CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

# DEEPFAKE DETECTION USING FEATURE EXTRACTION AND LSTM

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Krutarth Aghera

May 2024

The thesis of Krutarth Aghera is approved by:

| | |
|---|---|
| Dr. Wen-Chin (Amy) Hsu | Date |
| Vanessa Klotzman | Date |
| Dr. Rashida Hasan, Chair | Date |

California State University, Northridge

Acknowledgements

I would like to express my heartfelt gratitude to my advisor and committee chair Dr. Hasan Rashida for the continuous support of my master's study and research, for his trust, patience, motivation, enthusiasm, and immense knowledge. I would also like to thank my thesis committee members, Dr. Wen-Chin (Amy) Hsu and Vanessa Klotzman, for their support and guidance. Lastly, I would like to dedicate this thesis to my parents, family and friends for their continuous help and support throughout my education.

Table of Contents

List Figures

List of Tables

Abstract

DEEPFAKE DETECTION USING FEATURE EXTRACTION AND LSTM

By Krutarth Aghera

Master of Science in Computer Science

This paper delves into various deep learning models and how they might be applied to the problem of deepfake video detection. These models include hybrid architectures, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs). Using a variety of datasets, including FaceForensic++, DFDC, Celeb-DF, and others, we trained and evaluated our model to distinguish between real and fake videos. We obtain promising findings for our experiments: many models perform well in identifying deepfake videos. More specifically, training ResNext with LSTM across all datasets with an accuracy of 97.2% produces very good knowledge of spatial features related to deepfake manipulation, demonstrating CNN's capacity to extract features. Furthermore, by utilizing temporal and spatial features, the most competitive and accurate detectors at the moment are the state-of-the-art techniques, such as hybrid CNNs with LSTM networks. This work adds to the current body of research on deepfake detection, which benchmarks deep learning-based model performance on a variety of datasets. In order to achieve this, we examine several architectures and datasets to identify their advantages and disadvantages. We then attempt to utilize this information to drive future research efforts toward the development of a more reliable and effective deepfake detection system. Our results highlight the importance of diverse datasets for purposes beyond improving model generalization. As such, they should be at the center of future transdisciplinary and technological development initiatives aimed at addressing the collective issues presented by the manipulation of synthetic media. Finally, our goal is to use these better detection technologies to ensure the authenticity and integrity of digital material in the face of increasing media complexity.

# Chapter 1

## Introduction

Fast-developing Artificial Intelligence technology, among others, has enabled it to attempt to define the rather blurry line between reality and artificiality in the area of creating digital content. Deepfake technology has contributed to a number of technologies because of this very potent tool in creating counterfeit extremely efficiently videos and recordings of voices. The leading technology through which advanced machine learning and artificial technologies get developed, particularly for making audio and visual content that happens to provide high potential to the devices. New technologies and advancements like these have helped the public as a whole to have access to it other than the previous access, which was only by the domain of special effects artists in videos. This also raises the question:

The implications of deepfake technology go far beyond those for entertainment or even pranking people. It relates to the very basis of communication and even identity and credibility in modern communities. There is a very high possibility for deepfakes to be utilized in a malicious manner, especially in cases of political propaganda, financial fraud, and personal attacks, as they give the power to create highly persuasive videos of individuals of importance making or doing something quite different from what they said or did. The more these technologies spread, it is becoming difficult to demarcate which is real and which is not. This shows a growing need for sharp detection systems. Situated within the backdrop of technical innovation and social danger, this research aims to contribute to the important field of digital forensics.

### 1.1 Motivation

Literally, digital content production has been given a new face, and deepfake technology has opened the doors for a scope for greater reality in video and audio editing. The technological flair is so amazing that it really throws a serious gauntlet to the legitimacy and honor of the digital media. Deepfakes, which can delude the sharpest among us, blur the line of reality versus fabrication. This, therefore, poses

great threats against privacy, security, and even the veracity of information being propagated across digital platforms. The consequences are very manifold: from disinformation to identity theft, from undermining public confidence in the reliability of digital media. In fact, these problems of a very urgent nature outline the dire need for potent, viable methods of deep learning developed to spot and diminish the spread of deepfakes.

