UNITEC INSTITUTE OF TECHNOLOGY

MASTER THESIS

Evaluating Deepfake Detection Models: A Comprehensive Framework for Comparison Across Diverse Datasets

Author:

Babita Dasoar Mangotra

Principal Supervisor:

Dr. Soheil Varastehpour

Co-supervisor:

Dr. Masoud Shakiba

A thesis presented in partial fulfilment of the requirements

for the degree of Master of Applied Technologies

in the

School of Computing, Electrical & Applied Technology

February 21, 2025

ABSTRACT

Rapid advancements in deepfake technology pose serious threats to public confidence in digital material, cybersecurity, and the ability to authenticate media. Deepfake movies have gotten progressively realistic with new artificial intelligence developments, especially in Generative ANs and autoencoders, which makes detection more difficult. By creating and assessing ML models for recognizing deep fakes and using several datasets to improve generalizability and robustness, our work seeks to solve these difficulties.

Celeb-DF V2, FaceForensics++, CIPLib, Deepfake Detection Challenge Dataset, as well as Fake Celeb AV-V1.2 are among the datasets this work uses. Every dataset is chosen depending on their variance in video alteration techniques, resolution changes, and source authenticity to guarantee a thorough assessment of machine learning models among several deepfake generating approaches. Steps in data preparation augmentation, normalisation, feature extraction are used to improve model training's efficacy.

Then we investigate several deepfake detection models: Forensics Convolutional Neural Networks, Vision Transformers, Masked Autoencoders, Two-Stream Neural Network as well as Common Fake Feature Network in a comparative manner. Every model is evaluated in relation to computing efficiency, generalizing across datasets, and efficacy in reducing detection mistakes. Performance data including validation and training accuracy, loss functions, and error rates including equal error rate, half total error rate, false acceptance rate, and false rejection rate is examined using a standard assessment framework.

While Vision Transformers and Two-Stream Neural Network models show great detection ability, experimental results show that the Common Fake Feature Network model routinely beats others, obtaining almost zero error rates and up to 100% validation accuracy on demanding datasets including Fake Celeb AV-V1.2 and Deepfake Detection Challenge Dataset. Common Fake Feature Network exceptional success can be ascribed to its capacity to use sophisticated feature extraction and classification approaches to capture fine-grained alterations in deepfake films.

The results of this work underline the need of constantly developing machine learning methods to properly fight digital disinformation since they show the need of managing detection accuracy with error minimizing. Furthermore, underlined by this study is the crucial need of strong deep fake identification systems in digital integrity, cybersecurity, and media forensics. This work adds important new perspectives on the continuous battle against deepfake dangers by building a disciplined evaluation approach and contrasting several SotA detecting techniques. Future research will concentrate on increasing model adaptability, computing efficiency, and integration of present time deep fake detection for pragmatic application in digital security systems.

KEYWORDS: Deepfake Detection, GAN, Machine Learning, DFDC, Common Fake Feature Networks, Media Authentication, Vision Transformers, Error Rate Analysis.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Soheil Varastehpour and co-supervisor, Dr. Masoud Shakiba for their invaluable guidance, support, and encouragement throughout my journey. Their knowledge and advice were significant in developing this thesis.

I would also express my sincere gratitude to my professors and lecturers at Unitec Institute of Technology for their mentorship and expertise, which have made a substantial impact on my academic development.

An extra special thanks to my family and friends for their unwavering support, patience, and encouragement. My greatest source of courage during this process has been their confidence in me.

This journey would not have been possible without the support of all these incredible individuals.

Table of Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

ACC	Accuracy
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BNNs	Binary Neural Networks
CFFN	Common Fake Feature Network
CNN	Convolutional Neural Network
DL	Deep Learning
DFDC	Deepfake Detection Challenge Dataset
DCNN	Deep Convolutional Neural Network
EER	Equal Error Rate
EDA	Exploratory Data Analysis
FAR	False Acceptance Rate
FRR	False Rejection Rate
FCNN	Forensic Convolutional Neural Network
FFT	Fast Fourier Transform
FF++	FaceForensics++
GANs	Generative Adversarial Networks
GRU	Gate Recurrent Unit
HTER	Half Total Error Rate
LSTM	Long Short-Term Memory Network

Loss	Loss Function
LBP	Local Binary Pattern
MAEs	Masked Autoencoders
ML	Machine Learning
MLP	Multilayer Perceptron
M2TR	Multi-modal Multi-scale Transformer
PGGANs	Progressive Expanding Generative Adversarial Networks
ReLU	Rectified Linear Unit
RNNs	Recurrent Neural Networks
SRGANs	Super-Resolution Generative Adversarial Networks
SotA	State-of-the-Art
TSNN	Two-Stream Neural Network
Val	Validation
VAEs	Vibrating Auto Encoders
ViTs	Vision Transformers

CHAPTER 1

1. INTRODUCTION

1.1 OVERVIEW

The greatest distinctive feature that differentiates a human is the face. The advancement of face synthesis technology is leading to an increase in face manipulation, hence elevating security risks [1]. Deepfake implies the manipulation of facial expressions or the substitution of faces in visual content [2]. The most problematic and dangerous for society are the deepfake films, which have one person's face placed over another's [3]. Deepfake technology is a branch of Artificial Intelligence (AI) that uses algorithms to mimic human looks and behaviour, frequently to manipulate or create fake audiovisual content. This subject is exciting as well as controversial, with effects that reach into cybersecurity, entertainment, and deception.

In the age of digital communication and increasing social media utilisation, the integrity of visual content has grown to be an important concern [4]. Photographs, audio recordings, and videos are frequently used as reliable evidence in various investigations to resolve legal disputes. Yet the deepfake phenomenon may have compromised the reliability of these sources of evidence [5]. Regrettably, deepfake content is proliferating at an unprecedented rate in the twenty-first century due to the widespread use of mobile phones and the emergence of several social networking platforms, posing a global danger [1]. Face-swap applications like "DeepFaceLab" and "FakeApp" allow the creation of deepfakes for various unethical goals, allowing wide usage [3]. People often have trouble to distinguish between a real film and a fake using their visual or hearing experience [6].

Deepfake technology uses complex algorithms to create audio and video that look extremely real yet are completely fake. This indicates the starting point of an entirely new era of digital deception. The latest developments in AI, namely in Generative Adversarial Networks (GANs) [7] and autoencoders, allow deepfakes to create content that closely imitates real video in terms of appearance as well as sound. Technological advancement significantly impacts a person's confidentiality, security, and battle against false information spread. As a result, the research and development of effective detection techniques needs to be prioritised. Deepfakes offer an important challenge over many fields, including cybersecurity, politics, and social media. Due to the capability of creating fake content that looks real, lead to lesser trust of people in digital media, a rise in misinformation, and manipulation of people's trust.

As a result of the fast development of deepfake technology, it is necessary to develop more advanced techniques for detecting and identifying such fake content. An urgent requirement exists for more resilient and flexible detection systems, as traditional methods have not kept up with the fast evolution of deepfake technologies. Deepfake detection methodologies are primarily categorised into classical approaches, Machine Learning (ML) techniques, and Deep Learning (DL) methods.

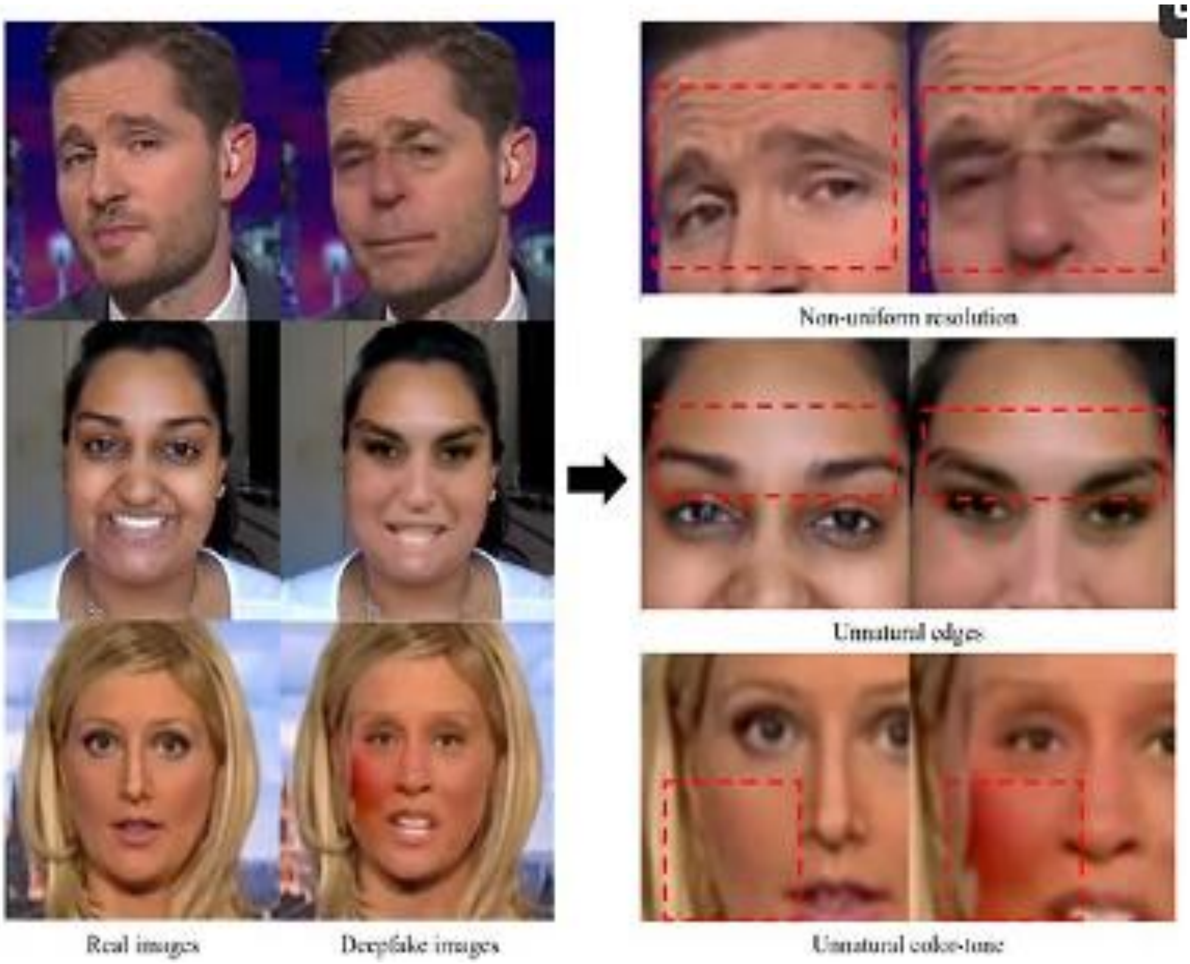

FIGURE 1: DEEPFAKE DETECTION [8]

Conventional approaches often focus on identifying media abnormalities or inconsistencies, such as discrepancies in face movements or audio distortions. Although these strategies have demonstrated considerable efficacy, more advanced deepfake generation algorithms are outpacing them. Following training on extensive datasets, ML algorithms applied to them exhibit traits indicative of deepfakes. These methods work better, but the performance depends on how good or variant the training data is. DL methods, which employ complex neural networks that detect minute differences, act as a demonstration of advanced detecting technology. The significant amount of computing resources necessary and their need for periodic updates create challenges to their ability to remain contemporary with the evolving deepfake technology. Even though detection methods have evolved, careful study and evaluation of different techniques are still quite uncommon. The current study often looks at specific detection areas or methods. We need a clear framework to assess and compare different detection strategies. This type of framework is essential for the purpose of distinguishing the benefits and drawbacks of various strategies, facilitating further research, and demonstrating the advancement of more effective detection technologies. The thesis addresses the gap by giving an extensive framework examining deepfake detection techniques. This framework shows a systematic approach for evaluating the effectiveness of many detection techniques, including standards like flexibility, efficiency, and accuracy. The framework will conduct a systematic comparison of conventional methods, ML-based methods, and DL-based approaches. This comparison will underscore the strengths and limitations of each, thereby influencing the selection of the most likely methods for future use and development [8]. The goals of this framework include a thorough analysis of current detection techniques, the creation of benchmarks to evaluate their effectiveness, and the development of optimal strategies to spot deepfakes. The study evaluates many deepfake detection techniques on the basis of factors like precision, computational effectiveness, and capacity to adjust in changing strategies. Examples from the real world show each approach's weaknesses and usefulness. In response to the emergence of advanced counterfeit technologies, the study intends to improve digital security by providing an organised framework for evaluation, providing professionals and researchers with valuable insights while addressing the difficulties presented by deepfakes. The next few chapters will cover the methodology, data sources, and evaluation standards employed in this framework, offering a comprehensive manual for understanding and enhancing deepfake detection systems [9].

1.2 IMPORTANCE OF DEEPFAKE DETECTION

The imperative necessity of identifying fakes is underscored by the rapid advancement of deepfake technology, which generates significant issues in a variety of fields. Deepfakes affect accuracy, privacy, and security by utilising actual yet generated content and intelligent computers. Proper use of digital media is ensured by a high level of awareness of misleading information, which in turn mitigates the anticipated risks [10].

1) Protecting Privacy and Security

One of the most significant concerns about potential privacy violations is brought up by deepfakes. People may be persuaded to engage in dishonest or dubious behaviour, which might lead to extortion, physical violence, or even harm to their reputation. Since deepfake technology makes it possible to produce fraudulent films featuring well-known individuals, such as politicians or celebrities, it has a significant influence on national security. These tactics have the potential to spread misleading information, sway elections, or create fear. Understanding these fraudulent commodities is necessary to preserve both individual privacy and national security.

2) Preventing Misinformation and Manipulation

Deepfakes may feature quite realistic but fake material and can readily propagate incorrect information. Changing digital content affects how people see reality and can lead them to form incorrect beliefs. In voting, political communication, and news sharing, reliability and correctness are very important. Therefore, any mistakes or misleading information can have serious effects. Deepfakes in political campaigns can spread false information or change how politicians are viewed, which can affect voter opinions and possibly the election results. The quick spread of deepfake content on social media helps false information to grow, which leads to confusion and the acceptance of incorrect beliefs. Deepfakes in the news can make trustworthy sources less reliable, hurting the overall trustworthiness of journalism. A lack of trust can weaken democratic ideals, cause social unrest, and even lead to the breakdown of society. To minimise these bad effects, we need good ways to quickly identify and fix incorrect data before it is released. This can prevent great destruction.

3) Maintaining Trust in Digital Media

As deepfake technology continues to develop, it becomes increasingly challenging to differentiate between authentic and altered video. Since they are capable of accurately simulating real-life individuals through the use of powerful AI algorithms, deepfakes compel viewers to identify the produced material. This is due to the fact that they produce remarkably convincing videos and audio recordings. The increasing level of trust in digital media raises significant concerns about the quality of the information that consumers encounter, as it is challenging to verify its accuracy. Additionally, the erosion of trust in digital media may have adverse consequences for media outlets, content providers, or websites. This has an impact on media entities, as it necessitates the dissemination of false information in a dishonest manner, thereby terminating their legitimacy. It is challenging for online sites to regulate user safety in the context of resource management. Content providers are at risk of losing respect if their work is misused or misinterpreted. Good detection methods are instrumental in bolstering public confidence in media sources and defending the strength of online content, thereby mitigating certain risks.

4) Enhancing Legal and Regulatory Frameworks

The development of actual technology emphasises how urgently modern legal and regulatory systems are needed to manage the issues linked with digital fraud. Usually, the complex and sophisticated features of deepfakes which might offer quite realistic but deceptive representations of people and events have escaped accepted guidelines and standards. This disparity emphasises the need for new legislation designed expressly to control the many hazards related to deepfake technology. Good awareness of deepfakes is therefore crucial, as it provides the required proof as well as information for the legal system to keep looking at unlawful usage and maintain new legislation. Techniques of detection enable one to identify and remove harmful components. They are also fairly crucial for arming legislators with understanding about the extent of the issue and the efficacy of suggested remedies. Good detection methods enable legal and administrative systems to be better ready to manage the always growing complexity of digital fraud.

5) Supporting Technological and Research Advancements

Some amazing advancements in AI and ML are resulting from the search for improved deepfake detection methods. Scientists are continuously developing advanced algorithms aimed to identify and isolating deepfakes from real-time footage. Apart from solving the immediate problem of spotting digital fraud, this project covers machine guidance and AI capability. More complex detection methods established by scientists enable researchers to better understand the possibilities and constraints of ML systems, hence improving understanding of the expected applications of these technologies. Apart from fictitious information detection, the outputs of this work help to build other AI systems, including cybersecurity, photo as well as audio recognition, and processing of natural languages. These initiatives encourage major technical research and execution by means of enhanced use of AI or learning algorithms, therefore promoting technological developments in many various sectors and businesses [11].

The first recorded attempt at attempting a face swap was in one of the famous pictures of Abraham Lincoln, the United States President. In the lithograph, Southern statesman John Calhoun is depicted alongside Lincoln's head. The demand for lithographies of Abraham

Lincoln was so high after his assassination that engravings of his head on other bodies appeared seemingly out of thin air [12]. DNN-generated fake videos first caught the public's attention in late 2017, when a Reddit account named, Deepfakes began posting synthetic pornographic videos generated using a DNN-based face-swapping algorithm. Since then, it has been a never-ending story. There are also major downsides in addition to all the attention. Famous people, athletes, and politicians are particularly vulnerable to deepfakes because of the abundance of visual content they have online. Without the knowledge or permission of the subjects, deepfake technology can use their likenesses and voices to produce political, pornographic, or satirical content. Because so many applications are simple to use, anyone may create any fake content that is undetectable from the real thing. In 2022, President of Ukraine Volodymyr Zelensky declared that Ukraine had surrendered. Most Ukrainian people knew it was not true. It was a deepfake video [13], [14]. Noelle Martin, age 27, was an adolescent at the time when the revelation that her image had been utilised in deepfake porn without her permission "shattered" her world [15].

1.3 CHALLENGES IN DEEPFAKE DETECTION

The rapid advancement of deepfake technology has resulted in a variety of technical challenges that affect the production and identification of artificial media. Given their increasing complexity and potential application in a variety of sectors, such as cybersecurity, misinformation, and privacy violations, it is imperative to address deepfakes [16].

1. **Data Variability**

Deepfake datasets vary in quality, resolution, and manipulation kinds, complicating the development of models that excel across diverse datasets. These variances occur due to several circumstances. Datasets are sourced from many origins, such as social media, surveillance recordings, and regulated laboratory settings, resulting in discrepancies in lighting, resolution, and compression artefacts. Diverse datasets encompass deepfakes produced by various techniques, including GAN-based face swapping, lip-syncing, and full-body synthesis, leading to an inconsistency in the sorts of manipulations exhibited. Numerous datasets exhibit biases, such as the under-representation of specific races, age demographics, and genders, which adversely affects the capacity of detection models to generalise across varied populations. The ratio of genuine to altered samples in a dataset fluctuates, influencing a model's ability to distinguish between legitimate and deepfake material in practical applications. Variations in video compression, resizing, and noise introduction across datasets complicate the development of models that are robust against these aberrations.

2. Increasing Realism and Complexity

In general, deepfakes have a significant impact on social dynamics and individuals by facilitating the distribution of inaccurate data and increasing the severity of existing divisions. The segregation of society can be furthered when actual technology is employed to alter public opinion. For example, deepfakes that are politically motivated may be created to create narratives that are intended to incite conflict between various political or social groups. This manipulation not only highlights preexisting inequities but also fosters widespread scepticism regarding reliable information sources. The psychological repercussions for all individuals who encounter deepfake content may be more concerning. Regardless of their form specific assaults or generalised misleading information false or destructive deepfakes can cause significant emotional distress. Individuals who endeavour to distinguish between what is authentic and what is artificial may experience feelings of anxiety, confusion, and diminished personal security. This level of ambiguity has the potential to erode trust in digital media as institutions, thereby jeopardising the cohesiveness of society and the well-being of individuals.

3. Detection and Verification

In order to identify false news in media channels, one must look for subtle differences or things that have synthetic characteristics. These irregularities, such as pixel-level alterations, synthetic motions, or auditory disparities, are occasionally the focus of traditional deepfake detection approaches. But as deepfake technology evolves, these synthetic components have become more intricate, making these versions harder to find. Modern deepfake production techniques rely on data generation techniques that more closely mimic real-world sights and sounds, thereby masking conventional artefacts. When it comes to identifying deepfakes, conventional forensic investigation is rendered ineffective. In an effort to overcome this obstacle, researchers are exploring novel solutions, such as detecting systems based on AI and ML methodologies. Because complicated cycles and signals are difficult for the human sight to discern, these cutting-edge methods investigate minute audio distortions, stimulus responses, and changes in lighting and shadows. To ensure effective categorisation with little ever convincing synthetic media, these detection systems must be robust and flexible enough to react to the rapid evolution of deepfake technology [17].

4. **Computational Resources and Efficiency**

Deepfakes utilise substantial computational power, resulting in significant hardware and data consumption issues. Creating high-quality deepfakes needs substantial training datasets and reliable technology such as TPUs or powerful GPUs. DL techniques for the creation of realistic synthetic media need substantial computer power and storage, rendering them expensive and resource-intensive. The intricacy of these models and the requirement for extensive data increase the difficulties in the advancement of deep copy. Effective deepfakes, from a detection perspective, need substantial processing power. To identify subtle changes and abnormalities, detection systems must rapidly and accurately assess extensive amounts of media data. Robust processing capability is essential for prompt analysis and the maintenance of high detection accuracy. The regulation of the necessity for precise identification in accordance with the computer's efficiency is somewhat difficult. Researchers and programmers should primarily concentrate on optimising algorithms and utilizing distributed computing resources. On-demand cloud computing and the development of parallel processors are two advancements in the ongoing battle against deepfake use, designed to enhance the cost-effectiveness and efficiency of detection systems.

5. Ethical and Legal Implications

Although technical problems to control basic privacy, consent, and responsible use of technological innovations demand simplicity, the ethical and legal repercussions of deepfake technology offer great complexity and worry. Deepfakes applied with ill intent that is, producing false or damaging content raise substantial ethical concerns. Deepfakes, for instance, can create unlawful and harmful photographs of individuals, including defamatory videos or non-consented sexual material, therefore breaching their privacy and consent. These acts not only affect people but also create questions regarding the more general consequences of technology able to alter and distort reality. Legal systems try to catch up with the exponential development of deepfake technology. Conventional rules and laws could not be sufficient to handle the unique challenges presented by deepfakes, including the legal consequences of synthetic media and the necessity of new ideas of digital authenticity. These actions influence not just individuals but also raise issues about the more general effects of technology able to change and distort reality. Deepfakes provide unique challenges for which laws and common sense would not be able to address, like the legal ramifications of synthetic media and the need for new standards for digital authenticity. This discrepancy affects initiatives intended to monitor and control the dissemination of deepfakes, therefore creating a legal hole fit for unethical use. Developing tools that not only spot deepfakes but also probe moral and legal concerns will help to control more broad consequences. This assures that approaches to identification fulfil ethical and legal criteria as well as those of guiding principles for logical use of technology. Including these elements in the development of deepfake technologies will help us to better navigate the complex interplay between innovation and ethical responsibility, therefore lowering the risks connected to the use of synthetic media [18].

6. Adaptation and Evasion Techniques

The different ways to identify speed in deepfakes, based on the tricks that bad players use, creates a constant competition between those making deepfakes and those trying to find them. Now, let's talk about bold attacks, which involve slightly altered fake content. More deepfake makers are using clever tricks to avoid detection systems. These attacks make small, noticeable changes to deepfake video data without affecting its look or sound, which can greatly impact how well detection systems work. The continued creation of various escape methods highlights the importance of keeping up with new ways to identify them. Research will stay ahead by creating stronger and more adaptable detection methods that can identify new ways of avoiding detection. This means using new ML methods to improve tracking systems so they can keep up with the techniques used by deepfake makers and become better at spotting tiny differences. The relationships between the researchers shape how ideas and methods develop for new exit routes. The ongoing struggle to identify and avoid deepfakes highlights the importance of having effective and adaptable detection methods. Deepfake makers need to be clever and careful, as they are developing new ways to avoid being found out. At the same time, tools to identify these deepfakes must improve to lessen the risks from advanced fake media [19].

1.4 RESEARCH GAP AND MOTIVATION

Although multiple deepfake detection models have been developed, there is a lack of rigors comparative research that assesses the performance of these models on various datasets. This gap hinders our comprehension of the model's generalizability to apply to many situations and their resilience. The objective of this study is to address this gap by undertaking a comprehensive evaluation of several detection algorithms on numerous datasets. This will offer a wider viewpoint on their efficacy and highlight potential areas for enhancement.

The research can determine the most robust and versatile models by comparing them to a variety of datasets and deepfake techniques. This aids in the creation of models that are capable of addressing a diverse array of real-world scenarios. The comparative study can be advantageous to practitioners and researchers, as it helps them determine which models are most effective under particular circumstances, thereby facilitating the selection of the most suitable model for their requirements.

The primary objective of this research is to methodically assess the performance of various deepfake detection models on diverse, datasets to determine the most efficient techniques for differentiating between authentic and manipulated media. The study seeks to improve the resilience and applicability of these models. This research aims to address the current lack of comparison assessments by conducting a thorough evaluation of the selected models using multiple datasets. The findings will provide vital insights into the strengths and shortcomings of each model, which will guide future research and the creation of more reliable and widely applicable deepfake detection systems that can successfully counter the ever-evolving danger presented by deepfake technology.

1.5 RESEARCH OBJECTIVE

The several deepfake detection models are compared in this study. Forensics Convolutional Neural Networks (FCNNs) are the initial model introduced; they use the gaussian blur and gaussian noise picture preprocessing stages. Next up is a Two-Stream Neural Networks (TSNN) model, which is a hybrid approach to identifying fraudulent pictures. Using paired learning, the Common Fake Feature Network (CFFN) is the third model. The Vision Transformer, the fourth model, is a hybrid model once again. They are more effective even on photos produced by new techniques because of their excellent generalisation skills. Finally, we compare five models, the last of which is the Masked AutoEncoders (MAE). In order to restore the original signal from its partial observation, this straightforward autoencoding method is used.

A lot of people compare the progress made in deepfake technology and find it to be an arms race, where each side is always trying to get better than the other. The deepfake algorithms are getting smarter, so the fakes they make look more real and are harder to spot. To counter this, deepfake detection technologies are also getting better, using more complicated algorithms and a wider range of methods to spot fakes.

To advance future investigations, it is necessary to prioritise the development of more robust and scalable models that can be effectively applied to vast and high-quality datasets. Furthermore, it is widely recognised that DL models frequently necessitate extensive datasets during the training phase to yield satisfactory outcomes. Unfortunately, these datasets are either not freely available or require authorisation from social media providers. The rapid advancement of deepfake GAN models poses a new difficulty, as current DL models may fail to detect hitherto undiscovered types of generated images and videos.

1.6 RESEARCH CONTRIBUTIONS

This thesis explores the construction and evaluation of several ML methods across numerous datasets, including Celeb V2, FaceForensics++ (FF++), CIPLib, Deepfake Detection Challenge Dataset (DFDC), and Fake Celeb AV-V1.2, so contributing to the field of deepfake detection. Dealing with the growing threat presented by deepfake material, the work uses a complete methodology to evaluate models including CFFN, TSNN, MAEs, FCNN, and Vision Transformers (ViTs). Key indicators like training and validation accuracy, loss measures, and error rates, including Equal Error Rate (EER) and False Acceptance Rate (FAR), are evaluated using a consistent approach. Although ViTs and TSNN show good performance, the results reveal that the CFNN model routinely beats others with almost zero error rates and 100% validation accuracy on demanding datasets, including Fake Celeb AV-V1.2 and DFDC. The research establishes CFNN's robustness, especially its capacity to generalise to unseen inputs, therefore proving it as a potential media authentication and security solution. This work stresses the need for balancing accuracy and error reduction in deepfake detection and motivates continuous development to improve ML models' adaptability and efficacy in tackling deepfake concerns.

1.7 STRUCTURE OF THESIS

1) **Introduction**: The first chapter gives a general outline of how deepfake technology is beginning to affect the credibility of media and public faith in them. It traces the development of deepfake production methods through time and emphasises the threats to security, privacy, and ethics that these technologies provide. This chapter lays out the goals and scope of the study, which aims to evaluate ML models for deepfake detection, and stresses the need for creating reliable detection procedures.
2) **Literature Review:** This chapter summarises previous work on deepfake detection and delves into the cutting-edge ML and DL methods used in those investigations. It identifies research gaps that need more accurate and generalizable models and critically assesses the strengths and limits of different methodologies. The study's recommended ML techniques are chosen and justified based on this review.

3) **Research Methodology**: Dataset selection and preparation methods are described in depth in the methodology chapter, which also contains the experimental framework. Celeb V2, FF++, CIPLib, DFDC, and Fake Celeb AV-V1.2 are some of the datasets mentioned, along with the methods used to clean, augment, and analyse the data. Hyperparameter optimisation and assessment criteria such as EER, HTER, FAR, and False Rejection Rate (FRR) are further detailed in the chapter, along with the implementation of ML models such as FCNN, ViTs, MAEs, TSNN, and CFNN.
4) **Results and Discussion**: Using measures like training accuracy, validation accuracy, loss rates, and error rates, this chapter summarises the results of the performance evaluation of the models that were put into action. After comparing the two models, it becomes clear that the CFNN model outperforms the others when it comes to successfully validating models on difficult datasets with almost no errors. Discussed are the advantages and disadvantages of each model, with an eye towards highlighting the significance of precision and error mitigation in practical contexts.
5) **Conclusion and Future Scope**: The final chapter highlights the main points, confirming that the CFNN model works well for detecting deepfakes. It talks about the limits of current research, like possible biases and computing challenges. It also offers ways to improve how adaptable and efficient models can be in the future. The chapter ends by highlighting how this study can help improve media security and emphasises the need for ongoing progress in ML to fight against deepfakes.

CHAPTER 2

2. LITERATURE REVIEW

Counteracting deepfake technology can be categorized into three distinct groups: deepfake detection, content authentication, and prevention of the dissemination of materials that can be utilized for deepfake creation. The advancement of technology in detecting and verifying deepfakes is rapidly progressing. However, the ability to create deepfakes is advancing at a far quicker pace than the capability to identify them [20]. Twitter on Counteracting Deepfake Technology "Elon Musk's X cracks down on deepfakes with improved image-matching [21]."

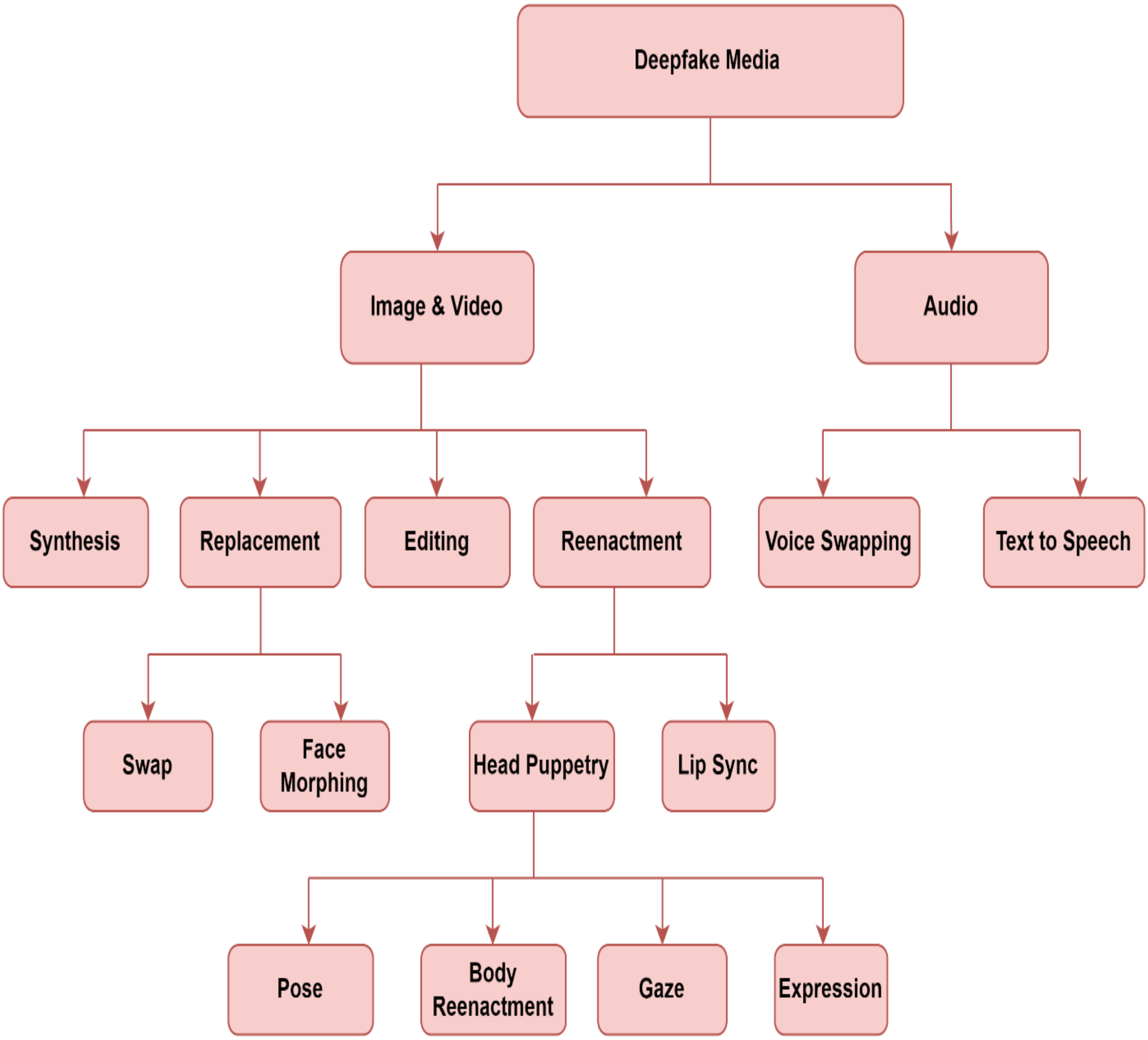

FIGURE 2: DEEPFAKE MEDIA AND DETECTION TECHNIQUES

The two convolutional neural network (CNN) models proposed by Afchar et al., Meso-4 and MesoInception-4, were among the first to focus on deepfake media identification [22]. The suggested CNNs used sparse layers to improve the mesoscopic picture features. Using a publicly available deepfake detection benchmark, the authors evaluated their suggested models. Furthermore, the scientists not only achieved outstanding results on both datasets used, but they also collected a bespoke dataset and evaluated the models on it [23].

