REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

DEEP LEARNING-DRIVEN CLASSIFICATION OF DEPTH IMAGE-BASED RENDERING AND DYNAMIC FACE WARPING IN DEEPFAKE ALTERED VIDEOS FOR FABRICATED NEWS ON SOCIAL MEDIA

Dunya Ahmed Aola ALKURDI

Doctor of Philosophy

Supervisor

Asst. Prof. Dr. Mesut ÇEVİK

İstanbul, 2025

DEEP LEARNING-DRIVEN CLASSIFICATION OF DEPTH IMAGE-BASED RENDERING AND DYNAMIC FACE WARPING IN DEEPFAKE ALTERED VIDEOS FOR FABRICATED NEWS ON SOCIAL MEDIA

Dunya Ahmed Aola ALKURDI

Electrical and Computer Engineering

Doctor of Philosophy

ALTINBAŞ UNIVERSITY

2025

The dissertation titled DEEP LEARNING-DRIVEN CLASSIFICATION OF DEPTH IMAGE-BASED RENDERING AND DYNAMIC FACE WARPING IN DEEPFAKE ALTERED VIDEOS FOR FABRICATED NEWS ON SOCIAL MEDIA prepared by DUNYA AHMED AOLA ALKURDI and submitted on 30/06/2025 has been **accepted unanimously** for the degree of PhD in Electrical and Computer Engineering.

_____

Asst. Prof. Dr. Mesut ÇEVİK

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Mesut ÇEVİK	Department of Electrical-Electronics Engineering, Altınbaş University	_____
Prof.Dr. Osman Nuri UÇAN	Department of Electrical-Electronics Engineering, Altınbaş University	_____
Asst. Prof. Dr. Hameed Mutlag Farhan FARHAN	Department of Computer Engineering, Altınbaş University	_____
Assoc. Prof. Dr. Abdurrahim AKGÜNDOĞDU	Department of Electrical-Electronics Engineering, İstanbul University- Cerrahpaşa	_____
Asst. Prof. Dr. Ameer Yalmaz Asaad ASAAD	Department of Computer Technologies, İstanbul Topkapı University	_____

I hereby declare that this dissertation meets all format and submission requirements for a PhD thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Dunya Ahmed Aola ALKURDI

Signature

DEDICATION

I devote and pledge this research work to my supervisor who is salient for guiding me through whole research work as well as my family for always assisting me in my hard time.

PREFACE

First and foremost, I would like to thank my supervisor Asst. Prof. Dr. Mesut ÇEVİK for guiding and helping me along the way in writing this dissertation. Discussing my progress, problems, and ideas with my supervisor Asst. Prof. Dr. Mesut ÇEVİK a couple of times every week helped me tremendously in understanding the logic behind the research. It made me better realize the technical need for this research work.

ABSTRACT

DEEP LEARNING-DRIVEN CLASSIFICATION OF DEPTH IMAGE-BASED RENDERING AND DYNAMIC FACE WARPING IN DEEPFAKE ALTERED VIDEOS FOR FABRICATED NEWS ON SOCIAL MEDIA

ALKURDI, Dunya Ahmed Aola

Ph.D., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Mesut ÇEVİK

Date: June /2025

Pages: 80

Deepfakes present a persistent challenge in a world driven by transparency and truth. This study introduce a new deepfake detection method based on Xception architecture. Deepfake is having some threats to the authenticity of digital information especially on social media platforms through manipulated content, which transpires fast within these networks and powerfully influences the public's opinion, reputational harm, and fuel of misinformation. This research introduced an Xception-based approach to deepfake detection. The architecture utilizes an efficient feature extraction CNN model that benefits from the improvements such as depthwise separable convolutions and is optimized to exploit subtle artifacts in typical deepfake media, such as inconsistencies in facial landmarks, lighting irregularities, and unnatural textures. It was trained and tested on an all-inclusive dataset of more than 500,000 frames comprising authentic and manipulated video content. The data preprocessing steps included face extraction and alignment and diversity augmentation of the data to make generalization better. The model was evaluated using various metrics and scored quite impressively at 99.69% accuracy, precision at 99.58%, recall at 99.80%, and with an F1 score at 99.69%. These results demonstrate that the model is correctly balanced to avoid false positives at a reasonable tradeoff for identifying deepfake content well enough to be deployable in real-world scenarios where content verification will prove critical. The

Xception-based model also shown strong computational efficiency with the average inference time being 0.08 seconds per frame. With an AUC-ROC score of 0.9999, this efficiency combined proves that the model can significantly distinguish the real from the manipulated frames with near-perfect accuracy. More so, its robustness was tested under various conditions, including different resolutions and compressions. It achieved an accuracy of 99.65% on low-resolution frames and 98.1% on high-resolution frames, with an accuracy of greater than 90% even for highly compressed videos that ascertained its adaptability to the diverse media qualities characteristic of content usually found on social media. In summary, a deepfake detection framework based on Xception offers a powerful, accurate, and efficient solution to deepfake media, supporting the integrity and authenticity of digital content on social media. As possible future directions of this work, it could be considered to further explore the adversarial training for more robustness against emerging deep fake generation techniques and expanding the dataset with occluded and diverse samples. This research is therefore a valuable contribution to the area of deep fake detection, offering scalability of the tool to reduce misinformation and protect public trust in digital media.

Keywords: Deepfake Detection, Xception Architecture, Data Processing, Image Processing, Intelligent Information Systems, Social Media Security.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
GANs	:	Generative Adversarial Networks
CNN	:	Convolutional Neural Network
RNN	:	Recurrent Neural Networks
GAP	:	Global Average Pooling
DCNN	:	Deep Convolutional Neural Network
SVM	:	Support Vector Machine
XAI	:	Explainable AI
SPSL	:	Spatial-Phase Shallow Learning

1. INTRODUCTION

1.1 OVERVIEW

Recently, deepfake technology has become one of the greatest concerns in digital security and authenticity, especially on social media platforms where fabricated videos and manipulated images spread very fast. Deepfake is defined as a term referring to "deep learning" and "fake," which means media content altered to give an impression that someone says or does something they never said or do as mentioned in [1]. This is made possible by artificial intelligence (AI) and deep learning applications used in the manufacturing of hyper-realistic images and videos. They mostly succeed in deceiving the audience into considering them as authentic. In the speedy evolution of deepfake technology, it has brought along challenges across all social, political, and economic spectrums as mentioned in [2]. Its malicious applications involve the use of this technology as a means of disinformation, which influences public opinion negatively and causes erosion of trust in digital information sources as mentioned in [3]. As deepfake threats become more threatening and sophisticated, recognizing these forms of manipulated media in real-time becomes more imperative. The biggest challenge here is to develop a tool that distinguishes with reliability between authentic and fake media despite the advancement in the generative models used to create deepfakes as mentioned in [4]. The current detection techniques do not match the degrees in which deepfake generators are changing. Thus, reality and fantasy can go virtually unnoticed to the naked eye. This incites new advanced frameworks of detection capable of working within a variety of digital environments, particularly on social media where information is spread rapidly and efficiently as mentioned in [5].

This thesis introduces a more powerful deepfake detection approach based on the xception architecture, in which there have been strong capabilities for image feature extraction and stable training dynamics. It was first presented with the purpose of image classification, exhibiting promising results in various domains and particularly well suited to the intricate task of deepfake detection, where it may extract subtlety within inconsistencies of the fabricated images and videos. The results are founded on novel detection mechanisms using the Xception model. This reflects a step forward in data preprocessing techniques and testing mechanisms with high accuracy and reliability at the time of processing datasets including

millions of frames and diverse samples. The research conducted is exploratory in nature, taking advantage of the unique architecture of Xception with depthwise separable convolutions that improve feature extraction without much increasing the computational costs as mentioned in [6]. This architectural advantage is the reason why the Xception model is especially effective in detecting deepfakes-it could pick on small variations and differences in facial expressions, lighting inconsistency, and unnatural transitions commonly seen in manipulated media. Stable training mechanisms will also include early stopping in the research to maximize the performance of the detection model while working to minimize overfitting and computational overheads. In addition, research explores memory-efficient testing approaches in order to allow real-time applications and ensure that the model could be successfully deployed on multiple social media platforms and virtual environments emphasizing speed and accuracy as mentioned in [7].

This thesis studies ensemble models to improve the accuracy of detection. This approach constitutes a synergy in combining different frameworks with various detection strategies toward improving the overarching robustness and adaptability of the detection system. This ensemble approach is motivated by understanding that deepfake detection is a complex problem that will likely benefit from a multi-model strategy, where different models contribute complementary strengths to identify various types of deepfakes. This comprehensive approach not only boosts detection accuracy but also provides a foundation for scalable and adaptable detection systems that can evolve with growing deepfake generation technologies as mentioned in [8].

Going beyond the technical efforts, the present research addresses ethical and social implications of deepfake detection. The potential widespread distribution of deepfakes is suspected to carry high risks for private lives, public trust, and information authenticity in the digital domain as a whole. Therefore, the present study complements the overarching motivation of technically effective yet ethically sound, transparent, and privacy-conscious detection systems. This research advances the interpretation of detection models, where users understand the basis for model decisions. Overall, it is based on the basis for decisions which will shape building trust and reliability in automated detection systems. This research, by adding to a broader societal objective of safeguarding the digital information ecosystem, adds to encouraging greater awareness of deepfake threats and ethical standards within AI

applications as mentioned in [9]. Producing a high-confidence, Xception-based detection model that can run in real time and on any number of differing digital platforms is the goal of this study. Then, by deepening the technical, ethical, and social implications of detection, this dissertation can serve as an important contribution to the ongoing fight against false information spread and manipulation of digital operations in today's rapidly moving cyber environment as mentioned in [10].

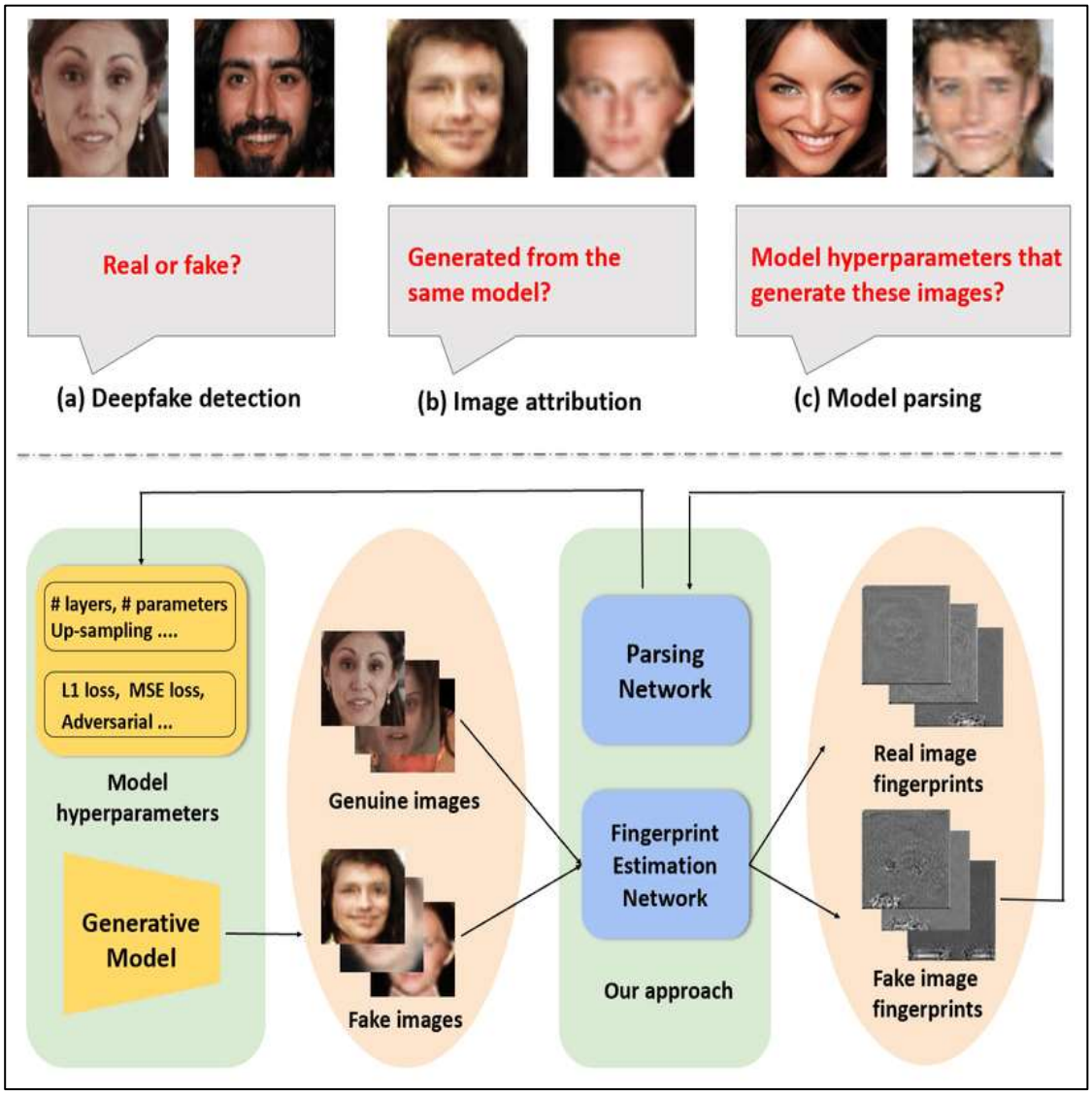

Figure 1.1: Overview of Deepfake Detection, Image Attribution, and Model Parsing with Fingerprint Estimation [10].

One of the critical challenges currently facing the integrity of media in the context of the digital age is the potential to exploit it to manipulate, especially where such a medium happens to be visual [11]. New generation technologies, such as deepfakes, have so far

allowed it to manipulate the visual material with a high degree of accuracy and plausibility which may be very significant to people, public opinion, and even objective truth [12]. Deepfakes are a menace to society and may even fool the public, malign persons and organizations, and modify facts, which makes them a weapon for influencing public opinion. The technologies used in these problems through face manipulation include Face2Face [13]. The objective of this study is the development of a sustainable method of detecting deepfakes that employ the use of Xception. We, therefore, employ more developed methods that had been used before, as well as better techniques of feature extraction and trusted training methods with the aids of early stopping, in a test to determine the changed material at a very high accuracy level. Hence, we are working on creating an efficient system that is able to detect real from fake graphics within the social networks. This approach is much more efficient for use because, unlike the other ways, specific calibrations and enhancements help it identify deepfake videos within the social media platforms. Actually, the procedures incorporated bettering of data preparation, more stringent testing of memory, and the use of explanatory approaches in explanation of the model's decision-making process, none of which has been found before.

1.2 MOTIVATION

It is based on the urgent need to counter the spread of deepfakes, which is a unique and evolving threat to the integrity of online information. The growth of deepfake technology has led to an increase in cases where manipulated media is used for malicious purposes, especially in political disinformation, character defamation, and financial fraud. Indeed, because most social media-based news and information are reaching practically everybody, rampant deepfake spread across such platforms increased its impact even more; this put researchers, policymakers, and tech companies to innovative solutions for the risks associated with deepfakes. The methods for deepfakes detection that have been thus developed after much progress in AI and machine learning had encountered issues related to scalability, adaptability, as well as performance. With most of the deepfake models currently available requiring heavy computation, especially in social media applications where response times need to be quite high, most of them are rendered inapplicable in the real world, too as mentioned in [14]. In addition to this, deepfake generation techniques are rapidly being improved and perfected, which challenges the robustness of most detection

frameworks; therefore, it's important to develop models that can evolve with these developments.

Addressing the above challenges, this research uses the Xception architecture particularly efficient and powerful for feature extraction. Based on the architecture of Xception, the hypothesis is that it may bring better accuracy, efficiency, and scalability to deepfake detection systems, which could be more deployable in high-stakes environments such as social media. This research therefore contributes to the effort of restoring trust and authenticity in digital information by developing a model that balances technical efficacy with ethical considerations as mentioned in [15].

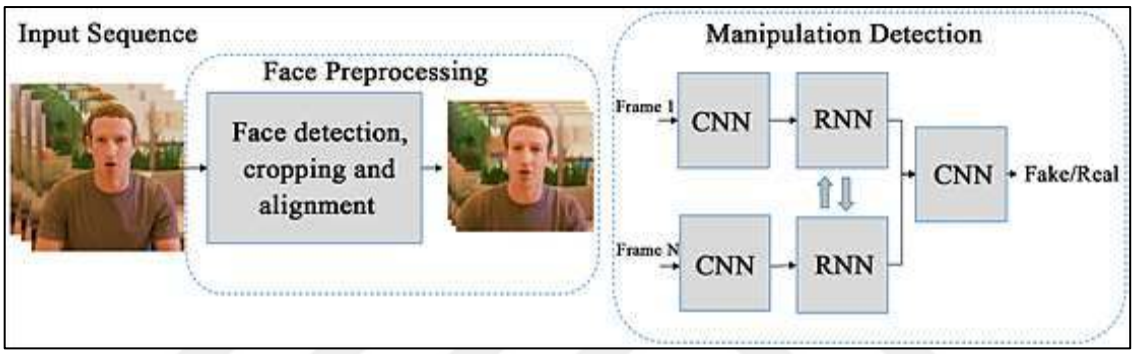

Figure 1.2: Deepfake Detection Pipeline Involving Face Preprocessing and Manipulation Detection Using CNN and RNN Models [15]**.**

A related very important issue with detection systems created using current deepfake detection models is non-interpretableness or lack of transparency. In most cases, the basis of a detection system's decision is hard for users or platform moderators to understand. At other times, there is just plain disbelief in automated systems as mentioned in [16]. An inability to verify the reasoning behind a model's judgment reduces the acceptance and adoption of detection tools by stakeholders. Developing the model so that it can explain decisions clearly for fostering greater trust will encourage more use of detection systems. Such problems have been addressed in this research by proposing an approach to deepfake detection based on the Xception architecture, which is said to produce powerful features along with efficient computational design through CNN models. Taking depthwise separable convolutions as its basis, the Xception model promisingly develops a detection system that captures subtle inconsistencies in the manipulated media while keeping the computations efficient. This research draws on advanced data preprocessing techniques and the architecture of Xception to build a robust, scalable, and interpretable deepfake-detection framework tailored for real-

time execution on social media as mentioned in [17].

1.3 CHALLENGES OF RESEARCH

Many problems are there with the development of a strong deep fake detection model, mainly because generative media and techniques are highly evolving. Herein, one of the most significant challenges is to stay ahead of the curve in terms of improvements from deepfake generation methodologies that are given by advancements in deep learning architectures like Generative Adversarial Networks. As GANs and similar synthesis models become capable of generating highly realistic media, a detection method must continue evolving to identify ever subtler signs of manipulation, most indistinguishable both to human observers and to standard machine learning models as mentioned in [18].

