Politécnico
de Viseu

Escola Superior
de Tecnologia
e Gestão de Viseu

Detection of fake images generated by deep learning

Stéphane Mesquita Monteiro

Dissertação

Mestrado em Engenharia Informática - Sistemas de Informação

Trabalho efetuado sob a orientação de
Professora Doutora Ana Cristina Wanzeller Guedes de Lacerda
Professor Doutor Filipe Miguel Simões Caldeira

Fevereiro de 2024

Politécnico
de Viseu

Escola Superior
de Tecnologia
e Gestão de Viseu

Detection of fake images generated by deep learning

Stéphane Mesquita Monteiro

Dissertação

Mestrado em Engenharia Informática - Sistemas de Informação

Trabalho efetuado sob a orientação de

Professora Doutora Ana Cristina Wanzeller Guedes de Lacerda
Professor Doutor Filipe Miguel Simões Caldeira

Fevereiro 2024

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my main advisor, Professor Ana Cristina Wanzeller Guedes de Lacerda, whose unwavering support, invaluable guidance, and scholarly expertise have been instrumental throughout the entire journey of this master's thesis. Prof. Cristina commitment to excellence and passion for the subject matter significantly contributed to the success of this research.

I am also indebted to my co-advisor, Professor Filipe Miguel Simões Caldeira, whose insightful feedback and constructive criticism have played a pivotal role in shaping and refining the content of this thesis. Prof. Filipe's dedication to academic rigor and his willingness to share his knowledge have been indispensable in enhancing the quality of the work presented here.

I also want to express my gratitude to every one of the instructors and staff for creating a supportive learning atmosphere that has encouraged critical thinking and thought-provoking conversations. Their enthusiasm and support have been invaluable to the advancement of this research.

I appreciate the companionship and collaborative attitude among my colleagues and other scholars. It has been stimulating and enlightening to share thoughts and passion for the topic of deepfake detection.

Finally, I would want to express my gratitude to my family and friends for their consistent support, tolerance, and understanding during this academic journey. Their encouragement has been the main source of my tenacity and accomplishment.

ABSTRACT

During the last few years, the amount of audiovisual content produced is continually increasing with technology development. Along with this growth comes the availability of the same information through numerous devices that any individual holds, including smartphones, laptops, tablets, and smart TVs, in an entirely free and open manner. These type of content are considered an authenticity element since they represent a reality record. For example, in court, photos frequently determine the jury's course of action since what is available is a recorded picture that validates a narrative and usually does not leave room for doubts. However, with the advancement of Deep Learning (DL) algorithms, a new and dangerous trend known as Deepfakes begins to emerge. For example, a deepfake can be a video or an image of a person on which their face or body is totally or partially modified to appear to be someone else. This technique is often used for manipulation, blackmailing, and spreading false information.

After recognizing such a dangerous problem, this study aims to uncover patterns that deepfakes show to identify authenticity as accurately as possible, using machine learning and deep learning algorithms. To get the highest level of accuracy, these algorithms were trained on datasets that included both real and phony photos. The outcomes demonstrate that deepfakes can be accurately identified and that the optimal model may be selected based on the specific requirements of the application.

Keywords: Machine learning; Cyber security; Deepfakes

ABSTRACT (PORTUGUÊS)

Nos últimos anos, a quantidade de conteúdo audiovisual produzido tem vindo a aumentar continuamente com o desenvolvimento da tecnologia. Juntamente com este crescimento, surge a disponibilidade da mesma informação através de inúmeros dispositivos que qualquer indivíduo possui, incluindo telemóveis, computadores, tablets e smart TVs, de uma forma totalmente livre e aberta. Este tipo de conteúdo é considerado um elemento de autenticidade, uma vez que representa um registo da realidade. Por exemplo, em tribunal, as fotografias frequentemente determinam a linha de ação do júri, uma vez que o que está disponível é uma imagem registada que valida uma narrativa e geralmente não deixa espaço para dúvidas. No entanto, com o avanço dos algoritmos de Deep Learning (DL), começa a surgir uma nova e perigosa tendência conhecida como Deepfakes. Por exemplo, um deepfake pode ser um vídeo ou uma imagem de uma pessoa na qual o rosto ou o corpo é totalmente ou parcialmente modificado para parecer ser outra pessoa. Esta técnica é frequentemente utilizada para manipulação, chantagem e disseminação de informações falsas.

Após reconhecer um problema tão perigoso, este estudo tem como objetivo descobrir padrões que os Deepfakes apresentam para identificar a autenticidade da forma mais precisa possível, utilizando algoritmos de Machine Learning e Deep Learning. Estes algoritmos foram treinados utilizando conjuntos de dados que contenham tanto fotografias autênticas quanto falsas, a fim de obter o melhor nível de precisão. Os resultados obtidos mostram bons resultados na identificação de deepfakes e que a escolha do melhor modelo pode ser ajustada às necessidades da aplicação em causa.

Palavras-chave: Machine learning; Cyber security; Deepfake

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

AI - Artificial Intelligence

ANN – Artificial neural network

API - Application Programming Interface

CIA - Confidentiality, Integrity and Availability

CNN - Convolutional Neural Network

CRISP-DM - Cross Industry Standard Process for Data Mining

DFDC – DeepFake Detection Challenge

DL - Deep Learning

GAN - Generative Adversarial Network

LSTM - Long Short-Term Memory

ML - Machine Learning

RNN - Recurrent Neural Network

TTS – Text to speech

VGG - Visual Geometry Group

1. INTRODUCTION

This chapter discusses the motivations for carrying out this work and describes and defines the recognized problem. It also details the main objectives and the document's structure.

1.1. Motivation

Our society heavily relies on audiovisual content that can be used for various objectives, from reestablishing or sustaining the truth to harmful scenarios like blackmail and extortion. Additionally, as technology advances, the methods for disseminating these items are becoming more widespread and quicker, enabling an image to be shared worldwide in seconds. This has several benefits, including fast information exchange across companies and the ability to access news from anywhere on the planet. However, the quick spread of this content has many downsides, including the dissemination of erroneous information and photos that result in well-known fake news.

Audiovisual information has always been prone to human manipulation, although these modifications are often easily discovered. However, as Machine Learning (ML) and Deep Learning (DP) advance, this material may now be transformed using a variety of algorithms, fostering the rise of Deepfakes. These Deepfakes tend to have much credibility since they are carefully manipulated to create a false situation. However, because of the widespread of this changed information, this false scenario can now be "sold" as the real deal.

This study aims to investigate Deepfakes detection techniques using machine learning to explore faster response times while maintaining the integrity and accuracy of audiovisual content. Although there are already specific algorithms to carry out this detection, their response times could be improved, making it easier to use in real-world scenarios, such as detecting a deepfake before sending contents over a chat service.

1.2. Contextualization

Although Fake news is not a new phenomenon, it is becoming more prevalent. Approximately 59% of respondents believe they have encountered such false news, with most people saying that most fake news is more successful on democratic concerns (Khan

et al., 2021). These sorts of news can be classified in several ways, one of which is visual based. Edited audiovisual content is employed in this form of fake news, generally through Photoshop (Manzoor et al., 2019). This sort of fraud has grown significantly with the development of audiovisual deepfakes because artificial intelligence-assisted editing of images, videos, and audio enables the contents to be modified with high quality, making it challenging to detect fraud (Molina & Berenguel, 2022).

Although the detection of deepfakes is new, and thus there are no benchmark values for the study of algorithms and detection approaches, there has already been a significant amount of research effort in the scientific community, and so several models that conduct the same detection exist. However, using these approaches is difficult since it requires computer skills and the usage of programming languages such as Python.

1.3. Problem definition

Most systems for identifying fraudulent audiovisual material rely on extensive image analysis to detect anomalous patterns in the image, such as pixel irregularity and interruptions between frames. Deepfakes, on the other hand, are a little more challenging to analyze. They are made using deep learning, which allows for much more credible results, and because the algorithms learn over time, the produced results are also improved. To detect these fake items, the same technology that created them (ML and DL) must be used to achieve more actual and tangible findings.

As stated in Table 1, the author (Kocak & Alkan, 2022) refers to several methods of identifying Deepfakes for distinct datasets and compares standard detection approaches and deep learning-based techniques.

Table 1 - Comparision o between detection methods (Kocak & Alkan, 2022)

Traditional detection methods		**Deep Learning-Based methods**	
Method	**Accuracy (%)**	**Method**	**Accuracy (%)**
Head pose	89	AlexNet	98.73
Deepfake TIMIT	98.7	Improved Xception	99.86
Hearbeat rhythms	96	XceptionNet	100
Mouth motion	71	VDTNet	90.2
DFDC	64.1	EfficientNet	92

However, because deepfake detection specialists perform these studies rather than end users who require speed in real-world scenarios, the majority of studies only address the accuracy of identifying deepfakes and ignore the rapidity with which this detection is performed. A detection system that is much faster but less precise would be more useful in a real-time environment than a model that is very precise but much slower.

1.4. Objetives

Considering the problem definition and contextualization, this study aims to achieve the following contributions:

- Reduce the problem by figuring out how to categorize phony photos and/or movies.
- Assist in the identification of deepfakes by offering resources to carry out this task, with a focus on end users.

The following activities will be implemented in accordance with these goals:

- Study of detection algorithms and analysis of input parameters to analyze which of these parameters allow achieving better results in different detection algorithms.
- Identify relevant deepfake datasets and real-life scenarios.
- A prototype application for identifying deepfakes is proposed. The implementation of this application will be based on a performance analysis of detection algorithms. The primary function of this application is to provide a user-friendly interface for uploading images and determining their authenticity. The study of the time performance of detection algorithms will be a critical factor in developing this prototype.

1.5. Expected results

According to the objectives and actions identified in the previous subsection, the following results are expected:

- Identify the most used deepfake detection algorithms and with better performance.
- Acquire parameters that allow for the best possible results in the detection algorithms: Given the different machine learning algorithms, it is intended to

analyze their performance in terms of accuracy and response times. It is also expected to be able to use different parameters in the algorithms and identify which parameters allow for a better result.

- Prototype application: After analyzing the algorithms, it is expected to create a prototype application that, in a simplistic manner, allows any user to identify whether the image they have is a deepfake or not. Once the use of machine learning algorithms is usually only accessible through programming and command line, this prototype is intended to make it possible for anyone, with or without computer knowledge, to use a simple detection tool.

1.6. Work Plan

This subchapter presents a detailed work plan outlining the tasks and activities to be undertaken in this thesis. The work plan is structured into several phases, including literature review, data collection and preprocessing, algorithm development and enhancement, evaluation and performance analysis, real-life application, and case studies.

A Gantt diagram is provided in Figure 1 to visualize the timeline and dependencies of the tasks.

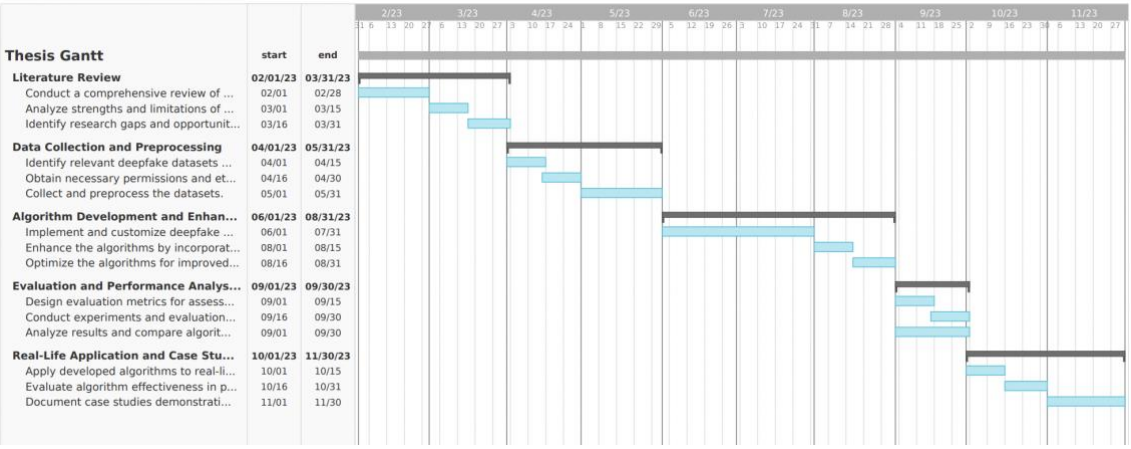

Figure 1 - Gantt Diagram

Below are the tasks and their corresponding subtasks that need to be completed:

- **Literature Review**
    - Conduct a comprehensive review of deepfake detection algorithms and real-life applications.
    - Analyze strengths and limitations of existing algorithms.
    - Identify research gaps and opportunities for improvement.
- **Data Collection and Preprocessing:**
    - Identify relevant deepfake datasets and real-life scenarios.
    - Obtain necessary permissions and ethics approvals.
    - Collect and preprocess the datasets.
- **Algorithm Development and Enhancement:**
    - Implement and customize deepfake detection algorithms (e.g., Xception, ResNet, VGG) for real-life scenarios.
    - Enhance the algorithms by incorporating domain-specific features.
    - Optimize the algorithms for improved accuracy and efficiency.
- **Evaluation and Performance Analysis:**
    - Design evaluation metrics for assessing algorithm performance in real-life scenarios.
    - Conduct experiments and evaluations using collected datasets.
    - Analyze results and compare algorithm performance.
- **Real-Life Application and Case Studies:**
    - Apply developed algorithms to real-life scenarios.
    - Evaluate algorithm effectiveness in practical applications.
    - Document case studies demonstrating successful deepfake detection.

1.7. Thesis structure

This study is divided into five chapters. The first provides the context, description of the problem, and study objectives. The second chapter dives into the study's primary subjects: cyber security, machine learning, deep learning, and deepfakes. This chapter also includes an examination of the existing works in the community. The technique and procedures employed are described in the third chapter. The fourth chapter explains all the labor done to get results. The fifth chapter contains an analysis and discussion of the collected results. The final chapter includes the closing notes and future work.

