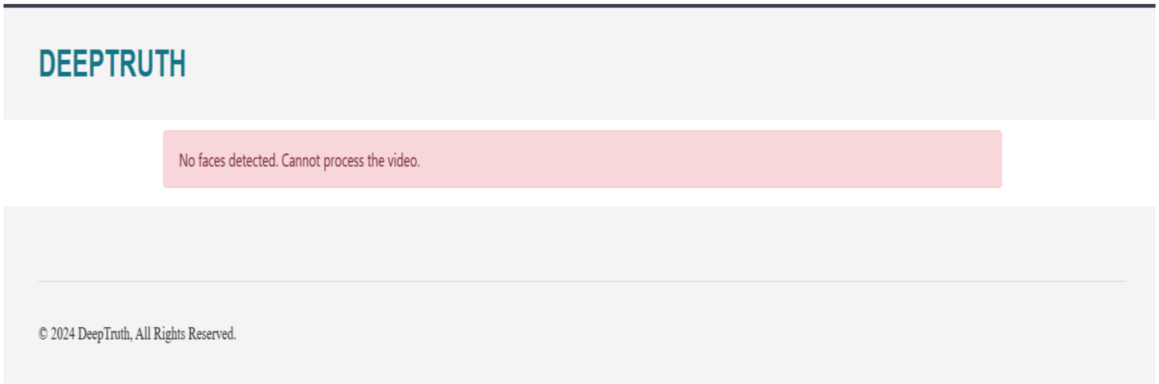
Our cutting-edge DeepFake detection tool uses state-of-the-art algorithms to safeguard your digital identity and reputation. With our intuitive interface and powerful features, you can trust that your content is secure from malicious alterations. Join us in the fight against misinformation and deception.

Maximum file size 100 MB

Sequence Length:60

Figure 5.4: Test Case 3

As shown in Figure 5.5: Test Case 4, if a user uploads a video which has no faces to detect than a error message is notified right away.

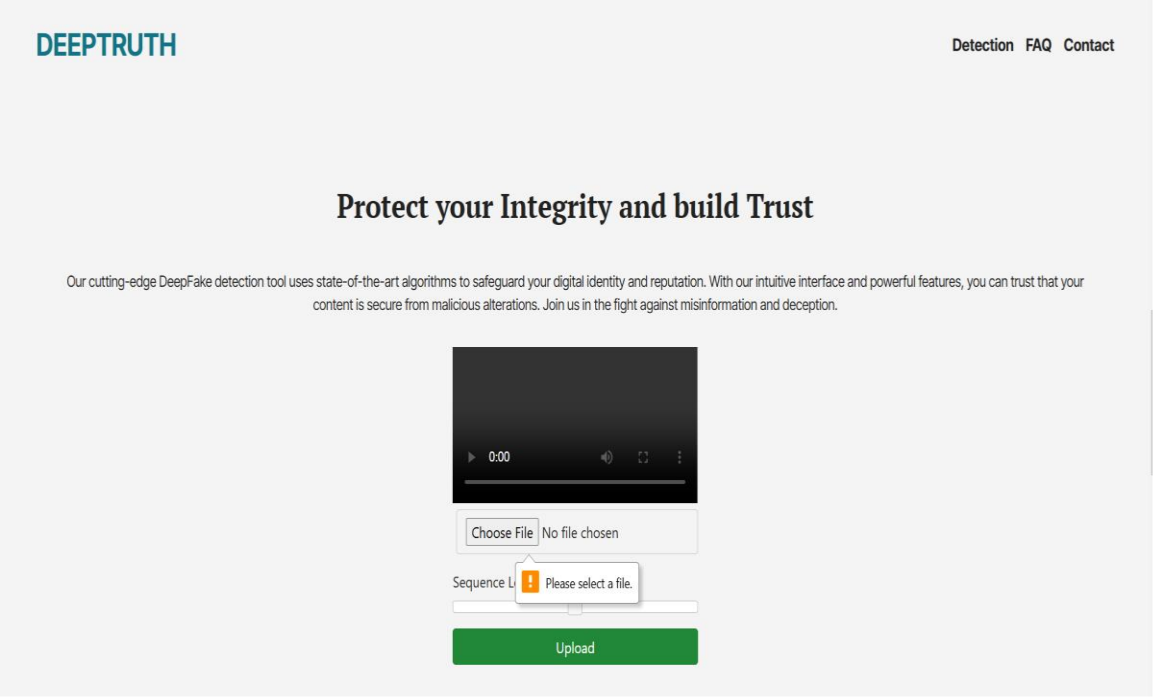


No faces detected. Cannot process the video.