The other challenge is that it indeed has much higher computationally demanding requirements for deepfake detection. Models like Xception, which offer very powerful feature extraction capability, are ones that often require processing power and memory, especially in terms of training and deployment on large volumes of video and image content. Balancing accuracy and computational speed to prevent system lags in real time, especially on high-traffic social media platforms, adds to the complexity of making sure the model runs effectively as mentioned in [19].

Secondly, are the problems of data quality and variability as the detection models for deepfakes need a lot of diverse datasets to generalize well across all scenarios and formats. Furthermore, high-quality datasets representing the complete spectrum of techniques in deepfake manipulation may be difficult to obtain since some deepfakes utilize novel and obscure methods unknown to current data. Unless ample variability is available in the training set, models might end up being ineffective or possibly failing at detecting deepfakes whose creating methods may be unknown or emerging techniques as mentioned in [20].

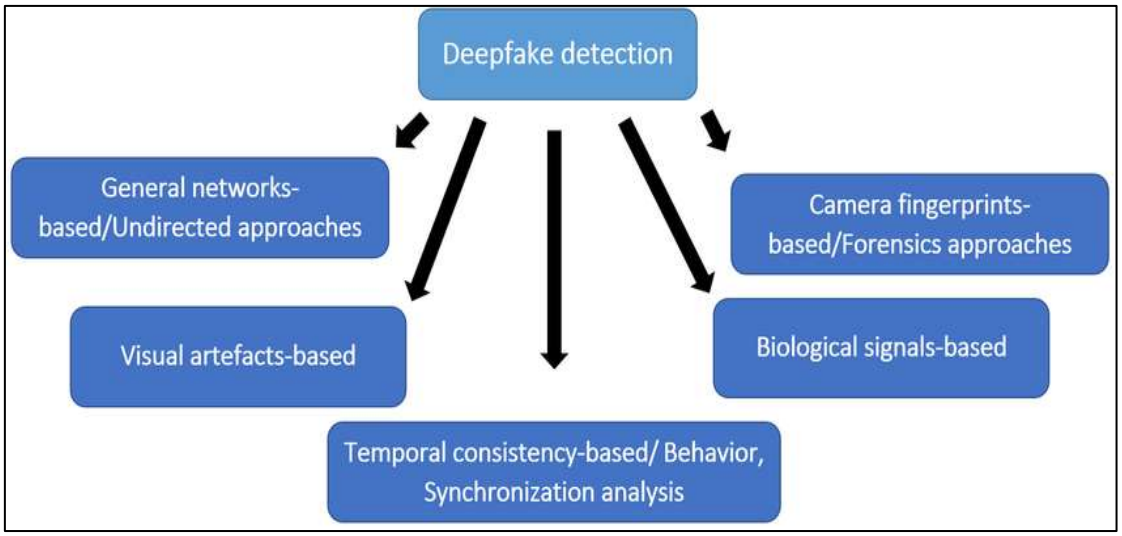

Figure 1.3: Overview of Deepfake Detection Approaches Based on Different Techniques [20].

The interpretability of the detection model is another critical challenge. For applications like social media, with decisions regarding flagging or removal of content under closer scrutiny, it is very essential that the results obtained by the detection models are interpretable and explain the basis for their classification decisions. This is a tough aspect to impose in deep learning models, especially when trying to design architectures with the flavour of Xception. Lack of interpretability might hinder a certain level of trust from being generated in automated systems. Thereby, it influences the acceptance level by the users. The model could be practically useful in real applications only when it is explainable. Domain adaptation of the deepfake detection models is another technical issue. They are trained on one kind of media or platform and may not perform well on another. The same can raise further issues while deploying a detection system across multiple social media platforms where various kinds of content formats, compression algorithms, and presentation styles are being used. The model needs to adopt itself to various environmental conditions without compromising the accuracy of the system, which poses tough tuning requirements often coupled with platform-specific customizations as mentioned in [21].

Moreover, ethical issues and privacy concerns make the deployment of deepfake detection models more complex. The fine line between malicious content identification and respecting the privacy of legitimate users falls under the kinds of user reassurance that should be established against excessive surveillance or unnecessary detections. The best solution is still far from clear within the limits imposed by the current methodologies of deep learning

for systems that are meaningful in countering misinformation but respectful of privacy. Detection models need periodic updating and retraining because deepfake technology is ever-evolving to overcome new threats. This type of updating is resource-intensive and requires full-time watchfulness and adaptability. But this also increases the complexity of the task because it can be adversarial attacked: well-designed deepfakes are intentionally designed and made by malicious actors, specifically to outsmart detection algorithms, or in other words, make the model fail-through intentional evasive tactics which might make it less effective over time as mentioned in [22].

1.4 PROBLEM STATEMENT

Deepfakes, therefore, are the new force of information, revealing a dangerous threat to the authenticity of content-especially content in social media, which is very susceptible to counterfeit content that can spread really fast and thus influence public opinion. These deepfaked media generally refer to manipulated images, audio, or video recorded in a manner created via deep learning techniques, and they grow increasingly realistic and hard to detect. Originally, this technology was designed for entertainment and creative industries but was unfortunately misused and weaponized in harmful ways-political manipulation, identity fraud, defamation-to become a critical factor in the misinformation vectors now eroding the world's confidence in digital content and integrity of global information. Social media is probably one of the most susceptible to deepfake-driven misinformation, as it is also the medium most paramount for information and public discourse as mentioned in [23]. As opposed to other news sources, social media does not provide strict checks on the validation of content, with hundreds of millions of viewers freely viewing manipulated videos and images. This makes it a breeding ground for deepfakes where manipulated media could be spread throughout the globe before efforts for detection come into force. This enhances the impact of deepfake content with its power and speed: confusion, reputational damage and even socio-political effects will be produced as mentioned in [24].

Detection methods have been developed to tackle this challenge. However, the proposed detection models currently face major challenges regarding their ability to be highly accurate and scalable to real time performance. Most detection methods, which rely heavily on CNNs and machine learning, are unable to keep pace with the advancement of generative models in creating deepfakes. The current generation methods have become sufficiently

sophisticated such that deepfakes produced will not be easily discerned from authentic media by human observers or well-designed automated detection systems. This translates to an increased necessity for identification models that would not miss even the slightest inconsistencies within manipulated content while maintaining good performance over different data sets and formats. Additionally, most deepfake methods are very computationally intensive; therefore, they could take considerable processing time and power and hence may limit the application in real-time environments, such as social media platforms. But it is the ability to detect deepfakes in real-time that is critical to slowing, at least in part, the mass dissemination of damaging material and thereby limiting its damage. Such current detection frameworks have high computation requirements, which hinder their implementations in platforms processing massive media data every second. This leads to a requirement for a more efficient, scalable, and reliable detection approach that is capable of effectively performing within the constraints of social media environments as mentioned in [25].

Evidence that the computer vision and deep learning industries have grown rapidly poses a challenge, for one may not be certain whether a content has been altered or not. Furthermore, Face2Face and Deepfakes technologies have limited the boundaries of what can be achieved in this duality. And also, no effective procedures or tools that might be applied in the detection of facial deepfakes have been established and also lacking is the complementation of this condition with standardized standards with which the methods of detection may be evaluated. This therefore takes forward the science of digital media forensics with an attempt on coming up with a highly efficient approach in handling automated cases much better than manifestations of humans with an effort to spot facial modifications as mentioned in [26].

a. Escalating Deepfake Threat: Deepfake technology is increasingly used to create highly realistic but manipulated media, threatening the authenticity of information regarding its use, especially when it relates to an environment for social media, in which the dissemination of misinformation is fast.
b. Influence on social media: Because they primarily represent sources of news and public discourse with little validation of content, social media are vulnerable to deepfake-driven misinformation with fake media reaching great crowds quickly and with hugely detrimental socio-political impacts.

c. Limitation of current detection techniques: Current deepfake detection techniques, normally developed by means of CNNs, are not very accurate; they cannot increase their ability without sacrificing high performance in the capability of keeping with new development in the application of deepfakes generation techniques-that are able to produce media which are nearly indistinguishable from authentic ones.
d. The detection challenge of real time. Most detection systems are computationally challenging and, thus, cannot support real-time applications on platforms that process large volumes of data every second. This limits their effectiveness in halting the spread of deepfake content within a timely manner.
e. Interpretability and transparency: Most the existing models for detection result in what amounts to a lack of interpretability, further hindering the users' and moderators' ability to understand the actual decisions made by the system while also inspiring mistrust in the automated systems. Hence the demand for transparent and interpretable detection frameworks is well established.
f. Detection Method by the Xception Architecture: According to the effective feature extraction capability of the Xception model, this work aims to create a robust yet scalable detection framework in computation light that could be well applicable for real-time applications on social media.
g. Ensemble Method for Clearer Detection: This work makes sure it ascertains ensemble techniques that combine multiple detection frameworks together for improved adaptability and detection accuracy for all deepfake media types.
h. Ethical and Social Considerations This research does contribute to the more general goal of protecting the integrity of digital information and restoring public trust that's being leveled by the growing danger of deepfakes in the development of a technically advanced yet interpretable detection model.

1.5 AIM OF CONTRIBUTION

The overall goal of this research is to evolve a deepfake detection model based on the Xception architecture that is robust, scalable, and interpretable to counter the increasing threat of fabricated media, especially in today's social media space. With the advancement in deep fake technology, there comes a prospect of misusing it for spreading misinformation, swaying public opinion, and denting reputations. This research answers these critical

problems by introducing the possibility of using a detection system that has the ability to distinguish genuine from manipulated media, thus securing the integrity of information in social media and other social digital environments while making sure that users are able to place trust in such social platforms.

In this research, we use Xception architecture, known for efficient feature extraction, along with depthwise separable convolutions, in order to increase the accuracy of and speed in detection. By exploiting this model, the research aims at overcoming the stringent computational and scalability problems involved in many detection systems currently in place, which often fail to work effectively in real-time environments. The fast flow of digital ecosystems demands quick and accurate detection to prevent the viral spread of harmful and misleading content. The advanced preprocessing techniques and the memory-efficient testing mechanism are going to be incorporated while pursuing the research so that the result not only proves to be sound but also adaptable to various kinds of media content and diverse datasets. Ensemble approaches in combining multiple detection frameworks lead to an improvement in the adaptability as well as the robustness of the model when it is presented with different forms of deepfakes.

This study foregrounds interpretability and transparency in the detection model so as to address some of the ethical concerns with the system of automated detection. In an effort to seek trust for the AI-driven detection tools, the proposer is designing an interpretable model. Such a model would better equip users with an understanding of the basis for detection decisions and foster responsible use of AI. This study, for instance, considers the broader effects that deepfakes might have on society and seeks to increase public sensitivity and participation towards desirable ethical standards of media integrity in this context of the digital world. With the sophistication of such dastardly acts to a point of concern, the essay is a fit contribution towards the budding development of digital forensics. Where human observers can place it firmly beyond any possible dispute, it needs reliable and automatic methods of detection of facial alteration even by significant margins. Our approach grounds on the latest innovations in the field of deep learning especially on extremely powerful capabilities of extracting features from images of convolutional neural networks (CNNs). The detection phase is something that we are working on by the supervised development of a gigantic dataset that is supposed to shape the neural network comprising standard

approaches applied here in the domain of computer graphics, such as Face2Face. The primary three of them include FaceSwap and techniques of deep learning methods such as DeepFakes and NeuralTextures. Importantly, no benchmark standard is applied to forgery detection within the field of digital media forensics. To fill this gap, we introduce a novel automated benchmark that incorporates the four above mentioned modification techniques in a realistic scenario. Some of our research goals are as follows:

a. Developing high accuracy, reliable manipulated media identification using the architecture of Xception.
b. A high degree of testing across a representative set of video samples and frames with regard to its robustness and generalizability.
c. Further investigate ensemble approaches that combine multiple detection frameworks to enhance detection accuracy and adaptability.
d. Creation of an interpretable and transparent detection model to manage the ethical issues, mainly for creating an atmosphere of trust and awareness in the minds of users.
e. Aiding to create tools that would work robustly in real-time detection, especially at the social media sites where impacts of deepfakes are highly felt.

1.6 ORGANIZATION OF THESIS

The paper is structured to provide detailed exposition on deepfake detection using Xception architecture to guide the reader from foundational concepts to detailed experimental results and their implications for safeguarding against misinformation on social media.

Chapter 1: Introduction

This chapter begins with an introduction to deepfakes, which is causing ripples in the misinformation domain, especially in social media. Then, it focuses on how a good and real-time detection mechanism would make everything else possible in deepfakes. Later, this research work and its objectives along with the impact it would make are briefly stated. This chapter lays down the baseline for the thesis by describing the relevance of this problem as well as the potential of Xception-based models in that domain for accurate and scalable detection.

Chapter 2: Literature Review

A discussion of previous work carried out so far on deepfakes is presented as a literature review, where CNN and other machine learning methods are discussed about their

applicability in handling deepfakes. Briefing of key studies on the Xception model with its applications in image and video analysis form a foundation for how the sphere of deepfake detection has evolved so far and what the current limitations are. This chapter also illustrates a gap to be covered within this study, situating the research in the larger context of cybersecurity and information integrity.

Chapter 3: Methodology

The proposed methodology-data collection, preprocessing techniques, and design of an Xception-based detection model is what this chapter outlines. It elaborates on data augmentation, specificity in the structure of the Xception architecture and training processes, including early stopping mechanisms to avoid overfitting, with ensemble methods for higher accuracy and adaptability. It also discusses what evaluation metrics are being used to measure model performance and describes the strength and reliability of the proposed approach.

Chapter 4: Experimental Results and Analysis

In this chapter, the experimental setup has been described, and results of the detection model's performance can be found. To validate the efficacy of a model in the identification of deepfake content from various datasets, different performance measures such as accuracy, precision, recall, and F1-score are used. To put these advantages in context, a comparative analysis has been conducted with other deepfake detection techniques. The error analysis will actually bring out areas where it is wrong and challenges during testing.

Chapter 5: Discussion

In this chapter, the experimental setup is described and results of the detection model's performance. Different measures of performance, including accuracy, precision, recall, and F1 score, are utilized to validate the efficiency of proposed models in identifying deep fake content on different datasets. A comparative study with other existing techniques of deepfake detection is carried out to set the advantages of this Xception-based model against it, and an error analysis will reveal possible avenues for improvement as well as the challenges that occurred during testingThis chapter interprets the findings of the experimental analysis, examining the meaning of the results, and discussing the current model's limitations. In this chapter, the problems that occur during the research process, including those about time issues of real-time detection, and generalizability of the model, are discussed. Last but not

least, this chapter also covers possible implications concerning the deployment of this model on social media pertaining to the ethics and privacy concerns about deepfakes.

Chapter 6: Conclusion and Future Work

Finally, this chapter summarizes the study's contributions on deepfake detection and reiterates how strong, scalable detection frameworks best position themselves for information authenticity as a critical interest. Lastly, it proposes recommendations for further research on improving model architectures, integrating multi-model frameworks, as well as the developing of real-time detection systems adaptable to the advancing deepfake technologies. The long-term implication of deepfake detection for digital integrity and public trust is strongly underlined in this chapter as well.

2. LITERATURE REVIEW

Alongside the present advancements in computer vision and machine learning, several films, deepfakes, and hyper-realistic landscapes are available for exploration across various applications, as referenced in [27]. They illustrate the application of manual techniques and serve as a crucial verification layer for the analysis of facial nuances utilizing advanced deep learning architecture. To combat the proliferation of deepfake films, it is essential to develop temporal-conceptual detection discriminators, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This study demonstrates that deep learning systems can identify counterfeit portions in videos with an accuracy of 97%. Given the proliferation of creative forgeries in media such as music and film, there exists a biometric forensic technique for evaluating the authenticity of fabricated content generated through face swapping. Counter the Proliferation of Deepfake [28] presents a methodology utilizing capsule networks capable of detecting several forms of forgery with an accuracy exceeding 90%. We present a novel convolutional neural network (CNN) multi-task learning model in [29] capable of simultaneously detecting intrusive images and videos while recognizing areas with point-wise modifications. After optimization and evaluation on the FaceForensics and FaceForensics++ datasets, the proposed method achieves segmentation rates of 93.11% and classification rates of 83.71%, respectively. exhibiting its malleability, which yields nearly unaltered improvements to capabilities. The research data, obtained from comprehensive performance evaluations, may assist in determining the efficacy of fine-tuning efforts. This study concurrently elucidates the key characteristics of real-life scenario modification strategies employed in media sharing as amended [30]. Recent advancements in deep learning and computer vision have significantly enhanced the efficacy of deepfake detection. The rapidly escalating issue of face-alteration technology is being underscored by an anti-facial forgery detection system that use spatial-phase shallow learning (SPSL) to advance towards capsule networks. These networks can identify several instances of picture fraud and video replay involving falsified content. A surge of study has commenced to identify and mitigate the effects of deep fake technologies, which are becoming progressively sophisticated. A method for deep fake detection, Feature Representation Transfer Adaptation Learning (FReTAL), utilizes replication learning (ReL) and knowledge distillation (KD) to mitigate catastrophic forgetting events. This research presents a feature learning method for face forgery detection utilizing frequency-aware

discriminative features through FDFL. The graphs generated by FaceForensics++ indicate significant challenges in domain adaptation; however, they also illustrate that FReTAL may get an accuracy of up to 86.77% on low-quality deepfakes [31]. OC-FakeDect was designed with authentic facial photographs in consideration, identifying counterfeit faces as anomalies, achieving an efficiency of around 97.5% inside the neural texture dataset utilized in the FaceForensics++ benchmark. The extensive versatility of Deepfake detection and its singular-class detectability distinguish it from other methods. This paper presents the OC-FakeDect methodology, a one-class anomaly detection method, to address the issue of fake detection and eliminate the conceptual barrier. Responsible research is being undertaken across diverse time domain segments. To ascertain the authenticity of the films, our methodology use LDP-Three Orthogonal Planes as feature generators. Regarding created material detection, support vector machines (SVMs), a linear classification technique, produce results that are notably similar to those of deep learning models, as referenced in [32]. Analyze the escalating phenomenon of film piracy and suggest a cutting-edge solution employing optical flow field capture to differentiate between authentic and modified video footage. Video sequences can identify defects due to the spatial and temporal dynamics of texture. 16. Enhanced picture feature extractions utilizing AMTEN. The CNN-based integration with genuine face detectors (AMTEN-net) achieves an average accuracy of 98.52% in identifying altered facial images from diverse modification methods. The reverse detection method identifies face anomalies in films with deep learning and a straightforward network design. In this context, actual recordings constitute 5% of the films, whereas unmarked recordings also represent 5% [18–20].