2. LITERATURE REVIEW

The Literature Review chapter covers the following research areas: Cyber Security, Machine Learning, and Deepfakes.

2.1. Cyber Security

In an era of extraordinary technical developments, the digital environment has grown into a dynamic ecosystem that involves every aspect of modern society. This widespread digitalization has given rise to a variety of emerging technologies, one of which is the increasingly sophisticated domain of deepfakes. Deepfakes, backed by artificial intelligence (AI) and machine learning algorithms, have emerged as a severe threat to the integrity and security of digital information.

As we engage this digital frontier, the importance of strong cybersecurity measures in protecting against the destructive consequences of deepfake technology becomes clear (Lezzi et al., 2018). The CIA triangle encapsulates the essence of cybersecurity, a basic structure comprised of three pillars: confidentiality, integrity, and availability (Sarker et al., 2021).

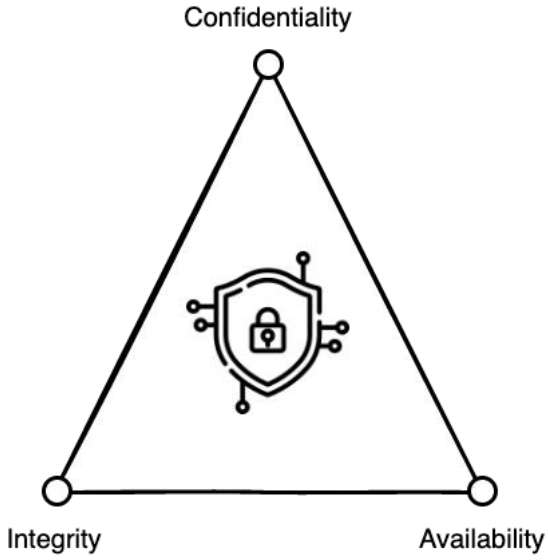

Figure 2 - CIA Principles
Source: (Sarker et al., 2021)

The first pillar of the CIA triangle is confidentiality, which ensures that information is only available to those with approved rights. Maintaining confidentiality is critical in the context of deepfakes to avoid unauthorized access to sensitive information, hence minimizing the possible exploitation of modified content for malevolent reasons. Robust encryption methods, access restrictions, and user authentication processes are critical components of a cybersecurity plan designed to protect data in the face of deepfake attacks (De Oliveira Albuquerque et al., 2014).

The second pillar, integrity, highlights the need to maintain the accuracy and reliability of information. Deepfakes substantially threaten data integrity by modifying digital material to trick human perception. Cybersecurity measures must be intended to identify and prevent illegal changes to digital assets, using techniques such as checksums, digital signatures, and tamper-evident technology. Cybersecurity frameworks can operate as a barrier against the spread of deceptive deepfake narratives by reinforcing the integrity of digital material (Samonas & Coss, 2014).

The third pillar, Availability, highlights the importance of having consistent and timely access to information. Deepfake assaults can cause important systems and services to fail, causing widespread disruption and weakening public faith in digital platforms. Measures to protect the resilience and continuity of operations in the face of deepfake-induced interruptions must be included in cybersecurity strategy. Redundancy, disaster recovery strategies, and resilient infrastructure designs all help to keep digital resources available even in the face of deepfake-related events (Samonas & Coss, 2014).

2.2. Machine Learning

Machine Learning (ML) is a critical paradigm in the field of artificial intelligence, allowing systems to learn from data and improve their performance repeatedly without the need for explicit programming (Shinde & Shah, 2018). The examination of ML is crucial in the context of this thesis owing to its substantial ramifications in both the development and detection of deepfake material. ML has a wide range of applications, from picture and audio recognition to natural language processing. Because of its adaptability, it succeeds in tasks like classification, regression, clustering, and pattern recognition. ML becomes essential in the development and detection of synthetic media material in the context of deepfake technology.

Deepfakes, a type of synthetic media produced by deep learning models, are an example of the symbiotic link between machine learning and misleading content production. By learning complicated patterns and characteristics from large datasets, machine learning algorithms contribute to the creation of hyper-realistic deepfakes. Concurrently, machine learning emerges as a critical tool for identifying and reducing the negative consequences of deepfake technology.

2.2.1. Types of machine learning

Machine learning includes several paradigms, each adapted to certain tasks and goals. Some of these paradigms are supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning, also known as supervised machine learning, is a machine learning and artificial intelligence subcategory. It is distinguished by using labeled datasets to train algorithms that properly categorize data or predict outcomes. As input data is imported into the model, the weights are adjusted until the model is well-fitted, which occurs as part of the cross-validation process (Nasteski, 2017). In machine learning, classification and regression are two fundamental types of tasks. The classification objective is to predict a categorical variable, such as identifying whether an email is spam. The model is trained on labeled data to learn the relationship between input features and output labels. Once trained, the model can be used to predict the class of new, unseen data (Aized Amin Soofi & Arshad Awan, 2017). Regression models, however, are used to forecast continuous variables such as cryptocurrency prices. In theory, a model will be trained using labeled data to learn about the link between data characteristics and the dependent variable. The model can forecast the result of new and unknown data by estimating this connection (Maulud & Abdulazeez, 2020).

Unsupervised learning is the process of training a machine learning model on an unlabeled dataset, where the input data is not paired with any output. The goal is to identify hidden patterns or data clusters without prior knowledge of the output. It is the best option for exploratory data analysis, cross-selling tactics, consumer segmentation, and picture identification because of its capacity to find similarities and contrasts in information (Dike et al., 2018).

Reinforcement learning is a paradigm in which agents learn to make decisions in a dynamic environment to maximize a cumulative reward signal. This type of learning is important in circumstances where an agent interacts with its surroundings, receiving feedback in the form of rewards or penalties based on its actions. The agent's goal is to learn a strategy that leads to optimal decision-making over time. Reinforcement learning found success in applications such as game play, robotic control, and autonomous systems, where adaptability and strategic decision-making are critical (Ebrahimi et al., 2022).

2.2.2. Deep Learning

Deep Learning (DL) is a subset of machine learning that utilizes artificial neural networks (ANNs) with multiple layers, commonly referred to as deep neural networks. The architecture of these networks, inspired by the structure and function of the human brain, allows them to learn and represent intricate patterns present in the data (Ravi et al., 2017).

In deep learning, algorithms use large datasets to train the network's layers. These layers extract features from the data to make predictions or decisions. The bottom layers learn simple features like edges and textures, while the top layers learn higher-level features like objects and scenes.

One of the key advantages of deep learning is that it can learn from data without explicit feature engineering. This allows the model to automatically discover relevant features from the data, which can be helpful in domains where the features are not known or difficult to define. However, deep learning also has some limitations and challenges. One limitation is that the models can be complex and difficult to interpret, making it hard to understand how they make predictions or decisions. Additionally, deep learning requires large amounts of data and computational resources to train, which can present a challenge in some domains (Li, 2018).

In the field of deep learning, recurrent neural networks (RNNs) are a significant breakthrough, especially when it comes to processing sequential tasks and data. RNNs, in contrast to conventional feedforward neural networks, have a special architecture that enables them to recognize temporal connections in a data series. Recurrent connections, which feed a neuron's output back into the network to create a dynamic, memory-like structure, are used to do this. Because of this feature, RNNs are excellent choices for

applications involving speech recognition, time-series prediction, and natural language processing. However, because of the vanishing or expanding gradient problem, conventional RNNs might have trouble capturing long-range dependencies. Several changes, including the introduction of Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) architectures, have been made to address these restrictions and improve the network's capacity to model and learn from sequential data (Tang et al., 2018).

Deep learning has made Convolutional Neural Networks (CNNs) a key component, transforming computer vision and image processing in the process. The capacity of CNNs to automatically learn hierarchical representations from input data sets them apart; this capability is especially useful for structures that resemble grids, like pictures. Convolutional layers are a tool that CNNs use to capture spatial hierarchies and relationships by methodically scanning and extracting local patterns or features. By down sampling the collected features, pooling layers additionally support translation invariance. Tasks including object detection, picture segmentation, and image classification are made possible by this architecture (Figure 3). Transfer learning, a widely used technique that highlights the adaptability and generalization powers of CNNs, involves fine-tuning pre-trained CNN models on tasks. The combination of RNNs and CNNs in different hybrid architectures has created new opportunities for solving challenging issues using both sequential and spatial information as deep learning continues to advance (Yin et al., 2017).

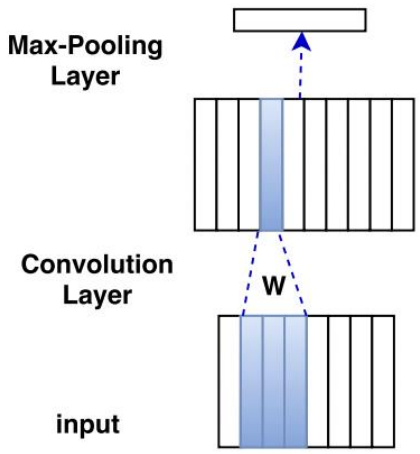

Figure 3 - CNN Architecture
Source: (Yin et al., 2017)

2.2.3. Training approaches for model efficacy

The efficacy of machine learning models, particularly in the complex area of deepfakes, is strongly dependent on the training methods used. Several techniques considerably contribute to model robustness and generalization, assuring their adaptability to different difficulties.

Transfer learning emerges as a powerful method that involves pre-training a model on a large dataset and then fine-tuning it for a specific job. Transfer learning is especially useful in the context of deepfake detection, where access to large, labeled datasets may be constrained. The pre-trained model captures general features and patterns, which are typically taken from a different but related task while fine-tuning tailors these learnt features to the intricacies of deepfake detection (Figure 4). This method improves model performance even when data is scarce, providing a realistic answer to the problems associated with training on limited labeled datasets (Vajpayee et al., 2023).

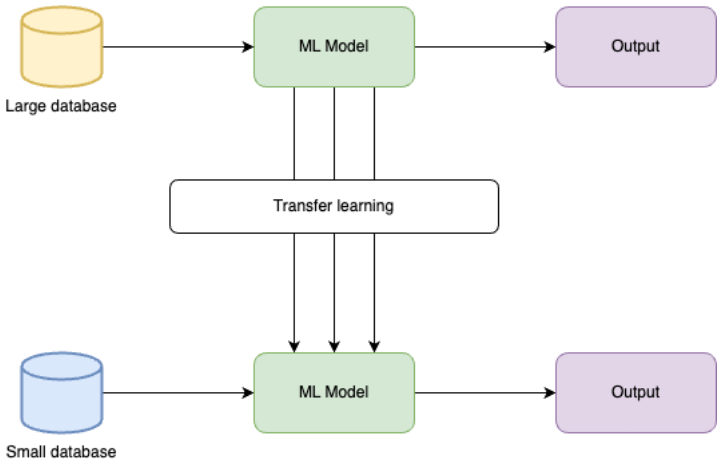

Figure 4 - Transfer learning
Source: (Niu et al., 2020)

Adversarial training incorporates adversarial examples into the training process to build a proactive defense mechanism. These examples are designed to test the model and teach it to resist manipulation efforts. Adversary training becomes a critical line of defense in the setting of deepfake technology, as malicious actors constantly enhance their generation algorithms. By exposing the model to adversarial cases during training, it

learns to recognize and resist manipulation, increasing its robustness in the face of emerging deepfake techniques (Ebrahimi et al., 2022).

Ensemble learning uses a collaborative technique to improve overall accuracy and resilience by mixing predictions from many models (Figure 5). When it comes to deepfake detection, where attackers use a variety of manipulation tactics, ensemble approaches offer a more comprehensive protection. Ensemble learning strengthens the ability to detect subtle patterns suggestive of deepfake content by utilizing the strengths of different models and limiting their flaws. This collaborative method is especially effective in traversing the complex world of synthetic media, providing a cohesive defense against the dynamic and ever-changing nature of deepfake production (Liu & Zhou, 2022).

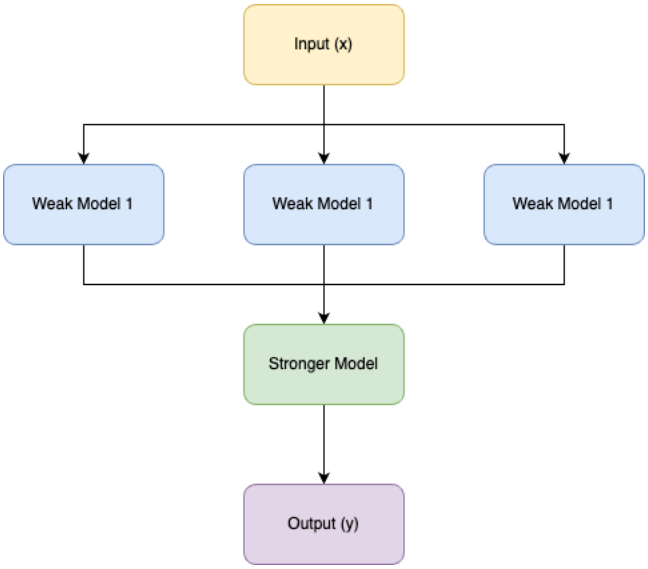

Figure 5 - Ensemble learning
Source: (Alam et al., 2020)

2.2.4. Evaluation metrics

In the enormous realm of machine learning, the goal goes beyond model training to ensure adaptation to previously uncovered data, and given that, evaluation metrics are critical in assessing this generalization capability since they provide insights into a model's performance and potential for real-world applications (Furnkranz & Flach, 2003).

At the heart of it all is accuracy, a fundamental parameter that assesses the overall validity of ML previsions. Accuracy, calculated as the ratio of accurately predicted occurrences of the total instances (Figure 6), gives a comprehensive assessment of model performance. Its sufficiency, however, may fade in settings characterized by class imbalances or varied misclassification costs across distinct classes (Handelman et al., 2019).