© 2024 DeepTruth, All Rights Reserved.

Figure 5.5: Test Case 4

If a user presses the upload button without selecting a file, an error notification will appear as visible in Figure 5.6: Test Case 5.



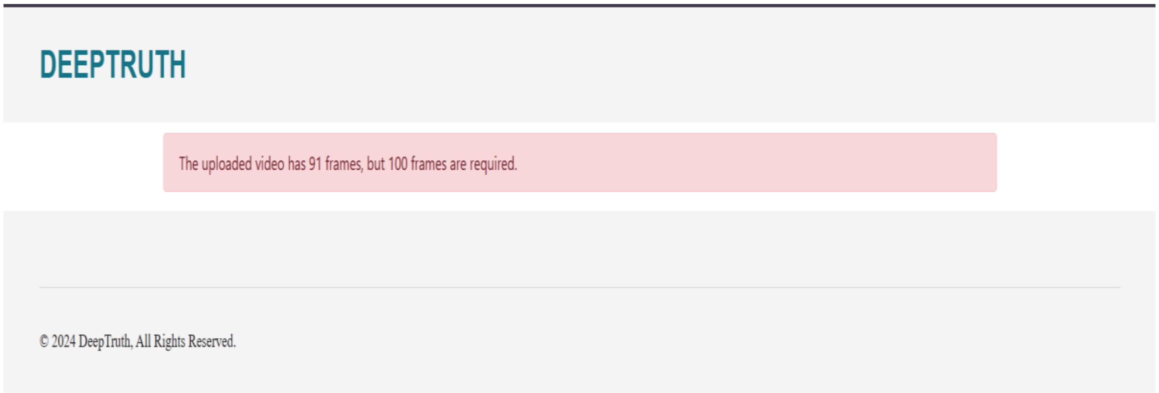
Protect your Integrity and build Trust

Our cutting-edge DeepFake detection tool uses state-of-the-art algorithms to safeguard your digital identity and reputation. With our intuitive interface and powerful features, you can trust that your content is secure from malicious alterations. Join us in the fight against misinformation and deception.

Please select a file.

Figure 5.6: Test Case 5

If a user uploads a video which does not have sufficient frames available, an error notification will appear as visible in Figure 5.7: Test Case 6.



The uploaded video has 91 frames, but 100 frames are required.

© 2024 DeepTruth, All Rights Reserved.

Figure 5.7: Test Case 6

5.6 Results and Discussion

Deepfakes videos were detected with the help of efficiency, due to the characteristics linked to the model described. It is also important to note that the presented model had low error on several datasets such as FaceForensics++, Celeb-DF and DFDC.

When the model is trained we have a chance of getting this confusion matrix as it was demonstrated above. Finally and lastly, the performance of the model is also demonstrated graphically using a confusion matrix. This is done to measure the effectiveness of the proposed Deep Learning model in detecting DeepFake videos as wells as to identify the cases of false positive thus giving way to further enhancing of the Deep Learning model if the need arises.

Our Model 1, based on 40 frames, returned Figure 5.7: Model 1 Confusion Matrix (40 frames) post model training.