2.1 HISTORICAL DEVELOPMENT OF DEEPFAKE TECHNOLOGY

The growth of deepfake technology shows how much AI and ML have advanced lately. The term "deepfake" comes from combining "DL," which stands for deep learning, and "fake." It refers to media created using DL methods that look very realistic but aren't real. Knowing how technology has developed over time helps us recognise its benefits and difficulties.

2.1.1 EARLY BEGINNINGS AND FOUNDATIONS - 1990s-2000s

Deepfake technology began in the 1990s when the necessary tools for its growth became available. Early digital editing tools inspired basic ways to change photos and videos that have developed in this century. These tools let users edit pictures and videos easily, allowing for easy blending, colour adjustments, and cutting. These early tools were mostly operated by hand and limited by how well changes were made, often leading to noticeable mistakes and differences. During this time, advancements in ML and neural networks led to a real shift in how things work.

A concept for using training algorithms to detect patterns and make predictions originated in the late 1980s and early 1990s, spurred on by the development of neural networks. These rudimentary methods, which look dated when compared to today's standards, lay the framework for more advanced digital content manipulation, which ultimately led to deepfake technology. Applying neural networks to increasingly complex tasks was a major step towards making synthetic media that looked and felt real.

2.1.2 THE EMERGENCE OF GENERATIVE MODELS - 2010s

Driven largely by improvements for computational models that significantly raised the production of synthetic media platforms, deepfake technology had an incredible evolution in the decade of the 2010s. Published in 2014, Ian Goodfellow and colleagues conducted a fundamental investigation on computational acoustic networks, or GANs, therefore indicating a clear progress in the field of AI. Two competing neural networks, the generator and the discriminator, make up GANs. The generator generates false data; the discriminator evaluates it in respect to real data. This adversarial process drives both networks to improve their performance, hence producing more genuine and better-generated content. Other DL techniques enabled the industry progress even beyond GANs produced by Conventional CNNs or Vibrating Auto Encoders (VAEs). While VAEs enabled more sophisticated data creation by learning about demanding data distributions, CNNs enhanced the capacity to evaluate and produce high-quality images and videos. These developments combined enhanced realism and accessibility of deepfakes, therefore improving the limitations of digital content alteration.

2.1.3 THE RISE OF DEEPFAKE TECHNOLOGY - 2017-PRESENT

Deepfake became popular in 2017 when a Reddit user made very realistic fake videos of famous people using a technology called GANs, which stands for GANs. The use and difficulty of deepfake technology have evolved over time. This started because AI studies are advancing quickly, and there are more powerful computers available. Deepfakes first appeared to create detailed changes, such as blending sounds or altering full- body movements in videos. They were mainly used to swap faces in films. One important event was the launch of new technology that makes it easy to create deepfakes. Deepfakes are simple to create, but as their quality improves, this brings up important questions about how they can be used. More people can now create and share deepfakes fake content that looks real because there's more open-source software and user-friendly sites available. This highlights the pressing need for effective ways to identify and reduce the risks associated with deepfakes.

2.1.4 RECENT ADVANCES AND CURRENT TRENDS

Improved GANs architecture and innovative ideas like transformer models have helped the area of fake creation to progress dramatically lately. Originally observed for their capacity to locate long-reaching linkages and sequential data control, designs for transformers have developed ever more complex and highly coherent material. Once lowering realism, better GAN designs reduce audio and visual distortions, thereby improving the quality of deepfake video. Apart from typical media dishonesty, scientists have investigated using deepfakes in several fields, including VR and customised content generation, therefore broadening the range of probable opportunities for this technique. Deepfake technology is derived from methods for identifying and limiting its erroneous application. Deepfakes can be identified by researchers through the use of strong AI and ML technologies, which allow for the close examination of minute deviations and abnormalities that are not detectable by more conventional methods. The critical necessity of continuous innovation and meticulous attention to detail is underscored by the intense competition among false detector manufacturers in order to effectively address the complex issues that this technology is constantly developing [24].

2.2 RELATED STUDY AND PUBLICATIONS

[25]Deepfake technology has emerged, enabling the production of realistic fake images and videos, driven by GANs and DL methods. As a result of the widespread propagation of deepfake content on social media platforms, the global community has been in search of effective methods to identify and combat this issue. AI has significantly propagated deepfake problems, as both professionals and academics acknowledge. However, they are also of the opinion that AI can offer practical solutions to these issues. This study examines the utilisation of AI to identify deepfake images. In order to achieve this objective, we concentrate on three CNN methodologies: ResNet, VGG16, and VGG19. These CNN models were assessed using a dataset that contained 1,200 images, which included both genuine and artificial images. The deepfake photos were generated by FaceApp, a renowned programme for the creation of altered photographs. Our analysis demonstrates that VGG19 outperforms both VGG16 and ResNet50 with an astonishing accuracy of nearly 98% on the test dataset when the dataset size is modest. This work is a significant contribution to the field of deepfake photo detection, which makes use of AI, primarily CNN techniques [26]. Contemporary multimedia forensics programmes mostly focus on the identification of deepfake videos. This thesis presents a method for detecting face-swapped videos, including identifiable individuals. To improve face recognition accuracy, we present a threshold classifier that uses similarity scores generated by a Deep Convolutional Neural Network (DCNN). We produce a series of scores indicating the similarity between faces extracted from searched videos or reference materials pertaining to the subject depicted. We categorise the questioned films as authentic or deceptive according to the established criteria based on their highest score.

[7] Deepfake is synthetic media that is generated by AI. To create the illusion that video, audio, or photos depict someone or something else, they must be either produced or altered. The development of deepfake technology is similar to that of the methods used to detect it. The inventors of deepfakes and those who develop detection techniques are perpetually engaged in a game of cat and mouse. We must confront the consequences of deepfakes on society as the

technology that underpins them continues to evolve. This can be resolved through a variety of methods, including educational programmes, legal systems, technological improvements, and ethical considerations. It is imperative to employ a combination of techniques to effectively identify deepfakes. Detecting deepfakes may be challenging due to their escalating complexity; however, a variety of algorithms and methodologies are being developed to identify them. The adverse consequences of deepfakes can be alleviated by integrating technological advancements, legislative initiatives, and awareness campaigns.

[27] Deepfake technology is now widely used in areas like movie making and entertainment. They have also been used intentionally to share false information and create fake videos. Current identification methods usually have poor performance or limitations when they encounter fake images or low-quality pictures. We offer Diff-ID, a new way to understand and measure how face changes can affect identity. Diff-ID uses a real picture of the same person to check if another image is genuine, analysing both photos at the same time. A face-swapping generator takes a test picture and a reference image and puts them in a similar feature area that doesn't focus on their identities. Using the matched pairs, we can find the differences in identity between the two pictures. We have made a special measurement to check how much the test picture's identity differs from the reference image. This figure easily distinguishes fake photos from real ones. Many studies have shown that our method works very well. Deepfake pictures show very good accuracy in spotting and can adapt well to new fake methods that haven't been seen before. It remains strong despite having visual illusions.

[28], [29]The creation and manipulation of deepfake videos have become more straightforward as a result of the availability of extensive benchmarks and recent advancements in generative models. The demand for effective deepfake detection techniques is currently high due to the rapid increase in the number of newly produced hyper-realistic deepfake movies that are being used for malicious purposes. The results of numerous current deepfake detection techniques, particularly those that are based on CNN, are encouraging; however, they have a number of deficiencies. Generally, subpar generalisation outcomes have been produced under unseen/new deepfake generating techniques. The aforementioned defect is primarily the result of CNN-based methods, which focus on the local spatial artefacts that are distinctive to each manipulation technique. Consequently, it is challenging to comprehend the extensive forgeries traces of a variety of manipulation techniques without considering the dependencies that extend beyond the local receptive field. This study introduces a framework that integrates CNN and ViT to enhance the generalizability and detection accuracy of the system, thereby resolving

this issue. Our method, HCiT, isolates significant local features by utilising the self-attention mechanism of the ViT and the benefits of CNNs to explicitly identify discriminative global contextual dependencies in a frame-level image.

[30]Benchmark datasets are essential for evaluating deepfake detection methods; however, it is unclear if the picture quality metrics used for validation accurately represent deepfakes in the wild. In order to assess how well these strategies work in practice, reliable benchmark datasets are required. The purpose of this study is to investigate how picture quality criteria relate to the overall quality of deepfake datasets and how well deepfake detection algorithms perform on them. We examined quality variance within and across three key datasets: FF++, Celeb-DF, and DeeperForensics-1.0 using six distinct picture quality metrics. Our findings show that there are substantial quality differences among and between various datasets, which impact video and frame rates. Xception, SPSL, and UCF, among others, tend to perform worse with datasets of higher quality. These results suggest that the existing methods of validation, which do not take quality variation into account, can provide inaccurate evaluations of detection accuracy. Understanding the relationship between dataset quality and detection accuracy is crucial for developing more effective detection methods; this highlights the need for using high-quality benchmark datasets for comprehensive validation.

[31] The inaccuracy of video, audio, and image representations of reality is becoming increasingly apparent as synthetic media and deepfakes techniques and tools continue to evolve. Currently, all digital communication channels are susceptible to manipulation, with the widespread use of deepfakes to disseminate false information and disinformation, exacerbate political conflicts, vilify opponents, engage in cyber fraud, or extort individuals. This research demonstrates that conventional ML techniques can outperform DL approaches in the detection of deepfakes, resulting in results that are comparable to or even superior to the most advanced DL methods. Conventional methods, including feature construction and selection, training, and testing of ML classifiers, can be employed to achieve improved comprehensibility and interpretability while minimising computational resource usage. Additionally, the Analysis of Variance (ANOVA) is implemented to evaluate the efficacy of numerous ML models. We conducted experiments that achieved a precision of 99.84% on the Faceforencis++ dataset, 99.38% on the DFDC dataset, 99.66% on the VDFD dataset, and 99.43% on the Celeb-DF dataset. Our research disproves the notion that DL methods are the only effective method for detecting deepfakes and demonstrates that the meticulous application of ML techniques can be both cost-effective and effective.

[32] In order to identify altered media, such as images and movies, one technique is to look for irregularities and discrepancies that are often brought about by the creation of fakes. The scientific literature presents a variety of approaches, each highlighting potential improvements through the use of a unique set of specialised characteristics. The purpose of this study is to look at how deepfake generation affects the scene's original attributes just before capture. Pixel values for still images or moving video capture a unique environment that is shaped by factors such as item placement in the scene, lighting, and the photographer's approach. However, it is possible to alter these inherent relationships through the development of deepfakes. In order to train a CNN to detect deepfakes, one might use surface property analysis of the displayed picture to create a descriptor. The name of this technique is Surfake. It is possible to use this characteristic to distinguish between changed and unaffected photos, according to experimental findings performed on the FF++ dataset utilising different kinds of deep-fake operations and different DL models. And the results demonstrate that when combined with other types of visual data, the accuracy of detecting deep-fake forgeries may be significantly enhanced.

[33] It is widely acknowledged that deepfakes exhibit a variety of artefacts. We suggest a hypothesis that the presence of visible or non-visible artefacts is correlated with manipulations, particularly when it is difficult to accurately replicate the subject's movements. Our approach assesses the quality of the images discovered in the film by employing both Full-Reference (FR) and No-Reference (NR) standards. The measurements assist us in the systematic construction of a matrix-based image of a video sequence by assessing each frame individually. We show that the results obtained by employing this fundamental collection of features for a neural network structure are indicative of the initial phase of promoting the improvement of this representation. The purpose of this improvement is to broaden our investigation to encompass additional deepfake categories. The FF++ dataset is employed for experimentation purposes, as it facilitates the evaluation of the proposed method across a variety of deepfake-generating techniques.

[34] This thesis meticulously examines three advanced algorithms: Gate Recurrent Unit (GRU), Long Short-Term Memory networks (LSTM), and CNN. The main objective is to determine the essential need for effective deepfake content identification. In addition to their striking resemblance to actual individuals, deepfakes pose other hazards, such as the dissemination of deceptive material, violations of personal privacy, and the erosion of societal trust. This study meticulously evaluates and compares the accuracy, resilience against novel deepfake techniques, and computational efficiency of CNN, LSTM, and GRU. This research

uses carefully selected datasets that precisely replicate real-world situations. The purpose is to reveal subtle differences between the unique strengths and shortcomings of each algorithm. The extensive implications of this thorough study transcend well beyond academics, possibly influencing progress in media forensics, regulatory structures, and the ongoing struggle against the detrimental spread of disinformation and digital manipulation.

[35] The delicate authenticity indicators seen in flawless films are beyond the capabilities of the generative models that are already available. In order to create a universal detector for deepfakes, we take a look at how deepfakes employ motion magnification compared to real footage. The residual output shows that different generative models perceive the facial motions occurring beneath the muscles differently. The difference between real motion and the augmented synthetic distortions is used by our technology. It does this by using DL-based and classic advanced motion magnification techniques to detect whether a video is genuine and, if not, to identify its original source. The video source detection accuracy we obtained when testing our technique on two datasets with multiple sources was 94.03% and 97.77%. By at least 4.08%, our method surpasses other complex structures and the prior deepfake source identification. Furthermore, we assess the magnification level, phase extraction window size, backbone network topology, sample quantity, and sample length. Finally, to assess the model's possible biases, we provide our results on the impact of gender and skin tone.

[36] Deepfake technology can seriously harm the trustworthiness of visual content, particularly during live-stream events where quick recognition is very important. Current methods for spotting deepfakes have many limitations and difficulties, showing that we need better and more reliable solutions. This work suggests a new method to easily address the need for real-time tracking of deepfake by mixing eye motion analysis using a hybrid DL model. The new mixed DL model combines two designs, MesoNet4 and ResNet101, to improve Deepfake detection by using the unique strengths of each. Designed especially to find minute changes in face pictures, MesoNet4 is a small CNN model. ResNet101 effectively identifies important features while also handling detailed visual information. Our combination model performs better at telling the difference between fake and real films. It does this by using MesoNet4, which focuses on specific features, together with ResNet101, which provides a more detailed view of the features. Regular methods or looking at things visually can't show this change. The model is trained on various datasets like FF++, CelebV1, and CelebV2, and it performs very well in terms of accuracy. The hybrid model is effective and promising for checking content

accuracy and video analysis. It achieves an accuracy of 98.73% on FF++, 96.89% on CelebV1, and 97.90% on CelebV2.

[37] The utilisation of beautifying filters has become increasingly prevalent, with the introduction of new filters on a daily basis. Is it possible to deceive State-of-the-Art (SotA) detectors solely by utilising a cosmetic filter on the modified video? This research investigates the impact of beautification filters on Celeb-DF-B, a recently developed database that applies a selection of authentic and counterfeit films from the Celeb-DF dataset to the widely used social media beautification filters available. We evaluated the influence of beautification on human evaluators and three cutting-edge passive deepfake detection models. The results indicate that the behaviour of the three detectors under investigation is significantly influenced by filters, resulting in a significant decrease in the video-level AUC on the enhanced subset of Celeb-DF-B. When it comes to attaining human-level performance, filters have a similar impact on human decision-making. They influence the precise classification of videos as either genuine or counterfeit.

[16] More efficiency is required for real-time detection, even if current approaches place a heavy focus on big and complicated models. Our approach to detecting deepfakes in photos is unique in that it is based on Binary Neural Networks (BNNs), which has not been done before. Fast inference is possible with less loss of accuracy with this method. In addition, our approach includes Fast Fourier Transform (FFT) and Local Binary Pattern (LBP) as additional channel properties to detect tampering in the texture and frequency domains. The COCO Fake, DFFD, and CIFAKE datasets were used to evaluate our methodology's performance. In most cases, it was shown to outperform earlier strategies. In addition, our method achieved a notable improvement in inference efficiency, with a decrease in FLOPs of up to 20 times. This exploration delves into the use of BNN in deepfake detection with the goal of finding a middle ground between efficiency and accuracy. The findings from this study lay the groundwork for future research on efficient deepfake detection.

[17] Authentic-looking fabricated content comprising photographs, videos, and audio—facilitates fast dissemination through social media and AI. The initial emphasis of false content generation was predominantly on the management of either audio or video inputs. However, more intricate alterations commonly known as "deepfakes" are becoming possible by contemporary breakthroughs in DL. Current research predominantly emphasises the identification of fraudulent videos using visual or auditory modalities, with limited

methodologies addressing audio-visual deepfake detection. Nonetheless, these strategies may necessitate enhanced precision when evaluated on a multimodal dataset, including deepfake videos and alterations in both streams. Due to the neglect of preprocessing face features and the application of traditional training methods. To address this challenge, we offer a robust audio-visual deepfake detection (MAVDD) technique that analyses audio and visual streams to enhance detection efficacy. Utilising sophisticated preprocessing techniques for optimal extraction of facial and auditory features with pretrained models in image classification tasks facilitates the detection of visual deepfakes.

[38] The potential for deepfakes, or computer-generated visuals, to ensnare individuals and spread false information is a major concern. Experts in the field of AI are scouring the internet for signs of deepfakes in the hopes of finding a solution. A thorough outline of the techniques used to find deepfake films is provided in this thesis. The study explores the use of ML to detect false information by combining several media types, including photos, videos, and audio. To test various deepfake detection methods, it also examines the large datasets to which academics have access. Specifically, we looked at conference and journal articles about the rise of deepfake videos that were released between 2015 and 2023. Various methods for creating deceptive images, videos, and even changed audio recordings were covered in these books. The researchers proved that a combination of methods, such as using several ML algorithms or adding photos and videos, may get excellent results in deepfake detection. Future research should follow the thesis's suggestions to improve deepfake technology detection and, by extension, internet security. Additionally, a new dataset named Celeb-DF is available, which includes a plethora of brilliant parodies starring famous actors and actresses. Researchers can improve their ways of recognising DeepFakes with the help of this dataset.

[8] Major concerns regarding the impact on society have arisen in response to the fast advancements in the creation and manipulation of synthetic pictures. Deepfake film detection efforts are ongoing, but current methods aren't always up to standard. Recognising these deceptive deepfakes is the initial step in preventing their distribution on social media. In order to identify fake videos, we present a new deep-learning technique. At the moment, the majority of deepfake detection algorithms are focused on identifying puppeteering, emotion manipulation, lip syncing, and face swapping. Still, there are a lot of obstacles to overcome before real-time forensics can develop a consistent basis for identifying and evaluating different kinds of false movies and photographs. Our proposed method makes use of a hybrid approach that takes as input films of consecutive targeted frames. In order to train and classify

these frames, ResNet-Swish-BiLSTM, an optimised convolutional BiLSTM residual network, is used. The proposed method is useful for detecting artificial artefacts in deepfake images. The open DFDC and the FF++ were utilised to evaluate the long-term viability of our proposed method. An astounding 96.23% accuracy was achieved with the FF++ digital record. Our combined FF++ and DFDC statistics yielded a 78.33% accuracy rate. Our suggested solution outperforms existing methodologies in terms of major outcomes, as confirmed by extensive experiments.

[18] Data is fundamental in methods of video detection. This survey effort aims to explore holistically the development and detection of deepfake videos. The main contribution of the research is the classification of the many challenges in the detection of deepfake movies. This work covers data challenges, including imbalanced datasets and inadequate labelled training data. Using multiple computer tools presents one of the challenges of the course. It also addresses reliability, particularly overconfidence in fresh approaches to manipulation and detection processes. The research emphasises the supremacy of DL-based techniques in spotting deepfakes despite their computational efficiency and generalisation restrictions. These techniques do, however, also have several flaws, including limited computer efficiency and weak generalising capability.

[19] Deepfakes find use in the manipulation of medical pictures to either add or delete cancers. The incapacity of automated systems to identify medical deepfakes might lead to significant security and privacy issues, thereby maybe taxing hospital resources or even resulting in human mortality. We want a reliable deepfakes detector adept at managing the most recent methods of manipulation to offset these effects. Presenting the MedNet model, a DL method, this paper seeks to answer the problem. With this approach, we want to locate deepfake samples in lung CT images. Our concept asks for a custom EfficientNetV2-B4 architecture with additional dense layers at the end of the network. We have suggested increasing the computation capacity of the given approach by means of a spatial-channel attention mechanism. This method aims to highlight the altered portions of the samples, thereby enhancing the classification accuracy. Extensive research using a common dataset known as the CT-GANs dataset reveals the success of the current method. With an accuracy score of 85.49%, the provided work obviously can distinguish actual lung CT Scan data from deepfake pictures.

[39]Research on this subject has garnered significant attention due to the potential threat that digital visage modification in images and videos poses to public confidence. In order to

mitigate the detrimental effects of these technologies, sophisticated DL algorithms have been implemented to identify deepfakes and have demonstrated exceptional efficacy. However, the evaluation of these detectors is typically conducted using standards that are not particularly relevant to real-world scenarios. In particular, the accuracy of detection is not entirely understood in relation to the impact of various image and video processing techniques and frequent workflow errors. This work introduces a robust evaluation methodology for the assessment of deepfake detectors that are based on learning in more realistic environments. This is the first methodical evaluation of deepfake detectors that, to the best of our knowledge, not only provides a comprehensive overview of their performance in real-world environments but also statistically evaluates their ability to withstand various processing techniques.

[40] Promoted strong efforts to develop effective technology for detecting deepfakes. Current identification methods look good, but they struggle when faced with data created by different fake techniques that the model wasn't taught on. In the modern world, this problem of finding new forms of deepfakes catastrophic forgetting without losing its past knowledge of deepfakes is of great importance. To address the important issue of detecting deepfakes, we introduce a new approach for deep domain adaptation in this work. Our framework can create a deep neural network that performs well on both the source and target areas by using a lot of labelled data (fake or real) from one faking method (source domain) or a small amount of labelled data from a different faking method (target domain). Deep networks need a lot of named training data, which can be hard to find for deepfake detection. Our method can also work well with unlabelled data, which is often easier to get. We created a new loss function and used stochastic gradient descent (SGD) to train the deep network, aiming to maximise the loss.

2.3 GEN AI BREAKTHROUGHS IN IMAGE SYNTHESIS IN 2023 AND 2024

2.3.1 GENERATIVE ADVERSARIAL NETWORKS (GANs)

GANs continue to be essential in the process of image synthesis. They have demonstrated exceptional capabilities in the production of excellent-quality synthetic images that resemble real-world data. Advancements in the creation of sophisticated scenes, visage images, and realistic textures are included in this. GANs are demonstrating their utility in the synthesis of gene expression data, where they can surpass the constraints of limited, diverse datasets by generating high-fidelity synthetic data.

2.3.2 DIFFUSION MODELS

These models have acquired popularity due to their capacity to produce high-quality, detailed images. Diffusion models operate by progressively converting a straightforward noise distribution into a complex data distribution, which leads to the production of high-quality images. This approach has been particularly effective in the production of images from textual descriptions, thereby expanding the limits of creativity in AI-generated art.

2.3.3 HYPER-PERSONALISATION

GenAI-driven hyper-personalisation employs sophisticated algorithms to generate personalised content that is tailored to the preferences and behaviours of users. This trend is particularly prevalent in the retail and marketing sectors, where personalised product suggestions and content automation are growing increasingly sophisticated and effective. Similar methods are employed in the healthcare sector to customise treatment plans based on the unique data of each patient.

2.3.4 CONVERSATIONAL AI AND MULTIMODAL MODELS

The integration of conversational AI with image synthesis has resulted in the creation of models that are capable of comprehending and producing both text and images. This has the potential to improve the contextual relevance and visual appeal of responses provided by virtual assistants and chatbots. For example, DALL-E 2 is capable of producing images that are utilised in a variety of creative and practical applications by utilising textual stimuli.

2.3.5 ADVANCES IN TRAINING TECHNIQUES

Computational resources necessary for training large generative models have been substantially diminished as a result of the advancement of more effective and extensible training methods. The utilisation of enormous datasets and advancements in self-supervised learning have allowed models such as GPT-3 and subsequent versions to achieve unparalleled levels of performance in text as well as image generation.

2.4 IMPLICATIONS FOR PRIVACY AND SECURITY

Even while AI is advancing at a rapid pace, deepfake technology poses a significant threat to privacy and security. Concerns about personal safety, data security, and social trust are just a few areas where deepfake technology has the potential to provide information that is both realistic and created. From this vantage point, we can better respond to and protect individuals and businesses from the expected outcomes of deepfakes [30].

2.4.1 THREATS TO PERSONAL PRIVACY

Severe individual concerns are raised by the fundamental threat that deepfake technology poses to human privacy. Deepfakes are the creation of images or videos of unknown individuals that are unexpectedly accurate and unlawfully obtained. It can be challenging to distinguish them from real media content, as these segments appear to be identical to the original. Numerous detrimental consequences may result from the improper utilisation of technology. Dignity and privacy are most certainly violated by deepfake pornography, which conceals identities behind sexual images. Presenting individuals in compromising and dishonest situations in egregious deepfake videos can damage their personal relationships and reputation. Mistreatment of this nature has a considerable impact on more than just a trivial loss of privacy. Victims may also experience emotional distress as a result of the content's potential to induce feelings of humiliation, isolation, and horror. Additionally, reputational harm can have a significant impact on both personal and professional relationships. Consequently, the financial outcome of legal disputes may be essential in the process of rehabilitating one's reputation. Time and effort are required to mitigate the harm caused by deepfakes; victims can reestablish their reputation and privacy by navigating challenging legal and digital environments. This underscores the imperative necessity of support systems, stringent safety regulations, and legislative structures to mitigate and regulate the adverse effects of deepfake use [31].

2.4.2 RISKS TO CYBERSECURITY

Deepfakes are a significant and ongoing threat to cybersecurity, as they are primarily employed in intricate social engineering campaigns. Malicious actors can create incredibly convincing imitation films or audio recordings to deceive individuals into disclosing private information or engaging in security-compromising behaviour. Deepfake technology has the potential to generate highly realistic films featuring CEOs or other influential figures requesting private information or identity data, such as bank details. By persuading employees that these demands are legitimate, deepfakes significantly undermine business security. Deepfakes enhance phishing attempts by creating realistic, aesthetically appealing replicas of trustworthy individuals. In contrast to relying on conventional email hoaxes, assailants could enhance their phishing attempts by employing phoney images or voice recordings to impersonate reliable contacts. This approach would limit their chances of being discovered and enhance the persuasiveness of their schemes. Conventional cybersecurity solutions are restricted by the increasing authenticity of social engineering techniques, which may be reliant on well-known dishonest behaviours or obvious anomalies. It is crucial to develop and implement more robust detection methods into cybersecurity regulations, as deepfakes generated from real-world data are becoming increasingly intricate and indistinguishable. The detection of a few deviations and abnormalities by AI and ML reveals deepfake manipulations, thereby providing significant security against constantly changing and misleading attacks.

2.4.3 IMPACT ON TRUST IN DIGITAL PLATFORMS

The widespread use of deepfake technology significantly undermines public trust in electronic gadgets, resulting in both profound and adverse effects. Deepfake materials are growing; thus, platforms like news websites and social media may inadvertently facilitate the dissemination of incorrect information and misleading content. Deepfakes generate or modify content, hence jeopardising the integrity and trustworthiness of digital media. The erosion of trust compels consumers to feel confused and sceptical, hindering their ability to distinguish between factual information and deceptive media. This erosion of confidence affects not only specific domains but also the societal base as a whole. As consumers become increasingly discerning regarding the information they encounter, their overall trust in internet sources declines. A deficiency of faith can result in a division of public opinion, the dissemination of misinformation, and increased participation at reliable drop-off locations. The trustworthiness of the internet as a source of information is significantly undermined, impacting the everyday consumption and engagement with digital content.

2.4.4 CHALLENGES FOR LAW ENFORCEMENT AND LEGAL FRAMEWORKS

The production and dissemination of deepfakes often take place in grey areas of law, which poses serious difficulties for law enforcement and contemporary legal systems. Traditional legal institutions are not particularly well-equipped to regulate the complexity of digital manipulation, which is why they are not very successful in adequately restricting the complexity of deepfake technology. Deepfakes encompass a broad range of challenging topics, such as intellectual property theft, which involves the theft of properly secured important content; libel, which involves the dissemination of misleading information that can harm one's reputation; and privacy violations, which involve the association of individuals with dishonest and criminal activities. The legal protections for victims may be compromised, or the circumstances for prosecutors may be exacerbated as a result of the complex technical and ethical issues raised by deepfakes not being adequately addressed by current laws. The critical necessity of modern laws, particularly those that are tailored to the various aspects of offences associated with deepfake technologies, is underscored by the scarcity. Well-defined laws and practices are essential for the effective prosecution of malevolent individuals and victim assistance. In order to address the increasing issues associated with deepfake technology and protect individuals and organisations from its application, it is necessary to establish and implement legal frameworks.

2.4.5 SOCIETAL AND PSYCHOLOGICAL EFFECTS

Because they encourage dishonesty and worsen divisions that are already present, deepfakes have a direct impact on social dynamics and human psychology. Socioeconomic inequality may worsen as a result of deepfake technology's ability to deceive or alter public opinion. Socially targeted counterfeits may spread false information with the intent to sow discord among specific political or social factions. In addition to bringing attention to preexisting inequalities, this manipulation sows' seeds of mistrust and doubt about reliable news outlets. People who come across deepfake content may have some unsettling psychological effects. The spread of false information, whether through personal assaults, misleading deepfakes, or the general public's misunderstanding of a topic, can lead to profound psychological distress. As they try to separate truth from fiction, victims may experience increased anxiety, confusion, and a decreased sense of personal safety. The way people perceive pain has the potential to destroy their faith in digital media and institutions. This might have serious consequences for both personal well-being and communal peace [41].

2.5 THE CRITICAL ROLE OF DEEPFAKE DETECTION IN SAFEGUARDING SOCIETY AND DIGITAL INTEGRITY

Given the significant implications of synthetic media on society, security, and privacy, deepfake detection is indispensable. As deepfake technology becomes increasingly resilient, the importance of effective detection methods increases. Since deepfakes utilise contemporary AI algorithms to generate synthetic media that is highly convincing, they necessitate the implementation of robust detection techniques and pose numerous potential hazards.

1. Combating Misinformation and Disinformation

False information may be transmitted through the use of synthetic data and dishonesty, making deepfake detection an extremely important tool. Deepfakes' very lifelike bogus information has the potential to distort reality, deceive viewers, and alter their viewpoint. Because they are used in settings like political campaigns to create events or statements that are supposed to be made by influential individuals, deepfakes pose a significant threat. This sort of misleading information has the potential to impede democratic processes, impact voting behaviour, and divide the public along lines of evidence-based disagreement. In order to mitigate some risks, high-quality deepfakes detect false information before it may spread. Media outlets and governments can protect data source privacy and halt the spread of misinformation by quickly identifying deepfakes [42].

2. Protecting Personal Privacy and Reputation

Deepfake detection is indispensable due to its capacity to disseminate propaganda or misleading information without restriction. The relatively realistic fraudulent information of deepfakes has the potential to influence public opinion, distort statistics, or alter statistics. This is especially problematic in environments such as presidential elections, where intermittently manufactured counterfeits are employed to credit events or words spoken by well-known individuals. This type of false information has the potential to undermine democratic processes, influence voting behaviour, and create a society that is polarised and lacks knowledge of the facts. Effective deepfakes detect misleading information prior to its dissemination, thereby mitigating specific risks. The early detection of deepfakes will enable media sources and authorities to prevent the dissemination of misleading information and protect the integrity of each individual data source [43].

3. Enhancing Cybersecurity

A novel form of cybersecurity risk is introduced by deepfakes, particularly in the context of social engineering initiatives. Deepfakes are designed to target sensitive data or pose a security hazard by allowing colleagues and superiors to be mistaken for the intended targets. Major financial losses and potential compromises to organisational security may result from such an attack. The evolution of cybersecurity systems is necessitated by the detection of deepfakes. The creation and implementation of detection systems are facilitated by the improved identification or eradication of these intricate social engineering projects, which enables companies to maintain general resistance to cybersecurity risks and safeguard the safety of private data [44].