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN}$$

Figure 6 - Accuracy formula
Source: (Dalianis, 2018)

Precision takes a very important spot by emphasizing the accuracy of positive predictions. Precision is calculated as the ratio of true positive predictions to the sum of true positives and false positives when the cost of false positives is significant for the use case (Figure 7). It emphasizes the need to accurately identify positive events (Handelman et al., 2019).

$$Precision = \frac{TP}{TP + FP}$$

Figure 7 - Precision formula
Source: (Dalianis, 2018)

Recall, on the other hand, assesses a model's ability to properly identify positive cases within a pool of real positives. Recall, defined as the ratio of true positive predictions to the total of true positives and false negatives (Figure 8), is critical in situations where missing positive cases have major repercussions, stressing the model's sensitivity to positive class identification (Handelman et al., 2019).

$$Recall = \frac{TP}{TP + FN}$$

Figure 8 - Recall formula
Source: (Dalianis, 2018)

The F1 score emerges as a standardized criterion, balancing accuracy, and recall. The F1 score, which is the harmonic mean of these two measures (Figure 9), provides a comprehensive measure of a model's performance. It is especially beneficial when both false positives and false negatives must be weighed equally (Handelman et al., 2019).

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Figure 9 - F1-Score formula
Source: (Dalianis, 2018)

These four main evaluation metrics, used together, provide a sophisticated insight into machine learning model performance. Their interaction reveals a model's strengths and limits, aiding practitioners in refining and improving algorithms for real-world applications.

2.3. Deepfakes

Deepfakes are an emerging technology phenomenon that has attracted a lot of attention because of its possible consequences. The modification of audiovisual content, mostly via the application of advanced machine learning algorithms, is at the heart of this digital world. Deepfakes are mostly found in the video and audio domains. They are the production of extremely lifelike, artificial content that closely resembles the appearance and actions of actual people. These components are built using the Generative Adversarial Network DL method (GAN) or autoencoders.

Deepfakes have the benefit of producing incredibly realistic photos and videos that are indistinguishable from the originals. It can change the entertainment and marketing sectors forever. For example, in the dubbing of a film, the actors' lips can be synthesized to appear in the original language. However, there are considerable worries about the harmful consequences of deepfakes. Deepfakes, for example, may be used to distribute fake news, extortion, revenge porn, tax fraud, and manipulate public opinion (Kocak & Alkan, 2022).

2.3.1. Deepfake generation

The GAN algorithm comprises two main components: a generator and a discriminator. The generator oversees the creation of new synthetic data, while the discriminator determines if the created data is true or false. These two components are trained in reverse, with the generator attempting to generate data that will trick the discriminator and the discriminator trying to detect fake data (Creswell et al., 2018).

Deepfakes are created in two steps (Figure 10): (1) training the GAN using a dataset of images and videos of the target person, and (2) using the GAN to generate new images and videos of the target person in various stances and expressions.

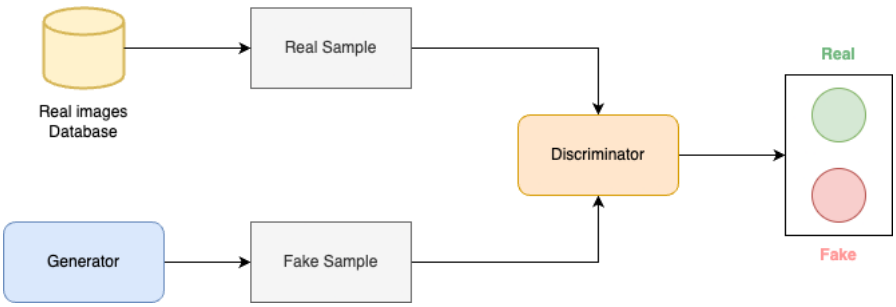

Figure 10 - GAN architecture
Source: (Jaleel & Ali, 2022)

An autoencoder's basic architecture consists of two interconnected components: an encoder and a decoder (Figure 11). The encoder compresses the input data into a latent space representation, and the decoder attempts to recreate the original input from this condensed form. The autoencoder's efficacy is judged by its ability to preserve the important properties of the data during the encoding-decoding process. The latent space,

which is frequently of lesser dimensionality than the input space, contains a distilled version of the input's intrinsic properties. This hidden representation is a compressed and abstracted version of the input that captures its essence in a more compact format (Katarya & Lal, 2020).

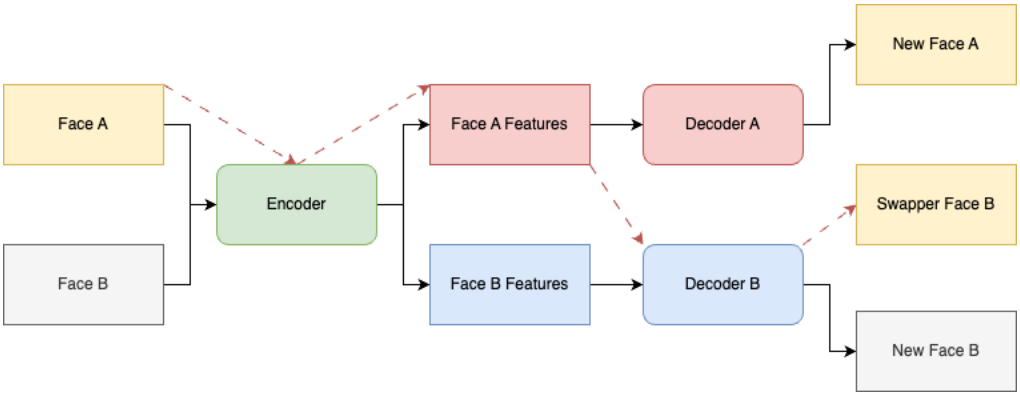

Figure 11 - Autoencoder architecure
Source: (Katarya & Lal, 2020)

2.3.2. Deepfake types

Deepfakes are classified into two types: visual deepfakes and audio deepfakes. Figure 12 highlights the five primary groups that stand out among the visual type.

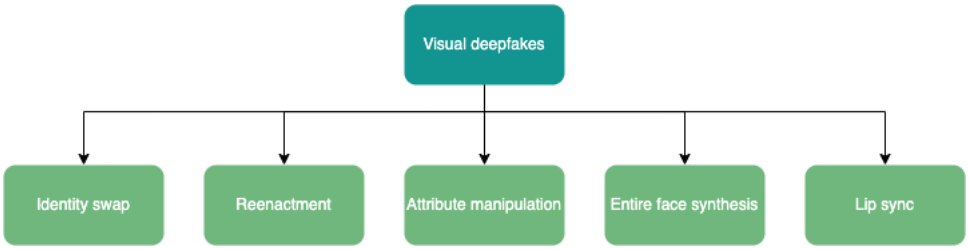

Figure 12 – Visual deepfakes
Source: (Dagar & Vishwakarma, 2022)

The primary visual deepfakes are Identity Swap, Reenactment, Attribute Manipulation, Entire Face Synthesis, and Lip Sync.

The identity swap approach (Figure 13) involves replacing a person's face in an image or video with the face of a completely new person, resulting, as the name suggests, in an identity swap (Ramachandran et al., 2021).

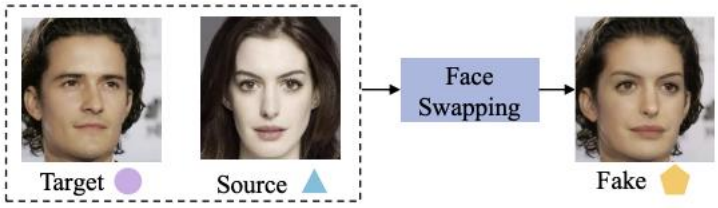

Figure 13 - Identity swap
Source: (Huang et al., 2023)

In the reenactment technique (Figure 14) the subject in the source (video) is manipulated like a puppet in this form of deepfake. These modifications might take the form of facial motions, head movements, or even bodily positions. This strategy is frequently used in film post-production to avoid having to re-record the whole sequence (Dagar & Vishwakarma, 2022).

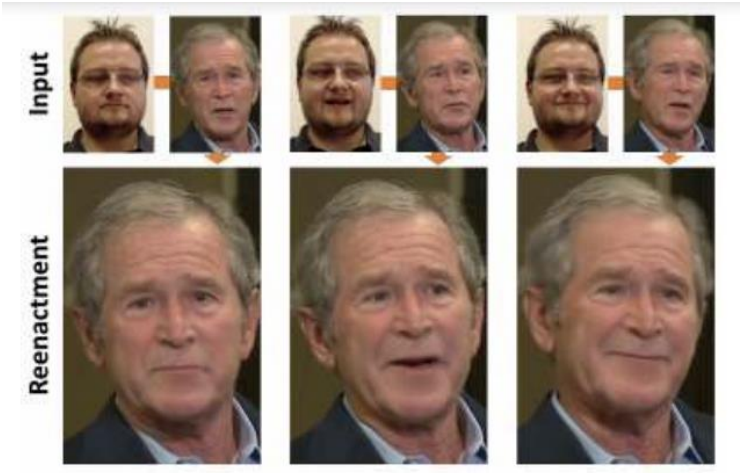

Figure 14 – Reenactment
Source: (Kumar et al., 2020)

Attribute manipulation (Figure 15) is a manipulation of qualities about a particular person. These characteristics include facial expressions, hair (size, shape, and color),

gender, age, clothes, and so on. That is, an initial picture featuring a 30-year-old youthful woman can be converted into an 85-year-old man while retaining the original lady's traits. This technology is often used in apps that replicate people's looks a few years later, as well as in film post-production when there is a time jump and the actor must seem with some indications of age (Dagar & Vishwakarma, 2022).

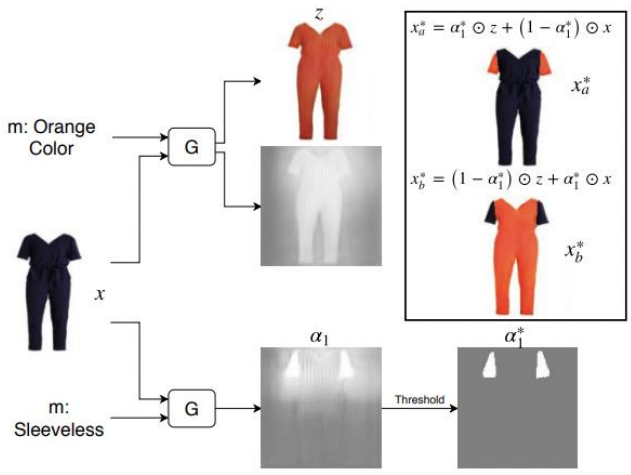

Figure 15 - Attribute manipulation
Source: (Ak et al., 2019)

The entire image synthesis approach (Figure 16) is meant to generate realistic but nonexistent things. In other words, the goal is to recreate a picture or video that replicates reality perfectly. Creating video games with realistic graphics that allow for a better level of player immersion is a practical use of this category. The ability to practically 100% replicate a person's look and behavior and force the presence of that same person to act in ways that he has never acted before is a clear hazard (Dagar & Vishwakarma, 2022).

Figure 16 - Entire image synthesis
Source: (Bansal et al., 2017)

The lip sync method entails synthesizing a video of a target identity so that the mouth area matches a particular audio input. This category has a variety of uses in the entertainment industry, including dubbing films into many languages so that the lips of the actors follow the various idioms and appear as realistically as possible (Masood et al., 2022).

The principal audio deepfakes are divided in two main types (Figure 17): text to speech synthesis and voice conversation.

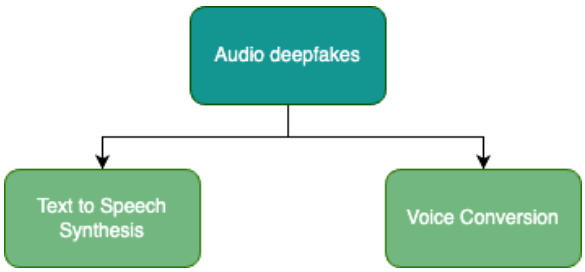

Figure 17 - Audio deepfakes
Source: (Dagar & Vishwakarma, 2022)

Deepfake text-to-speech (TTS) synthesis employs advanced artificial intelligence (AI) and machine learning algorithms to build very realistic and believable synthetic voices from text. On the obtained data, deep learning models such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs) are trained. The model learns the patterns and characteristics of the speaker's voice, such as pitch, rhythm, and pronunciation. When given new text, the trained deepfake TTS model employs voice

embeddings to generate speech that sounds like the target speaker. As a result, even when reading text that the speaker has never spoken before, the synthetic voice closely mimics the genuine speaker. Concerns have been raised regarding the possible misuse of deepfake TTS for impersonation, distributing misinformation, or creating fake audio content due to its capacity to convincingly mimic someone's voice (Salvi et al., 2023).

Voice conversion is a technology that converts the characteristics of a source speaker's voice into those of a target speaker while maintaining linguistic content. Voice conversion, as opposed to text-to-speech synthesis, which generates speech from written text, focuses on altering the features of an existing spoken utterance to mimic the voice of a different person. Voice conversion usually necessitates recording pairs from both the source and target speakers. These recordings serve as the model's training data. A variety of speech samples may be included in the data to capture various features of the speakers' voices. The trained model turns the acoustic features of a source speaker's voice into the corresponding features of the target speaker's voice throughout the conversion process. This change produces a synthesized voice that sounds similar to the target speaker while keeping the original linguistic information. Voice conversion can be used for a variety of purposes, including as dubbing in the entertainment business, creating voice alternatives for virtual assistants, and improving voice clarity in telecommunications (Khanjani, 2021).

2.3.3. Deepfake detection algorithms

There are already numerous deepfake detection algorithms available in the community. However all of them are variations of some core detection algorithms like, VGG, ResNet, XCeption, MesoNet and Inception (Jayakumar & Skandhakumar, 2022).