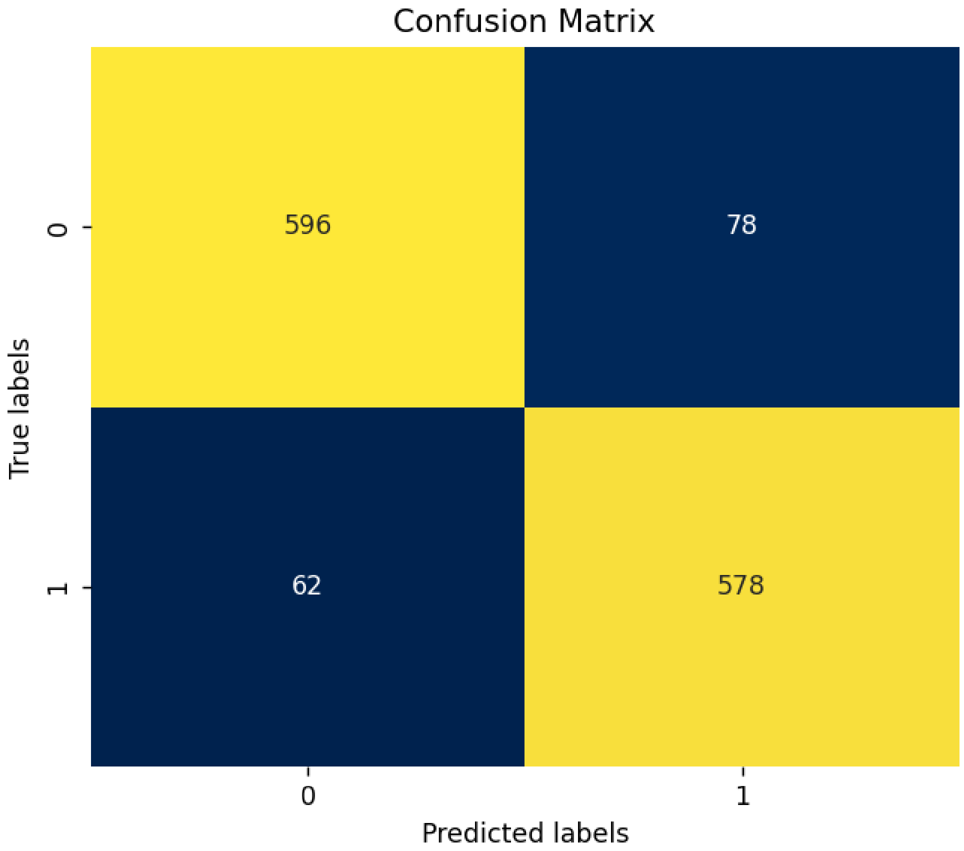


Figure 5.7: Model 1 Confusion Matrix (40 frames)

Model 1 results above assisted with calculating the evaluation metrics given in Table 5.2: Model 1 Evaluation Metrics (40 frames).

Table 5.2: Model 1 Evaluation Metrics (40 frames)

Evaluation Metrics	Value
Accuracy	89.35%
Precision	0.88
Recall	0.90
F1 Score	0.89
Specificity	0.88

Our Model 2, based on 60 frames, returned Figure 5.8: Model 2 Confusion Matrix (60 frames) post model training.