In this sense, our study targets suggesting up-to-date solutions for the detection of deepfake material using up-to-date deep learning techniques. Our goal is to improve existing forgery detection mechanisms, leveraging comprehensive datasets such as Face Forensics++, Celeb-DF, and the Deepfake Detection Challenge dataset, which have been made available to the public to tackle complexity that has never been presented before. Herein is a careful review and experiment through which our work seeks to improve the resilience of the deepfake detection system; thus, upholding the authenticity and integrity of digital media content. We, in this work, have sought to contribute to the efforts of strengthening defenses against the insidious threats posed by the intention of malicious actors to take advantage of vulnerabilities inherent in contemporary media technologies through a close analysis of the mechanics of deepfake generation and deployment. We work with rigorous innovativeness and interdisciplinary collaborations to make the veracity of digital content secure and uncompromised in the future. So that users across the globe can trust and have the confidence which they deserve.

**1.2 Research Questions**

Will it be possible to tune the deep learning models in such a way that they are capable of identifying deep fake material effectively at a really high level of accuracy over a wide range of complexity? This will attempt to find out how flexible and strong the deep learning algorithm is in the given scenario of the separation of manipulated content from the real one.

What is currently limiting the capability of up-to-date methodologies to detect deepfake videos? The discussion tackles the strengths and shortcomings of some of the methods that guide toward

unexplored areas which may be helpful for the researcher working in this field of detecting deepfakes.

Contribution to the model training and validation: how different datasets contribute to training and validation regarding the deepfake detection models. This will, in a later experiment, test the importance of a varied dataset to improve generalization and algorithm performance for the task.

Durability of Neural Network Topologies: How do different neural network architectures fair in recognizing the subtlest modifications, which would be characteristic of deepfake content? This study will thus critically measure the robustness and adaptability of the considered diverse neural network topologies to recognize small changes in media.

Suitability for Real-Time Detection: How useful will these provided neural network topologies be for real-time detection of deep fakes? This brings the question of its possibility of deploying the models into a real-world scenario while maintaining the perspective of computational efficiency and accuracy in a dynamic environment.

Our work was guided, therefore, by the following research questions: What are the effective means of detecting deepfakes? Developing and sharing state-of-the-art deepfake detection methodologies to further improve the understanding and development of more effectual strategies in combating the increased proliferation of synthetic media.

### 1.3 Research Contribution

This contribution is quite significant, since this new technique for the detection of deepfakes, brought in by means of the synergistic capabilities of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), is a great value addition to the area of digital forensics. In this work, both of the most powerful neural network architectures are combined with the following explanation: CNNs have been very successful in computer vision tasks and are good for extracting spatial features in the images or video frames, while RNNs are very effective in capturing the temporal dependency of sequences of data. In other words, it enables the approach to more expensively and in detail analyze deepfake content for higher

accuracy in detection and more robustness against evolving techniques in generating deepfakes.

The thesis will demonstrate how model performance can be enhanced and adapted to the dynamic of deepfake technologies by conducting rigorous evaluations across multiple datasets. Furthermore, the current study will help us better understand the function of various data in the model's accuracy and efficacy. It thus conducts a systematic study of model performance across multiple datasets, ranging from FaceForensics++ to Celeb-DF, as well as the Deepfake Detection Challenge datasets, to better understand the impact of dataset variability on detection accuracies and generalization capabilities. This comprehensive research not only emphasizes the necessity of diversity in training model and validation datasets, but it also provides essential insight into the strengths and limitations of current detection approaches.

These works thus provide an insight into the future of deepfake detection research that may guide more resilient and adaptive detection systems in mitigating the spreading of synthetic media and protecting digital content. Furthermore, the thesis recommends a new way of fusion using CNNs and RNNs, which are good at learning spatiotemporal features based on the literature. Meanwhile, the thesis discusses the effectiveness of integrating the pre-trained state-of-the-art deep learning models, Incv3 (InceptionV3), ResNext, along with LSTM networks for the task of deepfake detection. In these models, the feature extraction capabilities of pretrained models are combined with the temporal modeling prowess of LSTM networks. The new proposed architectures in the present study build on that with the aim of improving the detection accuracies further and perhaps being useful in real-time detection.

First, the research provides quite valuable insights into relative advantages, trade-offs, and model architectures systematically experimented with and evaluated for performance. Second, it systematically experiments and evaluates the performance of the architectures toward guiding researchers and practitioners who seek to devise deeper and better deepfake detectors. Related Work, has provided a deep review of existing literature and research relevant to the problem domain. The chapter is designed under

various sections as per the different approaches taken by past studies.

In Chapter 2, the thesis delves into the exploration of various deep fake detection techniques. The chapter begins with an analysis of CNN with LSTM networks for deep fake video recognition, highlighting the importance of reliable detection systems in combating disinformation. It further discusses eye blinking-based detection method, capsule networks approach, and biological signals framework. Additionally, the chapter comprehensively reviews the deep learning-based detection methodologies and systematic surveys of deep fake detection approaches, providing a comprehensive overview of the current landscape in deepfake detection research.