The XCeption architecture is a deep convolutional neural network (CNN) designed to detect complicated features in images while keeping model weight under control (Rismi et al., 2020). Deepfake detection and other image-related tasks lend themselves very nicely to XCeption. The depth-wise separable convolutions used in XCeption inception-style modules set it apart from other architectures. The network can capture characteristics at various sizes thanks to these depth-wise separable convolutions with fewer parameters (Pan et al., 2020). It is an improved version of the Inception architecture.

ResNet, or Residual Neural Network, is well known for its residual learning blocks, which allow for the training of extremely deep neural networks (Rismi et al., 2020). ResNet integrates the idea of residual blocks, which contain skip connections, into its architecture. Even in very deep networks, these skip connections allow the gradient to flow more smoothly during training. The number of residual blocks in these topologies is denoted by ResNet-50 and ResNet-101 (Mukti & Biswas, 2019).

The VGG (Visual Geometry Group) architecture is well-known for its simplicity and efficacy (Rismi et al., 2020). We picked the VGG16 architecture because of its proven performance in a variety of computer vision tasks. VGG networks stand out because of their straightforward design, which includes stacked layers and compact 3x3 filters. The specific model we used, VGG16, consists of 16 layers, and it is widely recognized for its performance in various computer vision applications (Tammina, 2019).

The MesoNet deepfake detection algorithm has gained prominence for its effective approach in discerning manipulated visual content. Inspired by its success, we selected the MesoNet architecture for its robust performance across diverse deepfake scenarios. MesoNet stands out due to its innovative design, featuring a distinctive focus on micro-texture analysis. The core of the model revolves around exploiting mesoscopic visual patterns within facial images. Comprising a network of carefully crafted layers, MesoNet demonstrates its efficacy through a meticulous examination of subtle details, contributing to its resilience against various deepfake manipulation techniques (Xia et al., 2022).

The Inception architecture, renowned for its intricate design and exceptional performance, has been a natural choice for our deep learning endeavors. Inception distinguishes itself through its intricate architecture, incorporating inception modules that allow for the simultaneous processing of features at varying spatial resolutions. This design choice promotes a holistic understanding of visual content, making it particularly effective for discerning complex patterns and structures within images. The model's depth and intricacy contribute to its versatility in handling diverse visual recognition tasks (Verma et al., 2021).

Our research coincides with three distinct architectures in the pursuit of a successful deepfake detection model: Xception, ResNet, and VGG16, each chosen for specific qualities mentioned in the comparison table (Table 2).

Table 2 - Deepfake detection algorithms comparison

Feature	XCeption	ResNet	VGG16	MesoNet	Inception
Architecture	Deep convolutional neural network (CNN)	Residual Neural Network (ResNet)	Visual Geometry Group (VGG)	Innovative design focusing on micro-texture analysis	Intricate architecture with inception modules
Design Focus	Detecting complex features with controlled model weight	Enabling training of extremely deep neural networks	Simplicity and efficacy in various computer vision tasks	Discerning manipulated visual content through micro-texture	Holistic understanding of visual content through inception
Key Components	Depth-wise separable convolutions, inception-style modules	Residual learning blocks, skip connections	Stacked layers, compact 3x3 filters	Micro-texture analysis, network of carefully crafted layers	Inception modules allowing simultaneous processing
Parameter Efficiency	Captures characteristics at various sizes with fewer params	Efficient training of deep networks with skip connections	Efficacy with straightforward design and compact filters	Robust performance with a focus on subtle details	Simultaneous processing of features at varying resolutions
Application Suitability	Deepfake detection and image-related tasks	Training deep networks, various computer vision applications	Proven performance in computer vision tasks	Robust performance across diverse deepfake scenarios	Versatility in handling diverse visual recognition tasks

We chose Xception because of its effectiveness in capturing complicated features. The architecture, which is distinguished by depth-wise separable convolutions and inception-style modules, exhibits proficiency in recognizing detailed visual cues, which is critical in the complex environment of deepfake detection. Because of its parameter efficiency, which allows for the recording of several attributes with fewer parameters, Xception is a viable option for real-time or resource-constrained applications. Furthermore, its demonstrated performance in image-related tasks, such as deepfake detection, provides a trustworthy baseline for our research.

ResNet was chosen because of its ability to handle deep networks. Residual learning blocks and skip connections make it easier to train extraordinarily deep neural networks, which is essential for deciphering the subtle manipulations inherent in deepfake content. ResNet's configurational versatility, as demonstrated by configurations such as ResNet-50 and ResNet-101, enables us to investigate multiple network depths while striking a balance between accuracy and computational efficiency. Furthermore, the deep learning community's extensive adoption provides compatibility and permits meaningful comparisons with established benchmarks.

VGG16 is used in our research because of its ease of use, efficacy, and proven adaptability. The simple design, which includes stacked layers and small 3x3 filters, fits with our emphasis on interpretability and ease of application. VGG16 is a trustworthy alternative for the versatility required in deepfake detection research due to its track record of performance across numerous computer vision applications. Furthermore, VGG16's status as a benchmark architecture in computer vision emphasizes its applicability for our comparative investigation.

2.3.4. Deepfake datasets

Deepfake datasets are essential for furthering the study of synthetic media mitigation and detection. Considering recent advancements, several noteworthy datasets have surfaced, each of which adds to our knowledge and assessment of deepfake-generating methods. DeepFake Detection Challenge (DFDC), FaceForensics++, CelebA-HQ, and DeeperForensics-1.0 are some of the most widely used deepfake datasets.

The DeepFake Detection Challenge (DFDC) is a prominent dataset in the industry, known for its extensive scope and widespread usage. DFDC, a collection of over 100,000 videos

of genuine and deepfake content produced by paid actors, was released by Facebook AI and its partners. The dataset ensures a realistic portrayal of the difficulties encountered in real-world applications by spanning a wide range of scenarios, lighting situations, and facial expressions. As a standard for assessing deepfake detection models, DFDC has encouraged healthy competition and creativity among researchers (Dolhansky et al., 2020).

Another well-known dataset used in deepfake research is FaceForensics++. It consists of an extensive library of movies produced with a range of synthesis techniques, such as Face2Face, DeepFake, and FaceSwap. FaceForensics++ offers an extensive collection of alteration techniques and real-world variations, making it an invaluable tool for testing and developing deepfake detection systems. The dataset has made a substantial contribution to our understanding of how detection models generalize to various manipulation strategies (Ramachandran et al., 2021).

The CelebA-HQ dataset was not designed with deepfake research in mind, yet it has been extensively used in the field, nonetheless. It is a valuable resource for training deep learning models for facial recognition and modification and contains high-quality photos of celebrities. Scholars frequently utilize CelebA-HQ to enhance their datasets and strengthen the resilience of deepfake detection algorithms (He et al., 2021).

A relatively recent addition to the deepfake dataset ecosystem is DeeperForensics-1.0. Tencent released this dataset, which aims to improve upon the shortcomings of previous datasets by including a wider variety of scenarios, resolutions, and compression errors. DeeperForensics-1.0 seeks to advance deepfake detection studies by collecting a wider range of deepfake variations (Rathgeb et al., 2022).

Choosing the DFDC as the primary dataset for research on deepfake detection is a judicious decision owing to several key factors, highlighted in the comparison table (Table 3).

Table 3 - Deepfake datasets comparison

Feature	DFDC	FaceForensics++	CelebA-HQ	DeeperForensics-1.0
Release source	Facebook AI and Partners	Technicolor, University of Erlangen	CelebA Dataset (not designed for deepfake)	Tencent
Total videos / images	100,000+ videos	1,000+ videos	30,000+ images	60,000+ videos
Variety of scenes	Diverse scenes and facial expressions	Various scenes and lighting	Celebrity photoshoots	Broad range of scenes and resolutions
Realism	Realistic content with paid actors	Various synthesis methods	High-quality celebrity images	Diverse scenes and compression artifacts
Benchmark recognition	Widely recognized as a benchmark	Commonly used in deepfake research	Often used for facial recognition	Emerging as a benchmark
Contribution to research	Benchmark for deepfake detection models	Generalization across manipulation	Augmentation for deepfake detection	Pushing boundaries in deepfake detection

First off, DFDC provides a significant scope with more than 100,000 videos that cover a wide range of scenarios, lighting, and face emotions. In addition to enabling reliable training of detection models, this variety replicates the intricacies found in real-world situations.

Another strong argument in favor of DFDC is the focus on realism. A high degree of realism is maintained in the dataset's videos since they are produced with the help of paid actors. The development of detection models that can successfully distinguish between real and modified content in real-world scenarios depends heavily on realism (Dolhansky et al., 2020).

DFDC is well recognized as a benchmark dataset for deepfake detection. Its acceptance by the scientific community and industry has made it a standard reference for assessing the efficacy of cutting-edge algorithms. Using DFDC as a benchmark promotes community engagement, healthy competition, and innovation in the research domain (Dolhansky et al., 2020).

To sum up, the DFDC presents an attractive blend of challenges that are thorough, realistic, scalable, and recognize benchmarks. Selecting DFDC as the main study dataset offers a strong basis for creating and assessing deepfake detection models that are sophisticated and suitable to the complexities of real-world situations.

2.4. Related Work

Deepfakes appear to be a new technology, but the scientific community has previously essayed some research and development in this field. This area has many challenges, particularly the poor quality of current data for algorithm training and the poor performance of these same algorithms (Zhang, 2022).

Most approaches for detecting deepfakes rely on characteristics and machine learning. For example, in this paper (Patel et al., 2020), the authors applied five distinct neural network models for deepfake recognition. They employed a dataset of around 7000 images for training and 2000 for the experimental setup. As shown in Table 4, they achieved precisions ranging from 0.862 to 0.902.

Table 4 - Accuracy comparison

Model	Accuracy
VGG16	0.893
Resnet50	0.89
InceptionV3	0.862
MobileNet	0.902
DenseNet121	0.897

Algorithms may consider many face and bodily characteristics in a deepfake. For example, in numerous images and videos, the position of the head is highlighted a decisive characteristic for detection. The authors of this work (Yang et al., 2019)

investigated this similar property using a face detector that extracts 68 facial expressions from each frame to train an SVM algorithm. Two datasets were utilized. The first, known as UADFV, achieved a recognition accuracy of 0.89, while the second, DARPA GAN, achieved a slightly lower accuracy of 0.843.

A combination of machine learning and deep learning can also be used as a recognition method. The technique of extracting facial expressions and identifying traits is done manually or with the assistance of a face extractor in the previously described experiments. On the other hand, deep learning allows these same features to be extracted in a more effective and practical manner and then handed to another algorithm responsible for recognition. That is an example from this study (Al-Dhabi & Zhang, 2021), in which the authors integrated Convolutional Neural Network (CNN) with Recurrent Neural Network (RNN), or Long Short-Term Memory (LSTM). CNN oversees extracting frames from videos and the face region from the same frames. The CNN output is then used to train the LSTM. The greater the number of frames tested, the higher the precision; for example, the authors obtained an accuracy of 0.998 for 100 frames.

Another method for detecting deepfakes is to use the same mechanism that generates them, namely the GAN. The authors of this study (John & Sherif, 2022) employ a variant of this GAN known as SGAN. The SGAN training procedure is identical to the GAN, except supervised model weights are updated using labeled data. Using the same dataset but various sizes, the authors concluded that precision grows with dataset size, achieving an accuracy of 86.7% with 10000 records and 92.3% with 40000 records.

In this study (Verma et al., 2021), the authors utilized a model known in the scientific community as Inception-ResNetV2 in their investigation. This model was trained using over a million photos from the ImageNet collection. The authors utilized a dataset from Kaggle DFDC that included 400 videos for training and another 400 for testing. The videos were separated into folders containing frames representing false and genuine data. 80% of the data was utilized for training and 20% for testing, yielding accuracy rates of 97% for training and 99% for testing.

The authors (Jayakumar & Skandhakumar, 2022) utilized a dataset comprising 6667 frames taken from videos during the data preparation phase for the study. Because it's a new sector, no one has established baseline values for benchmarking. The scientists did, however, compare their models to the three most prominent models in the scientific

community, namely Xception, ResNet, and VGG. The Xception model was 90% accurate, the ResNet model was 88% accurate, the VGG model was 56% accurate, and the authors' model was 92% accurate.

According to the review of related work, several models, especially Inception, Resnet, Xception, VGG and other upgraded versions of them, are almost always cited in research on the detection of deepfakes. Most of the research focuses on accuracy, but the time taken to achieve this precision has yet to be discussed. But to allow regular users to employ recognition techniques daily, the recognition speed must be as fast as feasible while retaining high levels of accuracy.

3. METHODOLOGY

This chapter describes the investigation methodology and the data extraction methodology used for this study.

3.1. Investigation methodology

Action research is a methodical technique to resolving real-world problems and improving procedures that involves a cyclical process of planning, acting, observing, and reflecting.

The key characteristics of action research include (Jrad et al., 2014):

- **Collaboration:** Action research is often collaborative, with researchers and practitioners working together to address a specific issue or challenge.
- **Cyclical Process:** It is a cycle (Figure 18) of planning, acting, observing, and reflecting. Based on constant feedback and learning, this cycle aids in the refinement and improvement of tactics.