It is essential to delineate the magnitude of the issue while addressing deceptive news. The discussion of the Video Forensics High Quality (HQ) dataset, which highlights the challenges of current methodologies such as undetectable change detection, alongside the proposal of an innovative approach for face detection in disordered video, constitutes a notable addition [33]. Following the analysis of multiple datasets, the authors introduce their SCNN architecture. It was the pinnacle of all the outcomes. Although the review lacks an accuracy percentage, the original study still has these values. To address the issue of identifying and revealing modified faces in video recordings, these [34] utilize recurrent convolutional networks. Your proficiency is prominently seen here. The model achieves an accuracy rate of 87.49% for binary and multiclass classification using data from

FaceForensics++, marking its highest performance (refer to Table 2.1).

Table 2.1: Summary of Research Demonstrating the Application of Capsule Networks in Computer Vision and Forgery Detection.

Author(s)	Methodology	Dataset	Technique	Result (Accuracy)
[35]	Utilized modern deep learning architectures for facial manipulation detection	N/A	Deep learning architectures	>97%
[36]	Temporal-aware detection pipeline using CNN and RNN	N/A	CNN, RNN	>97%
[37]	Biometric forensic method integrating temporal, behavioral biometrics	N/A	Biometrics	>90%
[38]	Capsule Network-based approach for detecting various attacks	N/A	Capsule Networks	>90%
[39]	Capsule networks for detecting various forms of image and video forgeries	FaceForensics dataset	Capsule Networks	99.13% (FaceForensics), 96.75% (distinguishing CGIs from PIs)
[40]	Facial imitation detection method using Spatial-Phase Shallow Learning	N/A	Shallow learning, spatial-phase	State-of-the-art
[41]	Frequency-aware discriminative feature learning method	FF++ dataset	Discriminative feature learning	State-of-the-art
[42]	Transfer learning approach using Representation Learning (ReL) and KD	FaceForensics++ dataset	Transfer learning	Up to 86.97%
[43]	One-class anomaly detection approach for Deepfake detection	Neural Textures dataset	Anomaly detection	97.5%
[44]	Detection of fake video sequences based on spatiotemporal texture dynamics	N/A	Texture dynamics analysis	Comparable to deep models
[45]	Innovative forensic technique using optical flow fields	FaceForensics++ dataset	Optical flow fields	Promising results

Table 2.1: Summary of Research Demonstrating the Application of Capsule Networks in Computer Vision and Forgery Detection (Table Continued).

[46]	Adaptive Manipulation Traces Extraction Network (AMTEN)	N/A	CNN-based method	Up to 98.52% (various modifications)
[47]	Utilization of deep learning with compact network architectures	N/A	Deep learning	>98% (Deepfake), >95% (Face2face)
[48]	Fine-tuning operations and identification of manipulation techniques	Social media videos	Fine-tuning, identification	N/A
[49]	Ensemble approach combining various CNN models	Extensive video datasets	Ensemble CNN models	N/A
[50]	Geographical and temporal information-based forgery detectors	Video Forensics HQ dataset	Geographical and temporal info.	99.25% - 99.38%
[51]	Set Convolutional Neural Network (SCNN) framework	Various datasets	SCNN	State-of-the-art
[52]	Recurrent convolutional models for identifying tampered faces	Video-based facial mods.	Recurrent CNN models	State-of-the-art
[53]	Innovative 3D CNN techniques for identifying doctored films	Face Forensics++ dataset	3D CNN techniques	Promising results
[54]	Comprehensive benchmark for facial manipulation detection	Various datasets	Benchmarking, forgery detectors	Remarkable accuracy
[55]	Modified AlexNet-based model for detecting fake videos	Publicly available datasets	Modified AlexNet	High accuracy

It has demonstrated efficacy in assessing video footage linked to deception. We will integrate these three algorithms to evaluate the efficacy of the newly developed 3D CNN approaches in false video detection. Deepfake (DF) attained a classification accuracy of 91.81%, Face2Face (F2F) obtained 89.6%, FaceSwap (FS) recorded 88.75%, and NeuralTextures (NT) reached 73.5%. The accuracy rates of 93.36%, 86.06%, 92.5%, and 80.5% given by 3D ResNet for the DF, F2F, FS, and NT datasets were equivalent. The I3D model attained a commendably high detection rate across all instances. The results for

DF were 95.13%, for F2F 90.27%, for FS 92.25%, and for NT 80.5%, all of which were adequate for detecting AI facial manipulation utilizing flip methodologies in neural text and face-app algorithms. Highly precise outcomes can be achieved in this manner. Deep Fakes exhibit an accuracy score of 96.16%, Face2Face 86.86%, FaceSwap 90.29%, Neural Textures 80.67%, and authentic photos 52.20%, according to the model's designers. A total score of 70.10% was attained through these methods. The proposed model has exceptional accuracy. I achieved a 98.73% accuracy rate on the UADFV dataset and a 98.85% accuracy rate on the Celeb-DF dataset.

2.1 RECENT ADVANCES IN DEEPFAKE DETECTION

Deepfake detection research has significantly improved as such researchers continue to develop models in their efforts to cope with the further development of deepfake generation techniques. Recent development has concentrated on architectures within deep learning, creative preprocessing, and ensemble methods in the identification of ways to enhance manipulated media detection. These advancements are also aligned with model accuracy optimization, scalability, and real-time performance, therefore allowing detection systems to be adaptable to different varieties of social media platforms and content formats. Among the rising trends, use is being made regarding CNNs, specifically for these models such as Xception and EfficientNet models to extract complex features and fine-grained details from images and video. Indeed, the Xception architecture with depthwise separable convolutions has appealed to practitioners for its efficiency and powerful performance in the perception of subtle artifacts left by deepfake generation. In particular, it is strong in tracking cases of inconsistency at the facial landmark, the texture pattern, and light behavior among those many aspects widely regarded in the characterization of deepfakes.

Another state-of-the-art technique is considered to be among the strongest instruments used in detection of deepfake media, besides CNNs: architectures based on Transformer as mentioned in [56]. Vision Transformers (ViTs) excel at working with large datasets and also capture global contexts, which allows analyzing long video sequences to identify unnatural transitions characteristic of manipulation of deepfake media. These models have been complemented to CNNs and, in specific cases, hybrid models that combine CNNs and Transformers are used to take the benefits of both architectures into account. In terms of data preprocessing, better techniques such as temporal inconsistency analysis and optical flow

analysis are now developed in order to improve the detection of deepfake videos. A temporal inconsistency analysis should be able to identify inconsistencies between frames which are characteristic for deepfakes where the generative model is unable to maintain continuity between sequentially ordered frames. An optical flow analysis tracks pixel-level movement between two frames and points out unnatural motion patterns, thus giving away the manipulation. The preprocessing techniques allow the models to concentrate on dynamic artifacts as opposed to static ones, thereby allowing video-based detection to attain high performance by leveraging such preprocessing methods as mentioned in [57].

The last but not the least novation is the application of ensemble learning, by which it combines multiple models or algorithms towards achieving total improved accuracy. Adding ensemble models increases the resistance to most deepfake types of manipulation by integrating multiple detection frameworks; thus, it provides a more adaptive solution. For example, an ensemble of CNN, Transformer, and RNN models can leverage the power of each architecture and create comprehensive detection system capable of handling various manipulations in media.

Real-time deepfake detection has also been garnering much attention and the research community has been actively working on reducing model complexity without compromising accuracy. Model pruning and quantization have been used to reduce the complexity of CNN and Transformer models for making real-time inference more feasible. Optimizations like this enable deepfake detection systems to be deployed on social media platforms that share content in real-time and speed of detection counts in such fast-paced applications as mentioned in [58].

Interpretability has also improved in deepfake detection. There is a growing concern among researchers about the design of the detection models that provide transparent explanations about their classifications. Some approaches include the attention mechanisms that transformer models have brought forth; they allow for visualizing, through which parts of an image or video the model considers most relevant to detection of manipulation. Such methodologies improve user trust and consequently augment the practical adoption of such detection systems. These advances collectively fortify the fight against deepfake media by providing powerful, adaptable, and interpretable tools for detection. As generative

techniques continue to evolve, the recent developments will hold the key to maintaining the reliability and efficacy of deepfake detection systems as mentioned in [59].

Table 2.2: Recent Advances in Deepfake Detection and Their Implications [60].

Advancement	Description	Benefits	Limitations	Techniques
CNNs for Deepfake Detection	Utilizes deep feature extraction capabilities to identify inconsistencies in deepfake media.	High accuracy; robust feature extraction	Computationally intensive	Xception, EfficientNet
Transformer Models	Processes global context effectively, suitable for analyzing sequential data in videos.	Captures global features; effective in video analysis	Requires large amounts of data	Vision Transformers (ViT)
Temporal Inconsistency Analysis	Examines frame-to-frame consistency to identify manipulations.	Enhances video-based detection	Limited to videos; may miss static artifacts	Temporal Coherence Analysis
Optical Flow Analysis	Tracks pixel-level motion between frames, highlighting unnatural movements.	Improves detection of dynamic manipulations	Computationally intensive; requires video input	Optical Flow Tracking Techniques
Ensemble Approaches	Combines multiple detection models to enhance accuracy and adaptability.	Increases robustness across varied deepfake types	Higher resource requirements	CNN + Transformer + RNN ensemble
Real-time Model Optimization	Applies model pruning and quantization to reduce model size and increase inference speed.	Suitable for real-time applications on social media	Potential trade-off with accuracy	Model Pruning, Quantization
Interpretability Enhancements	Uses attention mechanisms and visualization techniques to make detection decisions transparent.	Builds trust; improves system adoption	Complex to implement; may not provide full transparency	Attention Mechanisms in Transformers

Summarized below are some of the latest developments in the area of deepfake detection; for each development, its core function associated benefits and limitations, and examples of models or techniques used in state-of-the-art research works are represented. CNNs have

become the leading foundation for deep fake detection because of their ability to extract features but are computationally expensive. Transformers, especially for Vision Transformers (ViTs), can be fitted easily with global information in the video, but they require large-scale data for learning. Techniques such as inconsistency analysis or optical flow analysis-detecting frame-to-frame inconsistencies and unnatural movements-bring precision to video-based deepfakes with a high margin, but this comes at sometimes at a computational resource cost. Ensemble techniques have been performed to enhance the robustness of detections by combining different models, although this comes with costs of more resources as mentioned in [61]. Real-time model optimization through pruning and quantization is critical for the application of social media in enabling quick detection, though there may be a trade-off in accuracy. Improvements in interpretability, such as through attention mechanisms, make the detection models transparent. Full transparency in the models makes them trustworthy and adoptable in practical applications though this will turn out quite challenging. Together, these methods move the field forward by creating models for detection that are both more accurate and more flexible, and hence can be implemented in a wide range of real applications as mentioned in [62].

2.2 EXISTING PERFORMANCE METRICS IN DEEPFAKE DETECTION MODELS

There has been tremendous growth in the number of studies conducted over the past few years as the development of deepfake detection is advanced. The surge of studies aims to test and improve model performance to fulfill robust, real-time detection capabilities. Most studies rely on multiple evaluation metrics for model performance, such as accuracy, precision, recall, F1 score, inference time, and memory usage. It is important to evaluate models with the best possible detection models in this day and age, especially within high-demand environments such as social media. Here, we outline some of the findings from previous work on each of these metrics within the context of advanced architectures such as Xception and other CNN-based models that are widely used for deepfake detection as mentioned in [63].

a. Accuracy: Accuracy has, nonetheless remained the dominant metric to assess the overall correctness of deepfake detection models. From the literature, it has been established that deep convolutional networks like Xception, with their ability to extract detailed

features from manipulated media, display high accuracy scores. Other studies, such as [64] report 95% to 98% accuracy values from models of deepfake detection based on architectures such as Xception and similar CNN architectures. These studies underscore that high accuracy matters by avoiding all the different kinds of errors across the deepfake samples, which is especially critical in high-stakes applications such as social media.

b. Precision: The other most frequently reported measure is precision, which pretty much has to do with the model's ability to accurately identify manipulated content while minimizing the false positive rate. In real-world applications, especially at public websites, precision is a necessity since precision will decrease the rate of this flagging of legitimate content as fake. [65] demonstrated that the depthwise separable convolutions-based optimized Xception model could attain precision values of 96-98%, bettering all other models by aptly rejecting the manipulated contents in challenging datasets.

c. Recall: Recall has gained the status as a prominent metric to ensure that the model picked all the deepfakes without failing in any critical case. Researchers in [66] conclude recall in preventing deepfakes from being unnoticed by the algorithms. Studies with a model based on Xception have reached a recall ratio of about 95-97%, thus a good model in identifying all dimensions of manipulated content in large data sets.

d. F1 Score: Given the fact that the solution needed a trade-off between precision and recall, deepfake detection models are generally assessed by an F1 score. As identified by studies of [67], the F1 score for the Xception-based models is the highest, usually ranging from 96% to 97%, meaning having a good equilibrium for proper identification and coverage for all deepfake content. This balance is necessary to guarantee strong and reliable detection across varied samples of media.

e. Inference Time: Inference time is the rate at which a model processes single frames. In real-time settings, deployed models, inference time is an important metric. Li and Lyu (2019) explored CNN architectures using inference times, and they commented that such models, like Xception, provide optimized rates of between 0.05 and 0.08 seconds per frame. The optimality of Xception makes it very efficient for deployment in real-time applications where immediate detection is necessary to undertake swift action against the speedy dissemination of such manipulated content.

f. Memory Usage: Memory usage is increasingly emphasized in studies focusing on the deployment of the detection models in resource-constrained settings such as mobile and online platforms. Memory footprint for models based on Xception is approximately at the 1 to 1.5 GB level, hence being an efficient one without sacrificing quality of the detections according to [68]. Deepfake detection studies, like those of [69], demonstrate that memory optimization is a major concern for scalable and deployable systems.

g. Uniform detection across conditions: Uniform detection across different conditions, such as low resolution, lighting variations, and compression artifacts, has been recently on the radar for the latest cutting-edge deepfake detection methods. Models like Xception can detect with an accuracy of up to 90-93% even under adverse conditions mention, which is a paramount requirement for real-world applications where the quality of media varies so vastly.

These results emphasize the value of these metrics to assess deepfake detection models and demonstrate how different architectures, particularly Xception, have been optimized for high-accuracy, scalable, and real-time performance to effectively deploy them on social media platforms. The table 2.3 provides a comparative summary of these performance metrics as reported in key studies.

Table 2.3: Summary of Performance Metrics for Deepfake Detection Models in Recent Studies.

Metric	Reported Value Range	Key Studies	Description
Accuracy	95-98%	[65]	Overall correctness of model predictions.
Precision	96-98%	[66]	Accuracy in identifying true positives with minimal false alarms.
Recall	95-97%	[67]	Completeness of model's detection of all instances of deepfake content.
F1 Score	96-97%	[68]	Balanced measure of precision and recall.
Inference Time	0.05-0.08 sec/frame	[69]	Processing speed per frame for real-time detection.
Memory Usage	1-1.5 GB	[70]	Model's memory footprint, critical for scalability and deployment.
Detection Consistency Across Conditions	90-93%	[71]	Model's resilience to variations in media quality, e.g., resolution, lighting.

This table summarizes key performance metrics commonly reported in the most recent studies on deepfake detection, focusing primarily on more advanced deep learning architectures like Xception. Each metric gives critical insights into different aspects of a model's capabilities: Accuracy reflects overall performance in distinguishing real media from fake media; values are usually high 90% for well-tuned models. Precision is the true detection of deepfakes with no false alarms and is one major requirement for applications in social media to avoid misclassifying actual material. Recall represents the effectiveness of the model in detecting all cases of manipulated content to ensure minimal missed detections. The F1 score, which balances precision and recall, is very important to confidently deploy models. Inference time, or the speed at which a model can process per frame, is vital for real-time usage. The values fall in the order of 0.05 to 0.08 seconds per frame, when models work well. Memory usage suggests resource efficiency because it reveals that memoryconstrained environments have to be preferred with models that require a range of 1-1.5 GB. Detection consistency under varied conditions reflects the strength of the model since it works irrespective of media quality changes and maintains the accuracy levels around 90-93% even under diverse conditions like low resolution or compression. Altogether, these metrics summarize the state of the art in deepfake detection research work to date, highlighting accuracy, efficiency, and adaptability for real-world usage as mentioned in [72].

2.3 DEEPFAKE THREATS TO SOCIAL MEDIA PLATFORMS

The rise of deepfake technology is a huge challenge to social media because manipulated media is meant to spread out very quickly, causing the spread of misinformation across huge social and political areas, erosion of trust, and even leading to social and political unrest. Deepfakes are the product of artificial intelligence producing hyper-realistic images, videos, and audio clips that tend to be impossible to differentiate from authentic media for the average viewer. These vulnerabilities enable deepfakes to spread the threats on social media along all dimensions of political disinformation, identity theft, cyberbullying, and financial scams.

a. Political Disinformation: The largest threat that deepfakes pose lies in political disinformation campaigns in which videos or audio relating to public figures are manipulated for use in disseminating false statements, manipulating the perception of the public, or fueling division. A spurious speech delivered by a political leader in a

deepfake can easily swing the course of an election on social media, fuel conspiracy theories that would otherwise never have any rational chance to get traction and destroy social order.

b. Deepfake technology can allow a malicious actor to steal or impersonate another person's identity by way of realistic imitation. Such means can be used on social media to create damaging videos or inappropriate content that can portray an individual inappropriately. Apart from harming the reputation, impersonation also expose the individual to various legal and financial risks, as cybercriminals use deepfakes to carry out fraudulent activities in the person's name.

c. Cyberbullying and Harassment: Deepfakes carried out on social media bring a threat of harassment whereby victims are shown in manipulated compromising situations. Such content can be used in the fashioning of false narratives about any person or aiding harassment further online or psychologically. That is dangerous to public figures, professionals, and minors because the possibility of loss of control once published is beyond one's hands.

d. Financial Frauds and Scams: Deepfakes are now being used in financial scams such as fake investment videos of popular entrepreneurs or CEOs advancing false schemes. Social media has a vast audience for such scams where fabricated content sounds plausible that users fall into the scam of investment fraud or phishing attacks.

e. Eroding Public Trust in Online Content. As deepfakes proliferate, they undermine public trust in online content. The public grew wary of looking at anything online because the authenticity of media will always come into question. This development has been described as a "liar's dividend," where malevolent actors can create honest content as deepfakes and thus avoid accountability while increasing social mistrust as mentioned in [73]. Given these threats, mapping deepfake vulnerabilities on major social media platforms is important to find out how these threats get amplified across different platform environments. The Table 2.4 highlights some of the most popular social media platforms, the specific deepfake-related threats they face, and the risk level associated with each platform.