In Chapter 3, Proposed Methodology, the thesis turns its focus toward the presentation of the new proposed approach to the identified problem. Chapter 4, Experiments and Results, develop an empirical evaluation of the proposed methodology. The results later in this chapter are organized around the findings of the experiments, following some similarity in the sections whereby it categorizes and interprets the findings in light of deriving more meaning from the insights. Finally, Chapter 5 concludes with a summary of the main findings of the thesis and indicates contributions by providing a summary, future work, and limitation of our study.

## Chapter 2

## Related Work

In his paper "Deep Learning Technique for Recognition of Deep Fake Videos," Fahad Mira attempted to analyze current trends in the field of deep fake detection by performing an in-depth assessment of the currently used methods that employ this type of technology, deep learning. So, this study investigates the usefulness of CNN with LSTM network combinations in detecting actual versus fraudulent or manipulated information in a video. Mira emphasizes that deepfake technology poses insurmountable hurdles to disseminating disinformation, and that acute, reliable, and effective detection systems are the only assurance that will protect digital trust and security. The YOLO face detector is used with the CNN and XGBoost classifiers to give an excellent method for discriminating between actual and fraudulent facial frames. The current study makes three contributions to the current academic debate: a methodological proposal for an effectiveness assessment of different deep learning models, including against deepfakes, and, more importantly, providing additional evidence of the need to continue improving the overall level of cybersecurity in order to adequately counter deception in the digital age.

Li and Lu [13] created a groundbreaking deep learning technique for distinguishing AI-generated fake videos (DeepFakes) from real content. The methodology begins by establishing the warping procedure of the affine face, which is utilized to determine unique artifacts in the method used to create a DeepFake video. Such artifacts are caused by a resolution mismatch between the synthetic face and the actual video surroundings. Their technique takes use of such a resolution disparity. [13]. While previous systems have relied on large training sets containing both genuine and synthetic images, Li and Lyu use training by simulating adverse events with simple computational image processing techniques to circumvent the need for DeepFake to use its generated images as negative samples. Admittedly, their method is quite durable, as evidenced by DeepFakes' widespread applicability in a variety of settings. Their technique was proven

to be highly effective across a number of DeepFake video datasets, indicating significant progress in the ongoing fight against video forgeries.

Yuezun Li et al. [14] proposed a new technique for recognizing fully neural network-generated deep fake films using the impulsive physiological signal of eye blinking, which is typically neglected during deep fake synthesis [14]. This study focuses on distinguishing between deep fakes of very high quality and genuine videos by investigating the frequency and pattern of eye blinking, a physiological signal that is generally underrepresented in AI-generated videos, with the hope that researchers will continue to make progress with deep generative networks. They propose a Long-Term Recurrent Convolutional Network model for the temporal sequence of eye movements that outperforms the current approach. This technique outperforms established benchmark datasets, emphasizing the importance of using physiological signals to detect digital forgeries in videos. It is relevant because the quick expansion and relative ease of access that this deepfake technology has brought have had serious consequences for the security, confidentiality, and integrity of digital material.

Huy et al., [15] present a very fresh and innovative approach in the paper "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," which uses capsule networks for detecting a wide range of fakes from replaying attacks through static image or video playback to advanced computer-generated images and videos created with deep convolutional neural networks.[15] Clearly, this was very much off the beaten path of techniques used earlier, as here the use of capsule networks has been employed but had mostly been in computer vision challenges, and this would bring out its efficacy vividly in this fast-paced world of tampering in digital media. Leveraging the Gaussian random noise during training, this strategy provides robustness that leads to breakthrough efficacy across some datasets. On the other hand, the model's reliance on random noise at training time begs questions of whether the generalization and its operational efficacy in the real world, where data doesn't show equivalent noise characteristic, are tenable. This work lays the path for sophisticated neural architecture, such as capsule networks, in the near

future within the digital forensics area, where domain-agnostic, adaptive solutions are critical for combating digital media forgeries. In the paper "Fake Catcher: Detection of Synthetic Portrait Videos using Biological Signals" by Umur et al. [16], the authors create a pretty unique, odd way for handling the problem, a framework for identifying synthetic portrait movies based on biological signals.

Ordinary detectors of deep learning would ineluctably lose against the generative models capable of producing such sophisticated films. The authors introduce a classification system that would have the highest level of accuracy in distinguishing between authentic and synthesized information without depending on a model of generative design employed in the production of the false content, in the search of spatial coherence and temporal consistency in such signals with diverse transformations. They do this in a way not only indicative of great detection rates in many datasets but also in how they introduce an "in the wild" dataset to further show their method. This emphasizes their approach and further strengthens that their technique is strong and universally applicable in the development process of the battle against manipulation of digital video.