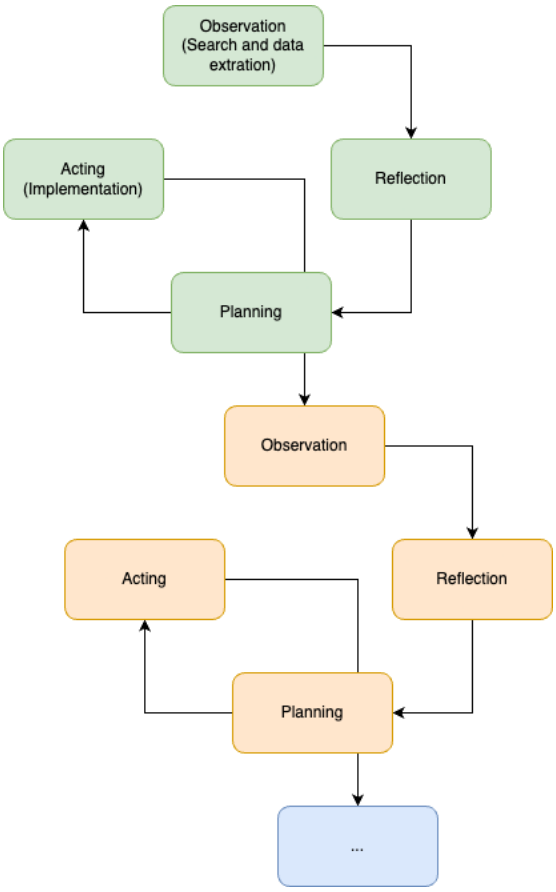

Figure 18 - Action research cyclical process
Source: (Saifudin et al., 2016)

- **Participatory:** Participants actively contribute to the study process, and their perspectives are appreciated. The goal is to empower individuals and communities to take control of their conditions.
- **Context-Specific:** Action research is frequently undertaken in specific corporate or community contexts, concentrating on practical concerns and solutions relevant to that context.
- **Problem-Solving Orientation:** The primary purpose is to address and solve the participants' practical problems or obstacles. It's a practical approach that focuses on practical, actionable solutions.
- **Continuous learning:** Continuous learning and adaptability are encouraged by action research. Each cycle's insights inform the following actions, enabling an iterative process of progress.
- **Ethical issues:** In action research, ethical issues are critical, and researchers must ensure that participants' rights and well-being are maintained.

A research plan was developed using this research technique to carry out this investigation comparing deepfake detection systems. The steps in the plan are as follows:

- **Identifying the problem:** The advent of deepfake technology presents a serious issue in a variety of domains, including journalism, politics, and security. To limit potential effects, reliable identification of deepfakes is critical.
- **Planning:** Preparation of a strategy for comparing existing deepfake detection systems. Consider accuracy, speed, and adaptability to various forms of deepfake content.
- **Acting:** Put the plan into action by applying several deepfake detection algorithms to a diverse range of deepfake videos. Gather pertinent data and create a baseline for comparison.
- **Observing:** We would collect data on the performance of each algorithm methodically, considering measures such as accuracy, false positives, false negatives, and other important parameters. The evaluation would include the algorithms' real-world efficacy as well as their adaptability to emerging deepfake approaches.
- **Reflection:** Each of the algorithms is compared, and a conclusion is reached as to whether or not some of them might be applied in real-time circumstances.

- **Iterative Cycles:** We would iterate the procedure if necessary, incorporating feedback and refining the comparative methodology. This could entail trying out new methods, tweaking settings, or increasing the dataset.

3.2. Knowledge extration methodology

The Cross Industry Standard Process for Data Mining was chosen for this study's research approach (CRISP-DM). This approach is extensively utilized and provides a well-structured alternative for analyzing and comprehending a business challenge and determining the best solution through data and analysis. As seen in the Figure 19, it comprises six critical points (Azevedo & Santos, 2008).

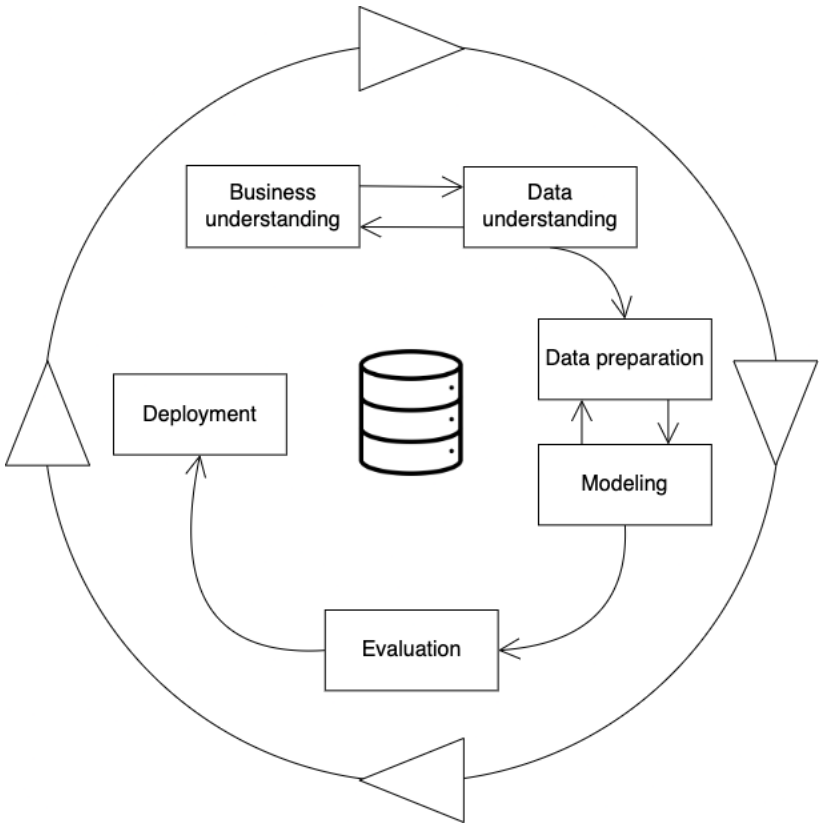

Figure 19 - CRISP-DM lifecycle
Source: (Azevedo & Santos, 2008)

- Business Understanding: Define the problem and objectives in this stage. The goal is to guarantee that the project is aligned with the objectives and feasible.
- Data Understanding: This stage emphasizes gathering and assessing existing data to ensure its quality and relevance to the problem. The objective is to obtain insight into the data and ensure it's ready for analysis.
- Data Preparation: This process aims to clean, convert, and prepare data for analysis. The primary goal is to ensure that the data is in the correct format for analysis and modeling.
- Modeling: Machine learning techniques are used at this step to create prediction models. The objective is to develop models capable of providing valuable insights and forecasts.
- Evaluation: Models are examined in this stage to establish their performance and quality. The goal is to find the optimal model that produces the most accurate results in the least amount of time.
- Deployment: The solution is deployed and monitored at this stage. The primary goal is to guarantee that the solution satisfies the envisioned goals.

Adapting the steps of the methodology above, the research work can be divided into the following steps:

- Business Understanding: Deepfake detection has emerged as a critical research area due to the increasing prevalence of manipulated media and its potential implications in various domains. The ability to accurately identify deepfake videos is crucial for maintaining trust, protecting individuals from harm, and mitigating the spread of misinformation. This thesis aims to contribute to the development of effective deepfake detection algorithms to address these concerns.
- Data Understanding: The DeepFake Detection Challenge (DFDC) dataset serves as the foundation for this research. It consists of a diverse collection of real and manipulated videos, with a particular focus on face-swapping deepfakes. By thoroughly exploring the DFDC dataset, we gain insights into its composition, class distribution, and potential biases. This understanding allows us to make informed decisions regarding algorithm selection and preprocessing techniques.
- Data Preparation: During the data preparation stage, we ensure that the DFDC dataset is suitable for deepfake detection. We address data cleaning to handle any missing or corrupted samples, ensuring the integrity and quality of the dataset.

Additionally, we employ feature selection or extraction methods to identify the most informative features that can effectively distinguish between real and manipulated videos. Data augmentation techniques, such as flipping, rotation, and cropping, are applied to enhance the dataset's diversity and improve the robustness of the deepfake detection models.

- Modeling: In the modeling stage, we select and apply four prominent algorithms for deepfake detection: Xception, ResNet and VGG. Each algorithm offers unique architectural characteristics that make them suitable for deepfake detection tasks. We provide a comprehensive overview of each algorithm, highlighting their strengths and relevance in the context of deepfake detection. Previous research studies have demonstrated the effectiveness of these algorithms on similar datasets, giving us a strong foundation to build upon.
- Evaluation: The evaluation stage focuses on assessing the performance of the selected algorithms on the DFDC dataset. We employ evaluation metrics such as accuracy, precision, recall, and F1-score to measure the effectiveness of the deepfake detection models. The experimental setup includes appropriate training-validation-test splits, and cross-validation techniques may be employed to ensure reliable and unbiased evaluation results. This stage allows us to compare the performance of different algorithms and identify the most effective approach for deepfake detection.
- Deployment: The deployment stage involves implementing the best-performing deepfake detection algorithm and integrating it into a practical application or system. This stage considers factors such as scalability, real-time processing capabilities, and user interface design. The aim is to deploy a robust and efficient deepfake detection solution that can be utilized in real-world scenarios to combat the spread of deepfake videos.

4. IMPLEMENTATION

To comprehensively investigate the efficacy of deepfake detection, a structured research architecture was designed and implemented (Figure 20). This chapter delves into the key components of the experimental framework, outlining the sequential steps undertaken for robust evaluation.

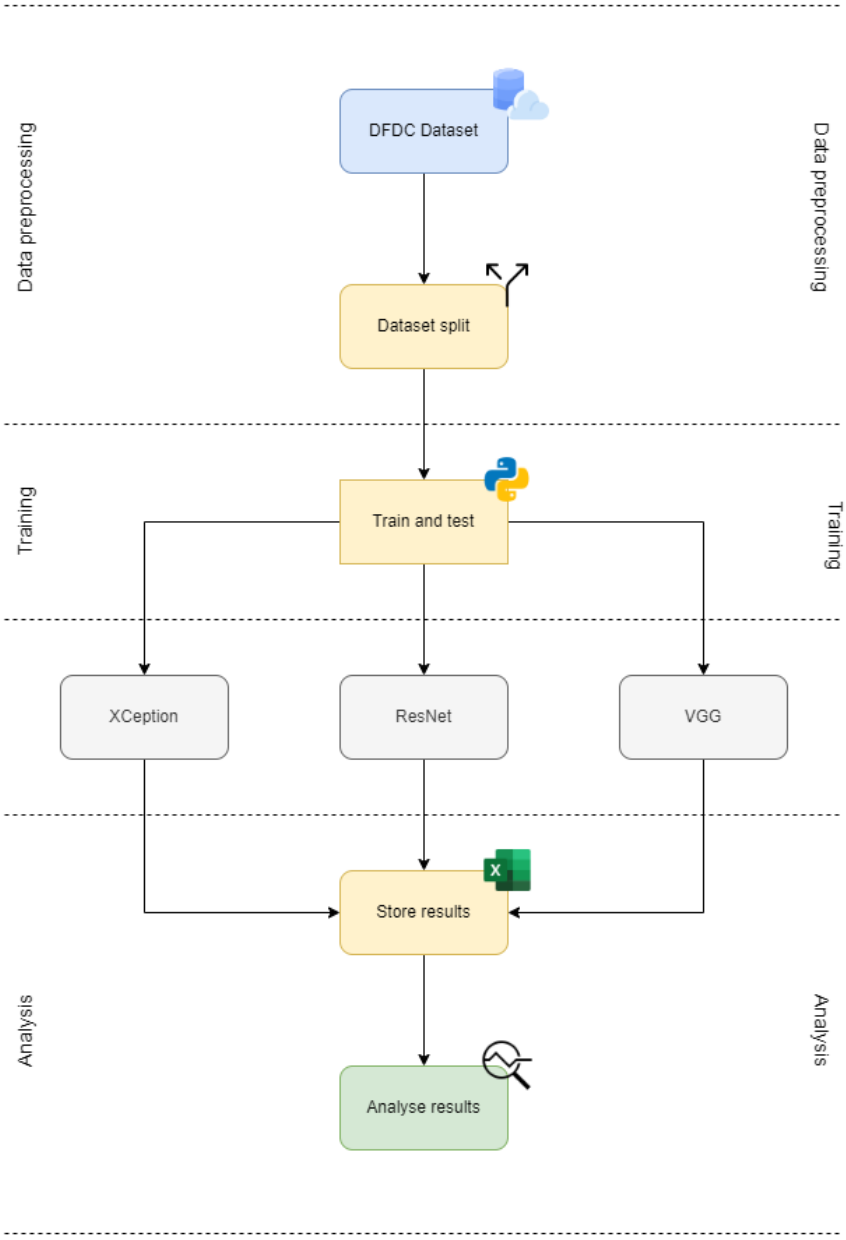

Figure 20 - Research architecture

4.1. Data processing

In this work, we provide an extensive description of the DFDC (Deepfake Detection Challenge) dataset, which forms the basis of our empirical investigation. For our research, it is essential to comprehend the properties and makeup of the dataset. At its foundation, the DFDC dataset is a professionally selected collection of videos covering a wide range of themes and manipulation techniques. This dataset is used to train and test our deepfake detection methods. It includes videos with actual, unedited material and deepfake versions. This pairing allows us to perform direct comparisons, allowing our algorithms to discriminate between genuine and altered content successfully. The dataset also includes several deepfake alteration techniques, such as face swaps, facial expression synthesis, and voice changes. This variety enables our detection algorithms to recognize a wide range of deepfake material.