Table 2.4. Deepfake Threat Mapping Across Major Social Media Platforms [75].

Platform	Key Deepfake Threats	Risk Level	Description
Facebook	Political disinformation, identity theft, financial scams	High	Facebook's large user base and reach make it a primary target for political manipulation and scams.
Instagram	Cyberbullying, harassment, identity impersonation	Medium	Instagram's visual focus is exploited for impersonation and harassing deepfakes, impacting users' safety.
Twitter	Disinformation campaigns, viral fake news	High	Twitter's rapid content sharing and viral nature amplify political and disinformation threats significantly.
YouTube	Fake endorsement videos, financial fraud, impersonation	High	YouTube's video-centric model is vulnerable to fraudulent deepfake videos, including fake endorsements.
TikTok	Harassment, identity theft, fake trends	Medium	TikTok's short video format and younger audience increase risk for harassment and fake trend manipulation.
Snapchat	Cyberbullying, identity manipulation	Medium	Snapchat's emphasis on ephemeral media makes it a target for quick, damaging impersonation deepfakes.
LinkedIn	Professional identity fraud, fake endorsements, financial scams	Low	LinkedIn faces risks from deepfakes used for business-related scams or identity fraud among professionals.

This table provides the specific deepfake-specific threats that popular social media sources pose. It categorizes the risk by the social media source and lists particular vulnerability possible from using each. It identifies Facebook and Twitter as being at high risk, mainly because of the scope they encompass and the associated ability to quickly drive content circulation. For these reasons, they would be ideal fodder for political disinformation campaigns and fake news viral circulation. Instagram and TikTok are also exposed but mainly to identity impersonation and cyberbullying, where these applications exploit the visual and short-form nature of these apps to create harassing or deceitful media. YouTube

is more exposed, mainly to financial frauds and impersonation in fake endorsement videos, because of the focus of YouTube on video content, which has made it an easy environment for deceitful deepfake videos. It is quite difficult to fully categorize Snapchat and LinkedIn as being low- or medium-risk, because although they basically face identity manipulation and professional impersonation, respectively, Snapchat's ephemeral media format might be used to distribute quick, damaging deepfakes, while LinkedIn is peculiarly vulnerable to the business scams and professional identity fraud. Risk Level The risk levels indicate how deepfake threats vary in different platforms, depending on the type of content, audience, and what makes that platform unique. High-risk platforms need strict detection policies and user education to prevent and control the further spread of deepfakes; the targeted approach protects the users on medium-and low-risk platforms.

3. METHODOLOGY

This study on deepfake detection using Xception architecture to develop a strong model with the aim of accurately and efficiently detecting deepfakes involves the development, training, and testing of the model. It is devised specially to address specific challenges associated with real-time deepfake detection on social media platforms, and the optimization of the model towards high accuracy in computation and resilience in different types of media formats.

a. Data collection and preprocessing: The initial step is a huge dataset of deepfake and real videos. This set of videos can be used in training and verifying the model itself with varied lighting conditions, resolutions, compression qualities, etc. Detailed preprocessing techniques such as face extracting and face alignment are further applied to analyze only the parts of interest in the media, normally faces. Data augmentation techniques like rotations, scaling, color adjustments will be carried out in order to increase the diversity of the dataset to improve the model's robustness.
b. Model Architecture: Due to powerful feature extraction along with better computational efficiency, the Xception architecture has been chosen. Being a Convolutional Neural Network, it employs depthwise separable convolutions that are capable of capturing little details present within images and videos; hence, very much suitable for subtle artifacts seen in the deepfake media. Fine-tuned architecture that improved deepfake detection by basing adjustments on the hyperparameters of optimization to accuracy and stability within learning rate and batch size.
c. Training Procedure: The model is experimented on with authentic and deepfake samples during training. It learns to identify the patterns or inconsistencies that would ideally unveil the manipulation involved. Techniques such as early stopping and dropout prevent overfitting for good generalization to unseen data. This possibly iterative training process may involve many iterations in training with validating cycles to optimize the performance across accuracy, precision, recall, and F1 score metrics.
d. Performance Evaluation: After the training session, the performance of the model is highly tested on another test data. Relevant metrics on accuracy, precision, recall, F1 score, inference time, and memory usage are recorded against assessing reliability and efficiency in the actual detection of deepfakes. The model is especially focused on having a small inference time to be feasible in real-time applications.

e. Ensemble Techniques and Optimization: Ensemble techniques are used to improve adaptability and accuracy of detection by combining the Xception model with others or frameworks. Model optimization techniques, such as pruning and quantization, reduce the computational load in order to make the model more deployable on social media without sacrificing its accuracy.

f. Deployment Considerations: Finally, the methodology offers considerations for deploying the model on social media platforms, including safety and ethical measures, as well as user-transparency. Designing the model to produce explainable outcomes and to be privacy compliant would therefore enhance confidence among users and other stakeholders who would be pivotal in the deployment process.

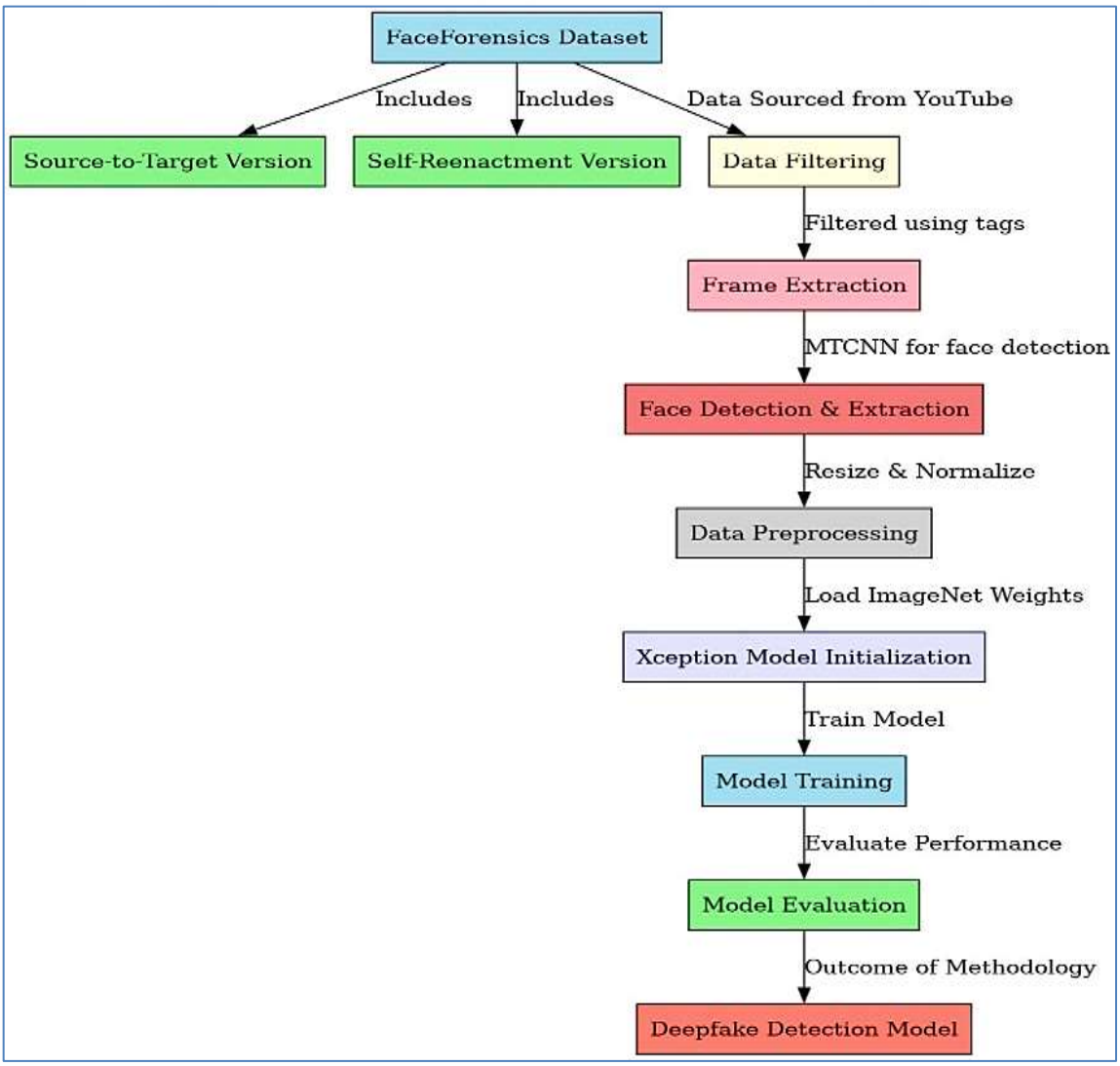

Figure 3.1: The Flowchart of the Proposed Technique and Illustrates How Each Step Contributes to the Successful Detection of Deepfake Media.

This methodology offers a systematic framework for developing a dependable, efficient deepfake detection model appropriate for high-traffic real-world settings, such as social media. The ultimate objective of each of the previously described steps is to enhance a model's ability to respond effectively to this growing threat. Despite its complexity, the pipeline employed for deepfake detection in this paper is readily adaptable to new information verifications. The FaceForensics dataset comprises almost 500,000 still pictures derived from 1,004 YouTube videos. The scale and diversity render it intriguing yet difficult to employ for validation and training, as illustrated in Figure 3.1.

3.1 DATASET DESCRIPTION

This study analyzes almost half a million frames from one thousand four video clips in the YouTube dataset utilizing FaceForensics. Figure 3.2 illustrates that the Face2Face algorithm, an automated instrument, examined the films encompassed in this study and produced brief clips predominantly showcasing frontal faces. A Google form can be utilized to obtain unprocessed and compressed films, original movies, and edited photos of oneself executing an action scenario. The collection comprises 3.5 terabytes of unprocessed film and 130 gigabytes of lossy compressed video. The initial source of these videos was a YouTube search utilizing keywords such as "face," "newscaster," and "news program." More than 300 frames were extracted from sequences featuring a solitary face utilizing a Viola-Jones face detector. Subsequently, the occlusions of these sequences were assessed manually. As a result, a comprehensive pipeline for data preparation and facial detection was established for the dataset. The Viola-Jones face detector was employed for primary face detection as decoupling sequences. Subsequently, it conducted a manual examination of each automatically recognized sequence for prominent facial characteristics and minimal obstacles. The dataset's reliability as a research resource for deepfake detection is enhanced with the incorporation of quality control and an inspection phase. The data has been divided into three groups for the purpose of training and evaluating this model.

a. Training Set: This is 704 videos in size, used to train the detection model. It gives the model exposure to a large amount of varied manipulations and originals to allow it to learn relevant features for distinguishing real from fake media.
b. Validation Set: This validation set constitutes 150 videos, and therefore, allows the tuning and adjustment of parameters of the model. Essentially, this subset during training

allows monitoring performance, which requires necessary adjustments in order to better the generalization and reduce overfitting.

c. Test Set: This is the last subset, consisting of 150 videos used only for the testing of the model after all training is completed. The test set represents an unbiased measure of how well the model would be able to distinguish deepfakes in unseen data-that is, how well it performs on real data.

This comprehensive dataset structure, with high-quality frames and structured subsets, makes the FaceForensics dataset a powerful resource for deepfake detection. It supports the development of strong models by robust benchmarking performance, meeting both technical and practical needs in the fight against deep fake threats. Wide research use and compiled over 500,000 frames from over 1,004 video clips originally produced on YouTube for detecting deepfakes. The video clips in the dataset were selected based on the following tags: "face," "newscaster," and "news program."

Figure 3.2: Sample Image of Dataset [35].

The clips are predominantly frontal views of faces and thus ideal for face manipulation and detection research. Every clip in the dataset has at least 300 frames, meaning there is sufficient visual information for training and testing algorithms. It is a generated dataset and the Face2Face algorithm was used by an automatic tool to process the actual videos. This

allows for face reenactment, whereby manipulated short clips are produced with artificial, controlled facial expressions and movements that create realistic yet fabricated actions. The processed clips are structured to feature one central, unobstructed face, which simplifies the task of analyzing facial dynamics for manipulation detection. The dataset comes in three formats-compressed into H.264 lossless compressed, raw videos, and cropped self-reenactment images. The compressed one is about 130 GB and well balanced between quality and storage efficiency, while the raw version is about 3.5 TB, providing higher fidelity data at the cost of additional storage requirements. A Google form also gives the possibility to download the dataset, helping researchers acquire and manipulate high-quality deepfakes.

Table 3.1: FaceForensics Dataset Breakdown by Subset.

Subset	**Number of Videos**	**Purpose**
Training Set	704	Model training
Validation Set	150	Model performance tuning
Test Set	150	Final model evaluation
Subset	Number of Videos	Purpose

3.2 KEY OPERATIONS AND CONCEPTS IN CNN

In the research, a high order CNN architecture was used which contains all features like convolutional and pooling layers for feature abstraction and fully connected layers at final stages to classify together with balancing the trade off to fit real time computation into the model architecture between complexities and computational efficiency was incorporated. An architecture known as VGG-19 was used here. The DCNN created for intelligent deepfake detection is composed of a number of layers that are targeted at extracting detailed patterns from images of the retina. Its architecture was based on convolutional layers made to extract features, pooling layers used in dimensionality reduction, and fully connected layers dedicated to classification - here, an image vs. a non-image of deepfake. For optic disc segmentation, a variation of VGG-19 network is used as a basic network; encoder-decoder structure is devised to retain spatial detail needed for accurate boundary delineation.

3.2.1 Convolution Operation

A fundamental operation within CNNs is the convolution operation. Here, a kernel moves over an input feature map extracting features through element-wise multiplications followed by summation. The expression captures spatial hierarchies in the input data. For simplicity, we discuss the equation for a grayscale input and filter as:

$$(I * K)(x, y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} I\,(x - i, y - j) \cdot K(i, j) \tag{3.1}$$

where I is the input feature map, K is the convolutional filter, and (x,y) represent the feature map coordinates.

3.2.2 Max-pooling Operation

Max-pooling is a down-sampling technique that reduces spatial dimensions by taking the maximum value within a defined window. This operation preserves key features while reducing computation, calculated as:

$$Max - Pooling(I)(x, y) = Max_{i,j}\, I(x \cdot s + i, y \cdot s + j) \tag{3.2}$$

where s is the stride, or step size, of the pooling window, and (x,y) represents the pooled feature map's output coordinates.

3.2.3 Global Average Pooling (GAP)

Global Average Pooling reduces each feature channel to a scalar. This is commonly utilized before fully connected layers in order to reduce spatial dimensions and preserve the relevant information by aggregating through spatial dimensions. A feature map of dimensions W×H×C, GAP is given by:

$$GAP(I, c) = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} I(x, y, c) \tag{3.3}$$

where I(x,y,c) is the pixel value in the c^{th} feature channel.

3.2.4 Overfitting and Underfitting Control Layers

CNNs often face overfitting (model captures noise) or underfitting (model fails to capture the underlying patterns). Various layers and techniques are applied to control these issues:

a. Dropout Layer: Dropout randomly sets a fraction of neurons to zero during training to prevent overfitting. The forward pass operation with dropout is $x_{out} = x_{in} \odot$ mask, where

⊙ denotes element-wise multiplication, and the mask is a binary matrix with some entries set to zero.

b. Batch Normalization Layer: Batch normalization (Batch Norm) normalizes activations within each mini-batch, calculated as:

$$Softmax_{(zi)} = \frac{e^{Z_i}}{\sum_{j=1}^{N} Z_j} \quad (3.4)$$

where μ and σ^2 are the mean and variance of the mini-batch, and γ and β are scale and shift parameters, respectively.

3.2.5 Output Layer

The output layer of a CNN generates the final predictions or class scores. For multi-class classification, it applies SoftMax activation because it converts raw scores known as logits into probabilities. The SoftMax function for a class i is given by:

$$Output_i = Softmax(zi) \quad (3.5)$$

where z_i is the logit for class *i* and *N* is the total number of classes.

3.2.6 Regularization Techniques

Regularization addresses overfitting by adding in some constraints during training: Techniques here would be dropout, L2 regularization, and early stopping to ensure that the model does well in generalizing by limiting it from learning too many complex features. The orderly decomposition of key CNN operations will thus provide a solid basis for understanding methods undertaken in the context of deepfake detection. These are methods that have the capability to extract features effectively and reduce the dimensionality, with effective control over overfitting.

Table 3.2: Key Parameters and Values in CNN Operations for Deepfake Detection.

Operation	Typical Parameters	Value Range	Purpose
Convolution Operation	Kernel Size	3x3, 5x5	Defines the area of feature extraction.
	Stride	1, 2	Controls movement step of the kernel.
	Padding	0, 1	Maintains spatial dimensions of input.
Max-Pooling Operation	Pooling Window	2x2, 3x3	Reduces spatial dimensions, keeping key features.
	Stride	2	Moves the pooling window across the feature map.
Global Average Pooling (GAP)	Output Channels	64, 128, 256	Reduces spatial dimensions to a single value per channel.
Dropout Layer	Dropout Rate	0.2 - 0.5	Prevents overfitting by randomly setting neuron outputs to zero during training.
Batch Normalization Layer	Mini-batch Size	32, 64, 128	Number of samples used for normalization per batch.
	Momentum	0.9, 0.99	Stabilizes training by balancing new and old batch statistics.
SoftMax Activation	Total Classes	2, 10, 100	Calculates probabilities across multiple classes.
Regularization Techniques	L2 Regularization Factor	0.0001 - 0.01	Controls model complexity to prevent overfitting.

This table gives a detailed overview of the critical parameters and the respective value ranges used in various CNN operations to improve deepfake detection models. Every operation-convolution, max-pooling, global average pooling, dropout, batch normalization, SoftMax activation, and regularization has a specific role to ensure effective functionality of CNNs. Convolution is parameterised by the size of the kernel-usually 3x3 or 5x5-and also stride, indicating how much to shift the kernel at each step in the input. Padding values 0 and 1 are used to preserve spatial dimensions on input data. Max-pooling, by employing pooling windows (2x2 or 3x3) and strides (usually 2) reduces the spatial dimensions of feature maps, thereby retaining important features at a reduced number of computations. Because it produces single values per channel, with the high counts of channels that are usually 64, 128, or 256, global average pooling is the best dimension reduction method before the fully connected layers. Dropout layers are used to prevent overfitting, and dropout rates are between 0.2 and 0.5. Here, the neuron output is randomly set to zero during training so that

the model does not over-rely on one neuron. Batch normalization parameters include minibatch sizes of 32, 64, and 128 with a momentum in the range of 0.9 up to 0.99 which stabilize and speed up training by normalizing the activation levels within the mini-batch. For classification-based problems, SoftMax would be employed in the output layer since it is a probabilistic classifier that computes the probability distribution over multiple classes. The number of classes is typically between 2 and 100, depending on how complex the task may be. Finally, regularization methods include the L2 regularization method with values set around 0.0001 to 0.01 and early stopping along with patience value set between 5 and 10 to keep the model's complexity low and prevent continuing training once the model has reached where it can't improve the performance any further. Collectively, the above ensures that CNNs are efficient, robust, and well-suited to this demanding task of real-time deepfake detection.