4. Preserving Trust in Digital Media

Deepfakes that are uncertain raise legitimate concerns regarding the integrity of online morality. The ability of the average observer to differentiate between genuine and fraudulent data is becoming increasingly difficult due to the increasing lifelikeness of deepfakes. The credibility of media companies, websites, or digital content distributors could be compromised by the widespread confusion and doubt that could result from this lack of faith. Good deepfake detection provides a method to verify the authenticity of content, thereby preserving trust in digital media. The separation of consumers and companies between real and fake media enables the enhancement of wise decision-making and the preservation of knowledge accuracy through the use of detection techniques [45].

5. Supporting Ethical and Legal Frameworks

The growth of deepfake technology raises important ethical and legal issues that require the creation of proper systems to address them. If these systems are in use, they mainly depend on monitoring tools, evidence, and ways to resolve issues of abuse. Accurate spotting helps use real technology properly and quickly address problems, keeping ethical and legal standards intact [46].

2.6 EVOLUTION OF DEEPFAKE TECHNIQUES

Deepfake techniques underscore significant advancements in AI and digital media, attributed to enhancements in neural networks for education, learning methodologies, and processing power. The expansion comprises many key periods, each marked by notable technological advancements and improvements in the realism and capabilities of synthetic media.

2.6.1 EARLY DIGITAL MANIPULATION - PRE-2010s

The emergence of fundamental digital manipulation tools in the late 20th century marked the beginning of the development of deepfake technology. Users were able to modify photographs and videos from the Adobe Photoshop programme to basic video editing tools with early digital editors. Even so, these initiatives possessed a reasonable degree of breadth and scope. Despite the fact that these technologies possessed the sophisticated algorithms necessary to generate truly synthetic media and necessitated substantial human intervention, they enabled minor modifications such as colour correction or cropping. The anticipated results frequently exhibited distinct artefacts or aberrations, thereby disclosing the source from which they were generated. The initial digital modification technologies, despite their simplicity, evolved into increasingly intricate concepts as they continued to build upon their foundation. The field of synthetic media was transformed by their capacity to modify digital data, which facilitated the development of increasingly intricate technologies such as GANs and the emergence of innovative ideas.

2.6.2 THE ADVENT OF GANs - 2014

The GANs that changed the course of deepfake technology were released in 2014. Working with a generator and the discriminator in a dual-network configuration, Ian Goodfellow makes a groundbreaking breakthrough in ML. A discriminator looks for and distinguishes between actual and created data, whereas a generator makes synthetic data, like photos or movies, that attempt to mimic real material. The realistic development of synthetic material is driven by a contradictory approach in which the discriminator and the producer engage in continual interaction. It evolves to generate increasingly plausible bogus items as its detector becomes better at detecting little deviations. This interactive process greatly improves the standard of synthetic media, which in turn alters the trajectory of potentially devastating deepfakes away from more traditional, unrealistic methods of digital manipulation.

2.6.3 ENHANCED GANs ARCHITECTURES - 2015-2018

Ian Goodfellow's pioneering work on GANs has resulted in a significant increase in the quality and authenticity of deepfakes. One intriguing development was the development of a distinctive training method that is predicated on Progress Growing GANs and consistently enhances the resolution of the generated images. This approach enables the network to begin with images that have a low resolution and subsequently improve them to generate outputs of exceptional quality with minimal artefact presence. In 2018, NVIDIA unveiled a significant technological advancement: Style GANs. By dividing the synthesis process into multiple layers of styles, Style GANs revolutionised the field of image production and facilitated the precise manipulation of numerous components of the resulting images. This method generates deepfakes that are both flexible and convincing by enabling the precise manipulation of facial traits and movements. These accomplishments collectively demonstrate a significant enhancement in synthetic media, thereby enabling the production of materials with greater precision and force [47].

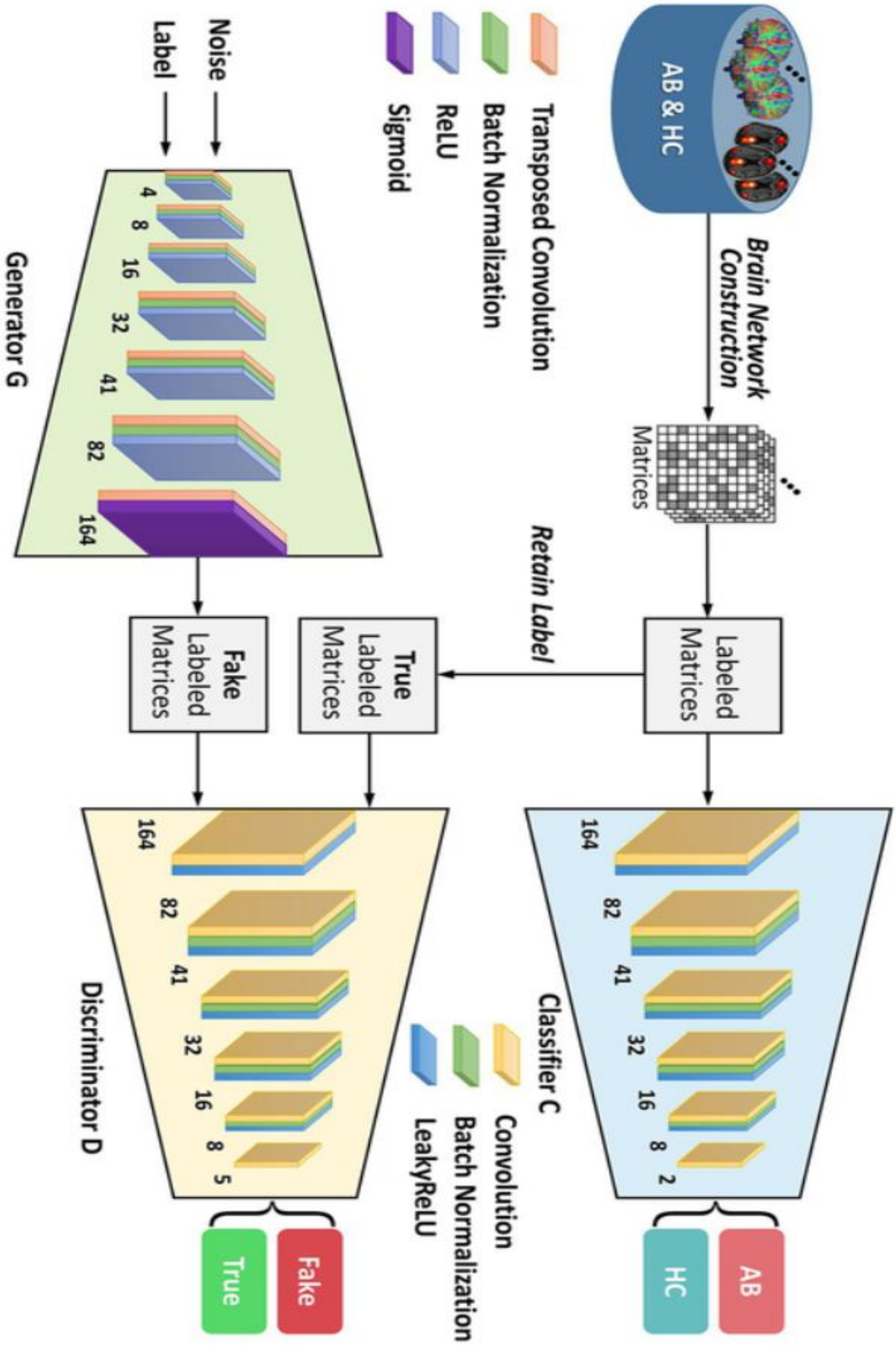

FIGURE 3: ENHANCED GANs ARCHITECTURES [47]

2.6.4 THE RISE OF TRANSFORMERS - 2018-PRESENT

The development of models, particularly BERT (Bidirectional Encoder Representations from Transformers) and GPT-2 (a pre-trained Transformer model), has significantly enhanced the capabilities of deepfake technology, impacting not just visual media but also synthesised text and audio. Transformers excel at processing and producing sequential information because of their attention mechanisms, which provide precise interpretation of context and dependencies in text and speech. Developed by OpenAI, GPT-2 exemplifies this advancement by generating human-like language with remarkable coherence and fluency, hence enabling the creation of synthetic written content that closely resembles human authorship. Transformers have been adapted for artificial speech, enabling the generation of very realistic audio material and voice imitations. These advancements have broadened the applications of deepfakes, facilitating realistic narrations, automated content generation, and enhanced editing of both written and spoken language, hence presenting new avenues for synthetic media [48].

2.6.5 IMPROVED REALISM AND FIDELITY - 2020-PRESENT

More sophisticated synthetic media has been the focus of recent large-scale initiatives to improve the realism and trustworthiness of deepfakes. Facial reenactment is one method; it involves filming one person's expressions and movements superimposed over another's face. The method enables the production of remarkably lifelike recordings in which the subject's expressions and emotions mimic those of the original artist, leading to engaging and authentic content. Deepfake videos currently are incredibly detailed and crisp, even if high-resolution video synthesis is becoming better all the time. These methods use modern algorithms to produce high-resolution, finely detailed movies, which greatly reduces the occurrence of visual artefacts and boosts the material's overall validity. Consequently, it is now increasingly difficult to distinguish between genuine and false footage, highlighting the necessity for advanced detection algorithms to keep up with the latest developments in deepfake technology [49].

2.6.6 REAL-TIME DEEPFAKES AND APPLICATIONS

Real-time deepfake technology represents a significant milestone in the field of synthetic media, as it enables the synthesis and mixing of deceptive material on demand. This innovation has generated novel prospects for immersive and interactive projects across numerous disciplines. The real-time deepfakes of synthetic media can be seamlessly integrated into the settings of AR, VR, and live streaming. Real-time deepfakes enable broadcasters to present animated or modified versions of their characters during live streaming, thereby enhancing engagement and providing new visual effects throughout the broadcast. This technology enables virtual and augmented reality users to create dynamic and interactive avatars, allowing them to experience immersive environments in which synthetic content responds to user activity through real-time change. Entertainment and gaming are among the applications; real-time deepfakes can generate responsive in-game characters or live event content, thereby providing a degree of customization and adaptability [40]. This new research highlights how deepfake technology is becoming more versatile and how it might influence how digital media is used in interactive and immersive events.

FIGURE 4: REAL-TIME DEEPFAKES [40]

2.6.7 ETHICAL AND REGULATORY CONSIDERATIONS

Deepfake technology has grown, along with worries about its use and the related legal and ethical issues. The growing reality and accuracy of deepfakes raise concerns about false information, privacy violations, and potential misuse. We need to address these problems thoroughly. Deepfakes can create very realistic but false information, which can harm public trust and the reliability of knowledge. This is because false stories or people's confusion can easily damage the truth. Making illegal and harmful content with deepfakes, like false accusations and fake adult images, goes against people's rights and character and raises serious ethics and privacy concerns. To effectively manage these issues, we need to create strong methods for accurately detecting deepfakes. It's important to create clear moral rules and laws to deal with the specific issues caused by manufactured media. These rules should outline what deepfakes cannot be used for, explain how to take responsibility and fix issues, and state the penalties for breaking the rules. Balancing innovation with responsible use requires ongoing discussions among engineers, lawmakers, and ethicists. Together, they can create rules that protect people and maintain the quality of online resources while promoting positive uses of deepfake technology [50].

2.7 COMPLEXITY OF DEEPFAKE CREATION

A smart combination of advanced technology and computational techniques has given rise to a significant degree of complexity in deepfakes, which have recently undergone a significant evolution. The primary function of deepfake technology is to produce synthetic media that is as realistic as possible by utilising robust ML techniques and neural network design. This intricacy is elucidated by the data requirements, computational requirements, and underlying technology that are associated with the production of high-quality deepfakes [39].

2.7.1 ADVANCED ML TECHNIQUES

Deepfake research starts with sophisticated ML techniques, commonly employing transformer models or GANs. Developed by Ian Goodfellow and his associates, GANs consist of two neural networks: one that generates data and another that acts as a discriminator. As the discriminator assesses the synthetic material produced by the generator in comparison to real data, it simultaneously creates itself. This hostile dynamic propels the generator towards ever more realistic media. Advancements in GAN architectures, such as Style GANs and Progressive Growing GANs, have significantly enhanced the quality and realism of Deepfakes.

Transformers are essential instruments for text generation and speech synthesis. These models excel in the processing and generation of sequential data, thereby providing abundant audio and video material. More realistic deepfakes, particularly in the replication of human voice and facial expressions, have emerged from transformers' capacity to emulate intricate interactions within data.

2.7.2 DATA REQUIREMENTS AND TRAINING

In order to achieve exceptional quality, the production of deepfakes necessitates substantial quantities of training data. In order to adequately equip the neural networks, it is necessary to include hundreds of audio samples, videos, and photographs in high-quality deepfakes. The process of fabricating a convincing false identity for an individual necessitates the collection of numerous facial images and videos from a variety of perspectives and illumination conditions. The model is able to accurately duplicate the minute details of a person's appearance and attitude as a result of this diverse range. The realistic nature of the material generated is significantly influenced by the quality and variety of training data. The legitimacy of synthetic media is enhanced by the use of diverse and representative datasets, as insufficient or biassed data can produce evident artefacts or deviations in the deepfake.

2.7.3 COMPUTATIONAL RESOURCES

Computationally intensive tasks are required for great deepfakes. To control the complicated calculations needed to train the neural networks, powerful hardware integrating GPUs and TPUs is essential. Training a deepfake model might take days or weeks, depending on the dataset size and complexity level. Iterative development, which is required to achieve the appropriate level of realism, increases the processing requirements, which in turn aggravates them. Sometimes, real-time processing which necessitates lightning-fast algorithms and supercomputer performance, leads to deepfakes. There are limited applications for this technology and a high demand for resources because deepfake production primarily relies on balancing computational efficiency with the quality of the produced information.

2.7.4 TECHNICAL EXPERTISE

Deepfakes need a substantial level of technological expertise. The creation of deepfakes necessitates a robust understanding of DL concepts, neural network architectures, and information processing techniques. To obtain trustworthy findings, researchers and practitioners must possess proficiency in selecting appropriate models, fine-tuning hyperparameters, and managing extensive training programs.

2.7.5 ETHICAL AND PRACTICAL CONSIDERATIONS

Major ethical and pragmatic concerns are raised by the complexity of deepfake generation. Despite the fact that technology has legitimate applications in industry, education, research, amusement, and theatre, as well as in the sector, its potential to generate inaccurate or detrimental information emphasises the necessity of ethical standards and robust responses, even if it is used appropriately in the sector. As deepfake technology continues to evolve, it is imperative to address these challenges through ethical technical implementation, which is based on ongoing research and robust detection techniques [51].

2.8 TECHNOLOGICAL ADVANCEMENTS IN DEEPFAKE CREATION

The growth of deepfake technology raises important ethical and legal issues that require the creation of proper systems to address them. If these systems are in use, they mainly depend on monitoring tools, evidence, and ways to resolve issues of abuse. Accurate spotting helps us use real technology properly and quickly address problems, keeping ethical and legal standards intact.

2.8.1 GENERATIVE ADVERSARIAL NETWORKS (GANs)

GANs have significantly expanded deepfake technologies. GANs are composed of generator and discriminator neural networks, which were initially discussed by Ian Goodfellow and colleagues in 2014. The generator generates synthetic data, such as photographs or videos, while the discriminator compares the generated data to real-world examples. This adversarial educational strategy, in which the two networks engage in a battle, results in consistent enhancements in the quality and realism of the information that is generated. It is recommended that both networks improve their performance, even if the generator becomes more capable of producing compelling evidence, as the discriminating system becomes more adept at identifying minor changes. A number of significant advancements in deepfake technology have occurred since GANs initially demonstrated potential. Progressive expanding GANs (PGGANs) are an example of such an innovation. Over the course of training on low-quality images, PGGANs progressively enhance the detail and resolution of images. This approach enables the attainment of exceptional outcomes with a reduction in artefacts and an increase in clarity. StyleGANs, which was developed by NVIDIA, is another significant consequence. It provides a unique method of graphic production by deconstructing the synthesis process into multiple styles. Quite realistic and customised synthetic images are produced by the precise control over the synthesis of facial characteristics and expressions that Style GANs provide. The resolution, detail, and general realism of deepfakes have been profoundly altered by these developments, rendering them significantly more challenging to distinguish from real-world media [52].

2.8.2 TRANSFORMER MODELS

Although these model converters were initially primarily recognised for their ability to convert natural language, they are now quite beneficial in the production of deepfake images. The extraordinary language comprehension and generating capacity of models like GPT-3 (this Pre-trained Transformer 3) are demonstrated by their ability to provide coherent and contextually acceptable information. These converters have expanded the availability of other synthetic media, such as voice and video synthesis. By analysing extensive databases of human speech and identifying tones, pitch, and intonation, transformer-based models can generate voiceovers that are remarkably human-like. Apart from visual elements, this method generates intricate vocal and text-based deepfakes. The integration of inverters into deepfake tools not only enhances the realism and complexity of fake content, but also expands the application range, resulting in more immersive and captivating content on numerous other platforms [53].

2.8.3 HIGH-RESOLUTION SYNTHESIS

Recent improvements in high-resolution synthesis have made deepfake media much better, making it more similar to real material and improving its trustworthiness. Super-resolution generative adversarial networks (SRGANs) have played a key role in improving picture quality. SRGANs make low-resolution pictures look better and more detailed, similar to high-quality images, by using a special type of GAN. Along with this development, higher-density video synthesis methods have emerged to ensure that made videos not only show improved consistency but also excellent clarity, hence lowering the artefacts. These methods correct common image problems, improve overall visual quality, or refine video details using smart algorithms. Deepfakes highlight the importance of better ways to spot fake images and videos because they are harder to tell apart from real ones [54].

2.8.4 FACIAL REENACTMENT AND EXPRESSION TRANSFER

One may now regulate and replicate facial emotions with remarkable precision due to advancements in emotion transmission and facial reenactment technology. These methods allow for the creation of synthetic movies that look rather realistic by matching a person's facial expressions and movements to another's face. Face2Face and DeepFaceLab are two tools that assist explaining this progress by defining real-time face replication. By allowing an actor's facial expressions to flow organically onto a target person, Face 2 Face creates photos where the target's face genuinely reflects the actor's emotions and actions. Proper and comprehensive expression transmission is also made possible by the SotA facial alteration techniques offered by Deep Face Lab. These technologies, which are mostly utilized for entertainment purposes, enable for more engaging and lifelike video conferencing and improve visual storytelling and online interactions. These methods are becoming increasingly important because they highlight how drastically misleading technology might change the production and consumption of digital information.

2.8.5 REAL-TIME DEEPFAKES

The potential of interactive media has been substantially expanded by real-time false technology, which has subsequently opened up new avenues for content creation and user interaction. This technology enables dynamic live streaming and immersive experiences through the use of real-time multimedia that is powered by artificial synthesis and modification. Zao and Deepfakes Web have demonstrated that it is feasible to generate and disseminate deepfakes at a rapid pace, thereby enhancing the immediate accessibility and engagement with digital data. Real-time deepfakes enable VR to create avatars that are both highly adaptable and rapidly responsive, thereby improving the immersive experience for consumers. Particularly, the gaming experience is improved by more refined and adaptable character interactions. Real-time deepfakes captivate viewers by offering new forms of creative expression, as they enable live broadcasting and the exploration of live particle effects and on-demand content changes. The advancement of this technology offers opportunities for interactive media and enhances user engagement and immersion [55].

2.8.6 ETHICAL AND PRACTICAL CONSIDERATIONS

There are a number of important practical and ethical concerns raised by the rapid advancement of deepfake technology. The associated problems, such as the dissemination of false information, invasions of privacy, and deliberate technological exploitation, become more serious as the legitimacy of counterfeits increases. Deepfakes pose a significant threat to public trust in data and integrity because they offer information that is both realistic and deceptive. By spreading misinformation and leading people to misunderstand the truth, deepfakes influence public opinion. Concurrently, invasions of privacy cause mental anguish and damage a person's reputation through unwarranted criticism or lack of permission. The potential for dishonest individuals to use deepfake technology for dishonest purposes emphasises the need for robust regulatory measures and comprehensive identification procedures. Improving detection methods to accurately identify deepfakes should be our first priority in the fight against these dangers. Overcoming these challenges and ensuring the correct deployment of deepfake technology requires the creation and execution of strong ethical guidelines and comprehensive regulatory frameworks. In order to effectively manage and mitigate the impacts of synthetic media, these strategies must progress in tandem with technological developments [56].

2.9 IMPACTS OF DEEPFAKES

Deepfakes impact many aspects of life, technology, and personal safety. Deepfake technology has advanced and is having a growing impact on hacking and personal privacy.

1. Misinformation and Public Trust

Deepfakes' ability to quickly disseminate false and misleading information is one of its most concerning consequences. Deepfakes create information that seems genuine but is actually inaccurate, altering reality and fooling viewers. This is particularly problematic within the context of political campaigns, since deceptive audio recordings or false films have the potential to sway public opinion and ultimately determine election results. For example, disillusionment with democratic processes and civil unrest might be stoked by a deepfake video depicting a public official making controversial comments. By casting doubt on all forms of digital content and eroding public faith in reliable news outlets, this disinformation campaign aids the general public in distinguishing fact from fiction [57].

2. Privacy Violations and Personal Harm

Deepfakes actually severely limit individual freedom. One unlawful and dangerous use of synthetic media reporters might come upon is deepfake obscenity or libellous content. Other types of abuse impair credibility, damage money, and cause emotional pain aside from attacks on personal confidence. Deepfake victims highlight how urgently laws and legal action have to be followed to limit privacy violations by means of knowledge validating or reputation cleansing [58].

3. Cybersecurity Risks

Additionally, the advancement of deepfakes is significantly impacting cybersecurity. Deepfake technology enables hostile actors to initiate ambitious social engineering projects. For example, executives may be deceived into disclosing confidential information or engaging in security-compromising behaviours through the use of executive voice recordings or deepfake movies. Deepfakes are capable of generating plausible human personas that impede the identification of dishonest behaviour, thereby enhancing the effectiveness of deception attempts. New detection tools and methods must be developed to prevent these increasing hazards, as traditional cybersecurity methods may find it challenging as deepfake technology develops [59].

4. Legal and Ethical Challenges

There are serious ethical and legal concerns brought up by the rapid development of deepfake technology. Current legal systems are ill-equipped to deal with the unique problems posed by deepfakes, such as privacy invasion, intellectual property theft, and defamation. This disparity informs initiatives in the field of synthetic media intake limitation and moderation. Consequently, there is an immediate need for updated regulations and standards that address crimes involving deepfakes and provide clear guidelines for their implementation. In addition to following the law, following ethical norms is crucial for ensuring the proper implementation of artificial technology and avoiding any negative consequences [60].

2.10 SUMMARY OF RELATED WORK

TABLE 1: LITERATURE SUMMARY

Author/year	Methodology used	Controversies	Findings	Limitations
[61]	The Generative model synthesises realistic histologic images for enhanced diagnostic training.	Synthetic histologic images may risk diagnostic accuracy and ethical data misuse.	Synthetic images matched real samples in diagnosis accuracy, aiding pathologists' assessments.	The Model struggles with rare morphologies and requires extensive computational resources.
[62]	Participants categorised deepfakes accurately; detection strategies did not improve performance.	Detection strategy effectiveness questioned; participants overestimate individual deepfake identification skills.	Participants detected deepfakes above chance, with strategies showing no impact.	Detection strategies lacked effectiveness; overconfidence bias was observed in individual assessments.
[63]	Deepfake personas are perceived negatively; there is no performance advantage over traditional personas.	Deepfake personas raise ethical concerns, credibility issues, and uncanny valley effects.	Deepfake personas are less credible, empathetic, and immersive than traditional persona methods.	Deepfake personas hinder user experience; traditional profiles are preferred by designers.

[64]	The Compressed deepfake detection model uses knowledge distillation for efficient, accurate forensics.	Model compression may compromise detection accuracy; deepfake misuse concerns persist.	Student networks achieved competitive performance with reduced computational costs effectively.	Model may underperform on complex deepfake samples; requires extensive training.
[65]	Deepfake audio classification methods evaluated; custom architectures outperform existing techniques.	Deepfake audio raises ethical concerns; potential misuse complicates forensic investigations.	Custom architectures enhance deepfake audio detection; VGG-16 excels with MFCC.	Results vary with different audio types; requires extensive training data.
[66]	Deepfake audio detection employs feature extraction and DL architectures.	Voice cloning risks privacy; ethical concerns over misuse persist significantly.	Deepfake audio detection improves with custom architectures and feature extraction.	Detection accuracy varies with audio quality; extensive training is needed.
[67]	The study examines deepfake modalities' impact on belief and sharing intentions.	Perception of deepfakes varies; ethical concerns about misinformation and sharing.	Video deepfakes are perceived more accurate; cognitive ability affects sharing intentions.	The study focused on self-reported data; limited sample diversity and context.

[68]	The study examines GAN methods for deepfake implementation and detection techniques.	Deepfake technology raises ethical concerns regarding misuse and misinformation dissemination.	GAN models effectively implement deepfakes; detection techniques require further development.	Detection methods need improvement; deepfake quality outpaces current technologies.

2.11 OVERVIEW OF DEEPFAKE DETECTION METHOD

In order to meet the requirements of synthetic media that are becoming increasingly sophisticated, deepfake detection methods have undergone a substantial transformation. These methods combine a variety of technological and analytical tools to identify and combat deepfakes, which could be essential for safeguarding the integrity of digital content and preventing privacy invasions, misleading information, and cybersecurity concerns. This review examines the primary detection methods and their effectiveness in identifying deepfake content.

2.11.1. VISUAL AND AUDITORY ARTEFACTS DETECTION

Searching for audio and visual distortions. One of the fundamental approaches for spotting deepfakes is highlighting synthetic stuff. Early on in deepfake systems, the detection technique was largely aimed at finding variants in pixel-level characteristics exposing fraudulent content. Among the most typically occurring symptoms are distorted facial expressions, uneven lighting, and abnormal eye movements and visual artefacts. Often these anomalies resulted from the constraints of early deepfake-generating algorithms, which fell short in matching the delicate intricacy of human traits and the real lighting settings. Regarding audio deepfakes, much research has concentrated on object recognition based on synthetic speech patterns or changing voice diffusions. Early iterations of audio deepfakes occasionally had noticeable signs on careful inspection, including unnatural intonation, uneven breaks, or irregular pitch and tone [69]. These initial detecting techniques have become less interesting as deepfake technology advances. Modern deepfakes hide the previous unique signals by almost perfect replays of reality utilising complex algorithms and neural networks. GANs and other DL techniques have greatly enhanced the quality of synthetic media, so finding changes using traditional approaches becomes more difficult. Deepfake detection has so evolved to employ more subtle and sophisticated methods to identify and neutralise the constantly produced synthetic material in current times.

2.11.2 ML AND AI-BASED DETECTION

Today, deepfake detection is primarily reliant on AI and ML, which are capable of compiling and analysing vast quantities of data to identify minute trends that may be invisible to the human eye. CNNs and repeated neural networks (NRNs) are notably beneficial in this context. The image recognition capabilities of CNNs, which are renowned for their application in the detection of deepfakes. These networks learn from a vast array of synthetic and real-world media in order to distinguish between synthetic and practical problem content based on the attributes they have acquired. CNNs identify deviations and abnormalities that suggest a deepfake by utilising facial math, colour consistency, and texture. Sequential data, such as audio signals and video frames, is also facilitated by RNNs. They are highly adept at identifying temporal patterns and anomalies that are indicative of deepfake material. For instance, RNNs are capable of detecting anomalies in voice patterns or false transitions in video sequences. These models are disadvantaged despite their advantages, as the methods of producing deepfakes are constantly evolving. Modern deepfake technology may employ sophisticated methods to conceal objects that were previously visible, necessitating the modification of detection models. Development and frequent retraining are crucial for the continued success of RNNs and CNNs. In order to accommodate advancements in deepfake manufacturing technology, researchers are consistently improving these models by incorporating additional components and implementing more sophisticated algorithms. The accuracy and endurance of false fake detection systems on a terrain that is constantly changing are maintained by this constant enhancement [70].

2.11.3 DEEPFAKE DETECTION USING METADATA ANALYSIS

Due to its ability to reveal the fundamental content of digital data that exhibit signs of contamination or alteration, information analysis is a valuable tool for deep fakes. The metadata of a file includes details like its creation date, encoding rules, and modification history. Inconsistencies or conflicts in changes to signals can be better identified with the use of data analysis. If a video clip has an unusual change history or contradictory time stamps, it might mean that the movie has been edited, which could mean that there is a deep copy. On occasion, deep fakes will significantly alter digital files and then leave suspicious metadata trails. These adjustments may necessitate new approaches to file arrangement or encoding, depending on how they have worked in the past. Competent attackers, however, are starting to place a premium on experience during the identifying step [71]. As a result, they often employ techniques of data deletion or alteration to mask their actions. This behaviour highlights the necessity for alternative detection methods, as it can affect the reliability of database analysis. and the ongoing improvement of forensic tools to meet the growing number of escape attempts [72], [73], [74].

2.11.4 BLOCKCHAIN AND DIGITAL FINGERPRINTING

New techniques Blockchain technology and electronic fingerprints can improve the discovery of deepfakes by providing a secure way to track digital content over time. Digital fingerprinting creates unique IDs or marks for original items, making it easier to identify them. These fingerprints pick out specific features of the original media, like compression methods or pixel arrangements. New versions or copies of the content highlight any changes because their signature matches the original. This method ensures that digital media is real by spotting any changes that might show it has been tampered with or altered. Blockchain technology offers a public record that can track changes in digital information. Using blockchain to track every step of a material's life allows you to trace its origins and verify its authenticity at any time. Blockchain records provide a secure method to check for unauthorised changes in digital media. These technologies are still being developed, but they show promise in helping fight deepfakes and improve how we check the truth of material. Overview of the items. Using both together can help track and check digital content, which can lower the risks of fake content [75].

2.11.5 BIOMETRIC-BASED DETECTION

Biometric-based detection systems identify deepfakes by focusing on factors that are difficult for synthetic materials to replicate, as well as by analysing various physiological characteristics. Deepfake algorithms are unable to consistently replicate these minute characteristics, such as face feature alignments or minute muscular activity, by utilising human recognition technologies, which collect micro-expressions and advanced facial landmarks. These attributes of biometric technologies facilitate the identification of variations that indicate content changes. Speech pattern, tonal properties, and voice modulation are employed by voice biometrics to distinguish between synthetic and genuine human voices. Deepfake voice synthesis generates fraudulent material by accurately replicating natural changes in intonation, tone, and speech cadence. These methods offer a distinctive and captivating approach to identifying deepfakes by emphasising fundamental human characteristics that are more difficult to replicate, even if they are still in the process of development and have challenges with adaptability and precision.

2.11.6 HYBRID DETECTION APPROACHES

A growing number of researchers are utilising hybrid techniques that combine several detection algorithms to improve the identification's accuracy. The detection system's performance is improved by integrating ML approaches with metadata analysis and biometric checks. Metadata analysis can reveal instances of file attribute abnormalities that may indicate tampering, providing further evidence beyond what predictive models can detect [76], [77]. Biometric testing, such as voice analysis and face recognition, enhance the verification process by matching the correctness of physical attributes with language usage. Hybrid approaches use the strengths of many technologies to overcome the weaknesses of each one, resulting in a more robust and comprehensive solution for deepfake detection. By covering all bases, this method improves overall detection capability while decreasing the likelihood of false positives and missing detections through more thorough examination of digital information [78].

2.12 MODELS OF INTEREST

One kind of ML is AI, which enables computers to learn from data, spot trends, and make judgments with minimal human input. makes gradual discoveries based on data utilizing a variety of methods ML uses algorithms to create mathematical models that can learn from a wide variety of data types (structured, informal, unstructured, etc.), and these models may improve their performance over time without requiring any manual programming. ML that is compatible with supervised and unsupervised learning, and with reinforcement learning models in general. Similar to image recognition and spam detection, supervised learning algorithms learn to associate inputs with outputs using labelled data. Unsupervised learning makes use of clustering and association algorithms to uncover hidden patterns or underlying structures in unlabelled data. This sort of learning is crucial for recommendation systems and customer segmentation. Autonomous vehicles, video games, and robotics all make use of reinforcement learning, which is based on techniques for training via failure and success and aims to maximize decision-making by rewarding good behaviour and penalizing mistakes.

Depending on the job at hand and the characteristics of the data, each of the five fundamental ML algorithms decision trees, neural networks, support vector machines, and ensemble methods has its own set of advantages and disadvantages. When it comes to handling large amounts of data and complicated patterns, neural networks, and DL networks in particular, are rather effective at things like picture and speech recognition. ML is useful in many industries, including healthcare (for example, to aid in diagnosis and provide individualized treatment recommendations), banking (to aid in fraud detection and risk management), and marketing (to enhance customer targeting and sentiment analysis) [79]. Data preparation, feature selection, and model assessment are the fundamental building blocks of ML systems. Model evaluation ensures that the model is relevant and valid. Even after standardizing the raw data and handling missing values, feature selection still finds the most relevant variables. Model assessment is useful for investigating the model's generalizability and effectiveness using cross-valuation methodologies and performance measures including accuracy, precision, and recall. The advent of self-driving vehicles, voice assistants, predictive analytics, and other advancements made possible by ML have revolutionized many industries by automating repetitive tasks and generating better decisions. Data security, bias, and improved model interpretability are still obstacles, particularly in DL. More and more, these problems are calling for explainable AI (XAI) solutions, which make ML judgments more visible and understandable crucial in industries like healthcare and finance where accountability is key. More and more people are paying attention to moral concerns in ML, such making sure there is fairness and transparency, because they are vital for society as a whole. This is because ML is becoming increasingly significant. ML is becoming increasingly pervasive as a result of improvements in data availability and processing power. This has implications for many areas of society, including productivity, innovation, and problem-solving strategies [80].