The DFDC dataset is huge, as it contains more than 100.000 videos, so given that we divided the dataset into three unique subsets in our research to enable training, validation, and assessment of our deepfake detection models (Figure 21). This separation is critical to assuring our models' robustness and generalizability. The following are the characteristics of this subset split, as well as the number of images in each subset:

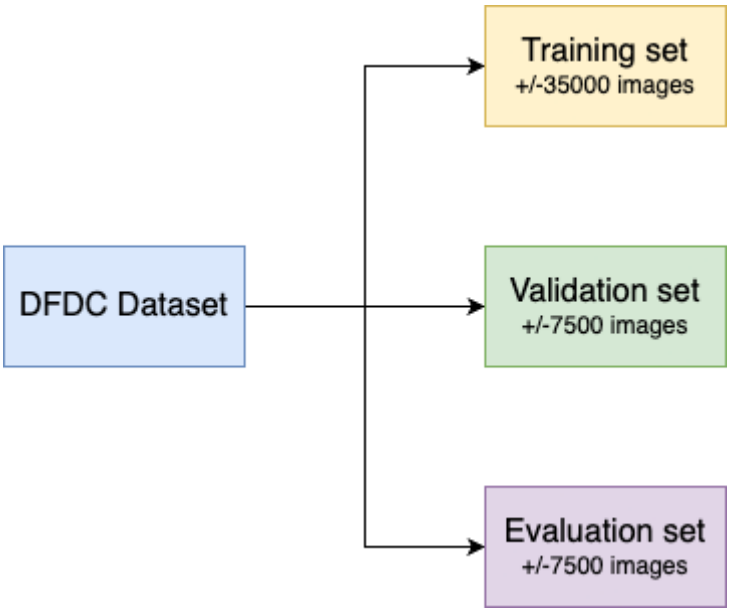

Figure 21 - DFDC dataset split

Our dataset's foundation is built on the training set (Subset 1, Figure 21). This subset covers a wide range of movies, totaling around 35,000 pictures. Each video includes a genuine, unedited version and a deepfake equivalent for various people. The training set is the foundation for our models to learn the differentiating characteristics and patterns that separate deep fake movies from real ones. Importantly, our models use transfer learning, which begins with pre-trained weights and is then fine-tuned on this training subset.

The validation set (Subset 2, Figure 21) is used to fine-tune and hyperparameter-tune our deepfake detection models. This subset comprises around 7,500 photos. It is used to track the performance and convergence of our models as they are being trained. It ensures that our models do not overfit the training data and provides a reference for adjusting hyperparameters such as learning rates and regularization approaches.

The assessment set (Subset 3, Figure 21) additionally includes about 7,500 photos. This data represents unexplored territory against which our deepfake detection methods are tested. In this assessment set, our models are tested by encountering deepfake and actual images they have never encountered during training or validation. The performance of our models on this subset provides an unbiased evaluation of their generalization capabilities.

4.2. Training algorithms

In our deepfake detection study, we employed three distinct deep learning architectures, each with unique characteristics and advantages. For training all of the three implemented algorithms have used a similar training process (Figure 22), passing by a process of pre training, and fine tuning. Some have also experiment transfer learning.

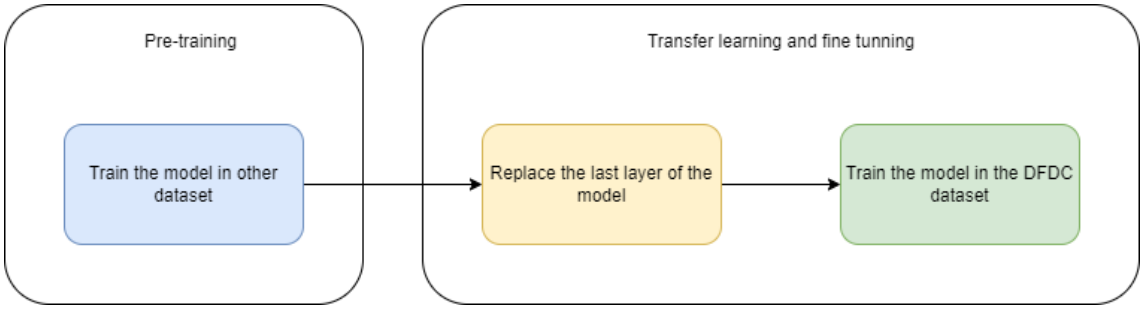

Figure 22 - Training process

A more detailed approach of each model is presented in the following points:

- **Xception model**
  - **Transfer Learning:** The core of our deepfake detection was an XCeption model that had already been trained. This signifies that we trained the model's weights on a sizable dataset before initializing it. Before fine-tuning the DFDC dataset, pre-training enables the model to pick up essential characteristics from various pictures.
  - **Fine-tuning:** We improved the XCeption model on the DFDC dataset after initializing it with pre-trained weights. The fine-tuning procedure entails modifying the model to the specifics of deepfake detection. Pre-trained weights give a robust foundation for the model, and finetuning fits the model to the job requirements.
- **ResNet model**
  - **Ensemble Approach:** Instead of picking a single ResNet design, we took an ensemble approach. This implies that we utilized ResNet-50 and ResNet-101 combined. Ensembling is the process of combining predictions from numerous models to improve overall performance. We want to adequately capture local and global properties by employing ResNet-50 and ResNet-101.
  - **Transfer Learning and Fine-Tuning:** As with XCeption, we initialized the ResNet models with pre-trained weights using a large-scale image classification dataset. These pre-trained weights provide a firm foundation for the models. On the DFDC dataset, the models were then fine-tuned to adapt them to the deepfake detection job.

- **VGG model**
    - **Pre-trained model:** We started with the VGG16 model that already had pre-trained weights. These pre-trained weights play a role, providing the model with feature representations needed for accurate deepfake detection.
    - **Transfer Learning and Fine-Tuning:** After pretraining, we fine-tuned the VGG16 model using the DFDC dataset. Fine tuning enables the model to adapt to the intricacies and complexities involved in deepfake detection, ensuring it performs optimally for this task.

4.3. Experimental setup

We used a MacBook Air with the groundbreaking Apple M1 processor, which has amazing processing capabilities, for our experimental setup. For data storage, the hardware configuration included 16 gigabytes (GB) of RAM and a 512-gigabyte (GB) Solid State Drive (SSD). This hardware choice was prompted by the M1 chip's well-known competency in AI and machine learning workloads, which ensured that our studies ran smoothly. The 16GB of RAM provided enough of memory capacity, allowing for quick data processing and model training. Meanwhile, the high-speed 512GB SSD offered quick data access and storage, which was critical when dealing with large datasets and testing findings.

In the software area, we used a Python environment coordinated by Anaconda version 2021.05, which allowed for the easy integration of multiple libraries. TensorFlow 2.5 was our core deep learning framework, offering a solid basis for building and training deep neural networks. Keras 2.4.3, NumPy 1.19.5 for numerical calculations and Pandas 1.3.3 for data processing were among the other libraries used. Jupyter Notebook 6.4.3 was used as the development environment of choice.

4.4. Results analysis

This subsection functions as a pillar of the thesis, connecting the theoretical foundations with real-world applications. Our objective is to offer significant contributions to the field of deepfake detection by conducting a thorough analysis of the data. Additionally, we aim

to contribute passing relevant information to practitioners, academics, and decision-makers about the difficulties that arise when manipulating synthetic media.

With an accuracy percentage of 87.5%, the data (Figure 23) shows that Xception is the most accurate of the three models. This implies that Xception is skilled at accurately identifying phony and genuine situations. Even though they function rather well, ResNet and VGG have lower accuracy rates—68.7% and 83.7%, respectively. These results highlight the significance of accuracy as a key performance indicator for assessing the overall effectiveness of deepfake detection algorithms.

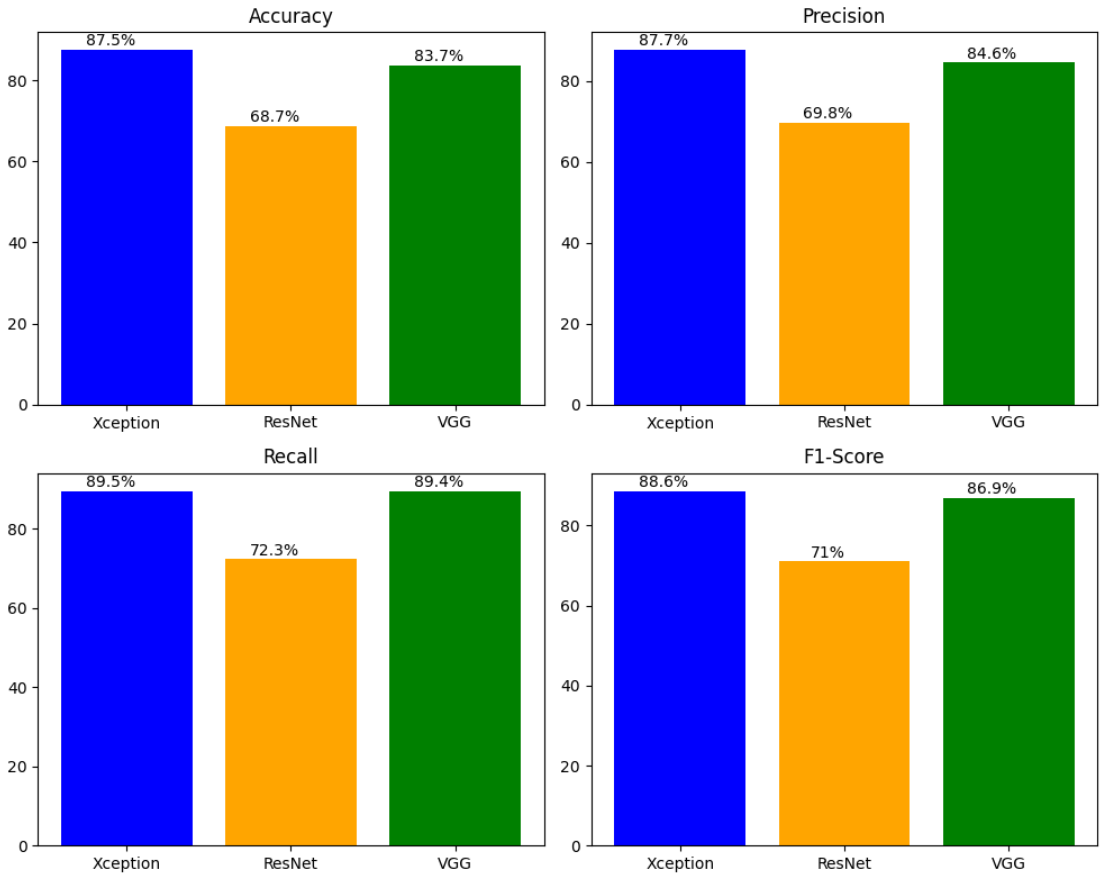

Figure 23 - Evaluation metrics

The precision of the model is determined by how well it can distinguish true positives from expected positives. Xception has the highest precision, with an accuracy rate of 87.7%. With accuracy ratings of 69.8% and 84.6%, respectively, ResNet and VGG trail

closely behind. This statistic is especially important for applications (like legal or security settings) where it is critical to minimize false positives.

The model's recall, or sensitivity, measures its ability to recognize every true positive case. With 89.5% recall, Xception stands out as having the highest recall rate, showing that it is effective in collecting a significant percentage of genuine positive cases. With a recall rate of 89.4%, VGG comes in close second, demonstrating its capacity to reduce false negatives. ResNet has a somewhat lower recall rate of 72.3%, but it is still functioning rather well.

The harmonic mean of accuracy and recall, or F1-Score, offers a fair evaluation of a model's overall performance. With an F1-Score of 88.6%, Xception stands out as the model that strikes the best balance between precision and recall trade-off. F1-Scores for ResNet and VGG are 71% and 86.9%, respectively. This measure is especially important for applications that want to strike a compromise between recall and accuracy.

The visualizations of the confusion matrix provide a thorough and insightful overview of the performance of the three deepfake detection models (Xception, ResNet, and VGG). The visualizations (Figure 24) illustrate the confusion matrices using heatmaps, making it easy to grasp the models' classification results. Several conclusions may be derived from the confusion matrix visualizations about the performance of the three deepfake detection models (Xception, ResNet, and VGG):

- Xception and VGG High True Positive Rates: In their confusion matrices, Xception and VGG both show high values in the true positive (TP) cells. This suggests that these algorithms are capable of accurately distinguishing between authentic and fraudulent instances of information.
- ResNet Demonstrates Balanced Performance: ResNet performs more evenly across true positive and true negative cases but has a lower true positive rate than Xception and VGG. This implies that ResNet accurately distinguishes between fake and genuine occurrences while maintaining a healthy balance.
- VGG Shows Higher False Positive Rate: Comparing VGG to Xception and ResNet, a greater proportion of false positives (FP) is observed. This suggests that VGG could have a propensity to incorrectly identify some real cases as fraudulent, which would raise the false positive rate.

- Consistent True Negative Performance: The three models demonstrate a high true negative (TN) rate, suggesting their accuracy in recognizing authentic examples of real material. Maintaining the validity of non-deepfake content depends on this.
- Xception and ResNet Exhibit Lower False Negative Rates: Lower values in the false negative (FN) cells are displayed by Xception and ResNet, indicating that they are effective in reducing the number of cases in which real material is mistakenly identified as fraudulent.
- Xception Shows Balanced Performance: Xception has strong accuracy, precision, and recall along with a well-rounded performance across several criteria. This implies that Xception is a good fit for applications that want to balance several elements of performance.

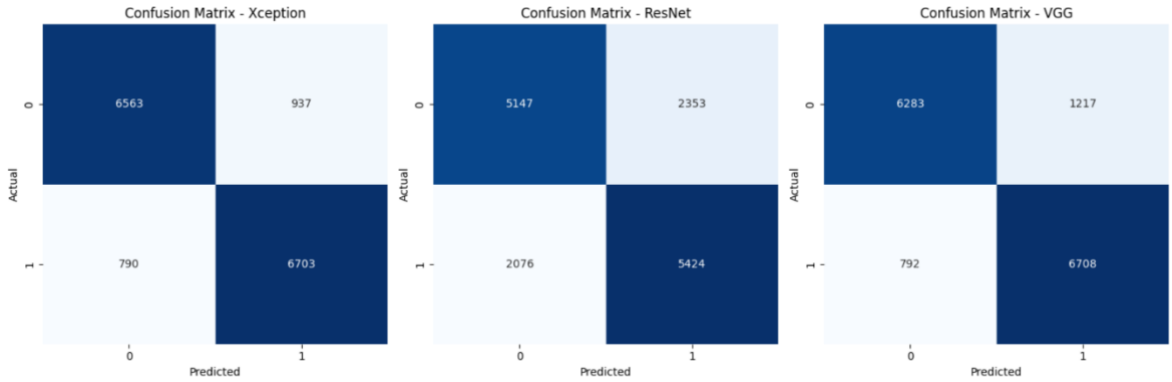

Figure 24 - Confusion matrices

But accuracy and the other metrics are not the only ones that matters, because in real world scenarios the response time plays a crucial role. With an 87.5 ms reaction time, Xception has the quickest response time out of the three models in Figure 25, whereas VGG has the shortest response time (1034 ms). The colors also serve as a good indicator of time efficacy; for example, green denotes a comparably low computational burden for Xception, whereas red denotes a considerably larger computational load for VGG.