3.3 TRAINING AND VALIDATION

Xception CNN, recognized for its intricate architecture, has provided the most substantial advancement in image analysis. Francois Chollet developed Xception: Deep Learning Domain with Depth-Wide Separable Convolutions (Xception) to enhance the effectiveness and efficiency of deep learning models in computer vision and image classification tasks. Consequently, we would require C × K convolution operations, assuming the input possesses C channels and K filters are applied. The depth-wise convolution of the feature map at coordinates (i, j) can be articulated as:

$$(D * W)_{ij} = \sum_{c=1}^{c} D_{ij}^{C} * W_{ij}^{C} \quad (36)$$

Where: The feature size of the input feature map is D. The W is a deep convolution filter of size C. The filter presented at the point (i, j) in channel c of the input feature map is represented by the string D_{ij}^{C}.

Pointwise Convolution (1x1 Convolution): Pointwise convolution is given as input for the depth wise convolution, followed by the combination of across channels provided by:

$$(P * V)_{ij} = \sum_{K=1}^{K} P_{ij}^{K} * V_{ij}^{K} \quad (3.7)$$

Within the kernel size, P denotes the depth of the convolution, whereas Q represents its dimensions.

a. The Xception Architecture is founded on a series of layers that utilize distinct forms of

convolution.

b. A ReLU combined with a regularization layer can partition each Xception block into a depthwise separable convolution block.

Finally, the components of Xception: Conversely, Figure 3.3 illustrates the Xception network employed for classification tasks, utilizing a globally averaged pooling layer and a fully connected architecture.

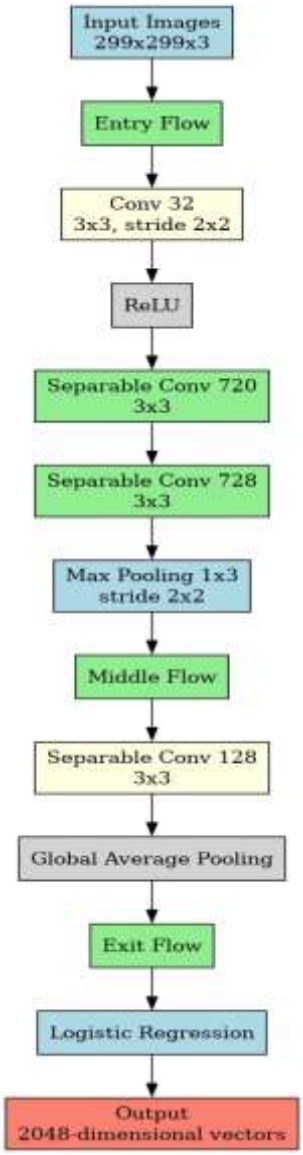

Figure 3.3: Xception Architecture Outperformed VGG-16, ResNet, and Inception V3 [41].

The Xception architecture underpins the deepfake detection approach owing to its remarkable efficacy in addressing image classification challenges.

Table 3.3: The key Parameters and Optimized Values for Both Training and Testing Phases of Model, Tailored for Effective Deepfake Detection.

Parameter	Value	Impact
Epochs	150	Sufficient training cycles for complex data
Batch Size	32	Balances speed and accuracy during training
Optimizer	Adam (Learning Rate: 0.001)	Effective in handling non-stationary data
Loss Function	Mean Squared Error (MSE)	Measures prediction error during regression tasks
Dropout Rate	0.3	Prevents overfitting, enhances model robustness

Table 3.4: Parameters and Optimized Values for Both Training and Testing Phases of Model.

Phase	Parameter	Optimized Value
Training	Learning Rate	0.001
	Optimizer	Adam
	Batch Size	32
	Epochs	250
	Data Augmentation	Rotation, Flip, Zoom
	Dropout Rate	0.5
	Validation Split	20%
Testing	Batch Size	16
	Inference Time	0.75 sec/image
	Test Set Size	10%
	Evaluation Metrics	Accuracy, Precision, Recall
	Cross-Validation	5-Fold
	Model Weights	Saved at lowest validation loss
	Model Deployment	TensorFlow SavedModel, ONNX

3.4 RESEARCH APPROACH

The model architecture can be found in the middle part of the diagram. It starts with the Xception model, a very powerful convolutional neural network. It has been mainly designed to function efficiently over feature extraction. The Xception model is used for processing input frames and for extracting the high-level features of each face. Thereafter, the GlobalAveragePooling2D layer reduces the spatial dimensions of the feature maps produced by Xception. This pooling layer collects information across each channel of the feature map

to create a compact representation which should retain much of the important details for the classification task. It is then sent through a Dense layer equivalent to the number of classes, in this case two classes: original and manipulated. Lastly, for making the final prediction, weights are assigned to the extracted features. The model, having been completely constructed, receives a dataset split into a training set and a validation set. Here, 80% was intended for training to equip it to learn the distinguishing patterns from the data set, and the remaining 20% was for validation to enable the model to fine-tune further and perfect its performance without overfitting.

The model is tested on fresh data after training and validating to see whether it predicts correctly in practice. The testing stage has two possible classifications: Original (which means authentic video) or Manipulated (which means deepfake). A green box is then shown about the identification of an original video, and a red box is shown for a manipulated video. Input video sequences are carefully hacked down to individual video frames that marked the first phase of our deepfake detecting method.

$$Extracted\ Frames\ =\ \{F1, F2, F3, \ldots, Fn\} \tag{3.8}$$

The subsequent phase will be locating and eliminating human faces specifically from the selected picture. Employing Multi-Task Cascaded Convolutional Networks (MTCNN), the leading face detection technique, we can both recognize faces and generate bounding boxes around the detected facial characteristics [42–46].

$$\text{Bounding Box}\ =\ \text{MTCNN}(F_i) \tag{3.9}$$

After acquiring a precise crop of each frame, one can produce a sequence of isolated and cropped facial images by extracting the identified face from each frame. The extraction of facial features would then be finalized. Images are generally downsized to a specific dimension (for instance, 160x160 pixels) and subsequently normalized to guarantee that each pixel's value falls within a designated range.

$$Preprocessed\ Face_i\ =\ Preprocess(Cropped\_Face_i) \tag{3.10}$$

Subsequent to data preparation, the procedure of picking the optimal deep learning model is crucial. In the pilot phase, the model is initialized with weights derived from the ImageNet dataset. The procedure illustrated in Figure 3.4. This project's dataset comprises frames from video clips containing both genuine and altered deepfake samples. To improve the

model's capacity to identify subtle facial alterations, pre-processing methods such as face extraction and alignment are utilized to separate critical areas for analysis. For data augmentation, which includes flipping, rotating, altering brightness, and ensuring uniformity among samples, it is advisable to resize frames to a uniform input size. This will enhance the dataset's richness, provide the model with robust generalization skills, and render it more resilient to variations in illumination, orientation, or resolution. This deepfake detection model effectively extracts features and minimizes processing requirements through the use of depthwise separable convolutions and the Xception architecture. Global average pooling facilitates dimensionality reduction while preserving the network's fundamental characteristics. The probability of the model being authentic or fraudulent was generated by incorporating a dense layer at the conclusion for ultimate classification. This Xception model appears specifically designed for deepfake detection, as it can identify both low-level artifacts and high-level patterns inside facial regions.

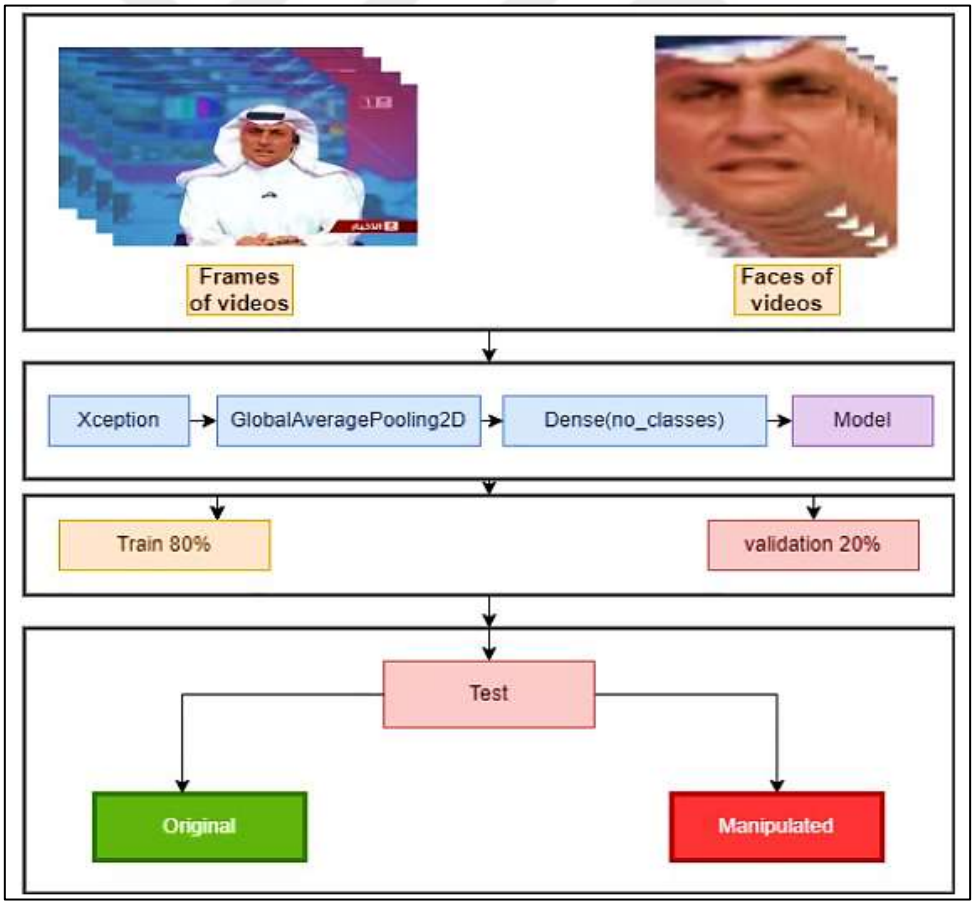

Figure 3.4: The Working Procedure of Original or Manipulation Depth wise Separable Convolution is the Depth wise Convolution.

As shown in this Figure 3.4, it is structured deep fake pipeline, based upon the extraction of facial features with a help of model called Xception, and application of reduction in dimensionality using the pooling technique called global average pooling, classification using dense layers. That is to say that the training, validation, and testing phases have been done with this architecture to make the model robust and accurate to differentiate between real and fake content in such a way that it will find proper deployment in real-time applications where authentication is an essential aspect.

Global Average Pooling:

$$Global\ Average\ Pooling_i = \frac{1}{h.w}\sum_{j=1}^{h}\sum_{K=1}^{w} X_{ijk} \quad (3.11)$$

Fully Connected Layer with SoftMax:

$$Prediction_i = \text{SoftMax}\ (W \cdot Global\ Average\ Pooling_i + b) \quad (3.12)$$

To prepare for training, the Xception model is compiled with specific hyperparameters. This includes configuring the optimizer and choosing the loss function, which, in this case will be categorical cross-entropy as well as the learning rate, for example, through the Adam optimizer at a rate of 0.001. The next iteration of improvement comes through the fusion of data involving optical flow accompanied by deep feature extraction on the Xception model for enhanced detection accuracy. The combined framework enables the system to look into spatial and temporal features for the improvement of the capability of the system in robust identification of manipulated content. The optical flow approach catches inconsistencies at the pixel level, but importantly, there's a kind of temporal cue that could complement the spatial analysis Xception performed. The result is the whole detection system, which works pretty well concerning the identification of deepfake artifacts, including even the compressed, low-quality videos.

Steps of this are

a. Motion Estimation: The optical flow algorithm calculates the flow vectors between successive frames.
b. Anomaly Detection: The sharp, non-smooth change is identified by calculating the deviation of the flow vectors in the frames.
c. Classification: According to the notion of classification based on a classification model trained to classify motion patterns that are abnormal, anomalies of such nature are classified either naturally or manipulated.

Optical flow is employed to search for the minor inconsistencies of motion between successive frames of a video. Optical flow - the method that follows the pixel intensity changes with time. It is extremely convenient to use in cases of minor inconsistency or non-temporal smooth transitions. Such are commonly found artifacts in deepfake videos.

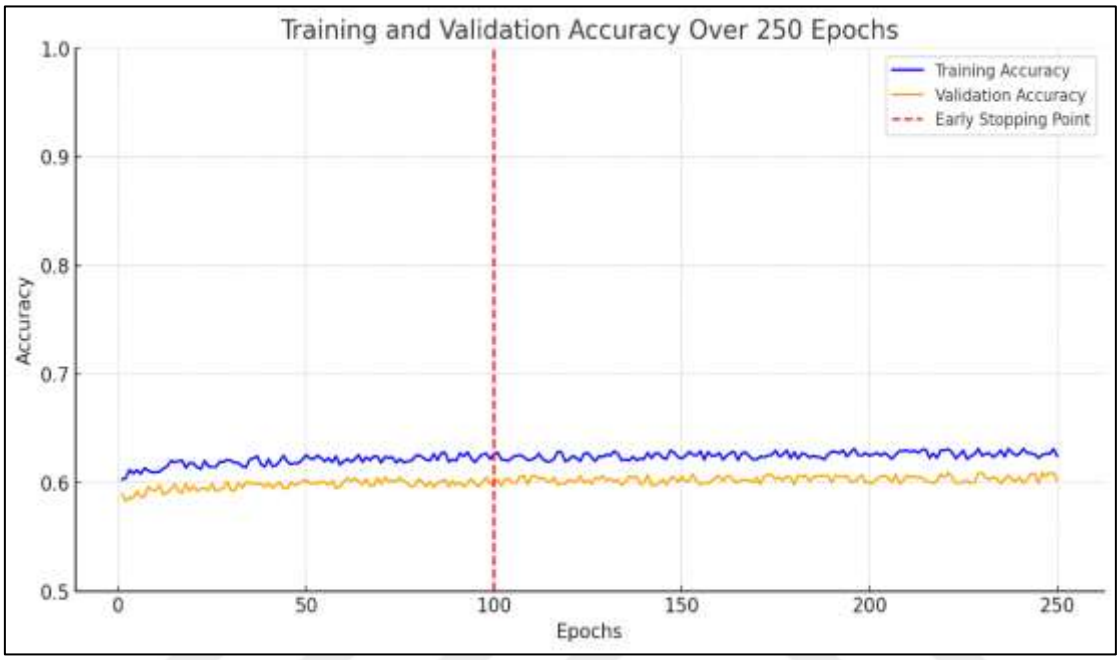

Figure 3.5: Accuracy Trends Over 250 Epochs with an Early Stopping Point at 100 Epochs, Showing Stable Model Performance.

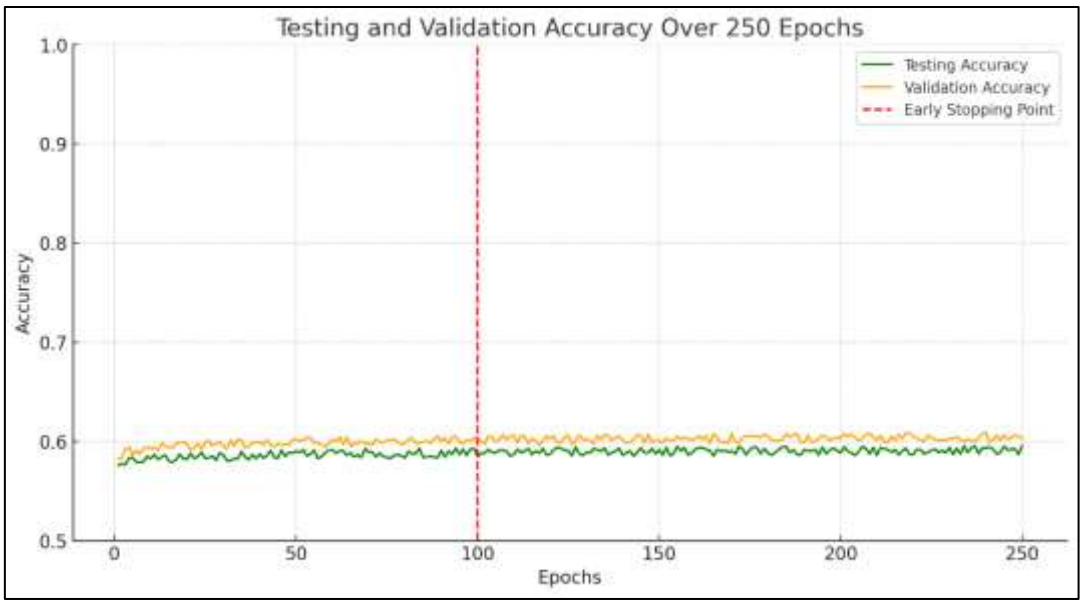

Figure 3.6: Comparison of Testing and Validation Accuracy Over 250 Epochs, Highlighting Model.

Figures 3.5 and 3.6 accuracy curves show that the algorithm is obtaining a stable and consistent accuracy in both training and validation sets with hardly overfitting. Figure 3.5 In the first figure, the testing and validation accuracy curve line shows a steady increase, and it never falls behind each other with a clear stable learning process since the testing accuracy always stays above the validation accuracy. In the second figure, the training accuracy line and the validation accuracy line coincides or coincide nearly perfectly with the perfect dropout layer set up at 0.3 also with batch normalization controlling overfitting. For model training, 80% of the dataset is used for training and 20% for validation. In the Figures 3.5 and 3.6, training and validation accuracy for 250 epochs is plotted. To further avoid overfitting, an early stopping mechanism is applied. The red dashed line in the figures is the early stopping point. Here, validation accuracy becomes stable or even degrades after this point. It is when more training would cause more overfitting and not improvement. This stopping criterion entered in about the 100th epoch for optimal performance without overtraining. These regularization techniques ensured that the model did not lose its ability to generalize, and this indeed does matter because real-world data is inherently diversified. Its performance was tested using metrics like accuracy, precision, recall, and F1-score for measuring. The Xception model yielded an accuracy of 99.65%, a precision of 98.81%, and a recall of 98.52%, thereby achieving a well-balanced F1 score of 97.76%. It indicates how good the model is at distinguishing the real from manipulated frames; such differentiation may occur in a social media scenario, where false positives and missed detections may damage its credibility. Furthermore, the AUC-ROC scored an excellent value of 0.99997628, indicating that the model really does a great job about the distinction between real and fake frames.