**2.2 Deepfake Detection Techniques**

As such, the interest in great interest in the academic and industrial domain has invited deepfake detection methods, motivated mainly by the increasing concern toward the authenticity and reliability of digital media. In the well-acclaimed work, Taeb and Chi [25] undertook a comprehensive review of methodologies to detect, deploying deep-learning-based paradigms. The researchers studied thoroughly by experiments and analyses of various deep learning architectures whether they are in fact effective for manipulated media content detection. This research, therefore, not only contributes to a better understanding of capabilities with deep learning-based approaches but also discusses their potential to mitigate challenges imposed by continuously increasing sophistication in deepfake generation. From this point of view, the study by Taeb and Chi could be referred to as the first more advanced one on the way to improving and reinforcing better mechanisms of detection of digital media integrity [25].

Simultaneously with him, Almars [26] also worked on discussing the problematics of deepfake detection through a comprehensive survey of deep learning methods. Almars rigorously categorizes detection approaches based on their reliance on deep learning principles in a systematic way that provides a panoramic view of the contemporary landscape of deepfake detection. Not to outline only the prevailing techniques using state-of-the-art, but also to explain the evolving role of machine learning algorithms in mitigating risk, which is inherent to synthetic media manipulation. The systematic review by Almaz et al. is a kind of compass for orientation in the thorny landscape of deep fake detection, not only for researchers but also for practitioners, bringing an impressive source of insights on the strengths and limitations of various deep learning-based approaches. The work of Almars, therefore, should be viewed as an essential contribution that lies in the effecting resilience of digital media ecosystems in the identified perils from the proliferation of deepfakes [26].

In contrast, Abdulreda and Obaid [27] depicted a large horizon of deepfake methods and detection tools for both traditional and cutting-edge. Placed against the larger canvas of deep fake technology and what it otherwise generally portends, the authors ultimately provide a big-picture view of the challenges and opportunities in the war against digital fraud. Abdulreda and Obaid provide a disciplinary interdisciplinary synthesis of detection methods as applied to the wide diversity of features associated with deepfake detection, while underlining the importance of disciplinary interconnection in the quest for their multi-reality and innovative research strategies [27].

Apart from this, Solaiyappan and Wen's [28] work is a comparative analysis with a focus on deepfake detection using machine learning in the domain of medical image analysis. Their work provides a great extension of the methodologies for deepfake detection beyond the conventional media format and underscores the relevance of the detection techniques in such a wide and diverse domain as healthcare. Solaiyappan and Wen add to the fast-growing literature on the domain of specific deepfake detection methodology by assessing the performance of machine learning algorithms in classifying manipulated

medical images. In many ways, their findings afford very rich insights into the applicability and effectiveness of machine learning approaches to safeguarding the integrity of medical imagery in effect, addressing authenticity concerns with the utmost degree of respect for healthcare data [28].

Besides, Mitra et al. [29] developed another method based on machine learning for application in the course of deepfake video detection in the purview of social media platforms. In so doing, this study is illustrative of a response to some of the unique challenges posed by the nature of content; it is wide and growing. This need for immediate detection underscores the importance within scalable and effective solutions that could be applied to mitigate risk with a sense of urgency at the back of deep fakes becoming pervasive in social media environments. The developed framework in this paper leverages advanced machine learning approaches to provide a pragmatic approach to organizations for addressing the evolving threat landscape in synthetic media manipulation, hence preserving the integrity of digital discourse for maintaining the trust between community members [29].

Thus, Solaiyappan and Wen survey the applicability of fake detection for the first time in a specialized domain of medical image analysis, therefore giving an utterly new application field for deep-fake detection [28]. The study evaluates machine learning algorithms on the performance to discern manipulated medical images and hence shows the promise that machine learning-based methods will improve security and reliability of health data. This work adds to a growing literature about domain-specific deepfake detection methodologies and brings underlined high demand for domain-specific detection for sensitive domains, such as health care. On a parallel note, the contribution of the work by Mitra et al. [29] also signals towards the urgent need for framing an effective deepfake detection mechanism in the changing scenario of social media platforms, if we are to take these platforms seriously. The present study proposes a pragmatic framework to detect deepfake videos in real time, using a novel machine-learning-based method for mitigating the risks that come along with the distribution of synthetic media content. The work thus underscores the pressing need for the development of scalable and robust

detection solutions to keep the integrity of digital discourse intact and make trust in online communities possible.