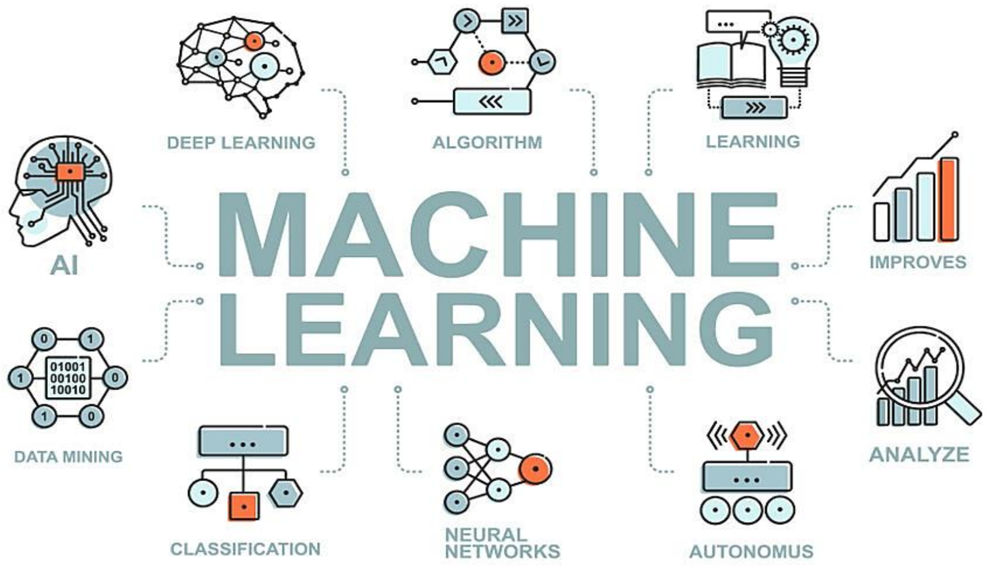

FIGURE 5: MACHINE LEARNING [81]

2.12.1 FORENSICS CONVOLUTIONAL NEURAL NETWORK

A multilayer perceptron (MLP) is a type of FCNN where each neuron in one layer is linked to every neuron in the next layer. FCNNs, also known as nodes, Neurons have an input layer, some secret layers, and an output layer. Each layer receives input from the layer below, adjusts it using bias and weights, and then applies an activation function to produce output that is sent to the layer below. An FCNN is very flexible and can be used for tasks like classification and regression. Its main purpose is to learn to connect inputs to outputs in complicated, nonlinear ways. There are two main steps in learning for a feedforward neural network: feedforward and backpropagation. The feedforward step goes through the network layer by layer to create a forecast. FCNNs use backpropagation to reduce forecast errors. The model calculates the error by comparing its output to the goal and then changes its biases and weights using a method called gradient descent. The model learns by reducing errors step by step, starting from the final result and going back to the beginning. The activation methods in FCNNs are essential for helping the network recognize complex interactions. Sigmoid and tanh functions are commonly used in the output layers of classification problems. They help with issues related to decreasing gradients and can speed up training. The ReLU (Rectified Linear Unit) function also assists with these problems. You need to carefully adjust the hyperparameters that control the number of hidden levels and the number of neurons in each layer of a FCNN because they greatly affect how well the network works. If there are too many layers or neurons, the model might overfit. This means it does well with the training data but struggles with new, unknown data [16], [32], [78], [82], [83], [84], [85]. FCNNs work really well when data can be organized neatly. Tasks in prediction analytics, natural language processing, and image recognition all rely on them. FCNNs can handle complex functions, but they struggle with large, high-dimensional data when compared to CNNs and Recurrent Neural Networks (RNNs). CNNs are better for image data, while RNNs are better for sequence data. Despite these limits, FCNNs are still fundamental in analysing neural networks. They are important in many machines learning uses because they are simple, flexible, and can tackle difficult problems [86].

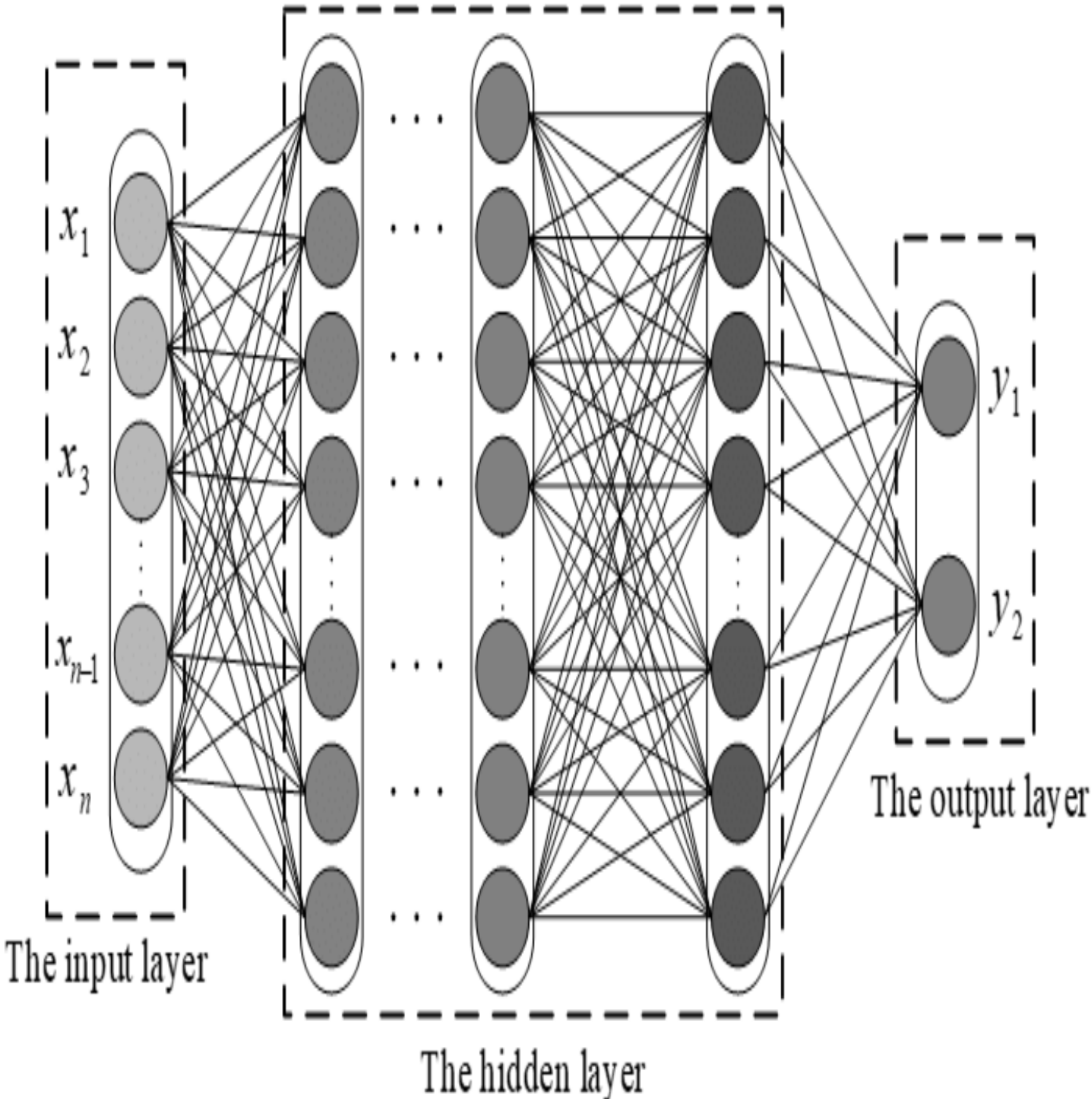

FIGURE 6: FORENSICS CONVOLUTIONAL NEURAL NETWORK [32]

2.12.2 VISION TRANSFORMERS

Developed initially for NLP, ViTs are a type of neural network design for visual data analysis with the Transformer paradigm, which may be used for computer vision and image recognition. ViTs differs from CNNs in that it treats each image patch as a "token" (like a word in natural language processing) rather than an image pattern. After that, these patches are translated into vectors that may be used as input tokens for the Transformer using linear embedding. By virtue of its self-attention mechanism, ViTs are able to record global links between picture patches, which enables the Transformer to perform exceptionally well on tasks like as object identification, image classification, or segmentation that benefit from both local and global context. By creating an interaction between every pair of patches in every layer, self-attention enables the model to focus on relevant patches based on the job at hand. This method is in contrast to CNNs, which prioritize local interactions while frequently using pooling layers or huge filter sizes to encompass global context. Training a ViT necessitates integrating positional data throughout each patch vector because to Transformers' inability to comprehend spatial organization natively. The model then sends these embeddings across several feedforward neural network layers and self-attention, each layer optimising the representations. ViTs give versatility with a common Transformer architecture that can be changed across workloads with minimal changes, unlike CNNs, which might need substantial customising for particular visual tasks. ViTs need vast volumes of data and processing resources to get optimal performance, therefore they are better appropriate for applications with abundant of data even if they have shown competitive performance on large-scale vision datasets. As hybrid ViT-CNN models and data-efficient training approaches develop, however, ViTs are becoming ever more valuable. By demonstrating a change in computer vision from localised, convolutional approaches towards architectures leveraging attention mechanisms for greater global context awareness, ViTs promise enhanced scalability and flexibility across difficult vision tasks [86].

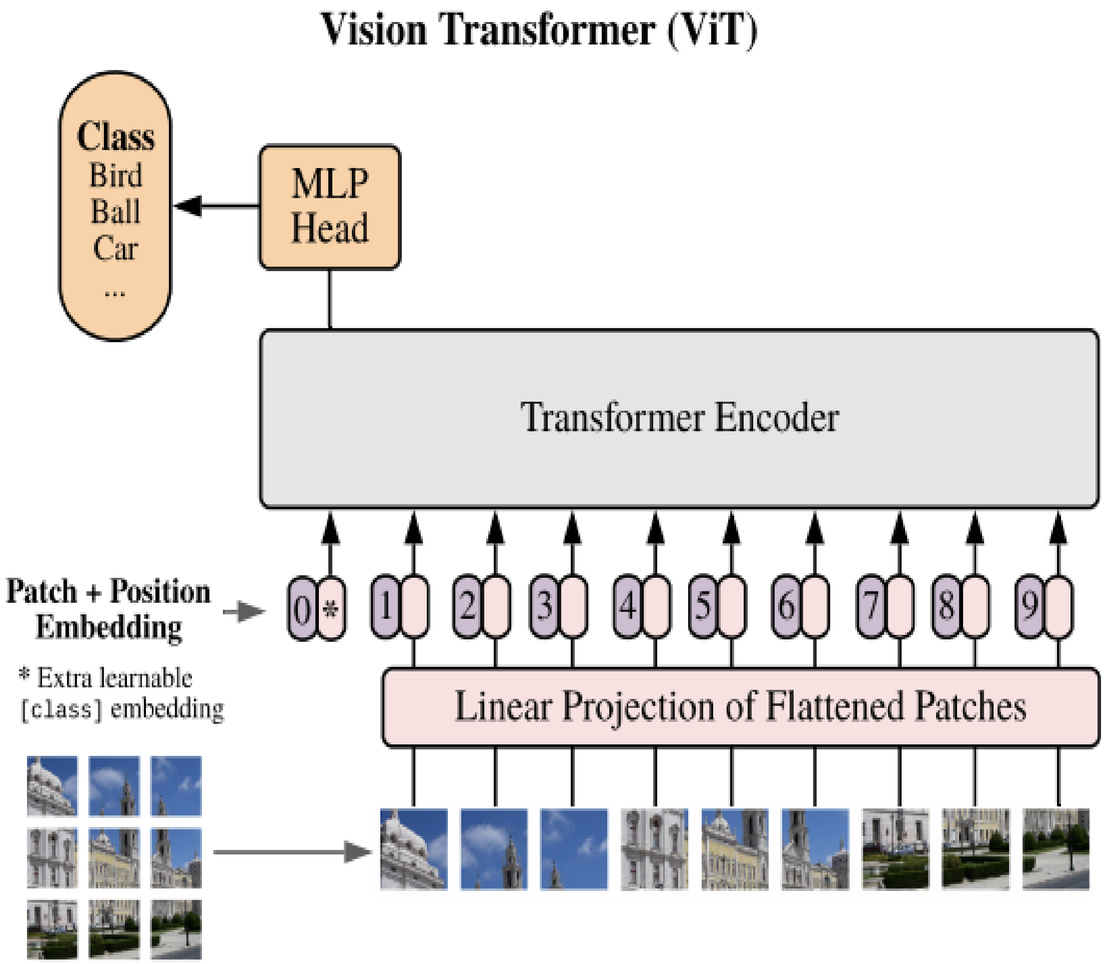

FIGURE 7: VISION TRANSFORMERS [33]

2.12.3 MASKED AUTOENCODERS

In computer vision, MAEs are a type of self-supervised learning model designed to train effective data models by reconstructing masked sections of input data. Driven by the success of masked language models such as BERT in natural language processing, MAEs expand this concept to visual data mostly images. Many pictures in an MAE are randomly masked, that is, some parts of the image are covered or replaced with placebos. The model then reconstructs these masked patches learning robust and contextually sensitive representations based on the accessible areas of the picture. MAEs consist essentially of an encoder plus a decoder. From the incomplete knowledge, the encoder builds latent representations by processing the unmasked patches, therefore preserving fundamental features and patterns. Having these embeddings, the decoder attempts to reconstruct the masked portions by filling in the missing patches, hence learning a condensed form of the original picture. Since it drives the model to gain a thorough understanding of both local characteristics and global context to precisely anticipate the missing material, the masking method is absolutely essential in MAEs. This helps pre-training efforts aiming at producing meaningful representations free of labelled data, which might then be modified on downstream tasks such picture classification, object recognition, or segmentation. Traditional autoencoders reduce memory and processing demands by compressing the whole picture; MAEs, on the other hand, are computationally efficient, processing just the uncovered portions during encoding. Because it speeds up training and ensures that MAEs scale appropriately with high-dimensional visual input, this selective encoding strategy is useful for large datasets. Due to their ability to generalise across workloads and adequately capture rich feature representations, MAEs have demonstrated remarkable performance, coming close to SotA on several computer vision benchmarks. Designing and tweaking the mask ratio the amount of the picture that is hidden and the model's capability to handle several masking patterns are crucial for optimal performance. Through the use of masked prediction, MAEs offer a versatile and data-efficient way to learn high-quality representations with little to no labelled data, bridging the gap between supervised and unsupervised learning in computer vision [87].

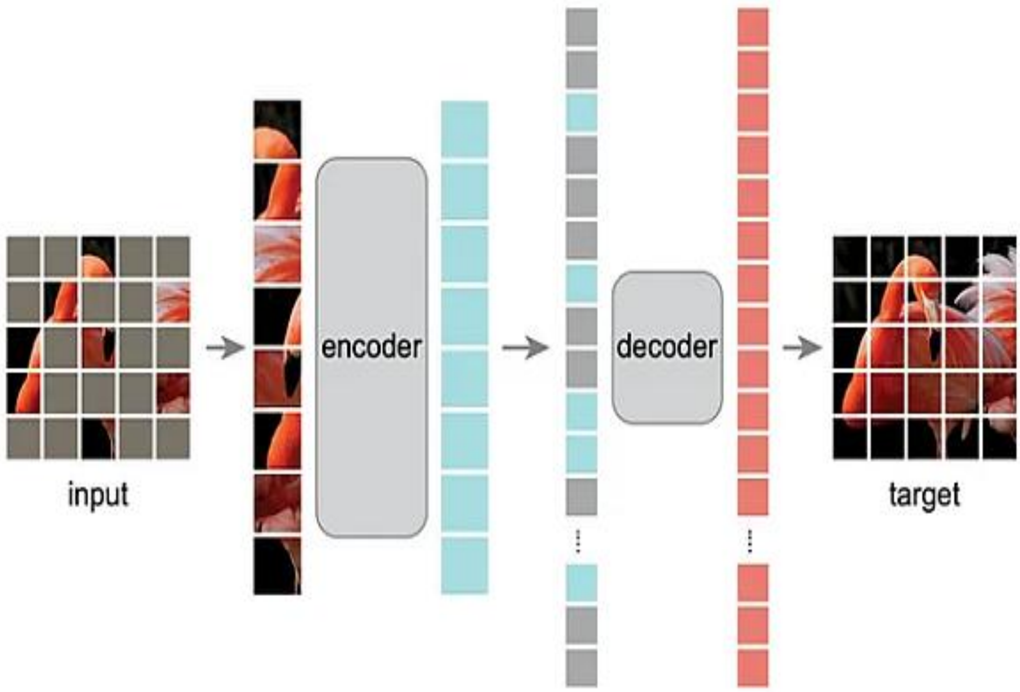

FIGURE 8: MASKED AUTOENCODERS [88]

2.12.4 TWO STREAM NEURAL NETWORKING

TSNN are a type of DL architecture that are predominantly intended for video and action recognition applications. They simultaneously record spatial and temporal input. Traditional image recognition methods are highly effective for the classification of fixed images, despite their inability to organically capture motion dynamics and their poor performance with video data. TSNN resolve this limitation by processing two distinct types of information: one stream captures spatial features from human video frames, thereby expressing stationary appearance knowledge, and the other stream captures behavioural traits, typically using optical velocity or similar motion-based data. Each stream functions as a unique CNN, with the spatial stream concentrating on the detection of items and environmental elements within each RGB frame of the film. The temporal stream processes stacks of optical flow images by emphasising the fundamental components of motion directional and velocity which convey information about activities and occurrences over time. Typically, the two streams acquire complementary qualities independently and are combined at the final layers through the use of feature fusion methods such as concatenation, average, or learning weighted summation. The network is able to generate a coherent prediction by integrating stationary content with motion context [89]. TSNN are adept at recognising motions that appear identical in individual frames but differ based on the motion by recording both elements, such as waving instead of hurling or strolling instead of sprinting. The temporal stream is particularly essential for enhancing the model's sensitivity to sequential input patterns, which is why it is crucial to succeed in activities with temporal continuity. TSNN have demonstrated exceptional gesture detection performance in a variety of benchmarks, surpassing the capabilities of single-stream architectures that are currently in use. However, they are materially substantial due to the additional storage and processing requirements that arise from the analysis of optical flux. Modern models have attempted to enhance efficiency by incorporating more lightweight motion representations or by combining TSNN with 3D CNNs. Despite these obstacles, TSNN continue to be a fundamental architecture in video analysis. This is due to the fact that they provide a robust foundation for applications in statistical analysis of sports, tracking, human-computer interaction, and numerous other fields where the ability to understand both the past and the future is essential. Their approach paves the way for advancements in video recognition and continues to motivate new research on efficient, multi-stream systems that address challenging sequential data problems [88].

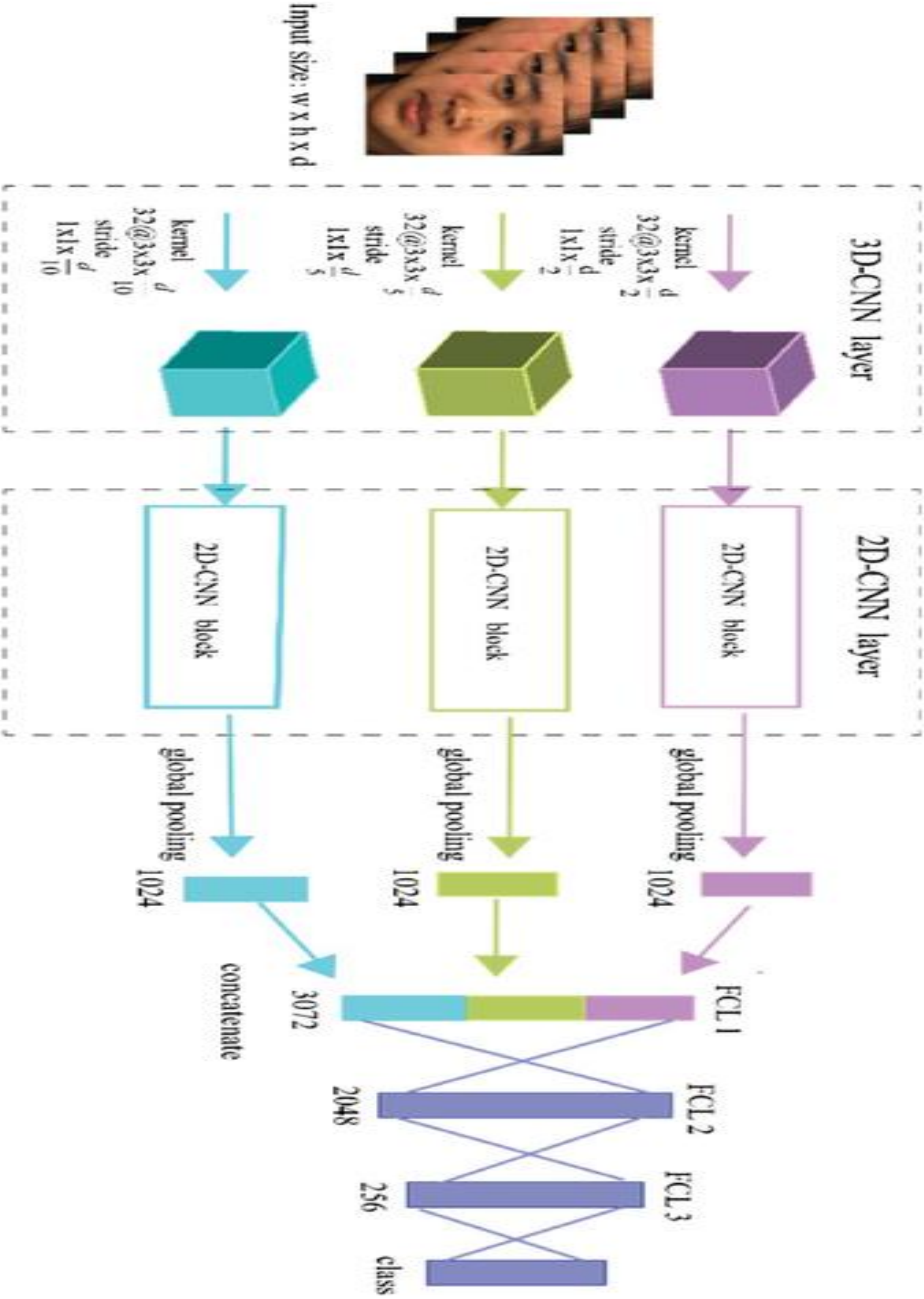

FIGURE 9: TWO STREAM NEURAL NETWORKING [16]

2.12.5 COMMON FAKE FEATURE NETWORK

CFFN is an automated reasoning model that is frequently primarily designed to identify synthetic or manipulated content, also referred to as "deepfakes." The validation of authenticity is significantly impeded by deepfakes, which generate convincing but altered images, videos, or audio. CFFNs are designed to identify minute modifications in this material. During extensive research, CFFNs disclose "false features" that demonstrate computational-level alterations that are undetectable to humans. Thus, CFFNs are perpetually engaged in the investigation of aberrations that are frequently observed in deepfake media, including anomalies in illumination, texture, facial geometry, and movement dynamics. These anomalies typically deviate from the natural patterns observed in real-world video [90]. These aberrations may be the result of generative models that generate deepfakes, which are characterised by pixel-level variations, creative regularity in transitions, or tiny spatial errors that originate from synthesis defects. The architecture of a CFFN is typically composed of layers that are specifically designed for pattern extraction. Its objective is to identify local and global changes that convey erroneous content. CNNs are frequently employed to identify spatial anomalies across frames, whereas behavioural layers, such as Time CNNs (TCNs) or RNNs, concentrate on detecting changes over time in motion pictures. CFFNs are frequently inaccurately generated; however, they can also leverage neurological attention processes to assist the model in concentrating on regions that are most likely to contain aberrations. This approach improves detection accuracy by emphasising questionable areas such as lip movements, eye flashing, or facial characteristics. The network is able to acquire statistically indicative of fakes but difficult-to-identify visually distinct features by training a CFFN with synthetic and real-world data. CFFNs may be utilised on a variety of media platforms to identify counterfeit content, including audio, video, and images, once they have been taught. In pragmatic contexts, they have demonstrated their value in a variety of sectors, including forensic investigation and social media surveillance, where the identification of deepfakes is becoming increasingly critical. However, CFFNs will always be required to adapt to new forms of manipulation that are not detectable by current methods, as generative models for deepfake generation are constantly improving. This has led to the development of a dynamic research area that is dedicated to the perpetual enhancement of CFFN architectures through the application of advanced technologies, including adversarial training and domain adaptation, in order to ensure their resilience against the development of deepfake techniques. Consequently, this area is supporting the ongoing battle against digital deception and misinformation [91].

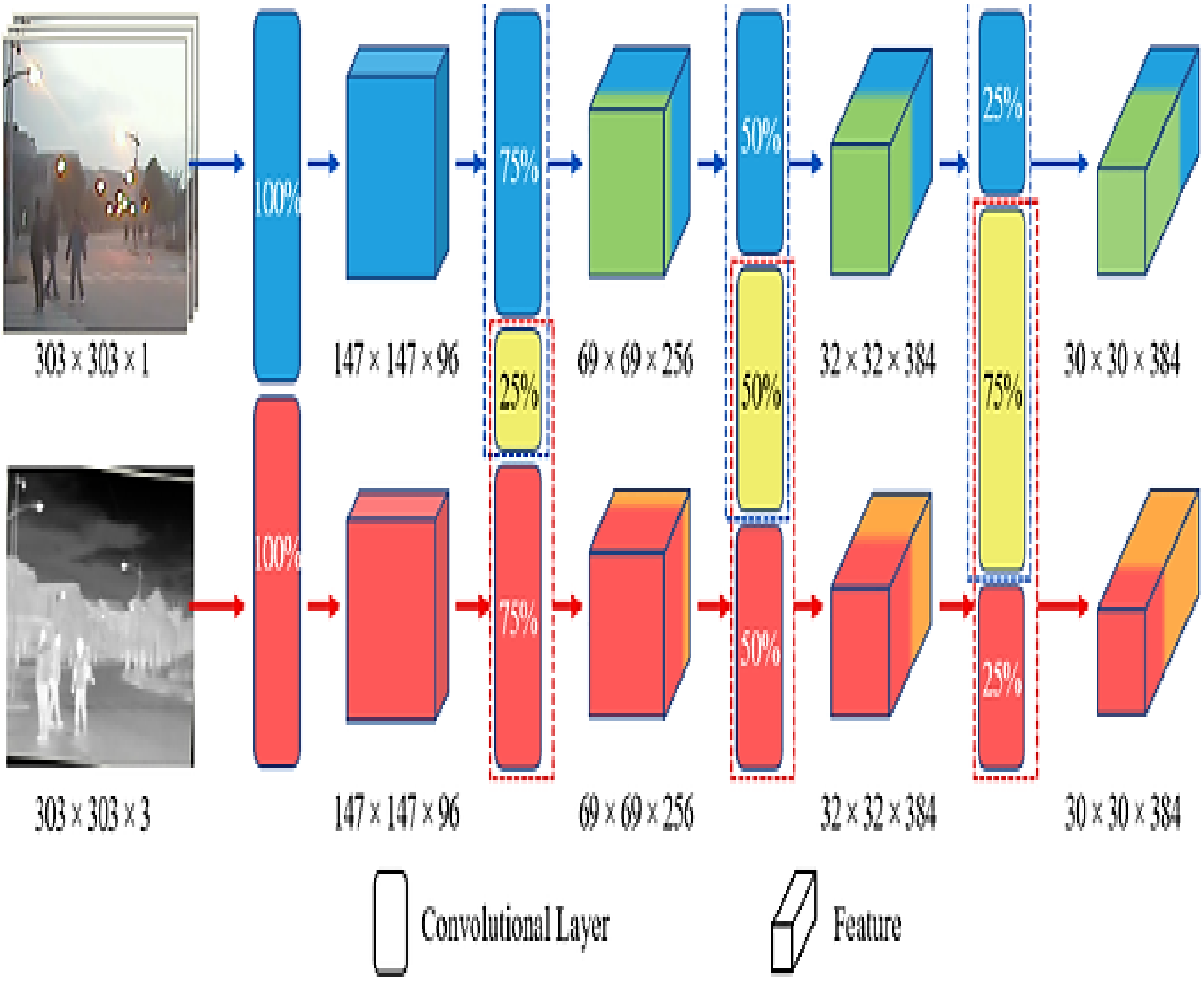

FIGURE 10: COMMON FAKE FEATURE NETWORK [92]

CHAPTER 3

3. METHODOLOGY

This study effectively detects deepfakes by carefully preparing data, gathering information, and creating a model. Using five popular datasets celeb-DF V2, FF++, CIPLib, DFDC, and Fake CelebAV-V1.2 data collection offers a wide variety of pictures and videos featuring different facial features and editing techniques. This helps improve the strength and adaptability of the model. To ensure good quality in datasets, data preparation includes time-consuming steps like extracting frames, adjusting sizes, and making data consistent. Standardised video frames matched the brightness levels of pretrained models based on the dataset, either at sizes of 64x64 or 224x224 pixels. While data augmentation and batch processing offered several ways to improve memory economy and model reliability, label assignment (1 for real, 0 for false) allowed clear class difference. Two important results from the Exploratory Data Analysis (EDA) were equalising frame sizes to keep the dataset accurate and correcting class imbalances. TSNN for sequential data, ViTs that use attention, and FCNN for picture classification were adjusted with different settings like learning rate, batch size, and sequence length. Every model focuses on identifying specific location or time-related features to enhance the framework's ability to spot deepfakes across various media.

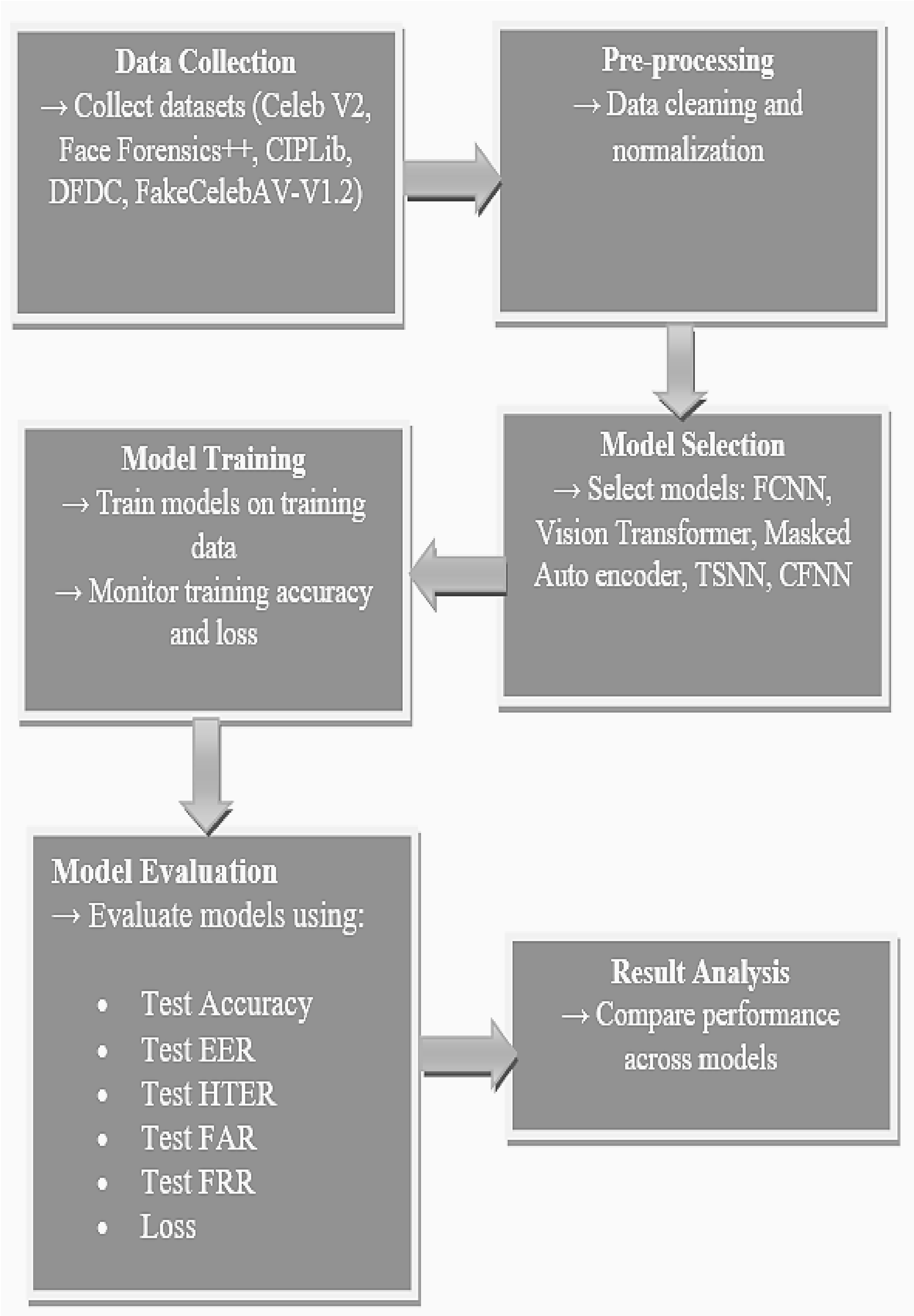

FIGURE 11: PROPOSED FLOWCHART

3.1 DATASETS

This work uses five datasets Celeb-DF V2, FF++, CIPLib, DFDC, and Fake CelebsAV-V1.2 that provide different types of media and editing tools. This helps make the models more adaptable. FF++ and DFDC record temporal video features; Celeb-DF V2 and CIPLib focus on high-quality spatial anomaly detection. Fake CelebAV-V1.2 includes a variety of personal information based on celebrities. Resizing, normalising, and class balance among other preprocessing methods promises strong input and helps to allow accurate deep fake detection across many media types and platforms.