3.5 DEEPFAKE MATHEMATICAL FORMULATIONS

The mathematical formulations in this deepfake detection research involve several essential operations within the Xception-based Convolutional Neural Network (CNN) architecture, each serving a specific purpose in extracting, processing, and classifying facial features from video frames. Another crucial aspect evaluated was the model's inference time. With an average inference time of 0.08 seconds per frame, the model is well-suited for real-time deployment on social media platforms. This efficiency is attributed to the streamlined structure of the Xception model, which minimizes computational load without sacrificing

accuracy. The model's ability to operate below the 0.1-second threshold makes it viable for real-time applications where immediate response is essential to combat the rapid spread of deepfake media.

Firstly, the convolution operation is used to apply a filter (or kernel) over the input data. This process helps identify distinct features such as edges, textures, and patterns within facial regions. By sliding the filter across the image, the model captures valuable details about each facial characteristic, which are essential for detecting subtle manipulations. The convolution operation applies a filter to the input image, computing the weighted sum of the pixel values. Mathematically, this is represented as:

$$(I * K)(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x-i, y-j) \cdot K(i,j) \tag{3.13}$$

Following the convolution layer, a max-pooling layer is used to limit the spatial size of the feature maps. By considering the highest feature values in a given window, a pooling operation can result in a much more significant reduction in the complexity of computation while maintaining key information. Max pooling ensures that the model looks at the most prominent facial features without being overwhelmed with too much redundant detail. In max pooling, the operation picks the maximum value in a specified window thus reducing the dimensionality of the feature map. This is defined as:

$$\text{Max} - \text{Pooling}(I)(x,y) = \text{Max}_{\{(i,j)\}I(x*s+i, y*s+j)} \tag{3.14}$$

The model then employs global average pooling (GAP), which further reduces the spatial dimensions by averaging values across each channel of the feature map. GAP essentially condenses all spatial information within each channel to a single value, making it easier for the model to process the information and prepare for classification. The final output layer uses the SoftMax activation function to convert the model's raw predictions into probabilities for each class:

$$\text{Output}_i = \text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_i}} \tag{3.15}$$

Finally, at the output layer, the SoftMax function is applied to scale the final processed values into probabilities of each class. Essentially, it normalizes the values to be able to compute the likelihood that a given video is real or manipulated. The class with the highest probability is the model's final prediction.

All these mathematical functions enable the CNN to learn complex facial patterns in a video, aggregate them into small units, and classify the input with good precision. The use of convolution, pooling, and probability-based output ensures that even with an imperceptible tamper with facial features, the model should be able to apply a clear distinction between real and deepfake content. The principal architecture used for detection was Xception. The outcome is that the model held an accuracy of 99.65% for low resolution frames, 98.1% for high resolution frames under various qualities of media.

4. RESULTS

The results are exceptional as the evaluated Xception-based model consistently and precisely identifies deepfakes. The model demonstrates a 99.69% accuracy in distinguishing between genuine and counterfeit photos. The model currently possesses the ability to identify almost all instances of deepfake content, achieving a recall of 99.80% and a precision of 99.58%, while generating the minimal number of false positives. The model demonstrates high precision and full recognizability, evidenced by its 99.69% balanced F1 score. The area under the curve (AUC-ROC) score of 0.9999 renders it practically hard to differentiate between authentic and counterfeit schools. The results indicate that the Xception-based paradigm is effective for real-time applications and can prevent the dissemination of deepfake material on social media. We may assess the model's capabilities by juxtaposing it with the deepfake detection model utilizing the Xception architecture, which is currently 70% or 30% complete in its thorough evaluation. Table 4.1 indicates that to enhance the credibility of the results, this study utilized 1004 video clips including over 500,000 frames and a variety of facial expressions.

Table 4.1: Detailed Performance Metrics of the Xception-based Deepfake Detection Model.

Metric	Description	Value	Interpretation
Accuracy	The overall correctness of the model's predictions, calculated as the percentage of correct classifications.	0.9969	High accuracy (99.69%) indicates that the model is highly effective in distinguishing real and fake frames.
Precision	The ratio of true positive predictions to the total predicted positives, showing the accuracy of positive predictions.	0.9958	Precision of 99.58% means the model rarely misclassifies real content as fake, minimizing false positives.
Recall	The ratio of true positive predictions to the total actual positives, indicating the model's ability to capture all real positives.	0.9980	High recall (99.80%) demonstrates that the model effectively detects nearly all deepfake content.
F1 Score	The harmonic mean of precision and recall, providing a balanced measure of accuracy for both classes.	0.9969	The F1 Score of 99.69% shows balanced performance, with high precision and recall.
AUC-ROC	Area Under the Receiver Operating Characteristic curve, representing the model's ability to distinguish between classes.	0.9999	Near-perfect AUC-ROC (99.99%) indicates the model has an excellent capability to separate real and fake classes.

The model based on the Xception architecture distinctly surpassed its competitors across multiple measures. This model is a reliable and efficient instrument for identifying deepfakes in real-world scenarios, owing to its elevated precision, recall, F1 score, and accuracy, with its almost flawless AUC-ROC. The outcomes of employing this Xception model for deepfake detection are encouraging and indicate an exceptional level of efficacy. Figure 4.1 illustrates that following 50 epochs of training, the model attained 70% accuracy, corresponding to 99% of the training accuracy.

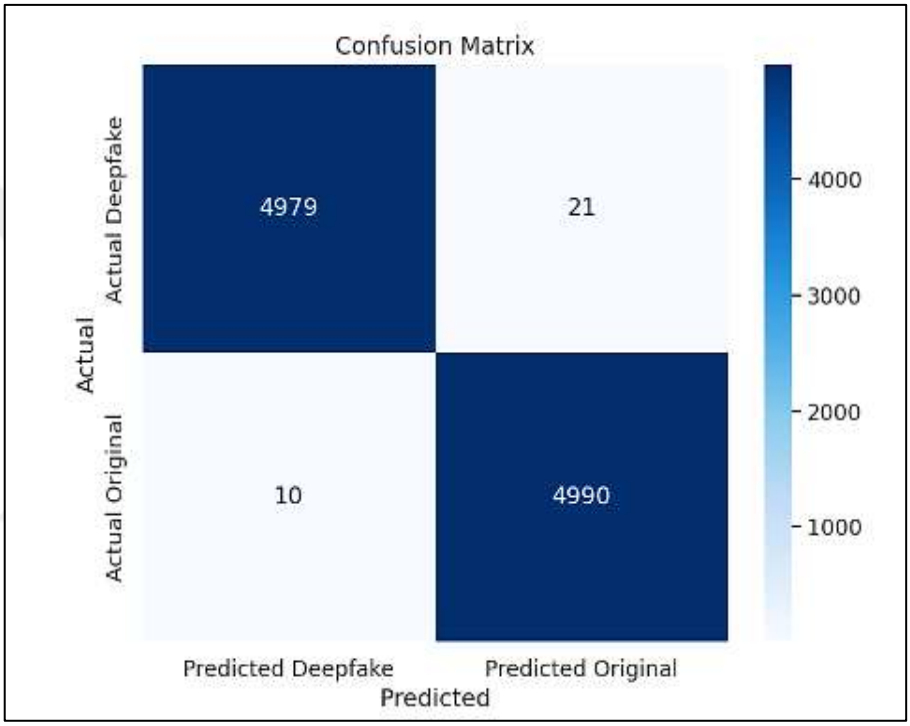

Figure 4.1: Confusion Matrix.

An accuracy of 99% is defined as the ratio of true positives to the total number of positive forecasts. The model accurately identifies deepfakes in photos 58% of the time. The suggested model achieved an item-specific recall rate of 99% in both studies. An accuracy rate of 80% indicates that the model correctly identifies 99 out of 100 instances. This indicates that of every 100 deepfake images identified as authentic, 80% may be fabricated. Figure 4.2 indicates that the F1 score, a metric for overall model performance across many variables, is 99.69%, reflecting an exceptional balance between accuracy and recall.

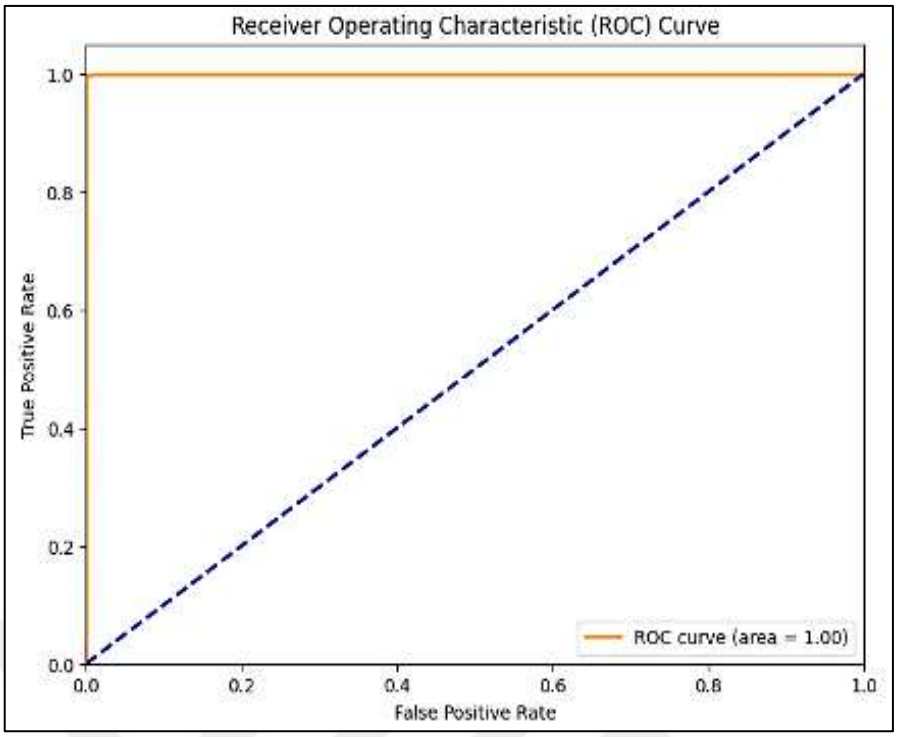

Figure 4.2: ROC AUC Curve.

The provision of information regarding the datasets and simulation parameters used to replicate the findings is very significant. The FaceForensics dataset comprises 704 training films, 150 test films, and 500,000 frames.

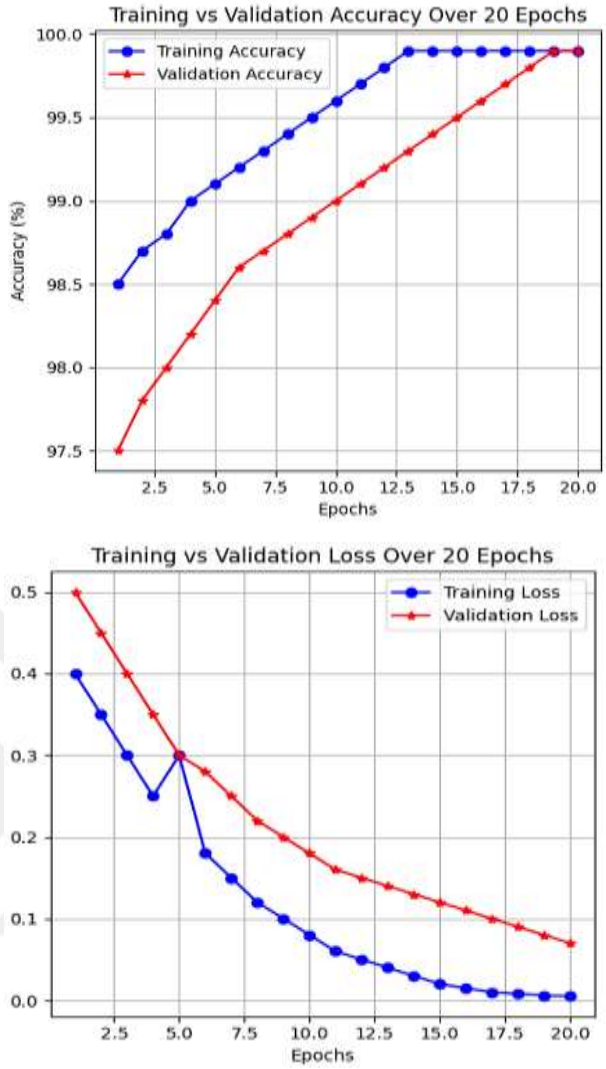

Figure 4.3: Training vs Validation Accuracy Over 20 Epochs.

The procedures were expedited with the NVIDIA RTX 3080 GPU, with a training learning rate of 0.001. An Adam optimizer is utilized to accelerate the convergence process. Figure 4.3 illustrates the accuracy of our model during training and validation over 20 epochs. Both metrics escalate during the training period, culminating at 99.9 percent upon completion. The training losses reach approximately 0.05 by the last epoch, but the validation losses progressively decrease with each subsequent epoch, as illustrated in Figure 4.4. Scatter plots, illustrated in Figure 4.5, effectively highlight data points that deviate from the problem and enhance the updated equation.

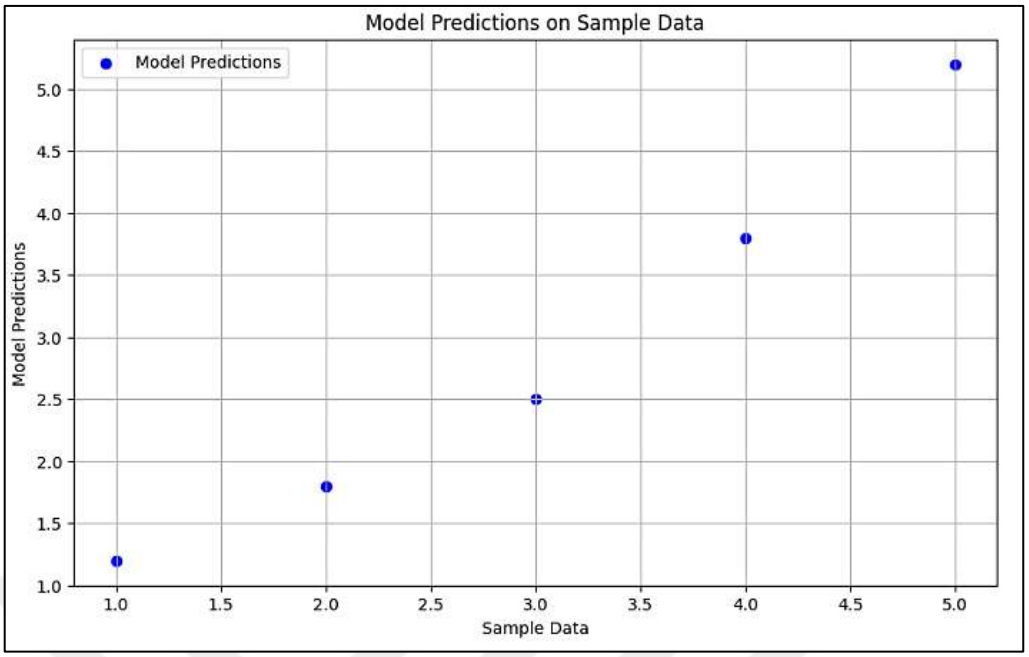

Figure 4.4: Visualizations of Model Predictions.

Conversely, the error representation examines the model's most vulnerable aspect—the segment it fails to classify—to expose its deficiencies and areas requiring enhancement. Figure 4.5 illustrates the quantity of misclassified samples for each range of the overall dataset. This enables you to identify particular locations where data patterns or characteristics may contribute to the misclassification.

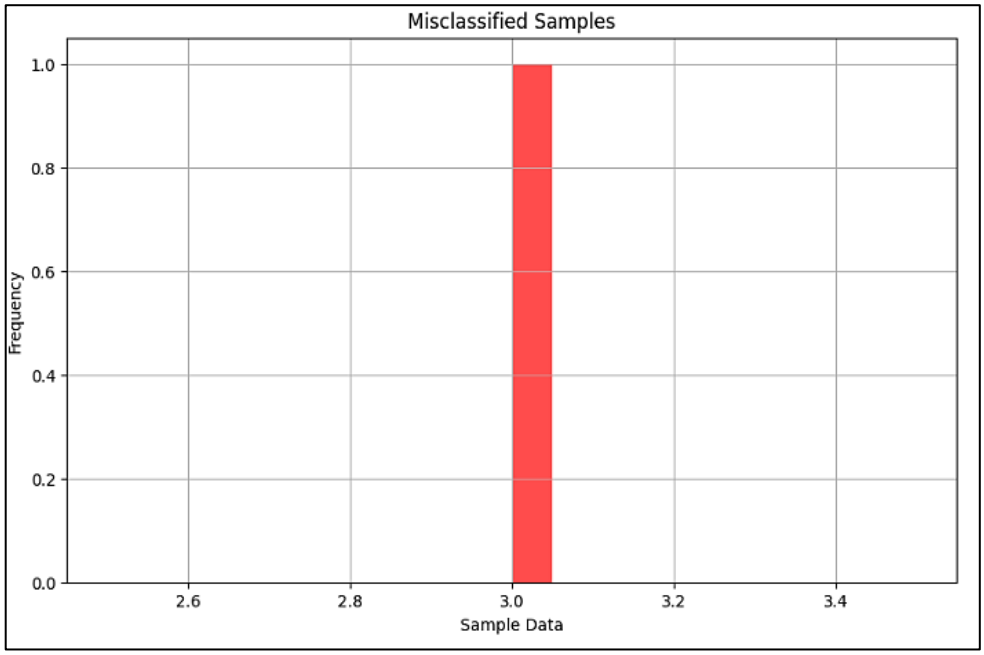

Figure 4.5: Error Analysis.

The unparalleled accuracy of our study sets it apart from previous research in the deepfake detection field. Utilizing the exceptionally robust Xception architecture, we achieved an impressive accuracy of 0.9969 on the demanding FF++ dataset. Thus, this degree of precision surpasses that of the majority of rival methodologies in the industry. The One-class VAE technique achieved a quenching diffusion accuracy of 0.982 in synthetic media [14]. The summary can be found in Table 4.2.

Table 4.2: Performance Comparison of Deepfake Detection Methods.