### 2.3 Convolutional Neural Networks (CNNs) in Deepfake Detection

Among the tools which have been developed in this field, Convolutional Neural Networks (CNNs) have come up as the most dominating ones, due to their exceptional performance in different computer vision tasks. It was found out that CNNs have been found well-suited at acquiring hierarchical features from images, and as a result, they are good at capturing subtle visual cues indicative of deepfake manipulation. CNNs work by the convolution learnable filters across input images that capture spatial patterns and code into feature maps. The CNNs applied in feature extraction in deepfake detection should apply to extract the features and the anomalies brought in or artifacts brought in by the manipulation in the course. The CNN-based detectors are thereby able to effect discrimination of the media content between real and fake.

Among the notable contributions, Zhao, Wang, and Lu [30] introduce a new method to detect deepfake videos with the application of a two-stream convolutional neural network (CNN) model. The first one is learning with two-level features and the combination of spatial and temporal information to improve discrimination ability between real and manipulated video. The training of the CNN on the dual-domain-features-extracted features of these two domains allows the model to pool even subtle artifacts and inconsistencies introduced in the deepfake generation process. The approach leverages spatial feature extraction capabilities of CNNs from image content and temporal information to reveal motion dynamics for the model to be able to find the synthetic media. Zhao, Wang, and Lu experimentally observed that deepfake video determination was best possible by their method through careful experiments and comparisons, so they showed the potential of the CNN-based architecture in robust detection of deepfakes. Their study apparently contributes to the state-of-the-art in deepfake detection methods and shows strong indication of the capabilities of CNN-based models against the increase in manipulated media content [30].

To complement the efforts of dealing with the challenges put forth by deepfake manipulation, Agnihotri [31] contributed in his own way to develop an advanced detection method in the area of deepfake by developing a deep neural network-based approach that would help in identifying manipulated media content. Therefore, in their work, the authors try to use a powerful tool just like convolutional neural networks (CNNs) for feature extraction and classification, which would have the potential to reveal deepfake manipulation even at the level of small visual artifacts. The approach used in the methodology of Agnihotri [31] is to train CNN to learn the unique features and anomalies presented by manipulated media, hence empowering the model to properly classify between genuine and synthetic media. Agnihotri thus realizes that with his deep learning-based approach, it realizes that CNN plays a very critical role in the identification and flagging of flagged content with regards to rigorous efficiency, thus underlining the critical role played by CNN in fighting against the spread of synthetic media. As such, the study by Agnihotri is a contribution not only to the great growing body of literature on detection but also demonstrates the necessity of availing increasingly sophisticated and advanced machine learning tools to preserve the integrity of digital media in an age of moving manipulation [31].

Lu et al., [32] introduced a new model for a deepfake video detection system called the 3D-attentional Inception convolutional neural network (CNN) system. They proposed a CNN architecture embedding an attention mechanism into the model, allowing selective focus on salient regions in the input data. Proposed in this method, the unification of the spatial and temporal attention mechanisms proposed has resulted in a much stronger discriminative model that the network can use in identifying anomalies and inconsistencies of features typical of a deepfake video. According to Lu et al., they have experimentally evaluated the effectiveness of their method on the benchmark of deepfake detection and showed its efficiency compared with traditional CNN architectures. They paid special attention to the attention mechanism as a very useful approach in methods for deepfake detection. This work sets one step forward for the development in the area of deepfake detection by providing a novel framework with attention

mechanisms for better performance and a more reliable detector [32].

In the same breath in the fight against deepfake content creation, Johnson [33] unpacked periocular-based deepfake detection with convolutional neural networks (CNNs). He noted that the fundamental focus of their research was to use the specific features around the eye for the sake of detecting synthesized media. Johnson showed that it is possible to train a CNN for extracting periocular features and making a decision based on those features that is, whether an image is genuine or manipulated. The very idea of making use of local features in deepfake detection could hence be applied. The approach here was, therefore, to develop a new, relatively simple state-of-the-art detector for this application, veering off from traditional approaches that focus on the use of global image features only, instead exploring the specialization in feature value might offer in improving detection performance. The contribution of Johnson's work is to further enlarge the existing body of literature on CNN-based approaches for deepfake detection. Particularly, it does offer very important insights on how useful and effective they can be when localized features are used to make robust deepfake detections. In main, Johnson's work is a huge stride in developing some of the newer strategies to fight the proliferation of deepfake content [33].