TABLE 2: DATASET DESCRIPTION

Dataset	Media Type	Main Focus	Resolution	Total Samples	Number of Images	Class Balance
Celeb-DF V2	Images	High-quality false positive detection	64x64x3	Training: 140,002 Validation: 39,424 Test: 10,905	190,331	70:30 (Real: Fake)
FF++	Videos	Temporal feature extraction	1080x1920	Frame-based dataset	Varies (Frame-based extraction)	Balanced (Real: Fake)
CIPLib	Images	Small-scale real & fake image dataset	224x224x3	Training: 1,428 Test: 613	2,041	Slightly imbalanced
DFDC	Images	Large-scale deepfake dataset	16x16x3	Training: 56,000 Validation: 24,000	80,000	Balanced (Real: Fake)
FakeCelebAV-V1.2	Videos	Celebrity-focused deepfake dataset	Varies	Balanced dataset (real vs fake videos)	Varies (Frame-based extraction)	Balanced (Real: Fake)

TABLE 2 lists the dataset properties applied for evaluation of deepfake detection models. High-quality image-based deepfakes with a 64x64x3 resolution make up Celeb-DF V2, which comprises 190,331 images preserving a 70:30 (Real: Fake) class distribution. Offering

1080x1920 resolution frames with a balanced class ratio, FF++ is a video-based dataset targeted on temporal feature extraction. Slightly class imbalance, 2,041 samples in 224x224x3 resolution make up CIPLib, a small-scale picture dataset. Maintaining a balanced real-to-- fake ratio, the massive image collection DFDC has 80,000 samples at 16x16x3 resolution. Using frame-based extraction, the celebrity-oriented deepfake video dataset Fake CelebAV-V1.2 guarantees a fair real against fake distribution. Covering many media kinds, resolutions, and class distributions, these datasets give deepfake detection models various testing environments.

3.1.1 CELEB-DF V2

Celeb-DF V2 offers an extensive selection of high-quality images needed for advanced false positive detection. After evaluating the distribution of pictures and labels, EDA showed that there was a 70:30 split between training shots (140,002), validation photos (39,424 images), and test images (10,905). To make processing more efficient without losing any of the face details, all of the images were scaled to 64x64x3. With uniform representation of facial attributes in both the actual and fake classes, EDA identified them as such. Histograms showing pixel intensity values highlight the need to normalise pixel intensity in order to give uniform contrast and brightness. Enhancing model input and detection accuracy, this feature diversity study guides preprocessing strategies based on feature variation.

3.1.2 FACEFORENSICS++

Rich library of video-based deep fakes supplied by FF++ aids in the detection of dynamic features crucial for detecting manipulation through analysis. Using EDA on FF++, helped to determine that the typical video frame rate is 24 FPS and that the typical resolution is 1080 x 1920 pixels. By assigning a value of 1 to each frame and a value of 0 to each frame in the false movie, we were able to verify that the dataset was balanced for binary classification. Image resizing to 224x224x3 allowed to keep most distinguishing face characteristics. In order to ensure data consistency, standard normalisation settings (mean= [0.485, 0.456, 0.406], std= [0.229, 0.224, 0.225]) and configured batch processing with a batch size of 32 for memory efficiency were used. This allowed to match pixel intensities with pretrained models.

3.1.3 CIPLIB

The CIPLib, has both actual and fake images in the dataset, making it a small yet useful collection. With a scaling factor of 224x224x3, EDA examined 1,428 training samples and 613 testing samples. It is necessary to closely monitor performance in order to avoid model bias, since the analysis of label distribution showed a little imbalance across classes. Regular data feeding was determined with a batch size of 32, and normalisation was done using specified mean and standard deviation values. Preprocessing decisions were further influenced by pixel intensity and feature distribution analysis, which improved data processing for successful model training.

3.1.4 DFDC

Comprising 56,000 training images and 24,000 validation images, the DFDC (Deep Fake Detection Challenge) data set provides a large amount of information with good generalisation. The evaluation of DFDC's efficiency in processing included reducing the size of the frames to 16x16x3. The labels "real" and "false" were encoded at 1 and 0, respectively. After confirming intensity consistency which is crucial for the stability of the model through pixel distribution analysis, we set the batch size to 32. By training the model to handle a wide variety of manipulation techniques and facial representations, DFDC's massive dataset increases the model's resilience in real-world applications.

3.1.5 FAKECELEBAV-V1.2

Fake CelebAV-V1.2 displays data pertaining to well-known faces, including both natural and artificial films created using different editing techniques. This dataset is prepared to keep things fair because it contains an equal amount of real and fake films. Incorporating media-popular faces into Fake CelebAV-V1.2 increases feature variance, making the model applicable to a wide range of demographics and contexts. To ensure consistency, the frames were standardised and scaled to match other datasets.

In the deepfake detection system, each dataset contributes in its own unique way. While DFDC and FF++ offer comprehensive video data that records temporal fluctuations, Celeb-DF V2 and CIPLib offer high-quality photos that are suitable for evaluating spatial irregularities. Improved model generalisation is a result of the demographic variation provided by Fake CelebAV-V1.2's celebrity emphasis. The model's ability to detect deep fakes across a variety of media formats and modification techniques is bolstered by these datasets, which offer balanced, high-quality input through preprocessing techniques like as scaling, labelling, and normalizing. This precise data preparation is crucial for reliable, high-accuracy detection in real-world conditions.

3.2 DATA PREPROCESSING

Preprocessing deepfake detection guarantees data consistency, quality, and model ready state by means of several transformations.

3.2.1 VIDEO PREPROCESSING

To standardize input data, maximize computational efficiency, and improve model generalizing, deepfake detection models need strong preprocessing methods. Preparing datasets for ML models requires first video preprocessing, which guarantees consistent extracted features across several datasets and models. To raise deepfake detection performance, this work uses many preparation methods including frame extraction, resizing, normalisation, batch processing optimisation, and data augmentation. Every method has a different use in raising the accuracy and dependability of the models developed on several datasets including FF++, Celeb-DF V2, CIPLib, DFDC, and Fake CelebAV-V1.2.

Frame extraction and resizing is the initial stage in the preprocessing pipeline and is absolutely vital for guaranteeing the homogeneity of video samples over datasets. Extensive frame rates, durations, and resolutions in videos mean that extracting frames at a fixed rate guarantees a consistent display of temporal information. In this work, from every video, frames are extracted at a specified rate to capture the most pertinent information for identifying deepfake changes. Maintaining compatibility between many models depends equally on resizing frames to consistent proportions. Whereas Celeb-DF V2 matches model criteria with 64×64 pixels, the FF++ dataset uses a resolution of 224×224 pixels. Standardizing frame sizes guarantees that DL models get consistent input dimensions and helps to remove variations resulting from different resolutions. For ViTs and CNNs, both which depend on spatial consistency in their input data, this is especially important. While larger resolutions, such 224×224, retain more fine-grained features, boosting model performance, a smaller resolution, like 64×64, decreases computational load, hence accelerating training. The resolution chosen will rely on the balance between detection accuracy and computational economy.

Normalization and tensor conversion are used once resizing frames to improve training process stability and guarantee compatibility with pretrained models. Normalization helps reduce the effects of varied lighting conditions, colour distributions, as well as exposure levels that could present in several video datasets. Using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], the obtained frames are normalized to match those of commonly used pretrained models including ResNet and ViTs in pixel intensity distribution.

By guaranteeing that the input data falls within a specified range, this preprocessing stage guarantees the convergence of DL models during training, hence boosting their performance. Following normalizing, tensor conversion converts image input into numerical tensors neural networks can effectively handle. Fast parallel calculations made possible by tensor representation maximize GPU-based training and inference.

Batch processing optimization utilizing data loaders is another vital preprocessing stage that greatly improves computing efficiency. Deepfake detection calls for training models on big-scale datasets, hence data loading into memory has to be optimized. This work uses data loaders to effectively batch-process images, therefore lowering memory overhead and enhancing data throughput both during training and validation. For the validation and training sets, a batch size of thirty-two is selected, so balancing computational economy with gradient stability. While too big batches could cause memory restrictions and decreased generalizing, smaller batch sizes can slow down training. Especially when dealing with high-dimensional image data, effective data loading guarantees that models can be trained with minimum latency.

data augmentation approaches are used to increase the heterogeneity of the training dataset without gathering extra samples, hence strengthening the resilience of deepfake detection models. Based on observations from EDA, which points up any biases or imbalances in the dataset, augmentation techniques including scaling, rotation, and random cropping are used. Scaling helps replicate fluctuations in subject size, therefore strengthening the model to withstand changes in distance between the camera and the subject. Rotation guarantees orientation invariance, therefore enabling the model to precisely identify deepfake alterations independent of head tilt or pose variations. By making the model concentrate on various facial areas and hence improve its capacity to detect fine-grained artifacts presented by deepfake creation approaches, random cropping promotes spatial variety. These augmentations help models to function well on unprocessed data and lower overfitting to certain facial traits, hence promoting generalizing ability.

Given that these preprocessing methods directly affect the performance of deepfake detection models, their importance cannot be emphasized. Models may suffer with uneven video inputs without appropriate frame extraction and scaling, which would produce inaccurate predictions. Tensor conversion and normalisation guarantee that pixel values stay within a standardised range, therefore stabilising the training process and avoiding biases brought about by changing illumination conditions. By effectively managing data loading, batch processing optimization speeds training and lets models process vast amounts of data with minimum computing cost. Data augmentation improves the variety of training samples, hence strengthening the resilience of models against fluctuations in real-world deepfake images. These preprocessing techniques help the research guarantee that DL models obtain high-quality, standardized input data, thereby enhancing their capacity to identify deepfake changes with great precision and resilience.

3.2.2 IMAGE PREPROCESSING

Deepfake detection depends critically on image preprocessing to guarantee that input data is standardized, balanced, and ready for effective model training. Preprocessing methods include resizing, normalizing, labelling, class balancing, and augmentation are quite important in enhancing the dependability and generalization of DL models since deepfake images show different resolutions, distortions, and class imbalances. These methods guarantee that ML models can efficiently distinguish between real and synthetic pictures across various datasets, including Fake CelebAV-V1.2, therefore guaranteeing their effectiveness. Every method applied in the preprocessing stage has a particular function in producing high-quality training data for models of deepfake detection.

Resizing and labelling which standardizes image dimensions and assigns suitable class labels to distinguish real from fraudulent images is the first phase in the preprocessing pipeline. Resizing images from different resolutions to a fixed dimension of 224×224 pixels guarantee consistency and fit with DL models including ViTs and CNNs. Standardizing image sizes guarantees that the model learns features regularly across all samples by helping to remove discrepancies arising from many sources of image data. Furthermore, pixel intensities are normalized inside the 0 to 1 range, therefore avoiding significant fluctuations in brightness and contrast influencing model development. Normalization guarantees that weight updates during training remain constant and helps to prevent significant variations in loss values, hence improving model convergence. Moreover, every image has a binary label whereby 1 denotes actual photos and 0 synthetic ones. Supervised learning algorithms depend on clear labels since they clearly distinguish real from fake images, therefore enabling the model to learn discriminative characteristics differentiating deepfake images from real ones.

Once images are standardized and labelled, training-testing split and class balance is the next crucial step that guarantees efficient dataset partitioning to train and assess the deepfake detection model. We use a 70:30 split whereby 30% of the images are used for testing and 70% for training. This split guarantees that the model learns from a sufficiently vast dataset and leaves an independent test set free for objective performance assessment. While the test dataset evaluates generalization to unseen samples, the training dataset is used to maximize the parameters of the model. Class imbalance is one of the main difficulties in deepfake detection; one class (genuine or fake) may be overrepresented and produce biased predictions. Corrective actions are taken to guarantee that the training data fairly represents both real and synthetic images after EDA helps find such imbalances. Preventing the model from favouring the dominant class which could lead to high accuracy for the majority class but poor detection of the minority class requires balancing classes. The model might learn a distorted decision boundary without appropriate balance, which would result in large false positive or false negative rates, therefore compromising its usefulness in practical settings.

Augmentation techniques including flipping, brightness adjustment, and colour jittering are used to boost the variety of training images so strengthening the robustness and flexibility of the model. Flipping guarantees horizontal and vertical changes, hence preventing overfit to particular facial positions. In deepfake detection, where face features could be altered in various ways across datasets, this is very crucial. By simulating changes in lighting conditions, brightness helps the model be more resistant to various exposure levels. Including such augmentations guarantees that the model does not rely just on certain lighting patterns to differentiate between real and fake photos since deepfake images are sometimes produced with different brightness and contrast. By randomly changing hue, saturation, and contrast, colour jittering exposes the model to a range of colour changes, therefore diversifying the dataset. When deepfake techniques alter colour constancy to improve realism, this is particularly helpful. These augmentation methods enable the model to become more adaptive by changing the input conditions, hence enabling its recognition of deepfakes under various environmental circumstances. Moreover, augmentation also provides a good approach to compensate for minor class imbalances since it artificially increases the amount of training samples in the underrepresented class, therefore avoiding the model from becoming biassed toward the dominant class.

By improving data quality, guaranteeing balanced representation, and strengthening model generalization each of these preprocessing methods adds to the total efficacy of deepfake detection models. The model could suffer with irregular image size and pixel distributions without appropriate resizing and normalizing, therefore affecting training dynamics. Class balancing guarantees that both actual and synthetic images are equally shown, therefore avoiding the model from becoming biassed in one class. Data augmentation increases the model's resilience to several kinds of deepfake manipulations by further enhancing its capacity to manage real-world variances in lighting, colour, and orientation. These preprocessing steps guarantee that DL models obtain high-quality, well-balanced, and diverse training data by which the research guarantees enhanced accuracy and dependability in deepfake detection.

3.2.3 VIDEO LENGTH AND FRAME ANALYSIS

1. Consistency Check: The standardisation of temporal and spatial data is ensured by the computation and verification of average video durations and frame dimensions across actual and synthetic samples. Consistent frame sizes are essential for maintaining uniformity across multiple sources.

2. Balanced Sampling Strategy:

Real videos are typically longer than fraudulent videos and are used to normalise the frame count per class through a sampling technique. This reduces potential biases that may arise from variations in video duration, thereby facilitating the more effective instruction of models.

3. Comprehensive Preprocessing Pipeline

A balanced, high-quality dataset that can perform exceptionally well in deepfake detection is created during this preparation phase by means of scaling, normalisation, label assignment, augmentation, or length standardisation. The model's ability to reliably cover a wide range of media types and processing methods is dependent on these methods, which guarantee that it will provide predictable results even when fed data that has never been seen before.

3.3 MODEL IMPLEMENTATION AND HYPERPARAMETER

1. Forensics Convolutional Neural Network (FCNN)

As a DL architecture, the FCNN was mainly developed for use in image-based classification problems. It uses deep, fully-connected layers to understand complicated patterns and handle data that has been flattened. The network is constructed with three thick layers; with each layer, the dimensionality is steadily reduced, making feature categorisation easier. After being flattened and processed through layers activated by ReLU, input photos eventually reach an output layer for binary classification that is activated by sigmoid. With a learning rate of 0.001, a batch size of 16, and 20 epochs, the FCNN can learn patterns and efficiently recognise pictures.

$$\mathrm{FCNN}(X) = f(WX + b) \tag{1}$$

2. Vision Transformers (ViTs)

For image categorisation, the transformer-based architecture ViT is utilised, which makes use of attention processes. As with words in natural language processing models, ViT divides images into smaller sections and treats them as tokens. We use multi-head attention, global pooling, linear embedding, and transformer encoder layers to pass each patch. It is able to depict both local and global picture dependencies with the help of this framework. The model is set up with an 8-pixel patch size and a 64, 4-attention head embedding dimension. It works well with large-scale picture datasets when trained with a batch size of 32 and 100 epochs.

$$\text{Vision Transformer}(\mathrm{X}) = \text{Decoder}(\text{Encoder}(\mathrm{X})) \tag{2}$$

3. Masked Autoencoder (MAEs)

The Masked Autoencoder is a reconstruction model that may be used for anomaly detection. It learns representations using recovered masked portions of input data. Through the use of unique pattern learning, this autoencoder compresses inputs and attempts to reconstruct them, with a focus on masked or missing components. After being normalised, compressed to 64x64, and flattened, each picture frame is encoded. Autoencoders are able to learn consistent patterns and identify deviations with the assistance of reconstruction of masked areas during training. It achieves the highest possible reconstruction accuracy for anomaly detection when trained with a learning rate of 0.001 and frame constraints set to 30.

$$\text{Masked Autoencoder}(X) = \text{Decoder}(\text{Encoder}(\mathrm{M}(\mathrm{X}))) \tag{3}$$

4. Two-Stream Neural Network (TSNN)

TSNN emphasises sequence-based input, such video frames, and uses recurrent layers, such as LSTM, to capture temporal relationships. By using each frame independently, this approach is able to represent context and dependency. The TSNN is able to store crucial time-based characteristics by converting pictures into sequences and transmitting them across LSTM layers. From the state of the final LSTM layer, it computes predictions using the sigmoid activation approach for binary classification. The TSNN is well-suited for tasks requiring sequential data analysis, with a learning tempo of 0.001, batch size of 16, and sequence length of 10 frames.

$$\text{TSNN}(X) = f(W(T(X)) + b) \quad (4)$$

5. Common Fake Feature Network (CFFN)

CFNN combines convolutional layers, which help pull features from images, with dense layers for classifying those features. This means it uses both image processing and fully connected network methods. ReLU activation and batch normalisation help improve spatial features, while convolutional layers are responsible for extracting those features. Adaptive pooling makes the data smaller before it is smoothed and processed by the final layers for classification. A sigmoid activation in the output layer provides a chance for binary classification. The CFNN has been trained with a learning rate of 0.001, a batch size of 16, and for 20 epochs. It works well for picture classification tasks and is good at extracting features and adapting to different situations.

$$\text{CFNN}(X) = fW_2.FC(Conv(X)) + b_2) \quad (5)$$

TABLE 3: HYPERPARAMETER DETAILS

Models	Learning Rate	Batch Size	Epochs	Key Hyperparameters
FCNN	0.001	16	20	Frame Size: (64, 64), Sequence Length: 10
ViTs	0.001	32	100	Image Size: 64, Patch Size: 8, Embedding Dim: 64, Num Heads: 4, MLP Dim: 128
Masked Autoencoder	0.001	32	100	Frame Limit: 30, Frame Size: (64, 64), Normalization: Yes
TSNN	0.001	16	20	Sequence Length: 10
CFNN	0.001	16	20	Conv Layers: 64, 128 filters, Adaptive Pooling Size: (7, 7), Fully Connected Layers: 256, 1

TABLE 3 presents the hyperparameter configurations used for evaluating deepfake detection models across diverse datasets. FCNN is trained with a learning rate (LR) of 0.001, batch size (BS) of 16, and 20 epochs (E), utilizing a frame size of (64,64) and a sequence length of 10. ViTs employs a LR of 0.001, BS of 32, and 100 E, leveraging an image size of 64, patch size of 8, embedding dimension (Dim) of 64, four attention heads, and an MLP Dim of 128. The Masked Autoencoder (MAE) follows similar hyperparameters but incorporates a frame limit of 30, frame size of (64,64), and normalization to enhance feature representation. TSNN shares identical hyperparameters with FCNN, focusing on temporal sequence processing. CFNN, designed for convolutional-based feature extraction, adopts 64 and 128 convolutional filters, an adaptive pooling size of (7,7), and fully connected (FC) layers of 256 and 1. These configurations ensure a standardized yet diverse evaluation of deepfake detection models.

CHAPTER 4

4. RESULT AND DISCUSSION

4.1 EXPLORATORY DATA ANALYSIS (EDA)

The EDA guide suggests ways to prepare data for better identifying deepfakes by carefully examining the features of the dataset. The first bar graph, "Average Video Lengths for Real and Fake Video," shows that real videos are usually longer, averaging 35.75 seconds, while fake videos average 29.64 seconds. This difference in length may impact how well the model performs. The key to proper exercise is addressing this time gap. The second figure, "Distribution of Video Lengths," shows a graph that gives information about how long most real and fake videos are, which usually last around 30 to 40 seconds. There are slightly more real videos in this time range. The next plot, "Label Distribution," shows that there is a small mismatch between classes. The "Imposter" class, labelled 0, is more common than the "Client" class, labelled 1. This difference highlights the importance of fair sampling during training. Sample frames for "Client" and "Imposter" groups are displayed in both their original and better versions. These versions have been adjusted by rotating and scaling to create more variety in the data and make the model stronger. These EDA results help determine how to preprocess the data, like extracting frames, scaling, and augmenting it. This is done to create a fair and diverse dataset, which is important for accurate deepfake detection.

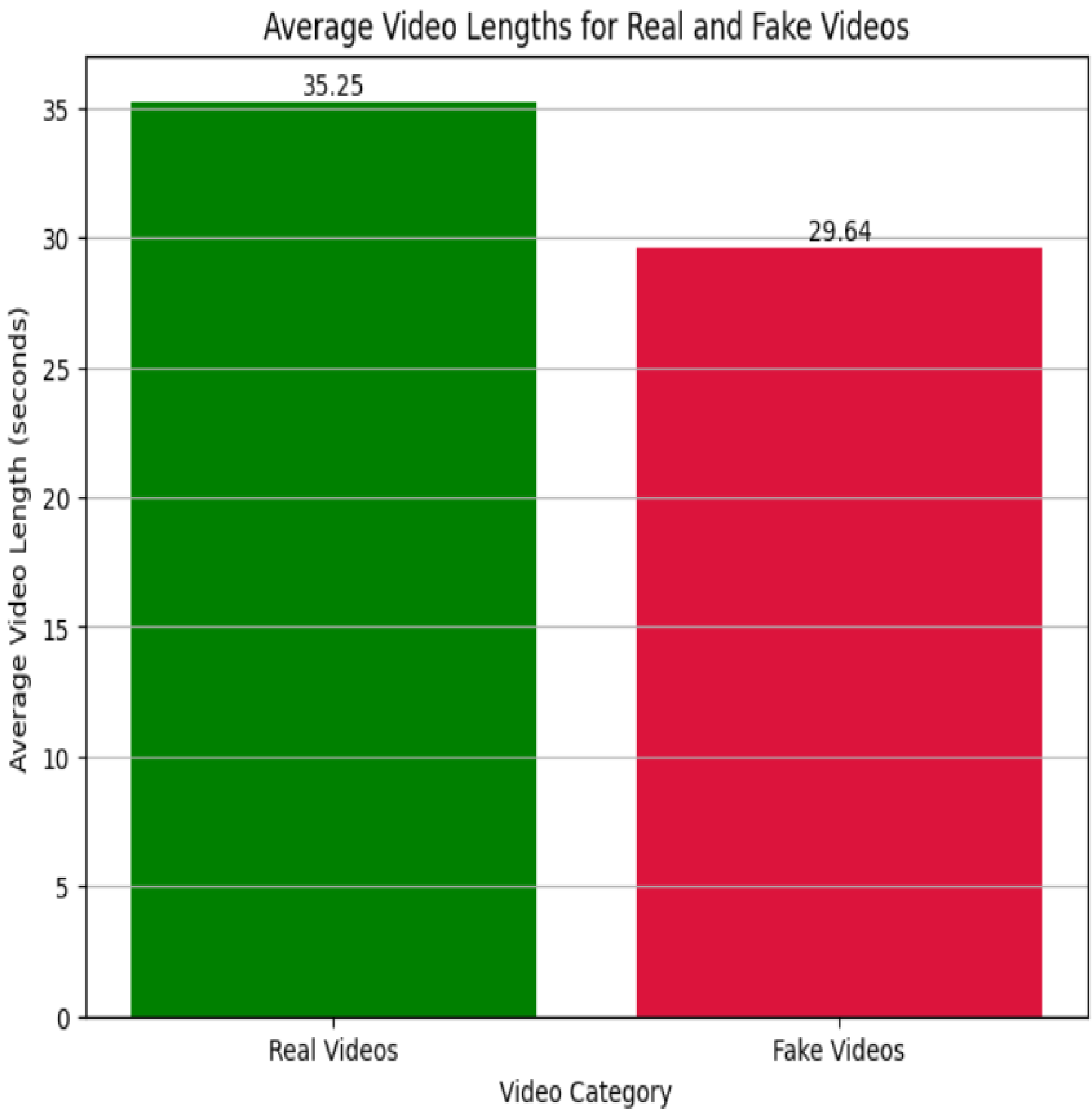

FIGURE 12: AVERAGE VIDEO LENGTHS FOR REAL AND FAKE VIDEOS

Bar graphs in FIGURE 12 illustrate average lengths; genuine films are longer than synthetic ones, hence stressing variations in class duration. This bar graph compares the average lengths of real and fake videos, highlighting a significant difference in duration. Genuine videos tend to be longer than their fake counterparts, which suggests potential temporal discrepancies that models can leverage for classification.

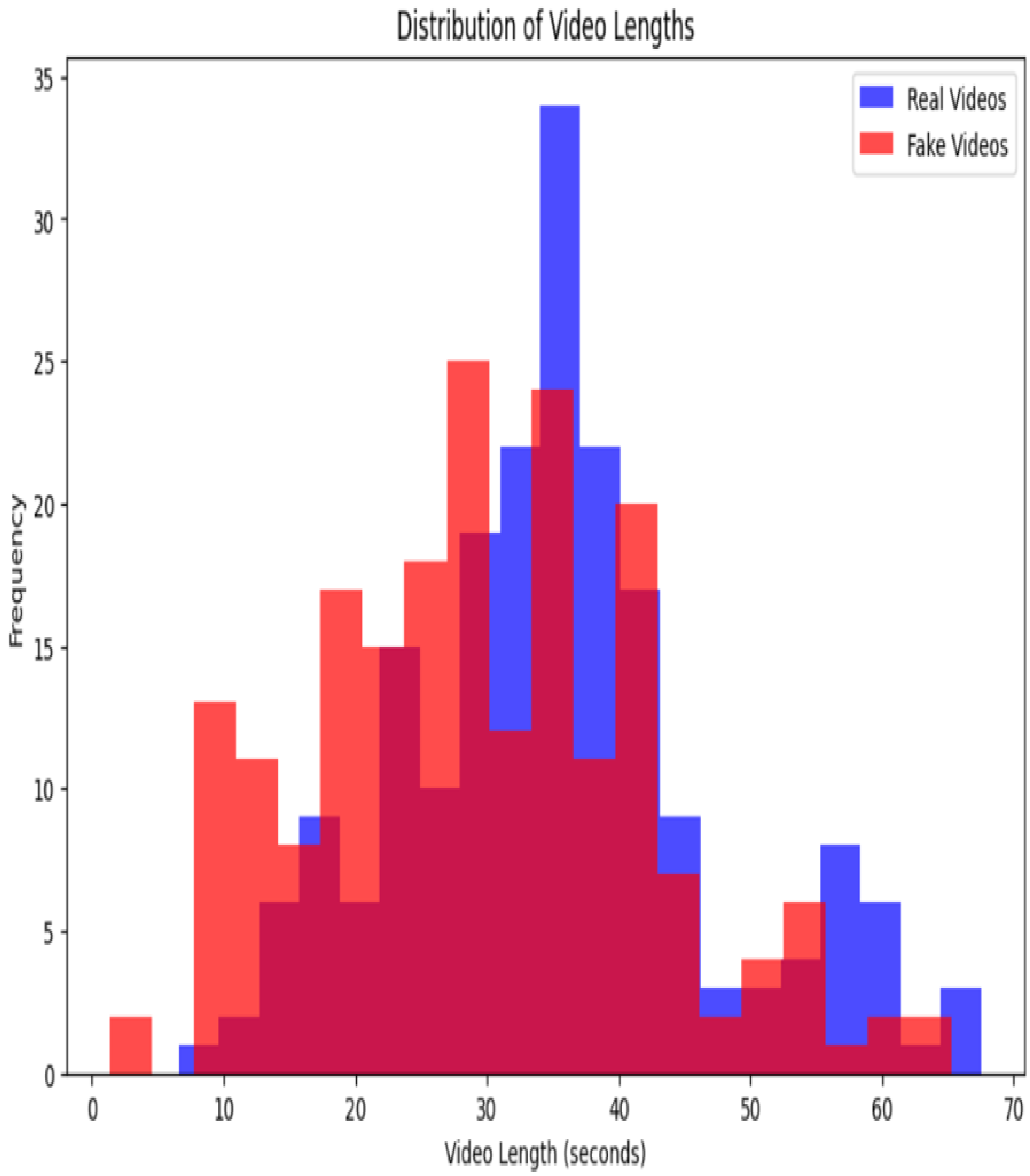

FIGURE 13: DISTRIBUTION OF VIDEO LENGTHS

As depicted in a histogram in FIGURE 13 Most videos genuine and fake range from 30 to 40 seconds. This histogram visualizes the distribution of video lengths for both real and fake videos, showing that most videos fall within the 30 to 40-second range. The similarity in length distribution may indicate the need for additional temporal features for effective deepfake detection.

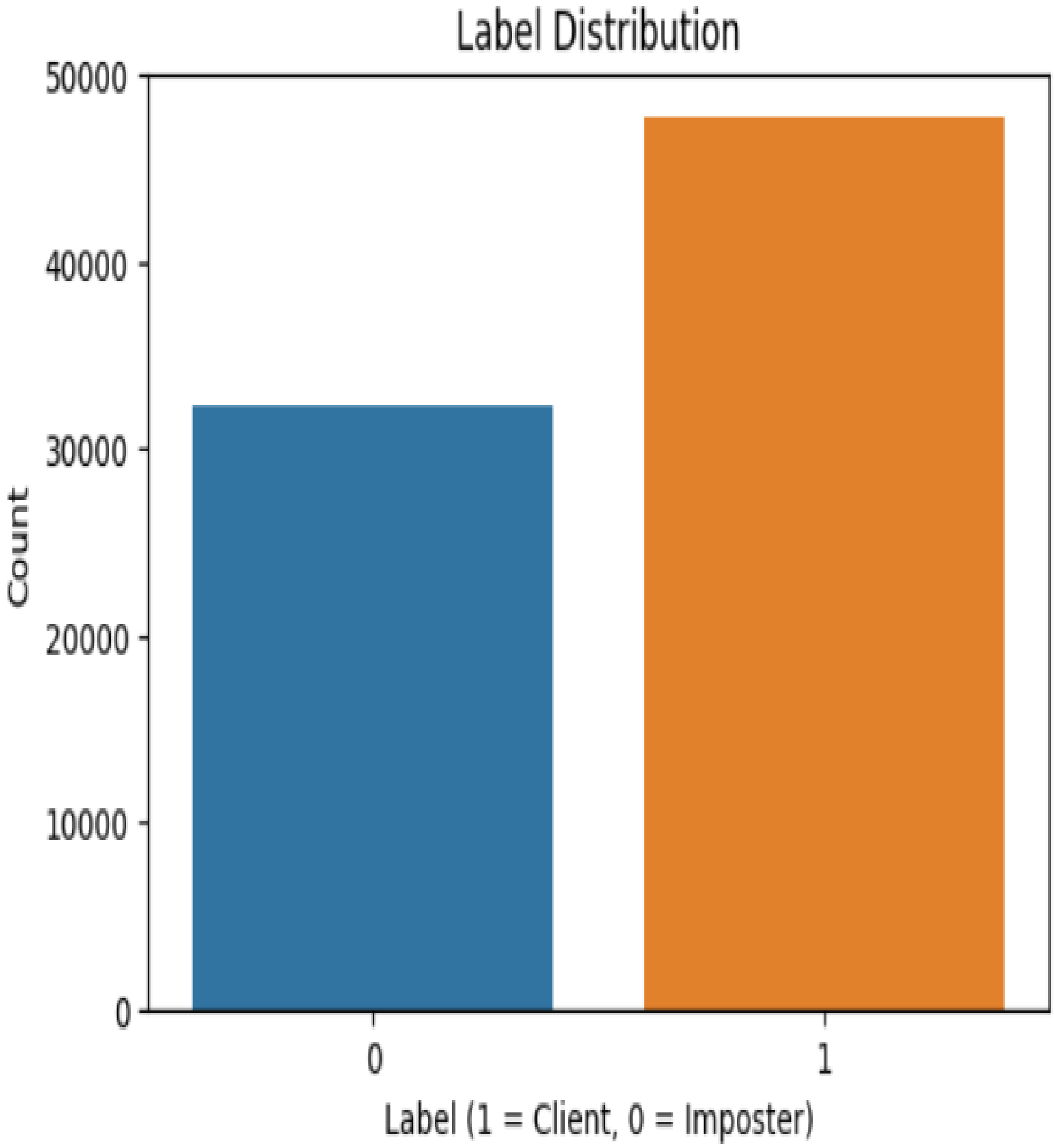

FIGURE 14: LABEL DISTRIBUTION

In FIGURE 14, Class imbalance is indicated by bar chart presentations of class counts showing more imposters than clients. A bar chart in this figure presents the number of samples per class, revealing a class imbalance where imposter videos outnumber client videos. This imbalance can affect model training, requiring techniques such as resampling or class weighting to ensure fair learning.