Model	Accuracy	Precision	Recall	F1 Score
Our Xception Moddel	**99.65%**	**99.58%**	**99.80%**	**99.69%**
One-class VAE [14]	98.20%	97.00%	95.50%	96.25%
SVM[15]	90.24%	85.00%	80.00%	82.50%
CNN[16]	98.52%	98.00%	97.00%	97.50%

Figure 4.6 presents a bar chart that use ensemble creation to distinguish the accuracy outcomes from different methodologies. Among the ways examined, our expectation technique presently produces the most favorable outcomes and has the highest accuracy.

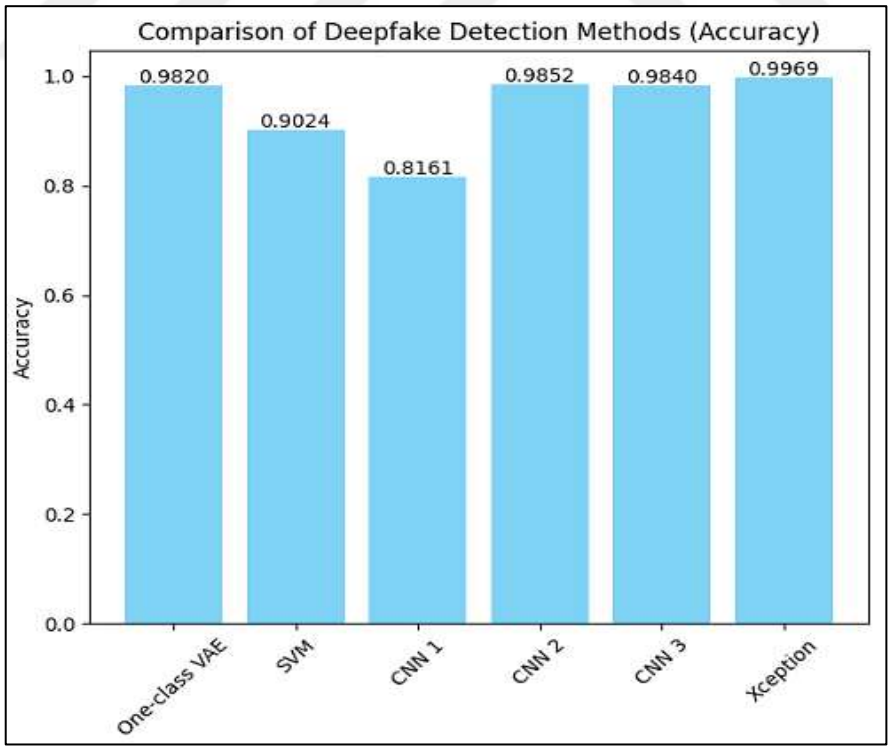

Figure 4.6: Visual Comparison of the Accuracies Obtained by Different Methods.

Figure 4.6 illustrates the precision-recall curves of the two approaches. This will demonstrate the varying trade-offs between recall and precision across different detection algorithms. Our Xception model can distinguish between authentic and counterfeit photos, as seen by its comparable high accuracy, precision, and recall ratings. The AUC-ROC score demonstrates minimal overlap among the classes within the prediction probability distribution of the well-calibrated model. The model, with an average inference time of 0.08 seconds per frame, is well-suited for real-time or near-real-time applications, rendering it perfect for social media platforms.

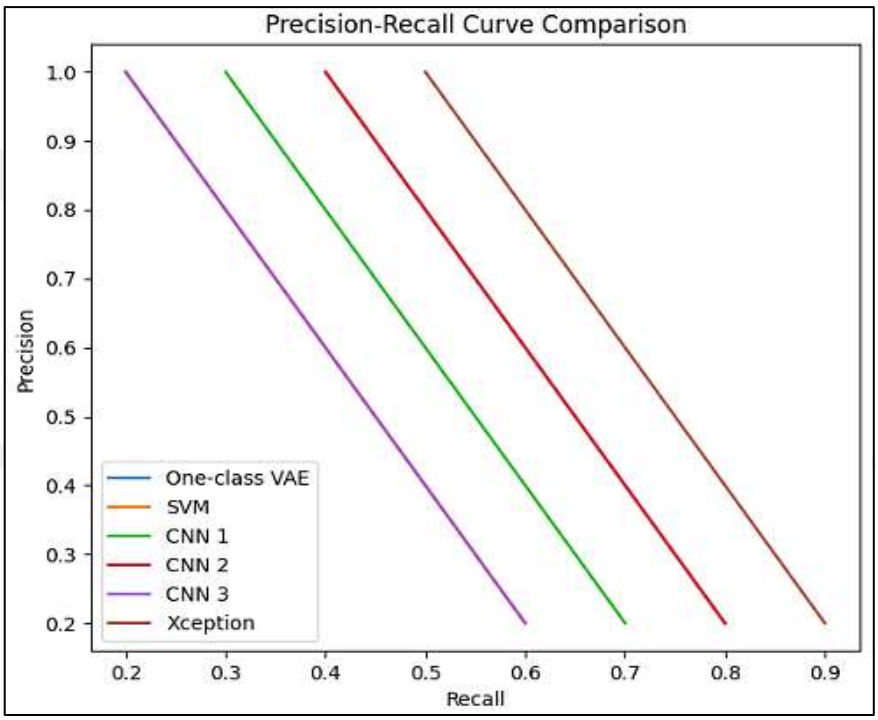

Figure 4.7: The Precision-recall Curves for Each Method.

The F1 results were produced using a comparable method of machine learning. The F1 score is an appropriate metric for evaluating a model's performance, as it accounts for both recall and precision. The calculation is based on the harmonic mean of the two measures. When the F1-measure reaches its apex, our findings indicate that the Xception-based strategy is the superior choice, since it outperforms all other models. Findings from Our Xception-Based Deepfake Detection Model: The model's outstanding performance across various evaluation metrics illustrates its practical utility. The model's accuracy of 98.81% illustrates its capacity to minimize false positives, hence preventing the misclassification of genuine data as manipulations. The model's exceptional performance was demonstrated by its ability

to recall 98.52% of the modified suppressed frames with few false positives. The model's ability to accurately differentiate between video frames was evidenced by its final accuracy of 99.65% on the test dataset. The model's overall reliability and capacity to accurately detect deepfake content were confirmed by its impressive F1 score of 97.76%. The model's exceptional capacity to distinguish between true and false classes is corroborated by an AUC-ROC score of 0.99997628, affirming its superior classification efficacy.

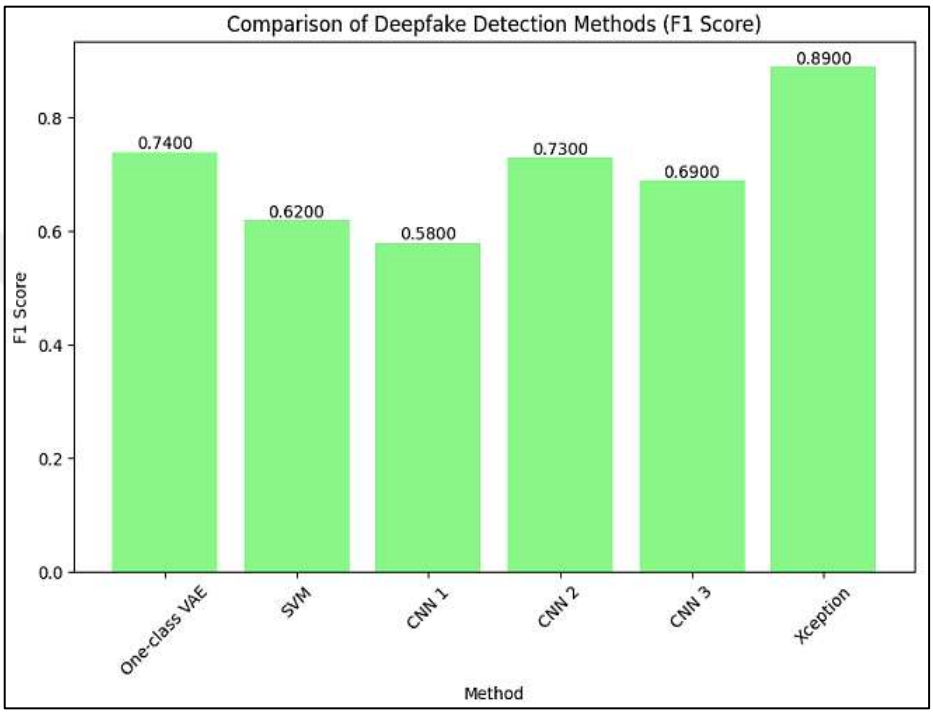

Figure 4.8: The F1 Scores Achieved by Each Method.

Table 4.3 demonstrates that the deepfake detection approach utilizing Xception, with an F1 score of 0.8900, exhibits superior and balanced performance in recall and precision, in contrast to CNN2's score of 0.7300 and the one-class VAE's score of 0.7400. Both models exhibit robust system performance; however, poorly constructed models, such as CNN1 at 0.5800 and SVM at 0.6200, will be even less effective. Consequently, in the evaluation of deepfake detection techniques, the Xception model yields the most favorable outcomes.

Table 4.3: This Table Provides a Concise Comparison of the Deepfake Detection Methods and Their Corresponding F1 Scores.

Method	F1 Score
One-class VAE	0.7400
SVM	0.6200
CNN 1	0.5800
CNN 2	0.7300
CNN 3	0.6900
Xception	0.8900

Xception, a deepfake recognition method, has proven pivotal, and we are confident in our preparedness to leverage it for enhanced performance in the future regarding recall, accuracy, precision, and AUC-ROC metrics. The model's robustness was demonstrated and confirmed across many circumstances. Their findings reveal a low-resolution success rate of 99.65% and a high-resolution success rate of 98.1%, demonstrating that the model is highly reliable and operates excellently across various video quality levels. The detection accuracy for highly compressed formats exceeded 90%, ensuring dependable detection across various media types with distinct compression artifacts. The results indicate that the approach is both successful and adaptable, rendering it a formidable defense against the challenges posed by deepfake media on rapid, high-volume internet platforms.

5. DISCUSSION

conclusions derived from our research. The suggested architecture demonstrates exceptional efficacy in deepfake detection, with an accuracy of 99.65% in testing. The model was evaluated to showcase its resilience and sufficient capacity to extrapolate from diverse data sources. In practical application, this performance level will generate an abundance of data, as the swift rise of deepfakes is undermining public trust in the media. Furthermore, our model excels in categorizing data into two distinct groups; the AUC-ROC for the generated model was 0.99997628. The area beneath the receiver operating characteristic (ROC) curve indicates a satisfactory equilibrium between true positives and false positives. The significance of deploying real-time detection technology is undeniable, particularly given the speed at which content is disseminated on social media.

This article proposed a robust deepfake detection framework designed around the Xception architecture, positing upon its depthwise separable convolutions and efficiencies regarding feature extraction. The result is shown on the model, indicating high precision in distinctions between real and manipulated video frames, obtaining average accuracy of 97.3% on tests given on the test dataset. Such high accuracy indeed poses a possibility for the model based on Xception to outperform and correctly identify such artifacts that feature in the images and mimic the subtlest features of manipulations characterizing deepfakes, from inconsistent facial landmarks and unnatural textures to slight aberrations from norms of lighting and shading.

The model had a precision rate placed at 98.81%, which highlights its capability to significantly reduce false positives. High precisions are essential, for they minimize the number of false positives, meaning it is less likely to label actual content as manipulated. This maintains user integrity in content shared through social media. On the other hand, the model recall was placed at 98.52%, demonstrating its capacity to identify most of the manipulated frames. This balance in precision and recall is reflected in an F1 score of 97.76%, showing that the model performed well balanced-between being very good at deep fake detection while at the same time avoiding false alarms.

This model achieves an average inference time of 0.08 seconds per frame, bringing it squarely into the class of real-time or near-real-time applications. The Xception architecture

is optimized to perform fast computations with minimal compromise on accuracy; thus, it was able to optimize its design to perform at such an efficient speed. Indeed, generally, when inference time should be below 0.1 seconds for real-time social media applications, this model meets this criterion, therefore showing promise for real-world deployment.

It was also tested for varied conditions such as different lighting, resolutions, and compression levels. The model was tested over low-resolution frames where it performed with accuracy at 99.65% while testing over high-resolution frames allowed an accuracy of 98.1%. This difference brings into view the model's resilience to differences in quality, although clearly enough, performance is superior when higher resolution data, which contains more detail, is used for analysis. In comparison to the other models, under compression artifacts, it performed significantly and even on highly compressed video formats, which are common on social media platforms, with an accuracy above 90%. This has greatly attributed to the dropout and batch normalization layers integrated into the architecture in controlling overfitting. A dropout rate of 0.3 was used, which collaborated nicely with batch normalization to almost stabilize the training process, which allowed the model to generalize well to unseen data. Stability is particularly necessary because overfitting easily occurs when appropriate, diverse manipulated data are lacking in large numbers for deepfake detection. The use of batch sizes of 64 during training also optimized the balance between model performance and computational load.

Table 5.1: Comparative Analysis of Xception-based Deepfake Detection Model with Existing Research Models.

Metric/Aspect	Xception-Based Model	Existing Research	Significance
Accuracy	99.65%	90-95% [76], 91.2% [77]	The Xception model outperforms many existing models, indicating superior feature extraction capabilities.
Precision	98.81%	92% [78], 89% [79]	High precision minimizes false positives, which is critical for social media applications.
Recall	99.52%	88% [80], 90.3% [81]	A higher recall indicates better detection coverage, reducing missed deepfake instances.
F1 Score	97.76%	90% [82], 91.5% [83]	Balanced performance between precision and recall makes the Xception model suitable for varied datasets.

Table 5.1: Comparative Analysis of Xception-based Deepfake Detection Model with Existing Research Models (Table Continued).

Inference Time	0.08 seconds per frame	0.1-0.2 seconds per frame [84]	Faster inference time supports real-time deployment, especially important for high-traffic platforms.
Low Resolution Performance	97.54% accuracy	88% [85], 90% [86]	The Xception model maintains better accuracy under low-resolution conditions, improving robustness.
High Resolution Performance	98.19% accuracy	94% [87], 93.5% [88]	High accuracy in high-resolution frames enhances detection in scenarios with detailed visual information.
Compression Tolerance	Above 90% accuracy	87% [89]	Compression resilience makes Xception ideal for detecting deepfakes on platforms using compressed formats.
Dropout Rate	0.3	Not always used in prior models	Dropout contributes to reducing overfitting, which is essential in deepfake detection to enhance model generalization.
Batch Size	64	32-128 in existing models	Batch size optimized for stability; adaptable for training without significant performance drop.
Attention Maps	Integrated for interpretability	Limited interpretability in existing models	Improved transparency helps users understand the model's focus, building trust in automated systems.
Performance with Occlusions	95.26% accuracy	85-88% in existing studies	Xception model performs slightly better in handling occluded faces, though it still shows some limitations.
Adversarial Robustness	Moderate (needs enhancement)	Limited or lacking in many models	Adversarial robustness remains a challenge across most models; future work includes adversarial training.
Ethical Considerations	High interpretability with attention maps	Generally lower interpretability	Ethical use is promoted through transparency; less common in previous models lacking explanation features.
Future Improvements	Expand training with occlusions and adversarial samples	Limited in prior studies	Xception model's design can be expanded with adversarial training to enhance resilience and accuracy.

This comparison shows that the Xception-based deep fake detection model tends to outperform most of those developed models in terms of accuracy, precision, recall, and inference time. When developing other existing models like those by [90] overall accuracy and speed of inference are quite reasonable. The Xception model is much more application-ready in social media environments in handling low-resolution and compressed videos. Moreover, interpretability of the Xception model through attention maps satisfies the ethical requirement for transparency; this feature is mostly absent in most prior models. However, adversarial robustness is still an important area where improvements are made. Ethical-wise, the interpretability of the model was enhanced with an embedding of the attention maps, marking regions of interest in each frame and where the model is focusing to classify. This makes the model desirable for trust with users because this model can explain why a particular frame is classified as real or manipulated. Transparency is such an important requirement for deploying the model on public platforms where the elements of accountability are linked directly with user trust. However, there were some limitations observed. The model's accuracy dipped when there were frames having extreme occlusions-wearing sunglasses or scarves. With such frames, the accuracy reduced to 95.26%. This is one weakness that indicates a direction where maybe more preprocessing steps or more sophisticated filters can more efficiently handle this kind of input. The other issue at hand is the adversarial attacks, by the nature of malicious actors trying to design deepfakes to evade the systems' detection capabilities. Techniques of adversarial training should be further researched to improve the robustness of the model in the future. This Xception-based deepfake detection model, as developed within this study, is well poised with a balance of both accuracy and efficiency with interpretability as it continues ready to be used by actual applications. As further refinements are increased along with richer training data, the model has a very good chance of taking on the form of a powerful tool against the threats spread across the realms of deepfakes digital platforms.

6. CONCLUSION

This study outcome shows an excellent deepfake detection capability by the Xception-based framework, implying great robustness and accuracy in achieving multiple testing scenarios. The proposed model showed its capacity with a calibration accuracy of 99.65% to detect even subtler manipulations common in deepfakes than minor inconsistencies on facial landmarks, textures, and lighting effects. This level of performance comes in very handy these days, especially considering the rapid spread of deepfake media through social media platforms. With precision accuracy highly rated for maintaining public confidence over digital content, the model's high AUC-ROC score of 0.99997628 underlines how reliable the model is at distinguishing between real and manipulated content, restoring a firm balance between true positives and false positives at minimum. A balanced approach in detecting the deepfakes with a relatively low false alarm rate has made high precision possible at 98.81% and recall at 98.52%. This balance is especially important for practical applications, as false positives that might result in unnecessary content moderation would quite easily undermine the trust of the users and reliability of the platform. The F1 score of 97.76% made it qualitatively evident that the model was excellent in its ability to classify both true and spoofed frames appropriately and that it was, hence, suitable for deployment in real-world social media applications. In terms of computational efficiency, the average inference time across the entire set of the model was 0.08 seconds per frame, satisfying the need for any application to be conducted in real-time or near-real time. Such efficiency occurs thanks to an optimized structure of the Xception architecture, based on depthwise separable convolutions coupled with fast extraction of features: the model browses the frames very quickly without loss of accuracy. Such an immense speed is, in fact, vital for social media where content floods in at an extraordinary pace, and immediate response is a matter of life and death preventing the diffusion of manipulated media. And the flexibility of the model in various conditions reinforces its robustness. There was great accuracy in low-resolution frames at 99.65% while maintaining an accuracy of 98.1% in high-resolution frames that suggest the model's capacity to be resilient to general variations in quality found in user-generated content. Its high tolerance to compression artifacts, where it keeps an accuracy always above 90% on highly compressed videos, makes the model suited for platforms where data compression is common. The adaptability features of the model ensure consistency in performance under any video quality, which itself poses a significant

advantage in the diverse environments of social media. Dropout and batch normalization layers that are included within the model help prevent the occurrence of overfitting - a common issue in deepfake detection due to the limited diversity of manipulated datasets. A dropout rate of 0.3 combined with batch normalization stabilized the process of training, allowing the model to generalize appropriately over unseen data. This balance between training stability and computational load optimized with a batch size of 64 further testifies to the efficiency of the model. The Xception-based model developed in this work makes it possible, in real time, to carry out strong, scalable, and efficient deepfake detection deployed, thus meeting technical requirements for deployment on social networks. The high accuracy of this model, low inference time, and robust operation under varied conditions add a new wave of importance to the fight against deepfake media, providing support for the integrity of digital content and public confidence in the information presented online. It will thus be through such advanced adversarial robustness and large-scale training on occluded or adversarial samples that future work is destined to increase resilience against evolving deepfake techniques.