**2.4 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks**

According to Grossberg [33] RNNs are a class of neural networks that are modeled to work on sequential data by keeping their state information. Unlike feedforward neural networks, processing inputs only in a single pass, R and N are connected forming directed cycles that could have temporal dynamic behavior. Article frames the basic principle of R and N together with their architecture, training methods, and the applications across different domains. This source is useful because it elaborates the dynamics of RNNs and gives a very strong founding of understanding the power of temporal modeling by RNNs and how they can be used for the purpose of deep-fake detection [33].

Further, from the paper "Deep Neural Networks in a Mathematical Framework," Caterini and Chang [34] derive the principles behind Recurrent Neural Networks (RNNs). It gives the mathematical

definitions to understand RNN, with a special definition of fundamental concepts such as hidden states with recurrent connections and gradient propagation through time. Caterini and Chang formalize the mathematical basis for RNN in order to make clear the theory of sequential data processing and temporal modeling. It, therefore, gives way to the study of comprehensive variants like the Long Short-Term Memory (LLSTM) networks, which have been introduced to meet the limitations that simple RNNs have in capturing or modeling distant dependencies from one single time step, like when processing time series data [34].

Kouziokas [36] discusses LSTM networks, with a focus on their application in energy appliance prediction. LSTMs employ memory cells in a specific type of RNN architecture capable of storing information over very extended sequences. As a result, LSTMs are ideal for applications that demand highly accurate interpretation of sequential data. Kouziokas investigates the temporal dynamic modeling capabilities of LSTM networks for electrical appliance use. Continuing with the use of the temporal modeling capabilities of LSTM networks, Kouziokas demonstrates that these techniques may be utilized to account for complicated temporal dynamics, resulting in correct predictions. This demonstrates that LSTM networks are highly adaptable when processing sequential data.

**2.5 Evaluation Metrics and Performance Benchmarks**

On the other hand, Rana et al. [37] carried out a systematic literature review on the detection of deepfakes, alluding to the kind of evaluation metrics and performance benchmarks that are applied. Against such a backdrop, the present study lists the commonly used evaluation metrics, such as accuracy, precision, recall, and F1-score. In fact, these would be the primary indicators of a detection system that correctly recognizes the media contents in authentic and manipulated forms. Also, Rana et al. talk about standardized benchmarks and evaluation protocols that are widespread in the field and insist on the use of uniform methods to carry out the evaluation of detection performance. This paper presents a summary of findings across current literature, challenges, and limitations for the evaluation of detection systems, hence possibly

guiding future research efforts to be more targeted in terms of what needs to be addressed [37].

The issue of evaluation metrics was compared with diversified performance of evaluation metrics that range from different deep learning-based detection methodologies in deepfakes [38]. In fact, it becomes a comparative analysis of diversified performance metrics ranging from different deep learning-based detection methodologies. In this view, Taeb and Chi compare the metrics of accuracy, precision, recall, and F1-score in a systematic way that will offer insights into the relative strengths and weaknesses in approaches to detection. Theirs, therefore, is a recommendation for the consideration of more than one evaluation metric towards making comprehensive judgments about the performance of the systems of deepfake detection and thus making informed decisions towards the selection of appropriate methodologies [38].

In a study by Shad et al. [39], a comparative analysis of CNN-based deepfake detection methods was performed. While in their work they only consider image-based detection, the observation underlines the important point for guidelines that need to be used for performance metrics more generally in the field. Shad et al. discuss the effectiveness of those methods by a comparative perspective of different metrics, such as accuracy, precision, and recall among various CNN-based methods in detection. This, in turn, will further the overall understanding of the evaluation metrics and performance benchmarks in the domain of deepfake detection, which will be beneficial for researchers and practitioners to provide guidelines on how to perform their system assessment in order to test their performance for the task of detection [39].

**2.6 Research gap**

State-of-the-art research in deepfake detection techniques toward its wide and broad perspective shows many big strides made toward robust detection using the latest and most advanced machine learning algorithms, more specifically deep learning models. However, critical research gaps remain to be filled in their avenue despite a lot of research on the subject. One of the important research gaps that need to be addressed is the standardization of metrics used for evaluating and benchmarking the performance of

deepfake detection systems. However, an agreement on standard metrics is yet to be reached, and most existing works commonly report accuracy, precision, recall, and F1-score in their evaluation, precluding direct comparison of detection methodologies. Besides, the diversity in the datasets used in the evaluation further complicates the comparison of performance, since the difference may be large between any two evaluations due to the selection of dataset characteristics.