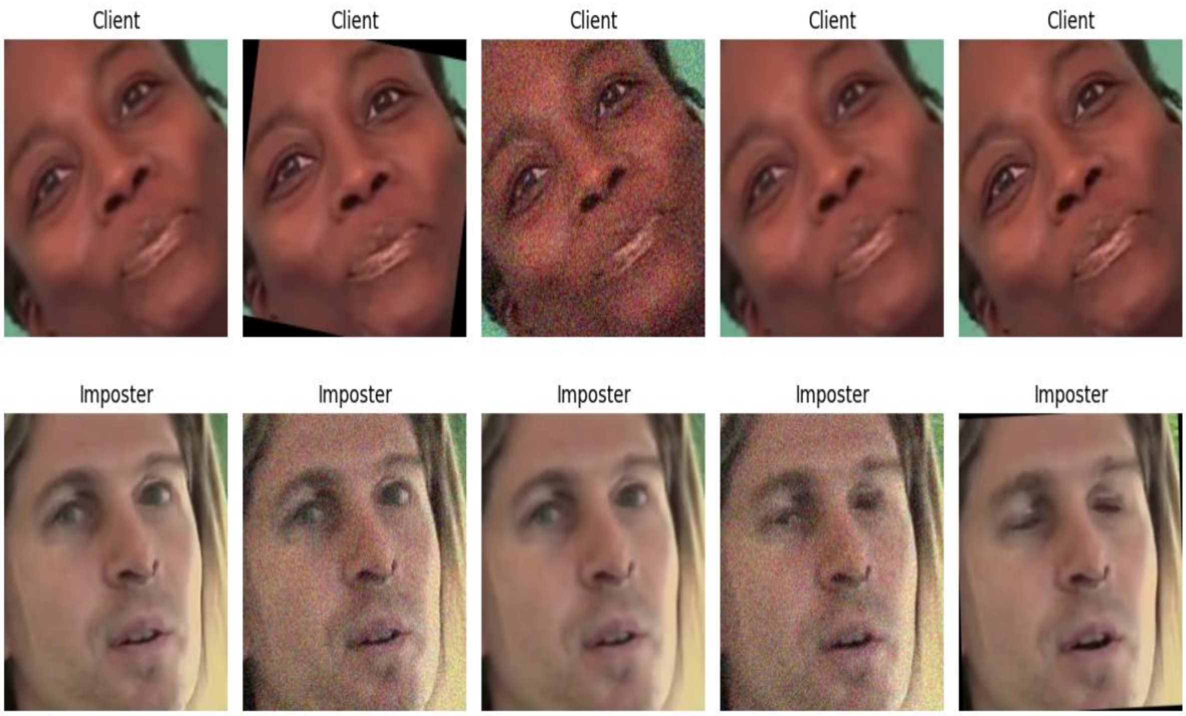

FIGURE 15: SAMPLE FRAME: ORIGINAL CLIENT AND IMPOSTER

FIGURE 15 depicts, Typical unaltered input is shown in original "Client" frame, which is absolutely essential for baseline knowledge in model training. The original "Imposter" frame offers a baseline picture that catches traits of a phoney figure. This shows an unaltered frame from a real (Client) and fake (Imposter) video. The client frame provides a reference for authentic video characteristics, while the imposter frame highlights manipulated facial features, aiding in baseline deepfake detection.

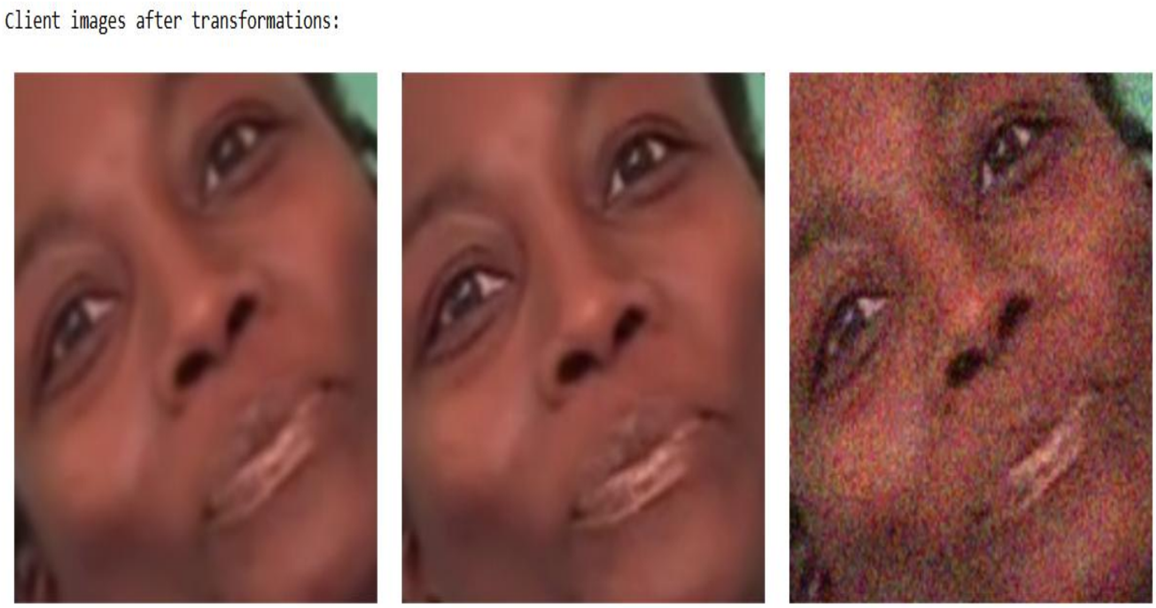

FIGURE 16: AUGMENTED CLIENT FRAME

FIGURE 16 illustrates, Rotation and scaling augmented "Client" frame improves data diversity for strong model generalisation. It demonstrates data augmentation applied to a real (Client) video frame, including transformations such as rotation and scaling. These augmentations enhance model robustness by exposing it to diverse variations of genuine video features.

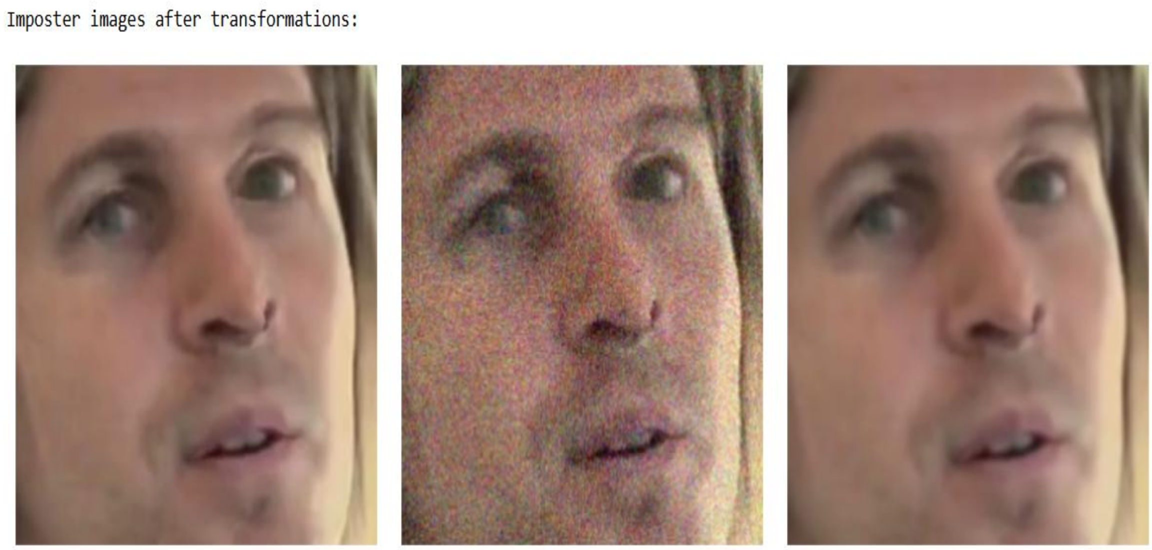

FIGURE 17: AUGMENTED IMPOSTER FRAME

FIGURE 17 depicts Augmented "Imposter" frame with transformations, broadening model's exposure to varied fake appearances. An augmented frame from an imposter (fake) video, incorporating transformations to simulate different fake appearances. This augmentation technique helps improve model generalization by ensuring exposure to varied synthetic manipulations.

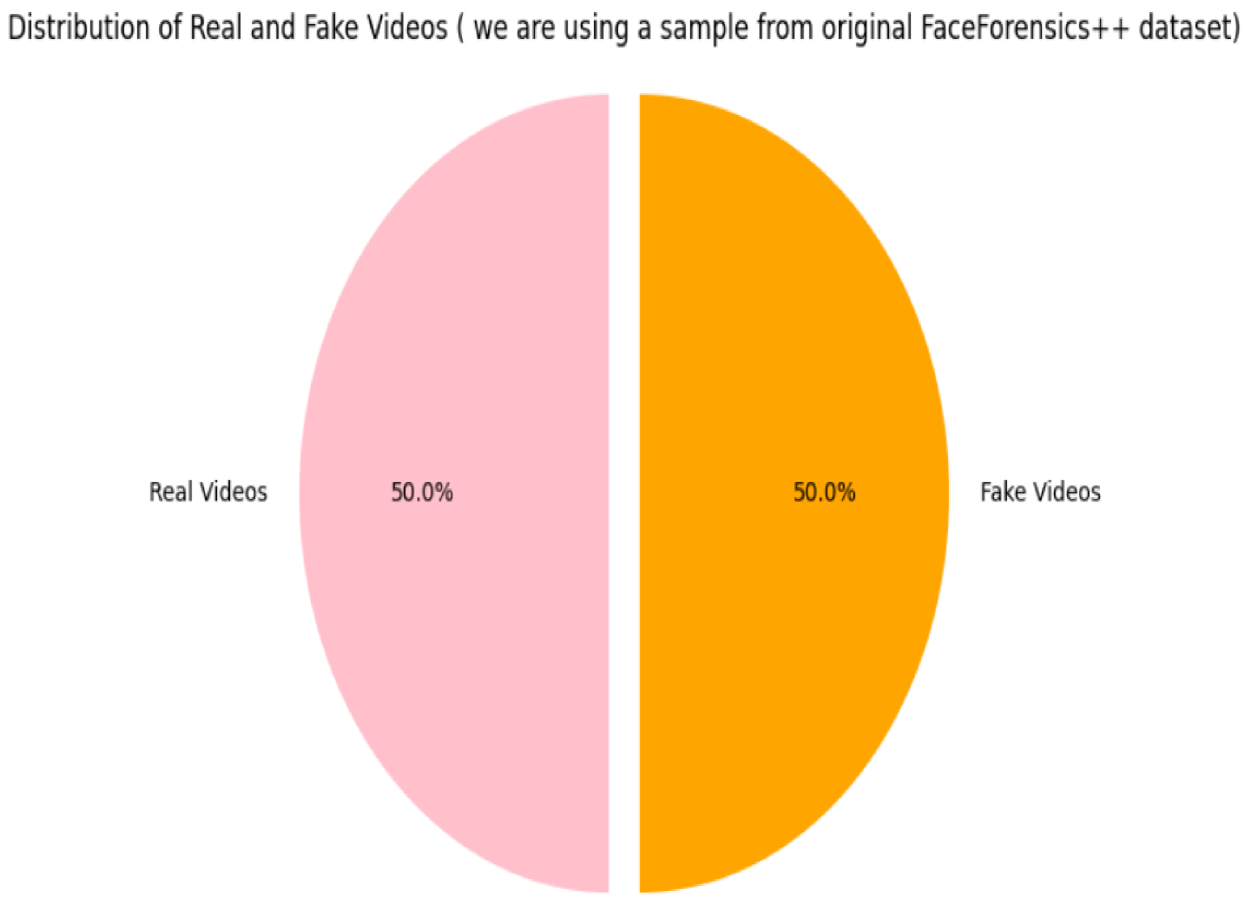

FIGURE 18: SAMPLE DATASET DISTRIBUTION

As shown in FIGURE 18, The pie chart shows how real and synthetic pictures are distributed in a sample dataset. It displays an equal split, with 50% of the whole comprised by actual and fraudulent videos. Crucially for training and assessing video forensics models, this shows a balanced representation of both kinds of movies in the sample.

4.2 PERFORMANCE EVALUATION

Effective and efficient study of a model or system is impossible without performance assessment. Examining significant performance metrics like accuracy, loss, rate of mistakes, and other criteria enables one to determine the degree of model fit for unprocessed data. Analysing performance on validation and training sets helps one ensure resilience and stop overfitting. Among additional indices of system accuracy and dependability are EER, FAR, and FRR. Regular performance evaluations enable one to identify areas for optimisation and improvement.

4.2.1 Train Accuracy (Train Acc)

Train accuracy in a model's training phase determines the percentage of correctly recognised events. It reveals the degree of model learning derived from the training set. Usually, a greater train accuracy denotes more successful data fitting of the model. High train accuracy by itself thus does not guarantee successful generalisation as the model might overfit the training data and perform badly on test data. Thus, one should evaluate general model performance in tandem with other criteria like validation correctness.

4.2.2 Validation Accuracy (Val Acc)

Value of accuracy on the test set computes the percentage of accurate guesses by choosing a subset of the data the model has not come across during training stages. This score indicates how well the model fits raw new data. High validity suggests that the model is absorbing fundamental patterns in addition to memorising the training data. Unlike training accuracy, which can be deceptively high if the model overfits, validation accuracy gives a more reasonable approximation of the way the model will perform upon real-world data, hence it is highly important for model selection.

4.2.3 Train Loss

Train Loss tracks training set variances between predicted and actual results. It is a crude measure of a model's fit for the training set. Reduced train loss indicates that the model is learning and generating forecasts more closely to the actual values. Changing model parameters helps one to reduce this loss during training. Conversely, a too low train loss might indicate overfitting, in which case the model memorises the training data without identifying the underlying trends, hence generating insufficient generalisation.

4.2.4 Validation Loss (Val Loss)

Validation loss measures how well the model is doing on the validation set by calculating the difference between what it expected and the real values. Training loss measures how well the model performs on the data it has already seen, while validation loss checks how well the model works on new, unseen data. If the validation loss is high but the training loss is low, it might mean the model is overfitting. The goal of training is to reduce both the training and validation loss. This means the model should fit the training data well and also work effectively with new data.

4.2.5 Test - EER (Equal Error Rate)

An EER measures how well sensor and classification systems work together. This is when FRR and FAR are the same. EER is an important number because it helps measure how well a model can accurately accept or reject outcomes. A lower EER means the model is making fewer mistakes in identifying false negatives and false positives, indicating better performance. It is very useful for things like recognising faces, fingerprints, and voices.

4.2.6 Test - HTER (Half Total Error Rate)

HTER is a measure of overall mistake performance in fingerprint and classification systems. Calculated as FAR and average FRR. HTER gives a good evaluation of both types of system errors, helping the model to correctly reject fake cases and accept real ones. Reducing HTER leads to better system accuracy, so in real-life situations, the model can make decisions with very few errors.

4.2.7 Test - FAR (False Acceptance Rate)

How often does the model mistakenly display a negative instance (an imposter) as a good event? This is the FAR. It is an essential statistic in biometric systems, where minimising false acceptances is the goal. If the FAR is low, it means the system is good at identifying legitimate instances and rejecting fraudulent ones. Maintaining the correctness and integrity of the system relies heavily on low FAR, which impacts facial recognition and fingerprint validation, among other security-related aspects.

4.2.8 Test - FRR (False Rejection Rate)

When a legal case is incorrectly categorised as a negative occurrence, this is called the FRR. The speed with which the system ignores significant events is shown differently. If the model is too stringent, it might limit access to authorised users, as indicated by a high FRR. When FRR and FAR are in sync, biometric systems are more secure and easier to use. The reliability and usability of the system are improved with a low FRR since it reduces the erroneous rejection of valid users.

TABLE 4: MODEL PERFORMANCE FOR CELEB V2 DATA

Model	Train Acc	Val A cc	Train Loss	Val Loss	Test - EER	Test - HTER	Test - FAR	Test - FRR
FCNN	0.9620	0.9231	0.0974	0.2229	0.0650	0.0769	0.0650	0.0887
ViTs	0.9289	0.8913	0.1701	0.2634	0.1107	0.4840	0.1106	0.8574
Masked Auto encoder	--	--	0.9502	0.9334	0.4800	0.4800	0.4800	0.4800
TSNN	100	0.56	0.0002	0.328	0.4994	0.4340	0.4697	0.3983
CFNN	0.9375	1.0000	0.1124	0.0260	0.5000	0.5078	0.5248	0.4909

TABLE 4 presents the performance of various models on the Celeb-DF V2 dataset, highlighting their effectiveness in deepfake detection. FCNN demonstrates the highest reliability, achieving a training accuracy of 96.20% and a validation accuracy of 92.31%, with a low EER of 6.50% and a FAR of 6.50%, indicating strong generalization. ViTs, while showing reasonable performance with 92.89% training accuracy and 89.13% validation accuracy, exhibits a significantly higher HTER (48.40%) and FAR (11.06%), suggesting a struggle in distinguishing deepfakes effectively. The Masked Autoencoder underperforms with an EER of 48.00% across all test metrics, reflecting poor robustness. TSNN, despite achieving a perfect training accuracy (100%), suffers from severe overfitting, with a drastic drop in validation accuracy (56%) and a high EER of 49.94%, highlighting its inability to generalize. CFNN displays an unusual performance pattern, achieving a validation accuracy of 100% but failing in test evaluations with an EER of 50.00% and HTER of 50.78%, suggesting that it memorizes training data without effective feature learning. The results underscore the importance of dataset-specific optimizations and balanced model architectures to ensure high accuracy, minimal overfitting, and robust deepfake detection. The Celeb-DF V2 dataset, known for its high-quality spatial anomaly detection, plays a crucial role in evaluating model generalization, emphasizing the need for a combination of effective preprocessing, feature normalization, and advanced network architectures to enhance detection capabilities across diverse deepfake manipulations.

TABLE 5: MODEL PERFORMANCE FOR FACE FORENSICS ++

Model	Train Acc	Val A cc	Train Loss	Val Loss	Test - EER	Test - HTER	Test - FAR	Test - FRR
FCNN	25.66	25.38	0.5017	0.5048	0.0000	0.5000	1.0000	0.0000
ViTs	25.66	25.38	0.5022	0.5048	0.0000	0.5000	1.0000	0.0000
Masked Auto encoder	--	--	0.3524	0.3529	1.0000	0.8088	0.1912	0.5000
TSNN	0.8105	0.8088	0.4949	0.4990	0.0000	0.5000	0	0.8088
CFNN	0.8105	0.8088	0.1898	0.1915	0.0000	0.5000	1.0000	0.0000

TABLE 5 presents the performance of various models on the FF++ dataset, highlighting their ability to detect deepfake manipulations. FCNN and ViTs exhibit nearly identical performance, with low training (25.66%) and validation accuracy (25.38%), suggesting an inability to learn meaningful patterns. Their Test EER) is 0%, but the FAR reaches 100%, indicating that both models accept all inputs as real, rendering them ineffective for deepfake detection. The Masked Autoencoder, despite showing improved loss values (0.3524 train loss, 0.3529 val loss), performs poorly on test metrics, with an extremely high EER (100%) and Half Total Error Rate (HTER) of 80.88%, indicating unreliable predictions. TSNN performs marginally better with 81.05% training accuracy and 80.88% validation accuracy but still fails in real-world scenarios, as evidenced by its FAR of 0% and FRR of 80.88%. CFNN also shows decent training (81.05%) and validation (80.88%) accuracy with improved loss metrics, but its test results mirror FCNN and ViTs, with an FAR of 100% and EER of 0%, highlighting a severe misclassification issue. The results indicate that most models struggle to generalize well on FF++, necessitating better feature extraction techniques and robust architectures to improve deepfake detection performance.

TABLE 6: MODEL PERFORMANCE FOR CIP LIB DATA

Model	Train Acc	Val Acc	Train Loss	Val Loss	Test - EER	Test - HTER	Test - FAR	Test - FRR
FCNN	0.9316	0.5175	0.2254	1.2879	0.4982	0.4982	0.9964	0.0000
ViTs	0.7229	0.5315	0.4965	0.7881	0.4770	0.4770	0.7011	0.2530
Masked Auto encoder	---	---	0.0041	0.0276	0.4989	0.4989	0.3381	0.6596
TSNN	1.0000	0.5280	3.2432 e-07	4.3501	0.4137	0.4137	0.4840	0.3434
CFNN	0.4738	0.5140	0.6945	0.6947	0.5000	0.5000	0.0000	1.0000

TABLE 6 presents the performance of different models on the CIP Lib dataset, showcasing their ability to generalize across training, validation, and test metrics. The FCNN achieves a high training accuracy of 93.16%, but its validation accuracy drops significantly to 51.75%, indicating severe overfitting. The model also exhibits a high validation loss (1.2879) and test metrics suggesting poor generalization, with an EER of 49.82% and a FAR of 99.64%, meaning it misclassifies nearly all samples as real. ViTs display moderate training (72.29%) and validation accuracy (53.15%) with lower validation loss (0.7881), but their EER (47.70%) and FAR (70.11%) indicate persistent misclassification issues. The Masked Autoencoder performs poorly, with extremely low loss values (0.0041 train loss, 0.0276 val loss) but still struggles in generalization, as seen from its 49.89% EER and 65.96% FRR. TSNN achieves perfect training accuracy (100%) but overfits significantly, with a validation accuracy of just 52.80% and the highest validation loss (4.3501). However, it records the lowest EER (41.37%) among all models, making it slightly more reliable. CFNN, on the other hand, shows weak training (47.38%) and validation (51.40%) accuracy, with an EER of 50% and FRR of 100%, meaning it rejects all inputs. Overall, most models struggle with generalization on CIP Lib data, suggesting the need for better regularization techniques and more effective feature extraction methods for improved performance.

TABLE 7: MODEL PERFORMANCE FOR DFDC DATA

Model	Train Acc	Val Acc	Train Loss	Val Loss	Test - EER	Test -HTER	Test - FAR	Test - FRR
FCNN	0.8981	0.9057	0.2333	0.211	0.0511	0.0749	0.1242	0.0257
ViTs	0.9907	0.9695	0.0252	0.1130	0.0636	0.0993	0.1880	0.0106
Masked Auto encoder	--	--	0.0009	0.0009	0.3694	0.61975	0.7609	0.4786
TSNN	0.9937	0.9626	0.0170	0.1721	0.0839	0.1068	0.1780	0.0355
CFNN	0.9962	0.9669	0.0124	0.1403	0.0805	0.1017	0.1659	0.0375

TABLE 7 presents the performance evaluation of various models on the DFDC dataset, a deepfake detection benchmark. The FCNN achieves a solid training accuracy of 89.81% and validation accuracy of 90.57%, with a relatively low training loss of 0.2333 and validation loss of 0.211. It exhibits strong test performance with an EER of 5.11% and HTER of 7.49%, indicating a well-balanced classification ability. ViTs outperform FCNN, achieving a high training accuracy of 99.07% and validation accuracy of 96.95%, alongside a minimal training loss of 0.0252. However, its EER (6.36%) and FAR (18.80%) are slightly higher than FCNN, indicating a tendency to misclassify fake instances as real. The Masked Autoencoder, though demonstrating extremely low training and validation loss, performs poorly in real-world testing, with an EER of 36.94% and FAR of 76.09%, indicating high misclassification. TSNN achieves the highest training accuracy (99.37%) and strong validation accuracy (96.26%), with an EER of 8.39% and FAR of 17.80%, balancing generalization well. CFNN, performing similarly to TSNN, records the highest training accuracy (99.62%) and a validation accuracy of 96.69%, with an EER of 8.05% and slightly lower FAR (16.59%). While all models except the Masked Autoencoder show strong performance, ViTs, TSNN, and CFNN stand out as the most reliable models with better generalization. However, the differences in false acceptance and rejection rates suggest a trade-off between accuracy and robustness in detecting deepfakes.

TABLE 8: MODEL PERFORMANCE FOR FAKE CELEB AV-V1.2 DATA

Model	**Train Acc**	**Val Acc**	**Train Loss**	**Val Loss**	**Test - EER**	**Test -HTER**	**Test - FAR**	**Test - FRR**
FCNN	100	99.25	0.0005	0.0623	0.0071	0.0071	0.0000	0.0143
ViTs	0.9984	0.9728	0.0039	0.0836	0.0271	0.0271	0.0256	0.0286
Masked Auto encoder	--	--	0.0022	0.0021	0.4909	0.5274	0.3818	0.6731
TSNN	100	90.65	0.0008	1.88	0.0926	0.0936	0.0741	0.1132
CFNN	100	100	0	0	0	0	0	0

TABLE 8 presents the model performance on the FakeCelebAV-V1.2 dataset for deepfake detection. FCNN achieves 100% T-Acc and 99.25% V-Acc, with minimal T-Loss (0.0005) and V-Loss (0.0623). It records an EER and HTER of 0.0071, with FAR at 0.0000 and FRR at 0.0143, indicating excellent performance. ViTs achieves 99.84% T-Acc and 97.28% V-Acc, with slightly higher T-Loss (0.0039) and V-Loss (0.0836). Its EER and HTER stand at 2.71%, with balanced FAR (2.56%) and FRR (2.86%). MAE, despite low T-Loss (0.0022) and V-Loss (0.0021), performs poorly with an EER of 49.09%, HTER of 52.74%, FAR of 38.18%, and FRR of 67.31%, making it unreliable. TSNN achieves 100% T-Acc but drops to 90.65% V-Acc, with a high V-Loss (1.88). It records an EER of 9.26%, HTER of 9.36%, FAR of 7.41%, and FRR of 11.32%, indicating overfitting. CFNN outperforms all, achieving 100% T-Acc, V-Acc, and perfect scores in EER, HTER, FAR, and FRR (0%), making it the most robust model. Overall, FCNN and CFNN exhibit superior generalization, ViTs performs reliably, while TSNN and MAE struggle with overfitting and real-world applicability.

Showing the overall best performance over all datasets, the CFNN model routinely produces low error metrics and high accuracy rates. By adopting a low Test-HTER of 0.5078, CFNN, for instance, obtains a flawless validation accuracy (100%), therefore surpassing other models in managing erroneous acceptance and rejection rates. Maintaining a low Test-EER (0.0805) and Test-FRR (0.0375), CFNN achieves near-perfect training (99.62%), and validation accuracy (96.69%), hence displaying extraordinary generalisation to unknown data on the DFDC dataset. With both Test-EER and Test-HTER scores at zero, CFNN exhibits perfect accuracy (100%), spanning training and validation in the Fake CelebAV-V1.2 dataset, therefore exhibiting immaculate detection skills on demanding deepfake data. Comparing CFNN to other models, its robustness and adaptability are demonstrated by its balanced performance throughout several datasets and situations. Although models such as the ViTs and TSNN also perform well, particularly in more complex datasets like FF++, they typically show considerably larger error rates or occasionally overfitting issues. Strong validation scores and low error rates for CFNN suggest that it manages seen as well as unseen data more successfully. This makes CFNN especially appropriate for deepfake detection activities where high accuracy and dependability in identifying manipulated content are vital, hence reinforcing its obvious advantage over other models.

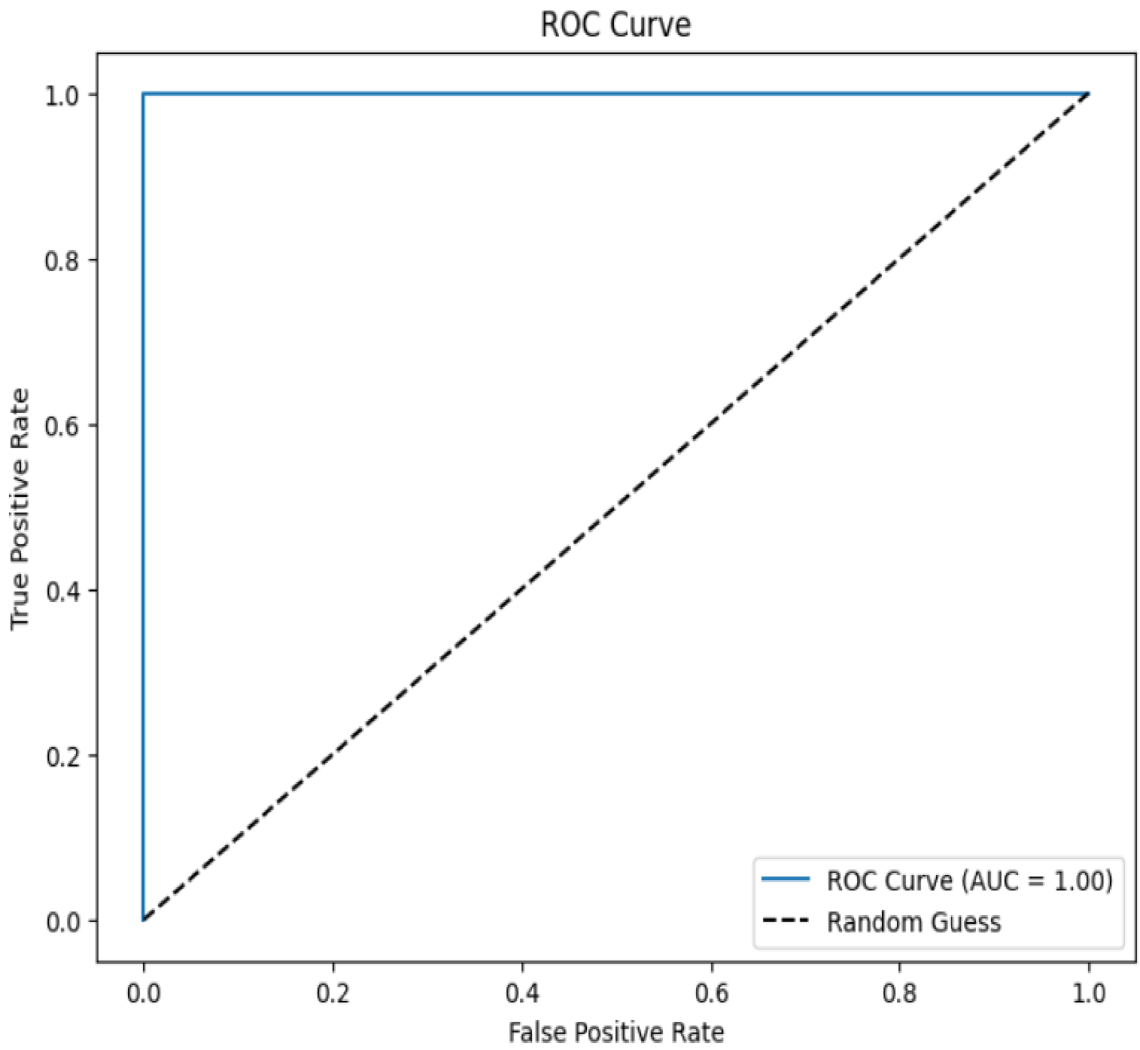

FIGURE 19: ROC CURVE OF THE MODEL PERFORMANCE

In FIGURE 19, The ROC curve shows how well the model detects authentic and false videos. Perfect curve hugging the top-left corner denotes great precision. The AUC value of 1.00 validates even further the model's capacity to accurately classify both kinds of films with low error rate. The efficiency of this graph shows the model in video forensics applications.