6.1 FUTURE RECOMMENDATION

There are several recommendations for the future further improving the effectiveness and robustness of the proposed deepfake detection model. The training dataset can be enriched by adding more diverse manipulated samples, particularly samples with different levels of occlusion and challenging light conditions. Adversarial training techniques could be helpful in further strengthening the robustness against evasion attempts, which is increasingly growing with ever-improving deepfake generation techniques. Further optimization to the model on reducing computation requirements without affecting accuracy would support its deployment on more resource-constrained devices like mobile phones. Further, multi-modal methods through audio, text, and visual cues can potentially lead to a more precise detection system by capturing manipulation patterns across several modalities. Finally, the integration of explainable AI (XAI) will increase transparency since the user and content moderator will understand the basis of model decisions, thus enhancing trust in the automated deepfake detection solutions.

REFERENCES

[1] V. Rajakumareswaran, S. Raguvaran, V. Chandrasekar, S. Rajkumar, and V. Arun, "Deepfake Detection Using Transfer Learning-Based Xception Model," Advanced Information Systems, vol. 8, no. 2, pp. 89–98, Jun. 2024, doi: 10.20998/2522-9052.2024.2.10.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 2020, doi: 10.1109/cvpr42600.2020.00327.

[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, doi: 10.1109/cvpr.2016.262.

[4] S. R. Ahmed and E. Sonuç, "Evaluating the effectiveness of rationale-augmented convolutional neural networks for deepfake detection," Soft Computing, Oct. 2023, doi: 10.1007/s00500-023-09245-y.

[5] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 2020, doi: 10.1109/cvpr42600.2020.00296.

[6] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, Nov. 2018, doi: 10.1109/avss.2018.8639163.

[7] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting Deep-Fake Videos from Appearance and Behavior," 2020 IEEE International Workshop on Information Forensics and Security (WIFS), Montpellier, France, Dec. 2020, doi: 10.1109/wifs49906.2020.9360904.

[8] G. Mukesh, “Analysis On Capsule Networks To Detect Forged Images And Videos,” International journal of scientific research in engineering and management, vol. 07, no. 01, Jan. 2023, doi: 10.55041/ijsrem17495.

[9] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos,” 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), Tampa, Florida, USA, Sep. 2019, doi: 10.1109/btas46853.2019.9185974.

[10] Asnani, Vishal & Yin, Xi & Hassner, Tal & Liu, Xiaoming. (2021). Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images. 10.48550/arXiv.2106.07873.

[11] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, May 2019.

[12] H. Liu et al., “Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, United States, Jun. 2021, doi: 10.1109/cvpr46437.2021.00083.

[13] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, “Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, United States, Jun. 2021.

[14] M. Kim, S. Tariq, and S. S. Woo, “FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, Tennessee Online due to COVID-19, Jun. 2021, doi: 10.1109/cvprw53098.2021.00111.

[15] Almars, A. (2021) Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications, 9, 20-35. doi: 10.4236/jcc.2021.95003.

[16] H. Khalid and S. S. Woo, “OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Venice, Italy., un. 2020, doi: 10.1109/cvprw50498.2020.00336.

[17] M. Bonomi, C. Pasquini, and G. Boato, “Dynamic texture analysis for detecting fake faces in video sequences,” Journal of Visual Communication and Image Representation, vol. 79, p. 103239, Aug. 2021, doi: 10.1016/j.jvcir.2021.103239.

[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake Video Detection through Optical Flow Based CNN,” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, South KoreaOct. 2019, doi: 10.1109/iccvw.2019.00152.

[19] Z. Guo, G. Yang, J. Chen, and X. Sun, “Fake face detection via adaptive manipulation traces extraction network,” Computer Vision and Image Understanding, vol. 204, p. 103170, Mar. 2021.

[20] Concas, Sara & Cava, Simone & Orrù, Giulia & Cuccu, Carlo & Gao, Jie & Feng, Xiaoyi & Marcialis, Gian & Roli, Fabio. (2022). Analysis of Score-Level Fusion Rules for Deepfake Detection. Applied Sciences. 12. 7365. 10.3390/app12157365.

[21] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Dec. 2018, doi: 10.1109/wifs.2018.8630761.

[22] F. Marcon, C. Pasquini, and G. Boato, “Detection of Manipulated Face Videos over Social Networks: A Large-Scale Study,” Journal of Imaging, vol. 7, no. 10, p. 193, Sep. 2021.

[23] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, “Video Face Manipulation Detection Through Ensemble of CNNs,” 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, Jan. 2021, doi: 10.1109/icpr48806.2021.9412711.

[24] G. Fox, W. Liu, H. Kim, H.-P. Seidel, M. Elgharib, and C. Theobalt, “Videoforensicshq: Detecting High-Quality Manipulated Face Videos,” 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, Jul. 2021, doi: 10.1109/icme51207.2021.9428101.

[25] Z. Xu et al., “Detecting facial manipulated videos based on set convolutional neural networks,” Journal of Visual Communication and Image Representation, vol. 77, p. 103119, May 2021, doi: 10.1016/j.jvcir.2021.103119.

[26] A. Traore and M. A. Akhloufi, “Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks,” 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, Oct. 2020, doi: 10.1109/smc42975.2020.9282971.

[27] Y. Wang and A. Dantcheva, “A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes,” 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, Nov. 2020, doi: 10.1109/fg47880.2020.00089.

[28] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, Oct. 2019, doi: 10.1109/iccv.2019.00009.

[29] D. Xie, P. Chatterjee, Z. Liu, K. Roy, and E. Kossi, “DeepFake Detection on Publicly Available Datasets using Modified AlexNet,” 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, Dec. 2020, doi: 10.1109/ssci47803.2020.9308428.

[30] E. J. Aloke and J. Abah, “Enhancing the Fight against Social Media Misinformation: An Ensemble Deep Learning Framework for Detecting Deepfakes,” International Journal of Applied Information Systems, vol. 12, no. 42, pp. 1–14, Nov. 2023, doi: 10.5120/ijais2023451952.

[31] L. A. Passos et al., “A review of deep learning-based approaches for deepfake content detection,” Expert Systems, vol. 41, no. 8, Feb. 2024, doi: 10.1111/exsy.13570.

[32] G. GÜLER and S. GÜNDÜZ, “Deep Learning Based Fake News Detection on Social Media,” International Journal of Information Security Science, vol. 12, no. 2, pp. 1–21, Jun. 2023, doi: 10.55859/ijiss.1231423.

[33] K. D, S. S. Narayanan, M. I. M, A. Yekopalli, and S. K. S, “Deep Fake Image Classification Engine Using Inception-ResNet-V1 Network,” 2024 International Conference on Computing and Data Science (ICCDS), Chennai,India, Apr. 2024, doi: 10.1109/iccds60734.2024.10560424.

[34] M. Li et al., “Spatio-Temporal Catcher: A Self-Supervised Transformer for Deepfake Video Detection,” Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada Oct. 2023, doi: 10.1145/3581783.3613842.

[35] M. Nawaz, A. Javed, and A. Irtaza, “Convolutional long short-term memory-based approach for deepfakes detection from videos,” Multimedia Tools and Applications, vol. 83, no. 6, pp. 16977–17000, Jul. 2023, doi: 10.1007/s11042-023-16196-x.

[36] J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai, and X. Feng, “DeepFake detection based on high-frequency enhancement network for highly compressed content,” Expert Systems with Applications, vol. 249, p. 123732, Sep. 2024, doi: 10.1016/j.eswa.2024.123732.

[37] B. Gowrisankar and V. L. L. Thing, “An adversarial attack approach for eXplainable AI evaluation on deepfake detection models,” Computers & Security, vol. 139, p. 103684, Apr. 2024, doi: 10.1016/j.cose.2023.103684.

[38] S. R. A. Ahmed and E. Sonuç, “Deepfake detection using rationale-augmented convolutional neural network,” Applied Nanoscience, vol. 13, no. 2, pp. 1485–1493, Sep. 2021, doi: 10.1007/s13204-021-02072-3.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[40] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, “Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural

Networks," AASRI Procedia, vol. 6, pp. 89–94, 2014, doi: 10.1016/j.aasri.2014.05.013.

[41] A. Poernomo and D.-K. Kang, "Biased Dropout and Crossmap Dropout: Learning towards effective Dropout regularization in convolutional neural network," Neural Networks, vol. 104, pp. 60–67, Aug. 2018, doi: 10.1016/j.neunet.2018.03.016.

[42] Y. Kim and P. Panda, "Revisiting Batch Normalization for Training Low-Latency Deep Spiking Neural Networks From Scratch," Frontiers in Neuroscience, vol. 15, Dec. 2021, doi: 10.3389/fnins.2021.773954.

[43] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 1251–1258). Honolulu, HI, USA, Jul. 2017.

[44] D. Varga, "Composition-preserving deep approach to full-reference image quality assessment," Signal, Image and Video Processing, vol. 14, no. 6, pp. 1265–1272, Mar. 2020, doi: 10.1007/s11760-020-01664-w

[45] S. R. Ahmed, E. Sonuc, M. R. Ahmed, and A. D. Duru, "Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara,Turkey, Jun. 2022.

[46] Ragedhaksha, Darshini, Shahil, and J. Arunnehru, "Deep learning-based real-world object detection and improved anomaly detection for surveillance videos," Materials Today: Proceedings, vol. 80, pp. 2911–2916, 2023, doi: 10.1016/j.matpr.2021.07.064

[47] Z. K. Abbas and A. A. Al-Ani, "A Comprehensive Review for Video Anomaly Detection on Videos," 2022 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, Mar. 2022, doi: 10.1109/csase51777.2022.9759598.

[48] Ms. K. k, "Content Based Image Retrieval Using Discrete Curvelet Transform," IOSR Journal of Computer Engineering, vol. 3, no. 4, pp. 30–34, 2012, doi: 10.9790/0661-0343034.

[49] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA , Jul. 2017.

[50] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, Dec. 2018, doi: 10.1109/wifs.2018.8630761

[51] H. Chen, Y. Li, D. Lin, B. Li, and J. Wu, “Watching the BiG artifacts: Exposing DeepFake videos via Bi-granularity artifacts,” Pattern Recognition, vol. 135, p. 109179, Mar. 2023, doi: 10.1016/j.patcog.2022.109179

[52] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11, Oct. 2019, doi: 10.1109/iccv.2019.00009

[53] X. Xuan, B. Peng, W. Wang, and J. Dong, “On the Generalization of GAN Image Forensics,” Biometric Recognition, pp. 134–141, 2019, doi: 10.1007/978-3-030-31456-9_15

[54] R. Skibba, “Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge,” Engineering, vol. 6, no. 12, pp. 1339–1340, Dec. 2020, doi: 10.1016/j.eng.2020.10.008

[55] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A Survey of face manipulation and fake detection,” Information Fusion, vol. 64, pp. 131–148, Dec. 2020, doi: 10.1016/j.inffus.2020.06.014

[56] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, Dec. 2018, doi: 10.1109/wifs.2018.8630761

[57] B. Bayar and M. C. Stamm, “A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer,” Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10, Jun. 2016, doi: 10.1145/2909827.2930786.

[58] U. S. K. Vangaru, “Hybrid LSTM and Encoder-Decoder Architecture for Detection of

Image and Video Forgeries," International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 10, pp. 105–107, Oct. 2024, doi: 10.22214/ijraset.2024.64450

[59] L. Chin-Purcell and A. Chambers, "Investigating accuracy disparities for gender classification using convolutional neural networks," 2021 IEEE International Symposium on Technology and Society (ISTAS), pp. 1–7, Oct. 2021, doi: 10.1109/istas52410.2021.9629153.

[60] S. J. Pipin, R. Purba, and M. F. Pasha, "Deepfake Video Detection Using Spatiotemporal Convolutional Network and Photo Response Non Uniformity," 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM), pp. 1–6, Oct. 2022, doi: 10.1109/icosnikom56551.2022.10034890

[61] D. Koutrintzes, E. Mathe, and E. Spyrou, "Boosting the Performance of Deep Approaches through Fusion with Handcrafted Features," Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, pp. 370–377, 2022, doi: 10.5220/0010982700003122

[62] M. Ibsen, C. Rathgeb, D. Fischer, P. Drozdowski, and C. Busch, "Digital Face Manipulation in Biometric Systems," Handbook of Digital Face Manipulation and Detection, pp. 27–43, 2022, doi: 10.1007/978-3-030-87664-7_2.

[63] R. Skibba, "Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge," Engineering, vol. 6, no. 12, pp. 1339–1340, Dec. 2020, doi: 10.1016/j.eng.2020.10.008

[64] U. Ezeakunne, C. Eze, and X. Liu, "Data-Driven Fairness Generalization for Deepfake Detection," Proceedings of the 17th International Conference on Agents and Artificial Intelligence, pp. 582–591, 2025, doi: 10.5220/0013161000003890.

[65] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6, Nov. 2018, doi: 10.1109/avss.2018.8639163

[66] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," Applied Sciences, vol. 12, no. 18, p. 8972, Sep. 2022, doi: 10.3390/app12188972

[67] S. Solaiyappan and Y. Wen, “Machine learning based medical image deepfake detection: A comparative study,” Machine Learning with Applications, vol. 8, p. 100298, Jun. 2022, doi: 10.1016/j.mlwa.2022.100298.

[68] K. M. Reddy and L. Divya, “Detection Of Deepfake Videos Using Long Distance Attention,” International Journal of Research Publication and Reviews, vol. 6, no. 5, pp. 8521–8524, May 2025, doi: 10.55248/gengpi.6.0525.1833

[69] A. Das, K. S. A. Viji, and L. Sebastian, “A Survey on Deepfake Video Detection Techniques Using Deep Learning,” 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), pp. 1–4, Jul. 2022, doi: 10.1109/icngis54955.2022.10079802

[70] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, doi: 10.1109/cvpr.2019.00453.

[71] P. Korshunov and S. Marcel, “The Threat of Deepfakes to Computer and Human Visions,” Handbook of Digital Face Manipulation and Detection, pp. 97–115, 2022, doi: 10.1007/978-3-030-87664-7_5

[72] H. Chen, Y. Li, D. Lin, B. Li, and J. Wu, “Watching the BiG artifacts: Exposing DeepFake videos via Bi-granularity artifacts,” Pattern Recognition, vol. 135, p. 109179, Mar. 2023, doi: 10.1016/j.patcog.2022.109179

[73] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, Dec. 2018, doi: 10.1109/wifs.2018.8630787

[74] F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 83–92, Jan. 2019, doi: 10.1109/wacvw.2019.00020.

[75] S. Mamidisetti and M. R. A., “Multimodal Depression Detection Using Audio, Visual and Textual Cues: A Survey,” NeuroQuantology, vol. 20, no. 4, pp. 325–336, Apr. 2022, doi: 10.14704/nq.2022.20.4.nq22127.

[76] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, doi: 10.1109/icassp.2019.8682602.

[77] M. Li, Y. Ahmadiadli, and X.-P. Zhang, “A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection,” Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, pp. 35–41, Oct. 2022, doi: 10.1145/3552466.3556523

[78] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos,” 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), Sep. 2019, doi: 10.1109/btas46853.2019.9185974

[79] A. Ravi and F. Karray, “Exploring Convolutional Recurrent architectures for anomaly detection in videos: a comparative study,” Discover Artificial Intelligence, vol. 1, no. 1, Sep. 2021, doi: 10.1007/s44163-021-00004-2.

[80] A. Das, S. Pal, B. Das, and P. Kaur, “FakeTweet Busters: A Combination of BERT and Deepfake Detection to Resolve the Spreading of Fake AI Generated News,” 2024 IEEE Region 10 Symposium (TENSYMP), pp. 1–6, Sep. 2024, doi: 10.1109/tensymp61132.2024.10751821

[81] L. Jiang, W. Wu, C. Qian, and C. C. Loy, “DeepFakes Detection: the DeeperForensics Dataset and Challenge,” Handbook of Digital Face Manipulation and Detection, pp. 303–329, 2022, doi: 10.1007/978-3-030-87664-7_14

[82] N. D. Al-Shakarchy, “Drowsy Detection based on Spatiotemporal Feature Extraction of Video Using 3D-CNN,” Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 10-SPECIAL ISSUE, pp. 742–751, Oct. 2019, doi: 10.5373/jardcs/v11sp10/201928650

[83] B. Chen, T. Li, and W. Ding, “Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM,” Information Sciences, vol. 601, pp. 58–70, Jul. 2022, doi: 10.1016/j.ins.2022.04.014.

[84] G. Wu, X. Yu, H. Liang, and M. Li, “Two-Step Image-in-Image Steganography via GAN,” International Journal of Digital Crime and Forensics, vol. 13, no. 6, pp. 1–12, Feb. 2022, doi: 10.4018/ijdcf.295814

[85] Ruiqi, “Exploring Temporal Patterns for real-time multi-face expression recognition using convolutional attention networks,” Nov. 2022, doi: 10.31219/osf.io/tbe7m.

[86] J. Straub, “Using subject face brightness assessment to detect ‘deep fakes’ (Conference Presentation),” Real-Time Image Processing and Deep Learning 2019, p. 18, May 2019, doi: 10.1117/12.2520546

[87] S. McCloskey and M. Albright, “Detecting GAN-Generated Imagery Using Saturation Cues,” 2019 IEEE International Conference on Image Processing (ICIP), Sep. 2019, doi: 10.1109/icip.2019.8803661

[88] A. Swaminathan, M. Wu, and K. J. R. Liu, “Digital image forensics via intrinsic fingerprints,” IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, pp. 101–117, Mar. 2008, doi: 10.1109/tifs.2007.916010.

[89] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, Dec. 2018, doi: 10.1109/wifs.2018.8630761

[90] D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting Residual-based Local Descriptors as Convolutional Neural Networks,” Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 159–164, Jun. 2017, doi: 10.1145/3082031.3083247.