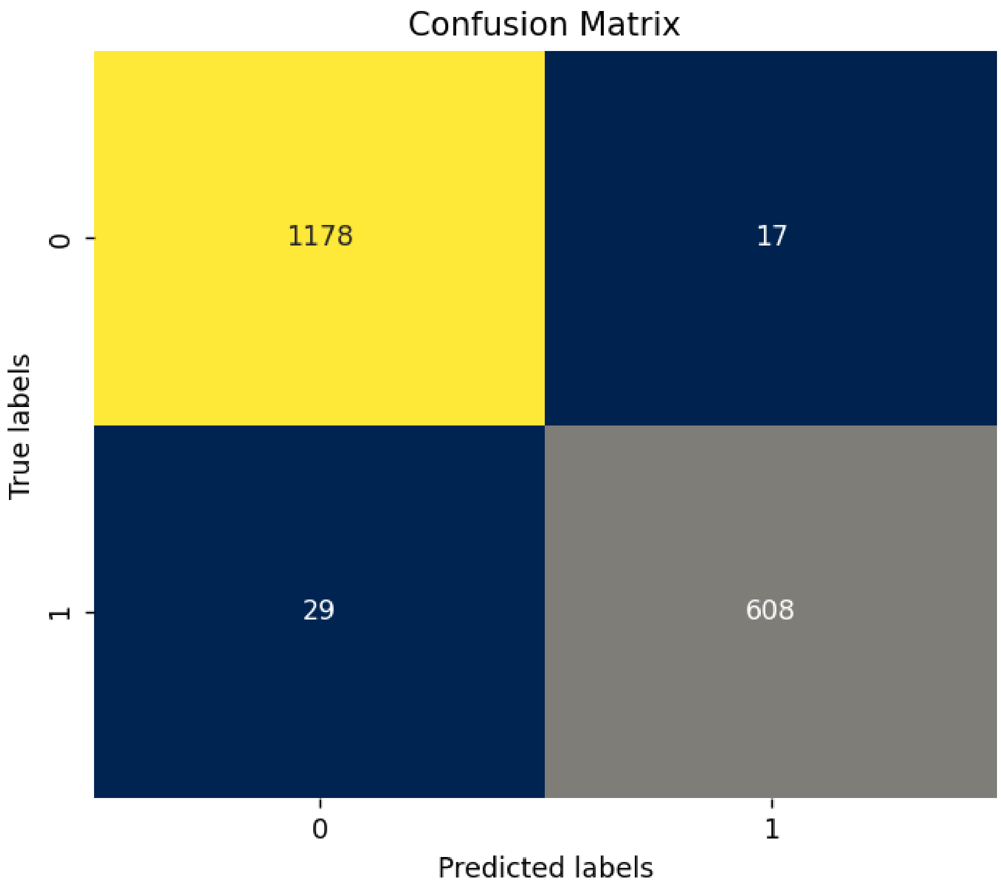


Figure 5.8: Model 2 Confusion Matrix (60 frames)

Model 2 results above assisted with calculating the evaluation metrics given in Table 5.3: Model 2 Evaluation Metrics (60 frames).

Table 5.3: Model 2 Evaluation Metrics (60 frames)

Evaluation Metrics	Value
Accuracy	97.49%
Precision	0.97
Recall	0.95
F1 Score	0.96
Specificity	0.99

Our Model 3, based on 100 frames, returned Figure 5.9: Model 3 Confusion Matrix (100 frames) post model training.