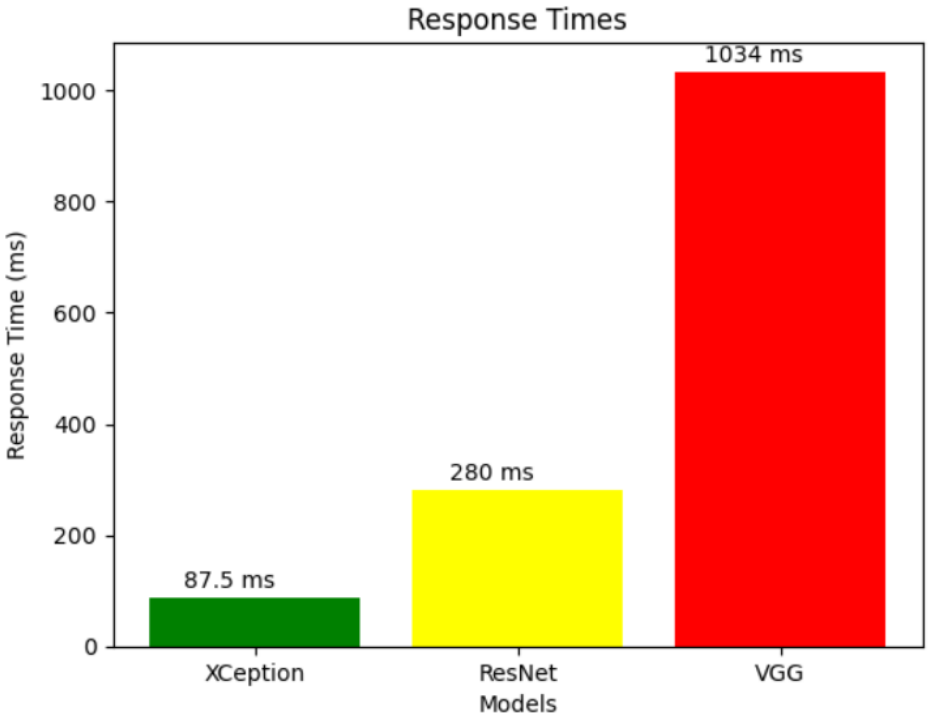

Figure 25 - Response times

By integrating assessment parameters (Accuracy, Precision, Recall, F1-Score) and reaction times, the graphic (Figure 26) offers a comprehensive perspective. This makes it possible for interested parties to evaluate each model's performance and computational effectiveness in a single display.

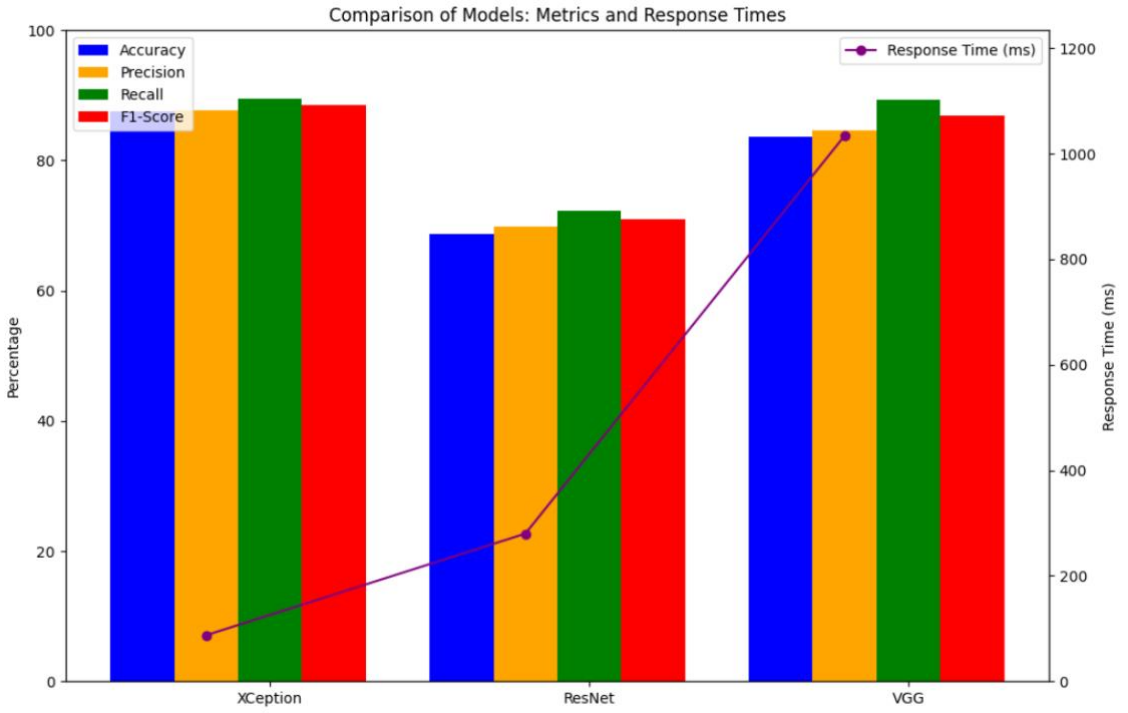

Figure 26 - Comparision of models

To sum up, a thorough analysis of three deepfake detection models—Xception, ResNet, and VGG—reveals unique performance traits that may be tailored to meet different application needs. As the best-performing model, Xception excels in both accuracy and precision, which makes it a great option in situations where reducing false positives is crucial. Its adaptability in attaining a harmonic trade-off between accuracy and recall is further shown by its balanced F1-Score. However, VGG also has a very good recall, which makes it especially useful in situations where reducing false negatives is the main concern. Its slower reaction time, however, implies a speed trade-off that may affect real-time applications. ResNet offers a modest compromise between speed and accuracy, placing it between Xception and VGG in terms of reaction time. This makes it appropriate for cases where speed and accuracy must be balanced. The best model choice ultimately depends on the particular priorities of the given application. Making a decision that is in line with the particular needs of their use case requires decision-makers to consider the significance of accuracy, speed, and striking a balance between false positives and false negatives. Concerning the particular objectives and limitations of the application, the provided visualizations provide a thorough overview, allowing for a more nuanced decision-making process.

5. USE CASE

We thoroughly tested several deepfake detection models in the preceding chapters, and the Xception model turned out to be the best performer with the lowest response times and highest accuracy. After demonstrating its superiority, this chapter explores the useful implementation of the selected model in an actual situation. The primary objective is to showcase the model's effectiveness and reliability when deployed in a user-friendly environment.

The main focus is on a user-centered use case in which users can utilize a web application to take advantage of the deepfake detection technique (Figure 27). By allowing users to add images to the site, a thorough review procedure that establishes the legitimacy of the supplied content is triggered. After that the Xception gives the output identifying the uploaded image as a deepfake or not. This use case reflects the practical utility of the deepfake detection model, addressing real-world concerns related to image manipulation and misinformation.

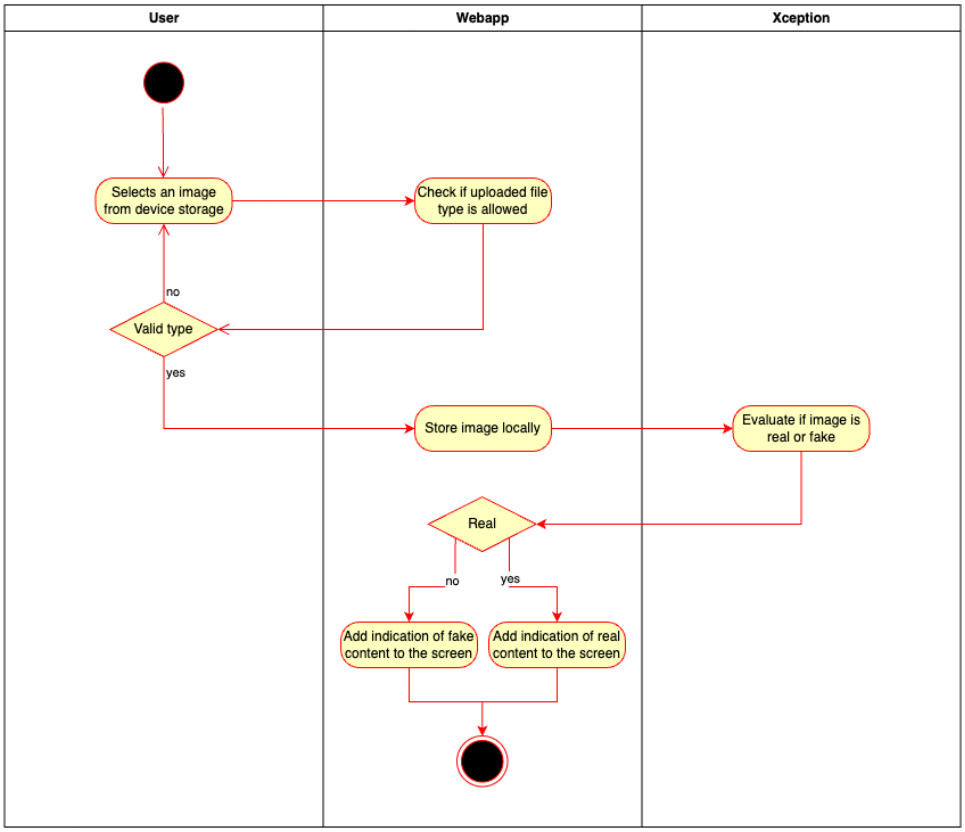

Figure 27 - Activity diagram

5.1. Webapp architecture

The architecture of the web application (Figure 28) comprises two essential layers: the Web App Layer and the Deepfake Detection Layer.

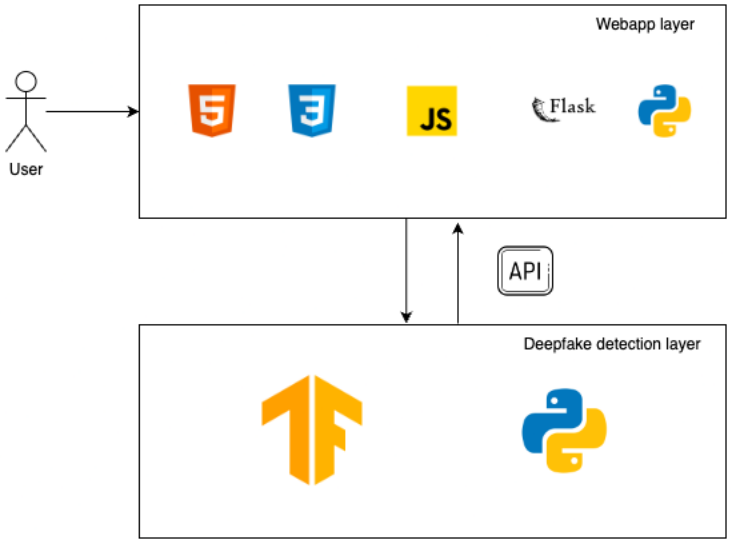

Figure 28 - Webapp architecture

This web app layer is the user interface, built with HTML, CSS, and JavaScript for front-end elements and Flask and Python for backend functionality. The Flask framework handles user requests and helps the client communicate with the deepfake detection layer. The Deepfake Detection Layer, which houses the complex Xception model, is at the core of the system. This layer, built with Python and TensorFlow, focuses on evaluating submitted photos, taking advantage of the Xception model's outstanding accuracy and efficiency in deepfake detection. The link between these two layers is made possible by an Application Programming Interface (API). The API acts as a communication bridge, allowing for easy data transmission and a consistent user experience. Users interact with the web app layer, initiating queries that are processed by the deepfake detection layer, and the findings are transmitted back to the user interface in real-time. This connection ensures that the deepfake detection procedure is efficient and effective within the user-friendly online application.

5.2. Use case demonstration

Upon deploying the developed application, users are afforded the capability to upload a photograph. If a genuine image is submitted by the user, the application subsequently generates the output as illustrated in Figure 29. Beyond confirming the authenticity of the uploaded image, the application additionally provides temporal information pertaining to the processing duration; in this specific instance, the processing time amounted to approximately 460 milliseconds.