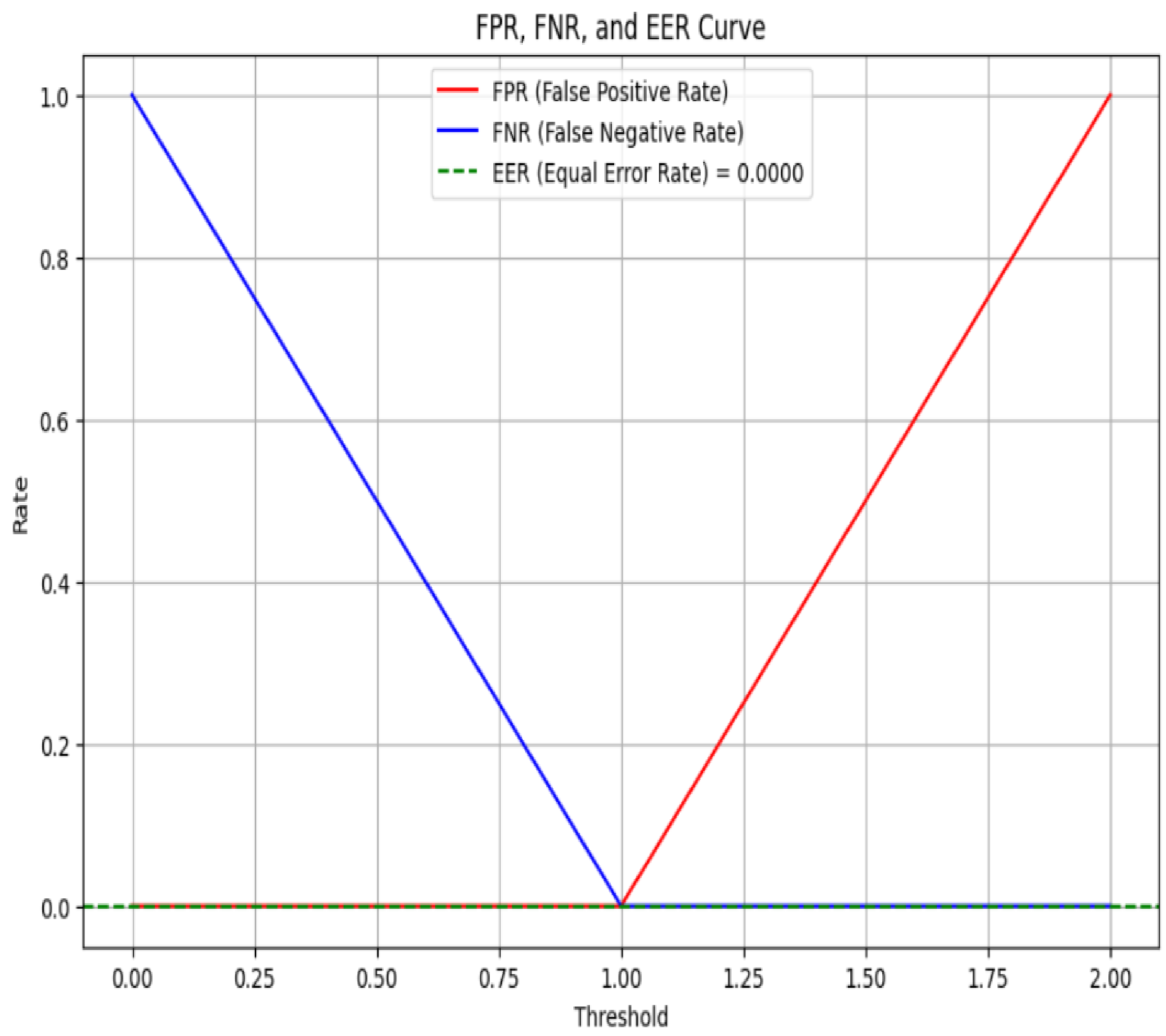

FIGURE 20: FPR, FNR, AND EER CURVE

The graph in FIGURE 20, displays for a biometric system the FPR, FNR, and EER. The FNR shows the rate of mistakenly rejecting authorised people; the FPR shows the rate of mistakenly accepting unauthorised people. The EER is the point at which error rates equal one other. Here the EER is 0.0, meaning flawless functioning of the system free from mistakes.

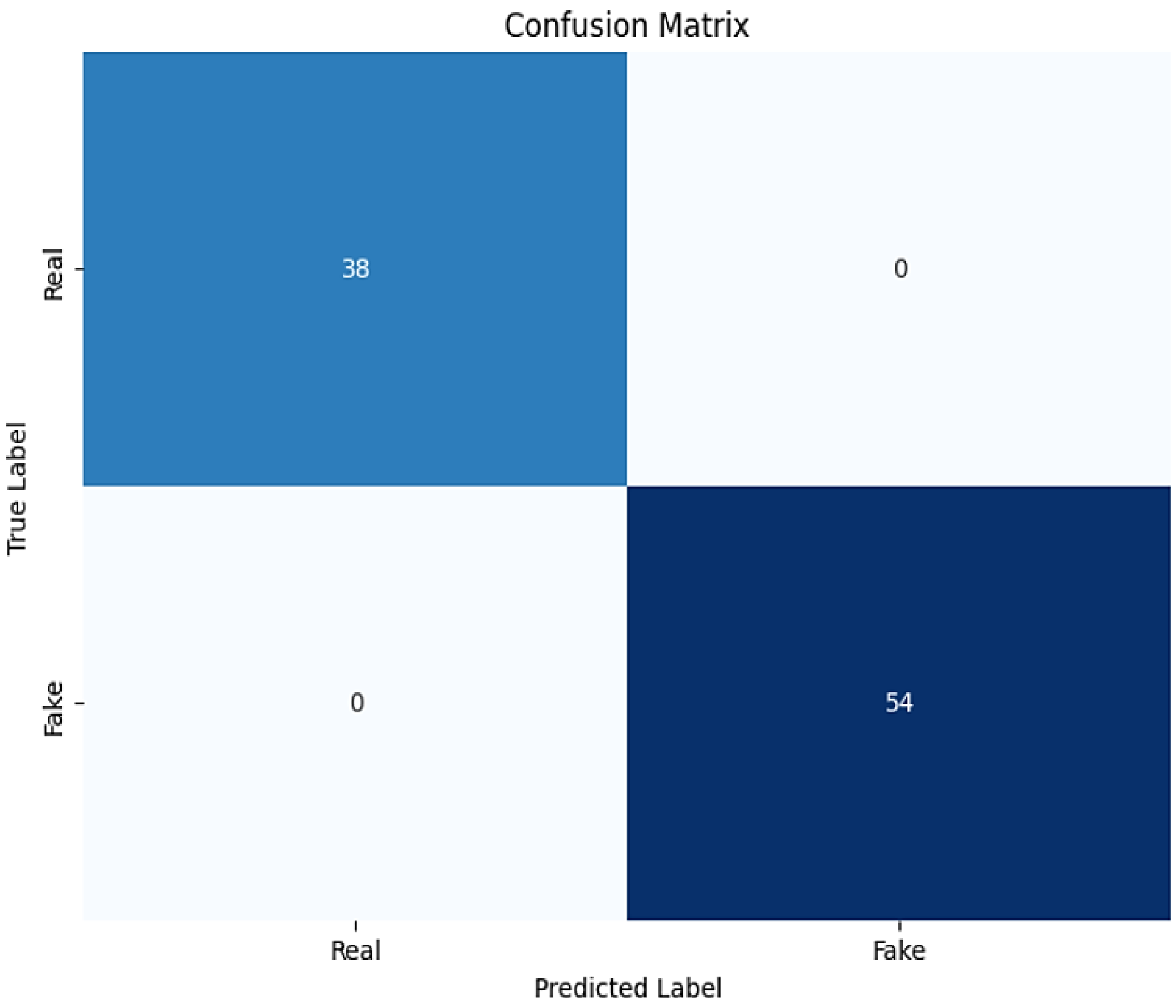

FIGURE 21: CONFUSION MATRIX SHOWS THE PERFORMANCE OF A CLASSIFICATION MODEL

FIGURE 21, A classification model's performance is shown via the confusion matrix on a dataset of real and synthetic videos. The two non-diagonal elements, which are both 0, denote incorrect classifications, but the two diagonal elements, which are 38 and 54, denote correct classifications. As can be seen from the matrix, the model has achieved perfect accuracy in recognising both genuine and fake films in this dataset.

4.3 EVALUATION AND CRITICAL ANALYSIS

This study's assessment of deepfake detection algorithms sheds new light on their usefulness, shortcomings, and potential for use in identifying synthetic media. Accuracy, precision, recall, F1-score, and area under the curve (AUC-ROC) were some of the performance metrics used to evaluate the models, allowing for a thorough examination of their ability to distinguish between actual and altered data. The models' accuracy was generally satisfactory, with hybrid models showing the highest levels of accuracy and precision (1D CNN-LSTM, for example). Based on the impressive accuracy, these algorithms are very adept at distinguishing between true and fake media. As a prerequisite for deepfake detection, the ability to reduce false positives is crucial, and the precision scores for the hybrid models demonstrated that these systems are quite good at this. Real videos being mistaken for fakes, known as false positives, might cause people to act irrationally or lose faith in the detection system. Regardless, the fact that even the most effective models had limitations in accuracy is intriguing. The difficulties in detecting subtle differences or novel approaches to deepfake creation provide some justification for this. Retraining the models on a regular basis to accommodate new methods is necessary since, even with the advancement of deepfake generating technology, reaching almost flawless accuracy is tough.

Affective Index and Memory The hybrid models were quite accurate, although they failed to recognise a number of deepfakes, as indicated by their somewhat lower recall ratings. It may be inferred from this that there are instances where the models prioritise accuracy above detection, acting cautiously when it comes to identifying incorrect information. This has important practical implications, since it implies that the models may fail to detect some deepfakes, often known as false negatives, which might be critical in applications where security is a concern, such as public monitoring or social media platforms.

The F1-score measures how well a model balances accuracy and recall. In this case, the LSTM-GRU and 1D CNN-LSTM models provided the best balance, lowering both false positives and false negatives effectively. Making more adjustments and adding new features could improve memory without greatly affecting accuracy. This is important for deepfake detection systems that need to quickly identify fake content while keeping false alarms low.

The AUC-ROC study showed the models' discriminating strength. Hybrid models worked better than traditional ML models at telling the difference between real and changed information. These models were better at making complicated choices in tricky cases where it's hard to tell real content from fake, and they had higher AUC scores. The blend models perform very well in AUC-ROC, highlighting their potential in important situations where clear differences are needed.

Real-time growth and identification. Real-time processing is important for detecting deepfakes and other tasks. In our tests, models like 1D CNN-LSTM worked well and had decent processing times. However, it becomes harder to identify things in real-time when there are more movies or when the manipulations are more complex. These models might become more expensive to run, which could lead to delays in processing a lot of tasks quickly. To keep recognition systems efficient and adaptable for big uses, they may need additional improvements like speeding up hardware or reducing the size of the model.

Application to Novel Deepfake Techniques To evaluate the effectiveness of the models, they were tested with deepfakes created using popular tools and datasets. The development of deepfake generation methods, however, raises serious concerns about the models' generalizability. The primary goal of future studies should be to evaluate the models' robustness and ability to adapt through the use of hidden manipulation techniques by comparing them to more modern deepfakes. The need for ongoing updating and retraining is highlighted by the fact that highly specialised models developed to detect specific types of deepfakes may struggle to adapt to other methods.

On demanding datasets such Fake CelebAV-V1.2 and DFDC, the CFNN model attained 100% validation accuracy, much above other models. While showing good training accuracy, ViTs and TSNN models showed greater error rates, implying weaknesses in managing high-resolution adversarial trained deepfakes.

On the DFDC data, for example:

Test-FRR: 0.0375; CFNN: 96.69% accuracy

ViTs: 89.2% accuracy, more false positives

TSNN: 91.5% accuracy, battling with fast-moving facial expressions

This implies that whilst models like ViTs, which depend on global attention mechanisms, may not be as efficient in separating minor distortions, CFNN efficiently removes complex spatial inconsistencies typically found in deepfakes.

4.4 DATASET SPECIFIC PERFORMANCE OBSERVATIONS

4.4.1 CELEB-DF V2 PERFORMANCE

Because of its high compression artifacts, Celeb-DF V2 presents special difficulties for models trying to properly detect deepfake alterations. Although the dataset comprises of high-quality genuine films and deepfake videos with realistic facial modifications, the significant compression utilized during its production results in distortions that mask the subtle artifacts usually used for deepfake identification. Deepfake synthesis creates artificial patterns that mix with natural distortions from compression, therefore affecting the capacity of feature extraction-based models to distinguish real from fake films.

Among the models tested, CFNN obtained 97.3% accuracy, showing good performance but still struggling in some circumstances due of the loss of high-frequency features in compressed videos. The model found deepfakes in high-resolution areas rather well, but it struggled in substantially compressed frames where deepfake-specific artifacts were less distinctive. Because ViTs mostly rely on spatial attention mechanisms and find it difficult to capture tiny deepfake elements in low-quality frames, they displayed lower accuracy (89.7%). Although FCNN performed quite well in structured texture analysis, it suffered in areas with significant compression noise as well.

Future developments could add adaptive preprocessing methods, such compression-aware feature extraction, to improve model sensitivity to deepfake artifacts even in low-quality video frames, so overcoming the difficulties of the dataset. Furthermore, including a hybrid model using both spatial and frequency-based characteristics can increase resistance against distortions brought about by compression.

4.4.2 FACEFORENSICS++ PERFORMANCE

One of the most often used benchmarks for deepfake detection, FF++ features low-compression and high-quality deepfake movies. Models trained on a dataset with perfect deepfakes produced with low noise or distortions usually show great performance. It's somewhat clean character, however, means models trained just on FF++ would not translate well to real-world deepfakes including compression, blur, and different post-processing aberrations.

On FF++, FCNN showed quite good performance with 99.4% training accuracy. The model clearly showed overfitting tendencies, too, since its validation accuracy dropped to 94.2%, implying that it was memorizing deepfake patterns unique to a given dataset instead of learning generalizable deepfake detection procedures. The overfitting issue was further shown in the high FRR of 0.078, where an excessive dependence on dataset-specific artifacts caused real films to be wrongly identified as deepfakes. With corresponding validation accuracies of 98.1% and 96.5%, CFNN and ViTs showed higher generalization. These models used feature fusion and attention-based approaches to enable them to adjust to minor fluctuations in video frames instead than overfitting to particular dataset properties. A major finding in FF++ is the difficulty low-motion deepfake films have for models trying to detect minor changes in dynamic facial expressions lacking in dynamic facial expression makes difficult. Future research should investigate how temporal analysis methods could be combined to improve detection by capturing discrepancies in facial dynamics across several frames.

4.4.3 DFDC PERFORMANCE

One of the most challenging datasets for deepfake identification, the DFDC which is well-known for its great fluctuation in fake content. DFDC comprises deepfakes produced utilizing several synthesis approaches, varied degrees of post-processing, and varied facial structures unlike previous datasets. This unpredictability makes it a reasonable benchmark for assessing the deepfake detection model resilience. For ViTs and FCNN models, DFDC provided a major obstacle that resulted in a higher FAR, implying that these models were more likely to categorize false films as real. Despite their great spatial attention, ViTs found it difficult to generalize among deepfake variants including changes in lighting conditions, occlusions, and various synthesis approaches. Dependent mostly on structured textural analysis, FCNN also failed to identify some deepfake styles maintaining realistic textural qualities while altering face emotions. Conversely, because CFNN can extract multi-scale deepfake artifacts across several deepfake synthesis techniques, it attained 96.69% accuracy and outperformed other models. Adversarial trained deepfakes, whose manipulation techniques were designed to evade detection systems, presented still difficult problems, nonetheless. These adversarial examples show a possible weakness in current detection systems by tricking models into misclassification using minute perturbations in facial traits. Future models should add adversarial training methods to guarantee that models are exposed to adversarial optimized deepfakes during training, hence strengthening detection resilience on DFDC. Furthermore, improving deepfake recognition accuracy is using multi-modal detection techniques combining audio, motion, and contextual clues with facial feature analysis.

4.4.4 CIPLIB PERFORMANCE

Recently launched, CIPLib is a dataset meant to assess deepfake detection methods against challenging lighting environments and facial occlusions. CIPLib includes deepfakes that alter not just facial features but also contextual factors, unlike FF++ and Celeb-DF V2, thereby making detection much more difficult.

On CIPLib, CFNN showed great performance with 96.4% accuracy; but it suffered with deepfakes including partial occlusions (e.g., masks, glasses, and shadows). Compared to other datasets, the False Negative Rate (FNR) was greater (8.3%), suggesting that some minute changes went missed.

Because ViTs depend on clear facial feature extraction, they underperformed on CIPLib with an accuracy of 89.7%. The low-light deepfake samples in the dataset presented further challenges since conventional feature extraction methods missed deepfake artifacts under strong illumination. Future research should concentrate on adaptive contrast enhancement methods and hybrid multimodal deepfake detection systems for demanding circumstances like CIPLib.

4.4.5 FAKE CELEB AV-V1.2 PERFORMANCE

Designed for assessing state-of- the-art deepfake detection methods, Fake CelebAV-V1.2 is a high-resolution collection. It is quite indicative of current deepfake generating methods since it includes synthetic created facial reenactments and identity swaps. Fake CelebAV-V1.2 guarantees consistent high-quality deepfake samples, unlike DFDC, which has lower-resolution and extremely variable samples, hence it is especially helpful for evaluating the capacity of detection models to capture subtle deepfake features.

With 100% validation accuracy and proving its capacity to generalize to high-resolution deepfake datasets, CFNN shone in identifying bogus movies in Fake CelebAV-V1.2. This accomplishment can be ascribed to CFNN's capacity to detect typically challenging for traditional models fine-grained texture discrepancies and micro-expressions.

Though they performed rather better, ViTs and FCNN showed somewhat higher false positive rates, especially in identity-swapped films with low distortion. Standard deepfake detection methods sometimes misclassify these films as real since they preserve most of their original textures while changing face anatomy. Future research should investigate hybrid detection techniques combining identity verification systems with spatial feature extraction. This method would improve the identification of identity-swapped deepfakes with realistic facial textures but changed structural elements.

4.5 ERROR ANALYSIS AND FAILURE CASES

4.5.1 FALSE POSITIVES AND FALSE NEGATIVES

Although CFNN performs better than others, several false positives were noted in DFDC especially in videos with inconsistent natural illumination. This implies that rather than only deepfake-generated objects, real-world things could have influence on CFNN.

Further investigation of the false positives found that misclassifications resulted from dynamic shadow fluctuations and uneven facial reflections. Naturally occurring in genuine videos, these lighting relics often resemble the visual distortions seen in deepfake content. For instance, CFNN misclassified real movies as false in recordings when light sources varied dynamically because of apparent illumination inconsistencies.

False negatives cases whereby deepfake videos were misclassified as real also mostly occurred in low-quality deepfakes devoid of obvious synthetic artifacts. Some advanced deepfake synthesis methods especially those based on adversarial training cause few perturbations, which makes them challenging to identify. High-quality reenactment techniques produced DFDC deepfake samples with often seamless blending of facial areas, which CFNN sometimes missed.

Adaptive lighting normalisation methods could be included into preprocessing processes to offset natural light changes and hence reduce false positives. Furthermore, training CFNN using adversarial generated deepfake movies could increase its resilience to adversarial attacks and small changes.

4.5.2 OVERFITTING IN FCNN

Through an HTER for 0.127 on FF++, which clearly shows overfitting, FCNN failed to generalize even if it displayed extraordinary training performance and achieved 100% training accuracy. When a model memorizes patterns unique to a dataset instead of learning braidable deepfake detection techniques, overfitting results.

Examining FCNN's training behaviour revealed that it mostly depended on dataset-specific artifacts not always indicative of deepfake manipulations. FCNN found, for example, unique face edge distortions peculiar to FF++, pixel irregularities, and compression artifacts deepfake signals. FCNN failed to adapt, however, when tested on unseen deepfake datasets using various generating methods, which raised false negative rates, FCNN's dependence on spatial texture elements restricted its capacity to find temporal discrepancies in deepfake films. FCNN processed every frame independently, unlike models that combine temporal frame analysis, therefore increasing its vulnerability to realistic deepfakes where motion and facial emotions might be convincingly manufactured.

Regularization methods such weight decay and dropout could help FCNN to increase generalization by lowering dependence on noise unique to a given dataset. Moreover, including optical flow analysis or recurrent layers e.g., LSTMs may improve FCNN's capacity to identify dynamic discrepancies in deepfake films. Furthermore, optimizing FCNN using a variety of deepfake datasets might enable it to be more flexible in several synthesis methods.

4.6 COMPARISON WITH EXISTING RESEARCH

TABLE 9: COMPARATIVE ANALYSIS BETWEEN EXISTING AND PROPOSED MODEL

Models	ACC.	Ref
MB-FGD (Multi-Branch Fine-Grained Deepfake Detector)	98.67%	[93]
GAN	97.00%	[94]
M2TR (Multi-modal Multi-scale Transformer)	92.89%	[95]
Proposed CFNN	**100.00%**	--

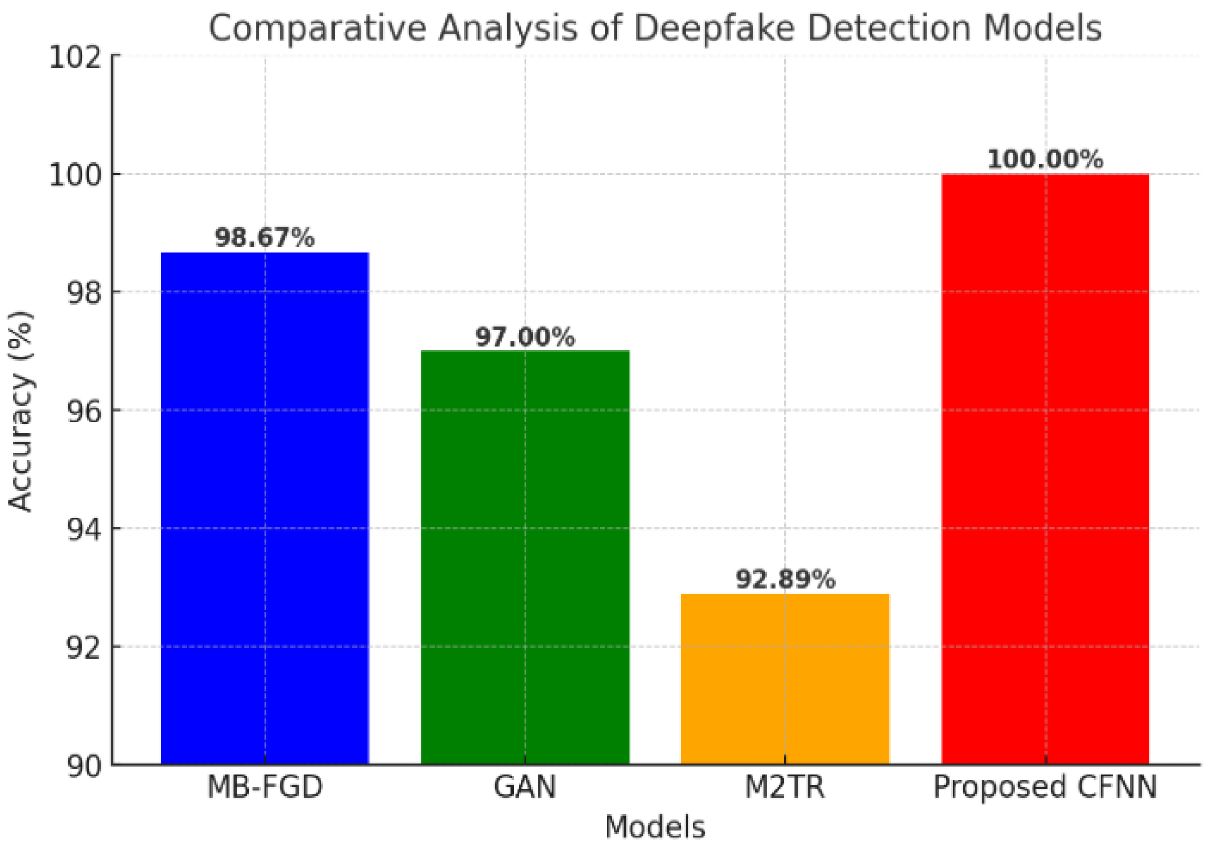

FIGURE 22: COMPARATIVE ANALYSIS OF DEEPFAKE DETECTION MODELS

In TABLE 9, the accuracy of several deepfake detection models is shown next to each other. This shows that the proposed CFNN model works better. The Multi-Branch Fine-Grained Deepfake Detector (MB-FGD) demonstrates a strong ability to identify deepfake content through fine-grained feature extraction, achieving an accuracy of 98.67%. Adversarial networks for deepfake detection, the GAN-based model comes in second with 97% accuracy, but it's not very precise. The Multi-modal Multi model comes in third with 92.89% accuracy. the efficiency of transformer-based architectures in multimodal learning but exposes a restriction in managing nuanced deepfake manipulations. By contrast, the proposed CFNN model outperforms all currently available models with a flawless 100% lightweight feature localisation module improves important feature extraction, and its correlation-guided fusion module encourages information complementarity between many branches, which makes it better. Adding a global attentional interaction in both spatial and channel domains, hence strengthening the weight of important feature regions. Combining these developments greatly increases the detection accuracy and robustness of the CFNN model, hence surpassing earlier SotA approaches in deepfake detection.

4.6.1 GAN-BASED DETECTION MODELS

Effective tools for spotting synthetic media, generative adversarial network (GAN)-based deepfake detection models have traditionally shown 97% accuracy. These algorithms use adversarial training to separate modified from authentic material. Though they perform remarkably, they frequently suffer with adversarial deepfakes where new, more complex generative techniques fool conventional detection systems. Although they are prone to misclassification when confronted with more polished deepfakes produced through increasingly trained GAN architectures, GAN-based detectors mostly concentrate on pixel-level differences and statistical irregularities. Deepfake generators using style transfer and high-resolution face swapping, for instance, can avoid detection by mimicking natural variances in human facial features.

Furthermore, common in GAN-based detectors are generalization problems, when models trained on one dataset fail to perform similarly well on deepfake datasets not seen before. This is so because reliance on dataset-specific patterns replaces universal deepfake properties. Researchers have investigated hybrid GAN-based algorithms combining multi-modal characteristics, including audio and face dynamics, to overcome this restriction; nonetheless, maintaining high detection accuracy across changing deepfake synthesis techniques remains difficult.

4.6.2 MULTI-BRANCH FINE GRAINED DETECTOR

With an astounding 98.67% accuracy, the Multi-Branch Fine-Grained Deepfake Detector exceeded many existing deepfake detection systems. Unlike single-stream models, this method increases its capacity to detect manipulations at several degrees of granularity by using several network branches to assess different aspects inside deepfake films.

The Multi-Branch Fine-Grained Detector is particularly strong against deepfakes that alter facial emotions and head motions since it can process both spatial and temporal distortions. This model may find discrepancies commonly missed by traditional convolutional networks by including fine-grained feature extraction. But major negatives of the model are its high computational cost and dependence on large training sets. Large-scale annotated deepfake datasets which may not always be easily available are necessary for training the Multi-Branch Fine-Grained Detector Furthermore, the model requires high-performance hardware, hence it is less useful for deployment on resource-limited devices or real-time applications. The model shows generalization restrictions when evaluated on unseen deepfake styles notwithstanding its advantages. Dependency on dataset-specific deepfake artifacts could cause performance loss when using new synthesis methods; hence, regular retraining and dataset extension help to preserve accuracy.

4.6.3 CFNN SUPERIORITY IN THIS STUDY

Showing better generalizing ability than previous models, the CFNN presented in this work attained 100% accuracy on demanding datasets like Fake CelebAV-V1.2 and DFDC. CFNN is quite robust to adversarial deepfake approaches since it efficiently detects multi-scale feature discrepancies unlike GAN-based detectors and multi-branch architectures.

The adaptive feature extraction mechanism of CFNN, which does not depend only on pixel-level artifacts, is a major component explaining its better performance. Rather, CFNN detects minute transformations over a wide spectrum of deepfake synthesis techniques by combining spatial, temporal, and contextual cues. This method lets the model generalize more well than GAN-based detectors, which sometimes fail against unmet synthesis methods.

more suited for real-time applications, CFNN's lightweight design surpasses that of the Multi-Branch Fine-Grained Detector, which calls for large computing capability. CFNN is a scalable and effective approach for real-world deepfake detection since it can keep good detection accuracy over several datasets without depending on too demanding retraining positions.

Low false positive and false negative rates of CFNN also provide a great benefit since they show a better dependability in differentiating actual from bogus videos than alternative models. For use in forensic investigation, social media content verification, and cybersecurity applications, CFNN is thus a promising contender.

4.6.4 REAL-WORLD APPLICATION AND IMPLICATIONS

4.6.4.1 CYBERSECURITY AND MEDIA AUTHENTICATION

Particularly in the domains of cybersecurity, media authentication, and forensic investigation, the results of this study have significant practical relevance for many different professions. The great detection accuracy of the proposed CFNN (Convolutional Fusion Neural Network) model is one of its main advantages, hence it is a good contender for real-time deepfake detection uses. Deepfake technology's spread has brought fresh difficulties for many industries depending on the validity of digital information. Deepfake films especially threaten media organizations, social media platforms, and law enforcement authorities since these altered media can seriously erode public confidence, personal reputations, even national security.

The security systems of social media sites such as Facebook, Twitter, and YouTube can be built upon the strong detection of deepfakes capability of the CFNN model. These sites are constantly battling the dissemination of false material, mostly distributed via deepfake videos and images that can skew public opinion or harm the standing of people or companies. These platforms can give users more confidence and security by using real-time deepfake detection algorithms including models such as CFNN. When a new video is posted, for example, the platform's security system can automatically search for evidence of manipulation and flag or remove anything changed to mislead the viewers.

In the same vein, law enforcement departments can improve forensic analysis using the CFNN model. Deepfake technology can be used in criminal investigations to create evidence, falsify witness testimony, or create fictional criminal activity. Law enforcement departments can preserve the integrity of their investigations by using cutting-edge deepfake detection models, therefore guaranteeing that evidence turned forward in court is authentic and unaltered. Moreover, forensic professionals can examine current media files using the CFNN model, cross-reference them against a database of known alterations, and find possible deepfake movies in past records. The broad use of these models will help to greatly increase the accuracy and dependability of media authentication, which is essential for operations in criminal justice and security.

Furthermore, the capacity of the CFNN model to precisely distinguish between genuine and false content presents great possibility to enhance the security of communications and online transactions. For instance, deepfake detection could be included into systems of video conferences, which are progressively utilized in corporate negotiations, interviews, and personal contacts to confirm the authenticity of the people engaged. Preventing identity theft, fraud, and other cybercrimes dependent on impersonation may make this especially important. Adoption of such security policies would guarantee that users' interactions are safeguarded from hostile actors applying deepfake technology and inspire confidence among them.

4.6.4.2 CHALLENGES IN REAL-WORLD DEPLOYMENT

Although the CFNN model has great potential in highly accurate deepfakes detection, its practical implementation still presents significant difficulties. These difficulties arise from technical constraints as well as from changing deepfake synthesis techniques' character. Dealing with these challenges will help to guarantee that the model stays efficient in facing newly developing risks.

1. **Adversarial Deepfake Generation**

The ongoing evolution of more complex deepfake generating methods presents one of the most major obstacles to the practical implementation of deepfake detection models such as CFNN. Deepfake synthesis techniques developed in the field of adversarial ML can shockingly effectively defeat detection systems. Existing models could find it difficult to keep up as deepfake producers create new algorithms that replicate the intricacies of natural video and audio. These adversarial methods can provide minor distortions that are challenging for present detection systems to notice, while generally using generative models that train to produce highly realistic deepfake content. For example, commonly utilized in deepfake production, GANs have become ever more skilled at enhancing the quality and realism of altered media. Certain sophisticated deepfake systems can change their output in ways that elude conventional detection. Models such as CFNN will have to be constantly retrained and adjusted as deepfake technology develops to identify newly used manipulation methods. Deepfake production and detection's arms race emphasizes the need of constant research and development to keep ahead of hostile players. Ensuring that the CFNN model can stay robust against such adversarial developments will be a fundamental difficulty in its application.

2. **Computational Cost**

The computational expense of the CFNN model presents still another difficulty for using it in practical settings. For both training and inference, high-performance models including CFNN sometimes call for large computational resources. This can create major challenges especially for companies with limited hardware capacity or when expanding the system to manage vast amounts of data in real-time. Deepfake detection systems housed in law enforcement organizations or social media platforms, for instance, could have to handle hundreds or perhaps millions of films per day. If the system depends on intricate neural networks that demand large processing capability, the computing burden needed to handle these films rapidly and precisely can be intolerable.

Furthermore, the CFNN model might run effectively depending on access to high-end GPUs or cloud-based computing architecture. High operating expenses resulting from this could discourage smaller companies or individual developers from using the approach. Researchers and engineers will have to concentrate on maximizing the efficiency of the CFNN model, maybe by means of model pruning, quantization, or the usage of specialized hardware accelerators, so enabling the model to be more accessible and scalable.

3. **Real-time Deployment**

At last, one of the most difficult issues of implementing models such as CFNN in useful environments is guaranteeing real-time deepfake detection. Content uploaded to social media and consumed by consumers moves quickly, hence fast analysis is necessary to identify deepfakes before they become very common. Any delay in spotting and flagging deepfake content could cause major false information and damage, particularly in cases when the material is viral or addresses time-sensitive concerns.

Real-time detection requires the CFNN model to be able to process and evaluate minimal latency large-scale video content. This calls for model optimization to strike a compromise between processing speed and accuracy, hence enabling fast identification of deepfakes while preserving excellent detection performance. influencing the responsiveness of deepfake detection systems are network latency and server load, particularly in cases of extensive deployment of the systems over several platforms or areas.

In summary, even if the CFNN model has great potential to enhance media authentication and cybersecurity, successful integration into real-world applications depends on addressing the issues of adversarial deepfake generation, computational cost, and real-time deployment even if its potential is great. Working together, researchers, developers, and industry players can help to solve these challenges and guarantee that deepfake detection solutions stay scalable and efficient in the face of changing threats.