This fills a much-needed gap in the standard evaluation protocols and benchmarks that would allow for not only fairness but also comprehensiveness in the assessment of deepfake detection systems across varied datasets and scenarios. The other significant research gap is the fact that there is very limited knowledge when it comes to generalizability and the level of robustness of techniques applied to detecting deepfakes from a broader perspective of different types of media content and manipulation methods. Most works in the domain do address only a few kinds of deepfake media, for example, video or images; hence, they might not capture other modalities or even the methods used in other media modalities. The rapid evolution in the methods of deepfake generation, therefore, demands that detection techniques are also constantly adapted and ratified for effective countering of the emanating threats.

Closing this research gap requires comprehensive evaluations of detection algorithms in different media types and under all critical manipulation scenarios, but also developing adaptive detection strategies that can alleviate new forms of manipulation. This further requires research into the ethical, social, and legal implications associated with the detection of deepfakes and its technology. As much as technology advancements in detecting deepfakes are really necessary to minimize risk from synthetic media manipulation, consideration has to be taken at the same time for the wider societal impacts such a technology would bring about. Nevertheless, other ethical issues, such as privacy, the spread of falsified information, and even possible detection algorithm bias, should be looked at more deeply. Beyond that, the legal framework in relation to the regulation of developed deepfake technology and its detection algorithm needs further elucidation towards appropriate governance and accountability.

# Chapter 3
# Proposed Methodology

In an attempt to recognize deepfakes, we propose a method based on a hybrid deep learning technique. We used it in both spatial and temporal data contained in videos, since deepfakes algorithms basically change the facial regions. Our detection focuses on efforts in the face region.

### 3.1.1 Frame extraction

The main step to any deepfake detection technology. The process splits the video down to each frame, making from dynamic video a set of static pictures for further review.

The detail in this idea comes from the fact that it provides us. By analyzing each frame, we could apply different types of advanced image analysis techniques that are not readily available when dealing with videos. Frame extraction in video is a well-designed approach for processing the videos in computer vision, which gives a prerequisite for the applications that may have ranges from object detection to activity recognition.

This is because the deepfake detection has to be done per frame; hence, in the light of visual content, it calls for an in-depth study to extract frames and look into each one of them, as the deepfake algorithm usually introduces small artifact or incongruity in the video at the frame level [4]. Technical Execution: It can be executed with the help of various tools and libraries, such as OpenCV, that allow analyzing images and videos to the maximum extent. Frame can be got back using OpenCV, while looping the video file and taking out each frame, thus turning the video stream into a series of pictures.

The approach eases the problem of detection, limiting it to picture analysis, but, on the other hand, gives way to the usage of strong image processing algorithms that had been researched and polished in the field of computer vision for centuries. Analytical Benefits: The dos in frame extraction in deepfake detection differ. First, it will use Convolutional Neural Networks (CNN) for the excellent performance they have exhibited in picture classification and analysis uniquely for every extracted frame. This becomes

very useful, bearing in mind that CNNs are able to pick out the intricate patterns and aberrations on images, even those that present minor artifacts forwarded during deepfake production processes.

Second, the temporal smoothness and consistency of the motions of the video in between the frames can be quantitatively assessed, which will aid in the detection of the cases where the deepfake has a temporal inconsistency.

### 3.1.2 Face Feature Detection

Most of these methods of deepfake detection are, in fact, focused on the facial features. As such, the identification and, more importantly, the assessment of face regions critical to deepfake detection becomes necessary for an overall effective process of detection.

In this stage, powerful face identification algorithms are used to accurately localize the faces in video frames, so the investigation can be applied over the most manipulation-prone parts of the content. The new face identification algorithms use deep learning methodology to identify the facial characteristics most accurately identified in many other situations that may include the variations in lighting, position, and facial expressions, etc. A striking example of this phenomenon is the use of Convolutional Neural Networks (CNNs), which have been trained on large datasets with immense accuracy in recognizing human faces.

The architecture will be an SSD Single Shot MultiBox Detector with the base of ResNext—an approach that currently demonstrates effective and accurate features of face identification required for other stages of the deepfake detection process [6]. Following the face regions' separation, it applies detailed research over these places for possible adjustments, analyzing the facial expressions, movements, and other small details that are never made by a deepfake video. Human facial expressions are rather complex, which actually makes them extremely useful for the genuineness test: even minor changes point to manipulation. Particularly, among such technologies, one of the key for the current study includes facial landmark localization of human faces. This technology helps in localizing the location of the key features of the

human face, such as the eyes, nose, mouth, and jawline, so that the full summary of a face structure will be given for further analysis. [7]

In fact, it reviews the face to identify tiny characteristics with deepfakes, such as differences in facial expressions, unanticipated motions, or abnormalities in facial landmarks among frames. It is improved in the capability for authenticity detection of the video material with the latest strength in face identification and analysis technologies, fighting major steps against digital impersonation and misinformation.