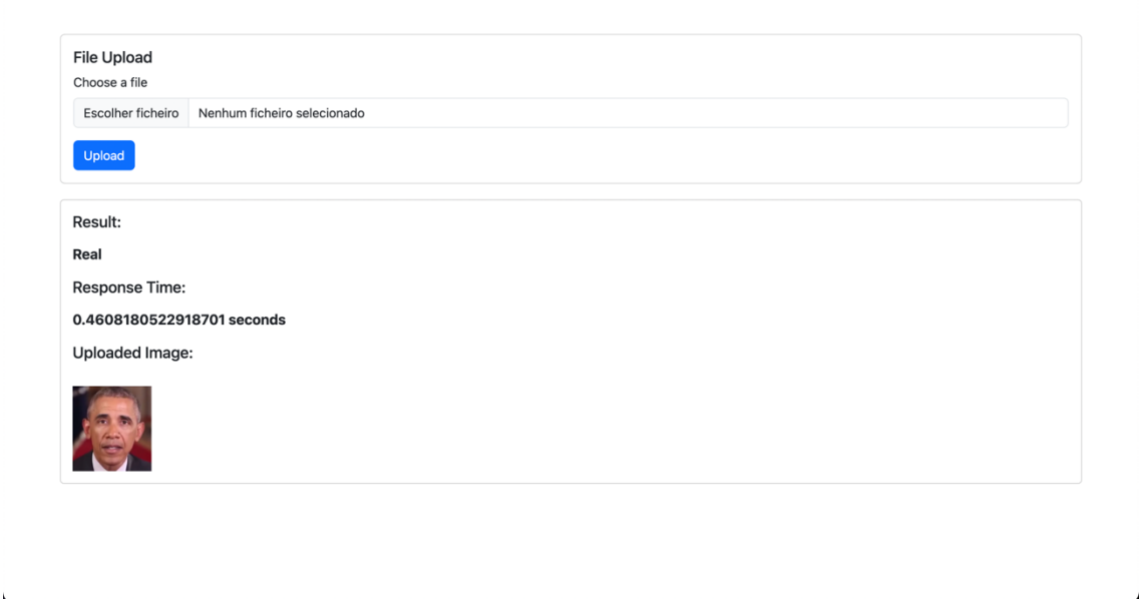

Figure 29 - Webapp - Real scenario

When a fake image is inserted, the web application notifies the user that the image is a deepfake (Figure 30), taking around 160 ms to reach this decision.

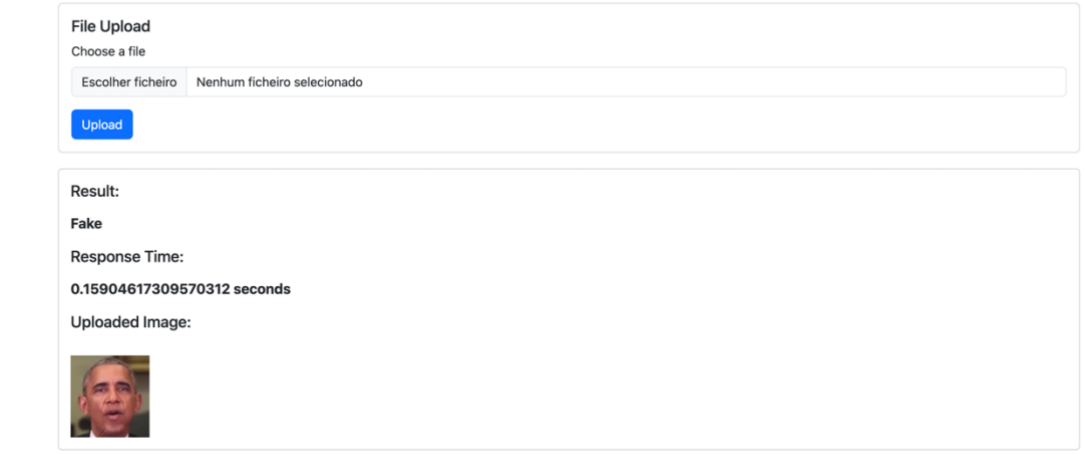

Figure 30 - Webapp - Fake scenario

With this application, despite its simplicity, we were able to conclude that deepfake detection can be done in real time to reduce the propagation of misleading information.

6. CONCLUSIONS AND FUTURE WORK

The journey undertaken in this master thesis traversed the intricate landscape of deepfakes, exploring their generation, the potential advantages they offer, and the inherent dangers they pose. A comprehensive literature review paved the way, shedding light on the evolution of deepfake technology and the diverse set of detection models developed to counteract its malicious applications. In the previous chapters, we carefully assessed Xception, ResNet, and VGG—three well-known deepfake detection models. Extensive testing showed that although every model had its advantages, the Xception model was the best performer when it came to all evaluation parameters (with an accuracy of 87.5%, a precision of 87.7%, a recall of 89.5% and a F1-Score of 88.6%). Its unmatched precision and quick reaction times highlight how effective it is against the constantly changing deepfake threats. The potential for useful deployment was demonstrated by the incorporation of the Xception model into an actual web application. It did, however, highlight the difficulties and areas that could use improvement in real-time situations. These results are compiled in this document, which also recognizes the pertinent contributions of ResNet and VGG and highlights the importance of the Xception model.

Even though this work is a big step forward in the field of deepfake detection, there are still a lot of unanswered questions that beg for more research and development:

- **Model enhancement:** Refinement and fine-tuning of the Xception model to enhance its performance and adaptability to emerging deepfake techniques. Exploration of ensemble approaches, combining the strengths of multiple models for improved detection accuracy;
- **Real-Time Implementation Optimization:** Optimization of the web application architecture to ensure seamless real-time processing. Addressing latency issues and exploring parallel processing techniques to enhance the responsiveness of the deepfake detection system;
- **User Interface and Experience Enhancements:** User interface enhancements to simplify the interaction with the web application. Incorporation of user feedback mechanisms and the integration of explainability features to increase user confidence in the model's predictions.

Finally, this master thesis has contributed with useful insights into the field of deepfake detection, establishing the Xception model as a relevant solution. Looking ahead, the proposed future work will serve as a road map for researchers and practitioners to continue improving the field and strengthening our defenses against the ever-changing world of deepfake threats.

REFERENCES

Aized Amin Soofi & Arshad Awan. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465. https://doi.org/10.6000/1927-5129.2017.13.76

Ak, K., Kassim, A., Lim, J.-H., & Tham, J. Y. (2019). Attribute Manipulation Generative Adversarial Networks for Fashion Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10540–10549. https://doi.org/10.1109/ICCV.2019.01064

Alam, K. Md. R., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*(12), 8675–8690. https://doi.org/10.1007/s00521-019-04359-7

Al-Dhabi, Y., & Zhang, S. (2021). Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, 236–241. https://doi.org/10.1109/CSAIEE54046.2021.9543264

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*.

Bansal, A., Sheikh, Y., & Ramanan, D. (2017). *PixelNN: Example-based Image Synthesis* (arXiv:1708.05349). arXiv. http://arxiv.org/abs/1708.05349

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: Generation, detection, and applications. *International Journal of Multimedia Information Retrieval*, *11*(3), 219–289. https://doi.org/10.1007/s13735-022-00241-w

Dalianis, H. (2018). *Clinical Text Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-78503-5

De Oliveira Albuquerque, R., Villalba, L., Orozco, A., Buiati, F., & Kim, T.-H. (2014). A Layered Trust Information Security Architecture. *Sensors*, *14*(12), 22754–22772. https://doi.org/10.3390/s141222754

Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised Learning Based On Artificial Neural Network: A Review. *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 322–327. https://doi.org/10.1109/CBS.2018.8612259

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The DeepFake Detection Challenge (DFDC) Dataset* (arXiv:2006.07397). arXiv. http://arxiv.org/abs/2006.07397

Ebrahimi, M. R., Li, W., Chai, Y., Pacheco, J., & Chen, H. (2022). An Adversarial Reinforcement Learning Framework for Robust Machine Learning-based Malware Detection. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 567–576. https://doi.org/10.1109/ICDMW58026.2022.00079

Furnkranz, J., & Flach, P. A. (2003). *An Analysis of Rule Evaluation Metrics*.

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *American Journal of Roentgenology*, *212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

He, Y., Yu, N., Keuper, M., & Fritz, M. (2021). *Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis* (arXiv:2105.14376). arXiv. http://arxiv.org/abs/2105.14376

Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., & Ye, D. (2023). Implicit Identity Driven Deepfake Face Swapping Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4490–4499. https://doi.org/10.1109/CVPR52729.2023.00436

Jaleel, Q., & Ali, I. H. (2022). Facial Behavior Analysis-Based Deepfake Video Detection using GAN Discriminator. *2022 International Conference on Data Science and Intelligent Computing (ICDSIC)*, 36–40. https://doi.org/10.1109/ICDSIC56987.2022.10075660

Jayakumar, K., & Skandhakumar, N. (2022). A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors. *2022 7th International Conference on Information Technology Research (ICITR)*, 1–6. https://doi.org/10.1109/ICITR57877.2022.9993294

John, J., & Sherif, B. V. (2022). Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection. *2022 Sixth*

International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 516–521. https://doi.org/10.1109/I-SMAC55078.2022.9987265

Jrad, R. B. N., Ahmed, M. D., & Sundaram, D. (2014). Insider Action Design Research a multi-methodological Information Systems research approach. *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 1–12. https://doi.org/10.1109/RCIS.2014.6861053

Katarya, R., & Lal, A. (2020). A Study on Combating Emerging Threat of Deepfake Weaponization. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 485–490. https://doi.org/10.1109/I-SMAC49090.2020.9243588

Khan, T., Michalas, A., & Akhunzada, A. (2021). Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, *190*, 103112. https://doi.org/10.1016/j.jnca.2021.103112

Khanjani, Z. (2021). *How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey*.

Kocak, A., & Alkan, M. (2022). Deepfake Generation, Detection and Datasets: A Rapid-review. *2022 15th International Conference on Information Security and Cryptography (ISCTURKEY)*, 86–91. https://doi.org/10.1109/ISCTURKEY56345.2022.9931802

Kumar, P., Vatsa, M., & Singh, R. (2020). Detecting Face2Face Facial Reenactment in Videos. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2578–2586. https://doi.org/10.1109/WACV45572.2020.9093628

Lezzi, M., Lazoi, M., & Corallo, A. (2018). Cybersecurity for Industry 4.0 in the current literature: A reference framework. *Computers in Industry*, *103*, 97–110. https://doi.org/10.1016/j.compind.2018.09.004

Li, H. (2018). Deep learning for natural language processing: Advantages and challenges. *National Science Review*, *5*(1), 24–26.

Liu, Y., & Zhou, Y. (2022). An Ensemble Learning Approach for COVID-19 Fact Verification. *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 383–387. https://doi.org/10.1109/ICBAIE56435.2022.9985887

Manzoor, S. I., Singla, J., & Nikita. (2019). Fake News Detection Using Machine

Learning approaches: A systematic Review. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 230–234. https://doi.org/10.1109/ICOEI.2019.8862770

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2022). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03766-z

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/10.38094/jastt1457

Molina, A. C., & Berenguel, O. L. (2022). Deepfake: A evolução das fake news. *Research, Society and Development*, *11*(6), e56211629533. https://doi.org/10.33448/rsd-v11i6.29533

Mukti, I., & Biswas, D. (2019). *Transfer Learning Based Plant Diseases Detection Using ResNet50* (p. 6). https://doi.org/10.1109/EICT48899.2019.9068805

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Niu, S., Liu, Y., Wang, J., & Song, H. (2020). *A Decade Survey of Transfer Learning (2010–2020)*. *1*(2).

Pan, D., Sun, L., Wang, R., Zhang, X., & Sinnott, R. (2020). *Deepfake Detection through Deep Learning* (p. 143). https://doi.org/10.1109/BDCAT50828.2020.00001

Patel, M., Gupta, A., Tanwar, S., & Obaidat, M. S. (2020). Trans-DF: A Transfer Learning-based end-to-end Deepfake Detector. *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, 796–801. https://doi.org/10.1109/ICCCA49541.2020.9250803

Ramachandran, S., Nadimpalli, A. V., & Rattani, A. (2021). An Experimental Evaluation on Deepfake Detection using Deep Face Recognition. *2021 International Carnahan Conference on Security Technology (ICCST)*, 1–6. https://doi.org/10.1109/ICCST49569.2021.9717407

Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., & Busch, C. (Eds.). (2022). *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*.

Springer International Publishing. https://doi.org/10.1007/978-3-030-87664-7

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Rismi, R., Endah, S., Khadijah, K., & Shiddiq, I. (2020). *Xception Architecture Transfer Learning for Garbage Classification* (p. 4). https://doi.org/10.1109/ICICoS51170.2020.9299017

Saifudin, A. M., Yacob, A., & Saad, R. (2016). The Facebook-in-action: Challenging, Harnessing and Enhancing Students Class Assignments and Projects. *Universal Journal of Educational Research*, *4*(6), 1259–1265. https://doi.org/10.13189/ujer.2016.040602

Salvi, D., Hosler, B., Bestagini, P., Stamm, M. C., & Tubaro, S. (2023). TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access*, *11*, 50851–50866. https://doi.org/10.1109/ACCESS.2023.3276480

Samonas, S., & Coss, D. (2014). *THE CIA STRIKES BACK: REDEFINING CONFIDENTIALITY, INTEGRITY AND AVAILABILITY IN SECURITY*.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science*, *2*(3), 173. https://doi.org/10.1007/s42979-021-00557-0

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Tammina, S. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(10), p9420. https://doi.org/10.29322/IJSRP.9.10.2019.p9420

Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2018). Deep Recurrent Neural Network for Intrusion Detection in SDN-based Networks. *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, 202–206. https://doi.org/10.1109/NETSOFT.2018.8460090

Vajpayee, H., Yadav, N., Raj, A., & Jhingran, S. (2023). Detecting Deepfake Human Face

Images Using Transfer Learning: A Comparative Study. *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, 1–5. https://doi.org/10.1109/InC457730.2023.10263216

Verma, A., Gupta, D., & Srivastava, M. K. (2021). Deepfake Detection using Inception-ResnetV2. *2021 First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT)*, 39–41. https://doi.org/10.1109/ICACFCT53978.2021.9837351

Xia, Z., Qiao, T., Xu, M., Wu, X., Han, L., & Chen, Y. (2022). Deepfake Video Detection Based on MesoNet with Preprocessing Module. *Symmetry*, *14*(5), 939. https://doi.org/10.3390/sym14050939

Yang, X., Li, Y., & Lyu, S. (2019). Exposing Deep Fakes Using Inconsistent Head Poses. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261–8265. https://doi.org/10.1109/ICASSP.2019.8683164

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing* (arXiv:1702.01923). arXiv. http://arxiv.org/abs/1702.01923

Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, *81*(5), 6259–6276. https://doi.org/10.1007/s11042-021-11733-y