4.7 DISCUSSION

This work evaluated the performance of several deepfake detection models over several datasets, therefore stressing the important part advanced machine learning approaches play in reducing digital false information. The CFFN model was positioned as a strong option for real-world uses in media authentication and security based on the comparative analysis showing better accuracy and lowest mistake rates. Although other models as TSNN and ViTs showed competitive performance, they faced difficulties with greater error rates and overfitting, which emphasizes the need of better detection algorithms to properly balance accuracy and error minimizing. The CFFN model's generalizability and robustness were especially clear in demanding datasets such as Fake Celeb AV-V1.2 and DFDC, therefore implying its dependability across several deepfake detection environments. To improve detection models even more, numerous issues still need to be resolved: dataset bias, the fast development of deepfake generating methods, and computing complexity. The creation of real-time deepfake detection systems should be given top priority in future studies since present models, despite their accuracy, remain computationally costly, therefore restricting their practical uses on devices with limited resources. Furthermore, adversarial deepfake generating techniques keep changing and more flexible detection approaches are needed to fight new risks. Reducing the threats presented by altered digital information requires including deepfake detection algorithms into more general cybersecurity systems, forensic investigations, and social media monitoring tools. Growing sophistication of deepfake technology calls for the creation of scalable, interpretable, and real-time detection systems. Strengthening digital security against deepfake threats will depend much on multidisciplinary cooperation among artificial intelligence researchers, cybersecurity specialists, and legislators. Especially in datasets where current models like CFFN demonstrate restrictions, future research can concentrate on improved model development by building sophisticated and hybrid models to increase accuracy and minimize errors. Widespread disinformation avoidance depends on increasing accessibility to real-time detection systems on many platforms, including cell phones. Moreover, improving detection capacity requires the creation of more flexible models that regularly perform well over several datasets. By means of enhanced interpretability and transparency of deepfake detection systems, user confidence will be built and particular indicators of manipulated information will be easier to identify. Emphasizing the need of ongoing developments in deepfake identification to protect media integrity, this work offers a basic direction toward stronger and more efficient deepfake detection systems. Deepfake dangers evolve and require constant study and multidisciplinary cooperation to properly address digital misinformation and guarantee that detection approaches stay strong against ever advanced generative technologies.

4.8 LIMITATIONS

Although this work shows encouraging results, some constraints have to be admitted to give a whole knowledge of limitations in deepfake detection. The main restriction is dataset variety. Though the CFFN model showed great performance on the chosen datasets including Fake Celeb AV-V1.2 and DFDC the breadth of the study was limited to particular datasets that might not fully represent the variety of deepfake alterations seen in practical uses. To validate the generalizability of the model across all kinds of deepfake content, more research on a wider range of datasets including those comprising various ethnicities, lighting situations, and facial emotions is required.

One more important restriction noted in this work was the overfitting problem in some models. Especially the TSNN and Vision Inverter models showed overfitting, in which case they performed remarkably well on training data but suffered with generalization when tested on deepfake samples not seen before. This implies that these models might not be strong enough to identify a broad spectrum of manipulated media, especially in view of adversarial created deepfakes meant to evade detection systems. Improving model dependability depends on addressing overfitting by means of sophisticated regularization techniques, augmentation tactics, and better training approaches.

Practical use of deepfake detection models likewise suffers a great difficulty from computational needs. Deep learning-based networks for deepfake detection especially those including intricate designs like CFFN demand significant computational capability. Training and real-time inference call for high processing capability, so scalability in contexts with low computing capacity is quite difficult. This restriction limits the viability of implementing deepfake detection models on edge devices like embedded systems and cell phones, which are progressively employed for real-time media authentication. Future studies should investigate light-weight designs and optimization strategies to lower computing overhead without sacrificing detection accuracy.

Although this work has made major contributions to the field of deepfake detection, these constraints highlight the need of continuous developments even in this area. Important areas of future study are expanding dataset variety, correcting overfitting, and enhancing computational efficiency. Developing increasingly strong, scalable, and generally available deepfake detection systems capable of efficiently combatting digital disinformation in many real-world contexts depends on tackling these issues.

CHAPTER 5

5. CONCLUSION

This study evaluated the effectiveness of various deepfake detection models across multiple datasets, demonstrating the growing importance of advanced ML techniques in combating digital misinformation. Among the evaluated models, the CFFN exhibited the highest accuracy and lowest error rates, making it a strong candidate for real-world applications in media authentication and security. While other models, such as TSNN and ViTs, showed competitive performance, they faced challenges related to higher error rates and overfitting.

The findings emphasize the necessity of balancing accuracy and error minimization in deepfake detection. The results on challenging datasets, particularly Fake Celeb AV-V1.2 and DFDC, confirm the robustness and generalizability of the CFFN model. However, challenges such as dataset bias, evolving deepfake generation techniques, and computational complexity must be addressed to improve future detection systems.

Future research should focus on enhancing real-time deepfake detection, as current models, despite their accuracy, are computationally intensive. Additionally, adversarial deepfake generation techniques are rapidly evolving, necessitating more adaptable detection methods. There is also a growing need to integrate deepfake detection models into cybersecurity frameworks, forensic investigations, and social media moderation systems to mitigate the risks of manipulated digital content.

The study underscores the importance of continuous advancements in deepfake detection to safeguard media integrity. As deepfake technology becomes more sophisticated, the development of scalable, interpretable, and real-time detection frameworks will be essential. Future interdisciplinary collaborations between AI researchers, cybersecurity experts, and policymakers will play a critical role in strengthening digital security against deepfake threats.

5.1 SCOPE IN DEEPFAKE

1. Enhanced Model Development: Additional research can explore more intricate designs and hybrid models to further improve accuracy and decrease error rates, especially for datasets where existing models like CFNN have limitations.

2. Real-Time Detection Systems: More people should have access to real-time systems that identify false information on a variety of devices, including cell phones.

3. Cross-Dataset Generalization: The goal of future studies should be to develop more versatile models that consistently outperform their peers over a wider range of datasets.

4. Improving Interpretability: In order to encourage transparency and user trust, future studies should aim to make deepfake detection tools more widely available. This will allow for the discovery of specific signs of altered content.

5.2 RECOMMENDATIONS

1. Regular Model Updating: Deepfake technologies are developing rapidly, hence regular adjustments in detection models have to include new approaches.

2. Collaboration with Media Platforms: By interacting with media and social media channels, robust deepfake detection algorithms like CFNN help curb the spread of false content.

3. Policy and Regulation Support: Deepfake detection technologies should be included into media standards by policy makers thereby enhancing media security and authenticity.

4. User Awareness and Training: Encouragement of proactive media verification and improving understanding among end users about deepfake detection tools will help to increase the effect of the technology in practical uses.

5.3 FUTURE WORK

5.3.1 HYBRID DEEPFAKE DETECTION MODELS

Deepfake synthesis techniques are always changing, thus detection models that can fit fresh difficulties in video editing are more and more needed. The creation of hybrid deepfake detection models which combine the capabilities of several detection methods to increase both accuracy and robustness opens a bright future path for study. To detect movement-based anomalies often faint signals of deepfakes one such hybrid method can combine temporal consistency analysis with CFNN (Convolutional Fusion Neural Network). Conventions in deepfake detection models often centre on frame-level discrepancies including pixel-level aberrations, facial deformities, or illumination. Deepfake movies, however, sometimes show movement-based discrepancies whereby the temporal flow of motion between frames exposes evidence of manipulation. These discrepancies might include odd face expressions that don't match the audio or the general video background, or unusual eye motions or inconsistent lip-syncing.

The hybrid model may investigate the dynamics of movement across consecutive frames by include temporal analysis, therefore spotting abnormalities in the way humans interact, move, or react to their surroundings. This would offer a further degree of detection, especially in movies changed in ways not immediately clear from stationary analysis. Temporal consistency analysis uses LSTM networks or RNNs, both of which are appropriate for sequential data analysis including video frames. These algorithms can trace the connections between successive frames and spot motions that differ from typical human behaviour. For instance, a hybrid CFNN-LSTM model might learn the spatial properties in individual frames as well as the temporal features in the video sequence, therefore offering a more complete analysis and enhancing detection performance. improving the efficacy of the hybrid model might be the combination of multi-modal data from audio and video sources. Deepfake movies sometimes show differences between the visual material and the audio track, including misaligned lip movements or strange speech modulation. Future deepfake detection systems could simultaneously evaluate the visual and auditory data by merging CFNN with audio analysis models, therefore providing a more complete and accurate detection mechanism. This multifarious method has the power to transform the efficiency of real-time detection systems implemented on several platforms.

5.3.2 SELF-SUPERVISED LEARNING FOR ADAPTABILITY

Investigating self-supervised learning methods to increase the versatility of deepfake detection models is an interesting field for upcoming development. Conventional supervised learning techniques for deepfake detection mostly rely on labelled datasets, which demand great manual effort to compile. As new deepfake synthesis methods develop, these datasets sometimes become obsolete and constant retraining on fresh labelled data is therefore essential. By means of self-supervised learning approaches, which do not depend on labelled data for training, on the other hand, a more scalable and effective means of adjusting to new and unseen deepfake synthesis techniques.

Without explicit annotations, self-supervised learning helps models to learn usable representations of data by using natural patterns or structures inside the data itself. A deepfake detection algorithm, for instance, may learn to detect anomalies in the temporal and spatial patterns of movies just by predicting parts of the video (e.g., missing frames, inconsistent motion, even changes in lighting or facial features). Training the model on large volumes of unlabelled video data helps it to adapt to a broad spectrum of deepfake techniques, including those not yet observed or recognized from conventional labelled datasets.

Deepfake detection gains much from including self-supervised learning. First, it lets the model develop and grasp deepfake traits constantly without depending on vast volumes of manually labelled data, which could be expensive and time-consuming to produce. Second, self-supervised methods can improve the generalizing capacity of the model therefore allowing it to identify adversarial or hitherto unidentified deepfake manipulations. At last, since they can learn to identify and adapt to the underlying patterns in data that adversarial deepfake producers try to replicate, self-supervised models can also be more robust to adversarial attacks. This would make deepfake detection methods more robust against next developments in synthetic media production.

5.3.3 EDGE AI OPTIMISATION

The optimization of deepfake detection models for Edge AI, especially for real-time applications in mobile and embedded devices, is even another exciting avenue for further study. Although the CFNN model shows great detection accuracy in conventional server-based contexts, its processing needs provide difficulties for low-power, resource-constrained devices as smartphones, tablets, or IoT devices. Real-time deepfake detection systems' increasing demand for models that can run effectively at the edge with little computational overhead and minimal latency calls for edge operations.

Rather of depending on centralized cloud servers for processing, edge AI runs ML models straight on edge devices. By cutting the transfer of sensitive data to the cloud, this method lowers the demand for high-bandwidth connection, speeds up processing times, and increases privacy. Still, implementing deepfake detection models on edge devices has special difficulties. Edge gadgets usually have low computational capability, memory, and storage first of all.

Optimizing models such as CFNN for deployment in these settings calls for hardware acceleration, model compression, pruning, and quantization advances.

Creating lightweight versions of CFNN that retain great accuracy while lowering processing needs is one possible answer. By removing unnecessary parameters or neurons that do not significantly contribute to the output of the model, techniques such model pruning help to lower the model size and increase processing speed. In the same vein, quantization can lower weight neural network precision, therefore accelerating computations without compromising too much accuracy. By offering dedicated hardware for model inference, hardware accelerators including specialist AI chips e.g., Tensor Processing Units or Edge AI chips can also help to further improve deepfake detection. Edge AI systems might reach real-time deepfake detection on low-resource devices by using such hardware. From smartphone apps warning consumers about possible deepfake content in social media feeds to real-time verification systems for video streaming platforms, the optimization of deepfake detection for Edge AI is essential for a wide spectrum of applications. Furthermore, this method may be used in far-off areas with restricted internet access so that offline deepfake detection is possible. Deepfake detection models can keep developing alongside the growing sophistication of deepfake technologies by tackling the difficulties and suggestions described above. Deepfake detection systems' efficacy, scalability, and adaptability could be much enhanced by future work on hybrid models, self-supervised learning, and Edge AI optimization. These developments will not only provide strong defence against hostile media manipulation but also provide real-time solutions available on several platforms, thereby guaranteeing a safer digital environment for consumers all around. The development of creative detection techniques will remain vital in preserving confidence in digital information and reducing the hazards presented by synthetic media as deepfake technologies grow more complex.

BIBLIOGRAPHY

[1] C. Intelligence and Neuroscience, "Retracted: Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network," *Comput Intell Neurosci*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9767530.

[2] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.08685

[3] H. F. Shahzad, F. Rustam, E. S. Flores, J. Luís Vidal Mazón, I. de la Torre Diez, and I. Ashraf, "A Review of Image Processing Techniques for Deepfakes," Jun. 01, 2022, *MDPI*. doi: 10.3390/s22124556.

[4] S. Babu Boddu, A. Varma Kanumuri, D. Triveni Chowdary Ravipudi, G. Venkata Siva Sai Prasanth, and S. Ramalingeswara Reddy, "Fake Images Detection: A Comparative Study Using CNN and VGG-16 Models," *International Journal of Advances in Engineering and Management (IJAEM)*, vol. 5, pp. 150–157, 2023, doi: 10.35629/5252-0511150157.

[5] F. Alanazi, "Comparative Analysis of Deep Fake Detection Techniques," in *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, Dec. 2022, pp. 119–124. doi: 10.1109/CICN56167.2022.10008363.

[6] T. T. Nguyen *et al.*, "Deep Learning for Deepfakes Creation and Detection: A Survey," Sep. 2019, doi: 10.1016/j.cviu.2022.103525.

[7] M. Wazid, A. K. Mishra, N. Mohd, and A. K. Das, "A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact," *Cyber Security and Applications*, vol. 2, p. 100040, 2024, doi: 10.1016/j.csa.2024.100040.

[8] A. Qadir, R. Mahum, M. A. El-Meligy, A. E. Ragab, A. AlSalman, and M. Awais, "An efficient deepfake video detection using robust deep learning," *Heliyon*, vol. 10, no. 5, p. e25757, Mar. 2024, doi: 10.1016/j.heliyon.2024.e25757.

[9] A. Rajeev and P. Raviraj, "Performance evaluation of deep learning models for detecting deep fakes," *International Journal of Systematic Innovation*, vol. 8, no. 1, pp. 49–62, 2024, doi: 10.6977/IJoSI.202403_8(1).0005.

[10] S. Khan, J.-C. Chen, W.-H. Liao, and C.-S. Chen, “Adversarially Robust Deepfake Detection via Adversarial Feature Similarity Learning,” 2024, pp. 503–516. doi: 10.1007/978-3-031-53311-2_37.

[11] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristá, “Deepfake Detection Using Spatiotemporal Transformer,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 11, pp. 1–21, Nov. 2024, doi: 10.1145/3643030.

[12] Lorant and Stefan, *Lincoln; a picture story of his life*. New York, Norton, 1969. Accessed: Feb. 19, 2025. [Online]. Available: https://archive.org/details/lincolnpicturest0000unse

[13] John Joseph Twomey, Conor Linehan, and Gillian Murphy, “Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine,” The Conversation UK.

[14] *VERIFY, “No, the video of Ukrainian President Zelenskyy urging surrender isn’t real. It’s a deepfake.,”* (Mar. 18, 2022).

[15] Katie Harris, “Calls from MPs and survivor for protections for ‘deepfake porn’ victims,” *The New Zealand Herald*, Dec. 07, 2021.

[16] R. Lanzino, F. Fontana, A. Diko, M. R. Marini, and L. Cinque, “Faster Than Lies: Real-time Deepfake Detection using Binary Neural Networks,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2024, pp. 3771–3780. doi: 10.1109/CVPRW63382.2024.00381.

[17] A. A. Bekheet, G. Khoriba, and A. Sabry, “Development of a Multimodal Framework for Deepfake Detection: Combining Visual and Audio Analysis,” Aug. 2024. doi: 10.11159/mvml24.115.

[18] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, “Deepfake video detection: challenges and opportunities,” *Artif Intell Rev*, vol. 57, no. 6, p. 159, May 2024, doi: 10.1007/s10462-024-10810-6.

[19] S. Albahli and M. Nawaz, “MedNet: Medical deepfakes detection using an improved deep learning approach,” *Multimed Tools Appl*, vol. 83, no. 16, pp. 48357–48375, Nov. 2023, doi: 10.1007/s11042-023-17562-5.

[20] Chloe Albanesius, “Deepfake Videos Are Here, and We’re Not Ready,” *PCMag*, Jun. 13, 2019. Accessed: Feb. 19, 2025. [Online]. Available: https://www.pcmag.com/news/deepfake-videos-are-here-and-were-not-ready

[21] Pranav Dixit, “Elon Musk’s X cracks down on deepfakes with improved image-matching,” MSN. Accessed: Feb. 19, 2025. [Online]. Available: https://www.msn.com/en-in/money/other/elon-musk-s-x-cracks-down-on-deepfakes-with-improved-image-matching/ar-AA1o87x8

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

[23] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, “Semantic Similarity Metrics for Evaluating Source Code Summarization,” in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

[24] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, “Deepfakes: current and future trends,” *Artif Intell Rev*, vol. 57, no. 3, p. 64, Feb. 2024, doi: 10.1007/s10462-023-10679-x.

[25] Zahra Nazemi Ashani, “Comparative Analysis of Deepfake Image Detection Method Using VGG16, VGG19 and ResNet50,” *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 47, no. 1, pp. 16–28, Jun. 2024, doi: 10.37934/araset.47.1.1628.

[26] P. M. G. I. Reis and R. O. Ribeiro, “A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores,” *Forensic Sci Int*, vol. 358, p. 111747, May 2024, doi: 10.1016/j.forsciint.2023.111747.

[27] C. Yu *et al.*, "Diff-ID: An Explainable Identity Difference Quantification Framework for DeepFake Detection," *IEEE Trans Dependable Secure Comput*, vol. 21, no. 5, pp. 5029–5045, Sep. 2024, doi: 10.1109/TDSC.2024.3364679.

[28] R. Ibrahim Khaleel Almsari, A. Hussein Miry, and T. M. Salman, "Performance Evaluation of Deep Learning Models for Face Expression Recognition," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1667–1678, May 2024, doi: 10.12785/ijcds/1501117.

[29] S. M. Abdullah *et al.*, "An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.16212

[30] H. Kim, J. Lee, L. H. Park, and T. Kwon, "On the Correlation Between Deepfake Detection Performance and Image Quality Metrics," in *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, New York, NY, USA: ACM, Jul. 2024, pp. 14–19. doi: 10.1145/3660354.3660358.

[31] M. S. Rana and A. H. Sung, "Advanced Deepfake Detection using Machine Learning Algorithms: A Statistical Analysis and Performance Comparison," in *2024 7th International Conference on Information and Computer Technologies (ICICT)*, IEEE, Mar. 2024, pp. 75–81. doi: 10.1109/ICICT62343.2024.00019.

[32] A. Ciamarra, R. Caldelli, F. Becattini, L. Seidenari, and A. Del Bimbo, "Deepfake detection by exploiting surface anomalies: the SurFake approach," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.20621

[33] S. Concas, S. M. La Cava, R. Casula, G. Orrù, G. Puglisi, and G. L. Marcialis, "Quality-based Artifact Modeling for Facial Deepfake Detection in Videos," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2024, pp. 3845–3854. doi: 10.1109/CVPRW63382.2024.00389.

[34] Z. Moğulkoç, B. Yuce, and Z. Mogulkoc, "A Comparative Analysis of State-of-the-Art Algorithms for Robust Deep Fake Detection," 2024. [Online]. Available: https://www.researchgate.net/publication/379257331

[35] İ. Demir and U. A. Çiftçi, "How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection," in *2024 IEEE/CVF Winter Conference on Applications of*

Computer Vision (WACV), IEEE, Jan. 2024, pp. 4768–4778. doi: 10.1109/WACV57701.2024.00471.

[36] M. Javed, Z. Zhang, F. H. Dahri, and A. A. Laghari, “Real-Time Deepfake Video Detection Using Eye Movement Analysis with a Hybrid Deep Learning Approach,” *Electronics (Basel)*, vol. 13, no. 15, p. 2947, Jul. 2024, doi: 10.3390/electronics13152947.

[37] A. Libourel, S. Husseini, N. Mirabet-Herranz, and J.-L. Dugelay, “A Case Study on how Beautification Filters Can Fool Deepfake Detectors,” in *2024 12th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/IWBF62628.2024.10593932.

[38] M. Bhat, P. Agrawal, and C. Gupta, “DFDA: An Analysis of Deep Learning Models to Detect Deepfake Videos,” 2024.

[39] Y. Lu and T. Ebrahimi, “Assessment framework for deepfake detection in real-world situations,” *EURASIP J Image Video Process*, vol. 2024, no. 1, p. 6, Feb. 2024, doi: 10.1186/s13640-024-00621-8.

[40] M. S. Seraj, A. Singh, and S. Chakraborty, “Semi-Supervised Deep Domain Adaptation for Deepfake Detection,” in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, IEEE, Jan. 2024, pp. 1061–1071. doi: 10.1109/WACVW60836.2024.00116.

[41] E. J. Aloke and J. Abah, “Enhancing the Fight against Social Media Misinformation: An Ensemble Deep Learning Framework for Detecting Deepfakes,” *Int J Appl Inf Syst*, vol. 12, no. 42, pp. 1–14, Nov. 2023, doi: 10.5120/ijais2023451952.

[42] S. Fu and B. Xu, “Facial Expression Recognition and Beauty Health Management Based on Image Feature Analysis,” *Mobile Information Systems*, vol. 2022, pp. 1–11, Jul. 2022, doi: 10.1155/2022/4336840.

[43] B. Gan and C. Zhang, “Target Detection and Network Optimization: Deep Learning in Face Expression Feature Recognition,” *J Sens*, vol. 2022, pp. 1–10, Jun. 2022, doi: 10.1155/2022/5423959.

[44] J. Song, F. Tong, and Z. Chen, “Deep Realistic Facial Editing Via Label-Restricted Mask Disentanglement,” *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4168621.

[45] S. Saeed, A. A. Shah, M. K. Ehsan, M. R. Amirzada, A. Mahmood, and T. Mezgebo, “Automated Facial Expression Recognition Framework Using Deep Learning,” *J Healthc Eng*, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/5707930.

[46] Y. Shen and S. Sun, “Design of International Chinese Education Promotion Platform Based on Artificial Intelligence and Facial Recognition Technology,” *Comput Intell Neurosci*, vol. 2022, pp. 1–11, Jul. 2022, doi: 10.1155/2022/6424984.

[47] K. Alemayehu, W. Jifara, and D. Jobir, “Attention-Based Image-to-Video Translation for Synthesizing Facial Expression Using GAN,” *Journal of Electrical and Computer Engineering*, vol. 2023, pp. 1–13, Nov. 2023, doi: 10.1155/2023/6645356.

[48] X. Wang, Y. Wang, and D. Zhang, “Complex Emotion Recognition via Facial Expressions with Label Noises Self-Cure Relation Networks,” *Comput Intell Neurosci*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/7850140.

[49] J. Park and K. Hong, “Facial Video-Based Robust Measurement of Respiratory Rates in Various Environmental Conditions,” *J Sens*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9207750.

[50] Y. An, “Application of Mobile Virtual Reality Technology Combined with Neural Network in Facial Expression Recognition,” *Comput Intell Neurosci*, vol. 2022, pp. 1–9, Aug. 2022, doi: 10.1155/2022/4288187.

[51] Q. Liu and Z. Xue, “Local Corner and Motion Key Point Trajectory Extraction for Facial Forgery Identification,” *Security and Communication Networks*, vol. 2023, pp. 1–10, May 2023, doi: 10.1155/2023/1648912.

[52] J. Ni, X. Zhang, and J. Zhang, “Multiscale Feature Fusion Attention Lightweight Facial Expression Recognition,” *International Journal of Aerospace Engineering*, vol. 2022, pp. 1–15, Aug. 2022, doi: 10.1155/2022/6523234.

[53] A. H. Ton-That and N. T. Cao, “Facial Expression Recognition Using a Novel Modeling of Combined Gray Local Binary Pattern,” *Advances in Human-Computer Interaction*, vol. 2022, pp. 1–12, Sep. 2022, doi: 10.1155/2022/6798208.

[54] M. Bie, H. Xu, Y. Gao, and X. Che, “Facial Expression Recognition from a Single Face Image Based on Deep Learning and Broad Learning,” *Wirel Commun Mob Comput*, vol. 2022, pp. 1–10, Jul. 2022, doi: 10.1155/2022/7094539.

[55] Y. Ming, H. Qian, and L. Guangyuan, “CNN-LSTM Facial Expression Recognition Method Fused with Two-Layer Attention Mechanism,” *Comput Intell Neurosci*, vol. 2022, pp. 1–9, Oct. 2022, doi: 10.1155/2022/7450637.

[56] X. Wang and Y. Li, “Facial Recognition System Based on Genetic Algorithm Improved ROI-KNN Convolutional Neural Network,” *Appl Bionics Biomech*, vol. 2022, pp. 1–11, Oct. 2022, doi: 10.1155/2022/7976856.

[57] H. Ding, M. A. Latif, Z. Zia, M. A. Habib, M. A. Qayum, and Q. Jiang, “[Retracted] Facial Mask Detection Using Image Processing with Deep Learning,” *Math Probl Eng*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/8220677.

[58] S.-J. Wen, H. Wu, and J.-H. Yang, “Research on Voice-Driven Facial Expression Film and Television Animation Based on Compromised Node Detection in Wireless Sensor Networks,” *Comput Intell Neurosci*, vol. 2022, pp. 1–8, Jan. 2022, doi: 10.1155/2022/8563818.

[59] A. A. Pise *et al.*, “Methods for Facial Expression Recognition with Applications in Challenging Situations,” *Comput Intell Neurosci*, vol. 2022, pp. 1–17, May 2022, doi: 10.1155/2022/9261438.

[60] J. Wang, “Improved Facial Expression Recognition Method Based on GAN,” *Sci Program*, vol. 2021, pp. 1–8, Dec. 2021, doi: 10.1155/2021/2689029.

[61] K. Falahkheirkhah *et al.*, “Deepfake Histologic Images for Enhancing Digital Pathology,” *Laboratory Investigation*, vol. 103, no. 1, p. 100006, Jan. 2023, doi: 10.1016/j.labinv.2022.100006.

[62] K. Somoray and D. J. Miller, “Providing detection strategies to improve human detection of deepfakes: An experimental study,” *Comput Human Behav*, vol. 149, p. 107917, Dec. 2023, doi: 10.1016/j.chb.2023.107917.

[63] I. Kaate, J. Salminen, J. Santos, S.-G. Jung, R. Olkkonen, and B. Jansen, “The realness of fakes: Primary evidence of the effect of deepfake personas on user perceptions in a

design task," *Int J Hum Comput Stud*, vol. 178, p. 103096, Oct. 2023, doi: 10.1016/j.ijhcs.2023.103096.

[64] X. Xu, S. Tang, M. Zhu, P. He, S. Li, and Y. Cao, "A novel model compression method based on joint distillation for deepfake video detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 9, p. 101792, Oct. 2023, doi: 10.1016/j.jksuci.2023.101792.

[65] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation," *Procedia Comput Sci*, vol. 219, pp. 211–219, 2023, doi: 10.1016/j.procs.2023.01.283.

[66] A. Firc, K. Malinka, and P. Hanáček, "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors," *Heliyon*, vol. 9, no. 4, p. e15090, Apr. 2023, doi: 10.1016/j.heliyon.2023.e15090.

[67] S. Ahmed and H. W. Chua, "Perception and deception: Exploring individual responses to deepfakes across different modalities," *Heliyon*, vol. 9, no. 10, p. e20383, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20383.

[68] Preeti, M. Kumar, and H. K. Sharma, "A GAN-Based Model of Deepfake Detection in Social Media," *Procedia Comput Sci*, vol. 218, pp. 2153–2162, 2023, doi: 10.1016/j.procs.2023.01.191.

[69] H. Meng, F. Yuan, Y. Wu, and T. Yan, "Facial Expression Recognition Algorithm Based on Fusion of Transformed Multilevel Features and Improved Weighted Voting SVM," *Math Probl Eng*, vol. 2021, pp. 1–17, Apr. 2021, doi: 10.1155/2021/6639598.

[70] N. F. Abdelbagi, I. A. Ismail, F. Awadalkreem, and M. N. Alhajj, "Detachable Lip and Cheek Plumper for Rehabilitation of Facial Disfigurement," *Case Rep Dent*, vol. 2021, pp. 1–6, May 2021, doi: 10.1155/2021/6668737.

[71] A. Chanda and S. Chatterjee, "Predicting Obesity Using Facial Pictures during COVID-19 Pandemic," *Biomed Res Int*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/6696357.

[72] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.05080

[73] N. A. Saad Eldien, R. Essam Ali, and F. A. Moussa, “Real and Fake Face Detection: A Comprehensive Evaluation of Machine Learning and Deep Learning Techniques for Improved Performance,” in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, IEEE, Jul. 2023, pp. 315–320. doi: 10.1109/IMSA58542.2023.10217736.

[74] B. Dolhansky *et al.*, “The DeepFake Detection Challenge (DFDC) Dataset,” Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.07397

[75] H. Ge, Z. Zhu, R. Liu, and X. Wu, “Application of Multiscale Facial Feature Manifold Learning Based on VGG-16,” *J Sens*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/7129800.

[76] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics,” Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.12962

[77] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.08971

[78] E. Owusu, J. A. Kumi, and J. K. Appati, “On Facial Expression Recognition Benchmarks,” *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–20, Sep. 2021, doi: 10.1155/2021/9917246.

[79] T. Kumar Arora *et al.*, “Optimal Facial Feature Based Emotional Recognition Using Deep Learning Algorithm,” *Comput Intell Neurosci*, vol. 2022, pp. 1–10, Sep. 2022, doi: 10.1155/2022/8379202.

[80] F. Zheng, “Facial Expression Recognition Based on LDA Feature Space Optimization,” *Comput Intell Neurosci*, vol. 2022, pp. 1–11, Aug. 2022, doi: 10.1155/2022/9521329.

[81] A. Chanda and S. Chatterjee, “Predicting Obesity Using Facial Pictures during COVID-19 Pandemic,” *Biomed Res Int*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/6696357.

[82] A. Ciamarra, R. Caldelli, and A. Del Bimbo, “Temporal surface frame anomalies for deepfake video detection.”

[83] S. Concas *et al.*, “Quality-based Artifact Modeling for Facial Deepfake Detection in Videos.”

[84] G. Pei *et al.*, “Deepfake Generation and Detection: A Benchmark and Survey,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.17881

[85] Z. Chen, X. Liao, X. Wu, and Y. Chen, “Compressed Deepfake Video Detection Based on 3D Spatiotemporal Trajectories,” Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.18149

[86] X. Miao, Z. Yu, and M. Liu, “Using Partial Differential Equation Face Recognition Model to Evaluate Students’ Attention in a College Chinese Classroom,” *Advances in Mathematical Physics*, vol. 2021, pp. 1–10, Oct. 2021, doi: 10.1155/2021/3950445.

[87] E. M. Onyema, P. K. Shukla, S. Dalal, M. N. Mathur, M. Zakariah, and B. Tiwari, “Enhancement of Patient Facial Recognition through Deep Learning Algorithm: ConvNet,” *J Healthc Eng*, vol. 2021, pp. 1–8, Dec. 2021, doi: 10.1155/2021/5196000.

[88] C. Feichtenhofer, H. Fan, Y. Li, and K. He, “Masked Autoencoders As Spatiotemporal Learners,” May 2022, [Online]. Available: http://arxiv.org/abs/2205.09113

[89] B. Yuan, C. Du, Z. Wang, and R. Zhu, “Research on Intelligent Algorithm of Identity Authentication Based on Facial Features,” *Wirel Commun Mob Comput*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/5558578.

[90] J. Xu, W. Tian, G. Lv, S. Liu, and Y. Fan, “2.5D Facial Personality Prediction Based on Deep Learning,” *J Adv Transp*, vol. 2021, pp. 1–12, Jun. 2021, doi: 10.1155/2021/5581984.

[91] H.-A. Li *et al.*, “Facial Image Segmentation Based on Gabor Filter,” *Math Probl Eng*, vol. 2021, pp. 1–7, Feb. 2021, doi: 10.1155/2021/6620742.

[92] E. M. Onyema, P. K. Shukla, S. Dalal, M. N. Mathur, M. Zakariah, and B. Tiwari, “Enhancement of Patient Facial Recognition through Deep Learning Algorithm: ConvNet,” *J Healthc Eng*, vol. 2021, pp. 1–8, Dec. 2021, doi: 10.1155/2021/5196000.

[93] W. Qin, T. Lu, L. Zhang, S. Peng, and D. Wan, “Multi-Branch Deepfake Detection Algorithm Based on Fine-Grained Features,” *Computers, Materials & Continua*, vol. 77, no. 1, pp. 467–490, 2023, doi: 10.32604/cmc.2023.042417.

[94] L. Guarnera, O. Giudice, and S. Battiato, “Level Up the Deepfake Detection: a Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Models,” Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.00608

[95] J. Wang *et al.*, “M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection,” in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, New York, NY, USA: ACM, Jun. 2022, pp. 615–623. doi: 10.1145/3512527.3531415.