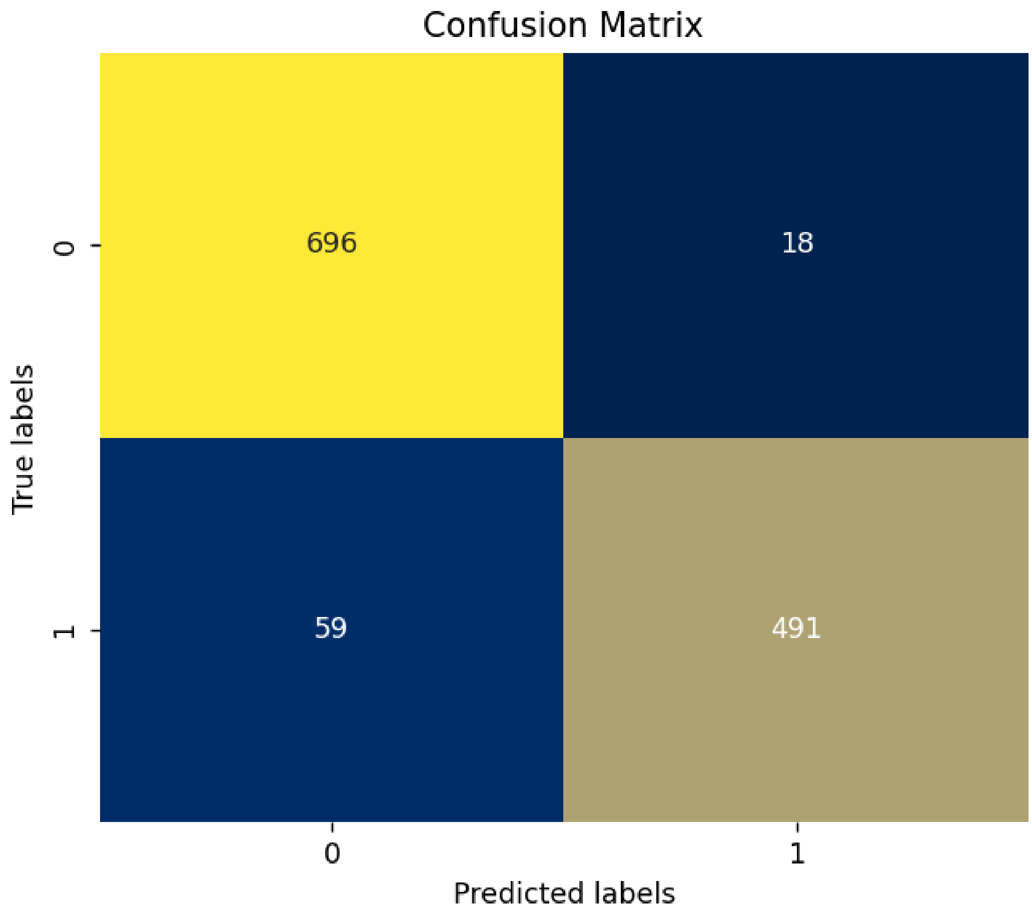


Figure 5.9: Model 3 Confusion Matrix (100 frames)

Model 3 results above assisted with calculating the evaluation metrics given in Table 5.4: Model 3 Evaluation Metrics (100 frames)Table 5.4: Model 3 Evaluation Metrics (100 frames).

Table 5.4: Model 3 Evaluation Metrics (100 frames)

Evaluation Metrics	Value
Accuracy	93.91%
Precision	0.96
Recall	0.89
F1 Score	0.93
Specificity	0.97

In arrangement three with 40, 60 and 100 frames used in the assessment of DeepTruth, performance measures increase gradually indicating that the model avails itself well in temporal data as more frames are evaluated. The performance of the model constantly rises with the best results achieved in the 60-frame capturing system of 97.49%. The downsides in terms of false negatives and false positives are controlled for through get large-valued recall and precision across the all arrangement and F1 score in the higher frame models hovering just slightly below 0.96. The efficiency of the model to distinguish between real and fake content is proved by a high value of specificity, which was up to 0.99, it brings down the probability of misclassification of the real videos into DeepFakes.

It also shows an ability to generalize to many different kinds of data, which is important for application use since the accuracy and consistency gradually increase across datasets. These results add credibility to the model and suggest that, although at some degree of computational costs, higher frame count might further enhance the detection performance. In conclusion, this study reveals that the proposed model is useful for accomplishing efficient DeepFake classification especially for high risk-high return scenarios such as checking media and cybersecurity.

5.7 Best Practices/ Coding Standards

To guarantee the code consistency, readability and reliability this section provides the information about the best coding practices and tabular conventions that is been used during the implementation of this project. Writing code, checking its correctness and remembering on some python standards, as well as versioning is good for now and the future.

5.7.1 Code Validation

The coding for DeepTruth adheres to the Python coding standards right from the choice of variable names for different modules to their structure.

5.7.2 Development Practices and Standards

For future changes, version control was adopted for managing change and the structural framework and content of models.

5.8 Chapter Summary

In this chapter, some main steps for the realization and the evaluation of DeepTruth are described. Some of the initial security measures are explained so as to let the system run with the desired level of confidentiality, and only for the required media. The chapter also discusses how to apt the system beginning with the environment to advanced algorithms like LSTM and ResNeXt. Setting up of the Database, defining the communication protocol and components integration to enable its operations. In extensive test cases, for instance, size and format of videos are tested and confirms that it has high accuracy across datasets. The feasibility of the model with high effectiveness is shown in specific markers and the confusion matrix: the F1 score is 98%, the specificity also is 98%. Lastly, a brief on coding standards and best practices is made in view of providing for current as well as future development efficient and effective code executions.

Chapter 6:
Conclusion and Future Work

Chapter 6: Conclusion and Future Work

6.1 Introduction

This final chapter summarizes the contributions of the project to the relevant field and assesses the methods and results employed in the study. It examines the strengths and weaknesses of the method in context with the project’s impact on the lineup of DeepFake detection. This part also involves the evaluation of the model’s efficiency with regard to such aspects as strong and weak indications. The chapter also identifies the further research directions The chapter also gives the potential improvement of the detection approach along cardinal lines which may help to make it more robust, efficient and effective. It discusses how these improvements can benefit DeepFake detection techniques to keep evolving due to the ever-growing dynamic of synthetic media.

6.2 Achievements and Improvements

Furthermore, by comparing ResNeXt CNN and LSTM RNN a powerful DeepFake detection model was created, which was proved to be very accurate in a range of datasets as FaceForensics++, Celeb-DF, and DFDC. With all the functional and efficiency enhancing algorithms and using Django for an on-line interface, the system has created a dependable and easily navigated platform to keep DeepFakes identification in real time. The ability to model temporal data using the framework was demonstrated using frame counts and higher frame setting allowed for better accuracy and much fewer false positives. Such outcomes prove that the model has applicability for critical uses including cybersecurity and media. However, this has been enhanced by better model tuning and data preprocessing in the recent past as from the case of iterative optimization.

6.3 Critical Review

However, it is essential to acknowledge a couple of points, to my mind, regarding the model’s high effectiveness in detecting DeepFakes. The evaluation of high frame counts is a resource-consuming function for definite tasks which is one significant challenge that can heavily affect computational profiling. analyzing high frame counts is one significant obstacle that can significantly burden computational power. This problem highlights the importance of optimization approaches to increasing the rate of work without decreasing the quality, for example, increasing the computational capacity or reducing the model’s size. Furthermore, to improve security in the development process, little live security testing was conducted in the real world and this means there may be areas that may be exploited when the system has been implemented. This creates an opportunity for additional research to work on the improvement of the model to counter the emerging security threats. Lastly, while the model achieved commendable detection efficiency, it remains imperfect in the best sense, and could be refined in the future, including its adaptation to new more intricate DeepFake techniques. The model's ability to adapt to new threats is likely to be maintained by its further extension and additional refinements in terms of the ability to handle qualitatively new and more diverse, and hostile, forms of synthetic media.

6.4 Future Recommendations

There are several critical domains, which should be optimized at all costs to enhance the efficiency of the model and its scalability in future experiments. The future work should therefore attempt to improve the feasibility of the model in handling big datasets possibly using existing technologies such as pruning, quantization, or other complex neural protocols. Another would be to increase the size of the training across more DeepFake techniques and real context along with the improvements. However the efficacy of the model when implemented in real-use operating conditions would be understood further by implementing it in a simulated live system for real-time evaluation of security threats. This would help in a timely chance and development to the threat resistant capacity of the model. The candidate, who will lead the project, will have to guarantee the proper user privacy and data security while the project will grow up in the live environment and exclude dangerous leaks with personal data without violating the legislation in the field of privacy. However, before implementing the model in practice to make utilization of modern technologies ethical, the following factors will be paramount.

6.5 Chapter Summary

In this chapter, an overall summary of the major milestones achieved in the course of the project is presented, an evaluation of the strategies used, and a final assessment of the overall contribution towards DeepFake detection is provided. The chapter also mentions benefits such as high accuracy of detection, and ability to perform real time detection and drawbacks such as its resource needs. This approach shows that optimization is important and that further experiments should be conducted in real-time environments to further enhance the model against the new changing threats. Several directions of work are proposed, including the enlargement of sets of training images, augmenting the quantity and thus the quality of available computational power, and using the proposed systems in real-life environments for security purposes. The above recommendations are aimed at improving and consolidating the capacity of a model to adapt with the fast evolving synthetic media space.

References

[1] 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

[2] https://www.kaggle.com/c/deepfake-detectionchallenge/data

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies,Matthias Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images” in arXiv:1901.08971.

[4] Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., & Alshehri, A. H. (2023). Deep fake detection and classification using error-level analysis and deep learning. Scientific Reports, 13(1), 7422.

[5] Kumar, M., & Sharma, H. K. (2023). A GAN-based model of deepfake detection in social media. Procedia Computer Science, 218, 2153-2162.

[6] Sun, F., Zhang, N., Xu, P., & Song, Z. (2021). Research Article Deepfake Detection Method Based on Cross-Domain Fusion.

[7] Gandhi, A., & Jain, S. (2020, July). Adversarial perturbations fool deepfake detectors. In 2020 International joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

[8] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5001-5010).