### 3.1.3 Temporal Feature Analysis:

The main differences among the various deepfake detection are in the understanding of temporal data in videos. Temporal differences that are not apparent in separate frames but get revealed when frames are arranged in a sequence. The difference arises due to frame-by-frame changes that are inherent to deepfake-generating algorithms, which, when and if not done correctly, could bring about incongruities between the appearance and movements of the persons over a series of videos.

However, despite being powerful enough to produce high-quality lifelike photographs, the deepfake algorithms still remain inconsistent with the facial dynamics over some time. This is further compounded with the very delicate and pinpoint movements of human facial expressions, many muscles of the face work in unison. Even the slightest change in this movement shows that the movements are not natural and could indicate the presence of a deepfake. Temporal feature analysis is hence a very key aspect in the detection of deepfakes.

Many networks have been designed to tap into this temporal information, some of them including the Long Short-Term Memory (LSTM) networks. LSTMs, especially, are powerful in learning trends from a data sequence; therefore, they serve as a very crucial model in the prediction of data. [9] This makes the LSTMs be capable of finding the irregularities in the sequence of facial reactions and motions. Consequently, it is a very accurate method of determining changed information.

This method will help in detailed spatial feature analysis of video data since the emphasis is on the temporal component of the video data, hence making video data assessment more accurate. The combined model, integrating Convolutional Neural Networks (CNNs) for spatial analysis and Long Short-Term Memory (LSTMs) for temporal analysis of features, is therefore placed in an ideal position to offer a complete solution to the deepfake detection problem.

### 3.1.4 Hybrid Model Integration:

In this paper, a hybrid deep-fake detection model was presented that efficiently couples Convolutional Neural Networks (CNNs) with Long Short-Term Memory-2018 LSTMs, and hence, extracts the benefits provided by spatial and temporal data present inside the videos. The chosen type of CNNs is very good at detecting fine visual components from images, hence perfect for detecting small anomalies presented in deepfake manipulations very frequently. On the other hand, LSTMs do a pretty good job in sequence analysis, which implicitly captures the consistency of it in facial expressions across video frames [9]. Their potential of recognizing long-term dependency is important in case of an anomaly pointing to a certain manipulation done intentionally.

This employs our approach that covers both the parts of CNNs and LSTMs, making it an effective architecture that largely advances the scope of deepfake detection by considering the visual and behavioral part of video data. Such combined models are thus promising to the task of video classification and deepfake detection on account of being more effective at differentiating between real and manipulated data. [8] [10] This model, therefore, improves the detection accuracy by further adapting to the more complex nature of deep fakes through combining spatial analytic skills of CNNs with the sequential data processing abilities of LSTMs.

### 3.1.5 Tool used in creation of deepfakes

***Faceswap:*** Faceswap is an open-source deepfake tool that allows users to swap faces in images and videos. It utilizes deep learning techniques, particularly Generative Adversarial Networks (GANs), to

generate realistic facial swaps. Faceswap typically involves several steps: face detection and alignment, feature extraction, face swapping, and blending. The tool automates much of this process, making it accessible to users without extensive deep learning expertise. Users provide source images/videos containing the face they want to swap and the target face onto which they want to swap it. Faceswap then uses its trained models to perform the face swap, adjusting features such as expression, lighting, and angle to match the target face.

***Faceit:*** Faceit is another deepfake tool that specializes in face swapping. It offers features similar to Faceswap but may have different algorithms or user interfaces. Like Faceswap, Faceit employs deep learning techniques such as GANs to generate realistic face swaps. Users provide input images/videos of the source and target faces, and the tool handles the rest of the process. Faceit may offer additional features or customization options depending on its specific implementation and development.

***DeepFaceLab:*** DeepFaceLab is a popular deepfake software suite that provides a comprehensive set of tools for creating high-quality deepfakes. It supports various deep learning architectures, including GANs and autoencoders, for generating realistic face swaps. DeepFaceLab offers advanced features such as model training, fine-tuning, and post-processing to enhance the quality of generated deepfakes. Users can train their own deep learning models using DeepFaceLab or utilize pre-trained models for face swapping. The software provides a user-friendly interface for performing complex tasks such as data preparation, model training, and face manipulation.

***Deepfake Capsule GAN:*** Deepfake Capsule GAN is a deepfake tool that specifically utilizes GANs for generating realistic face swaps. GANs consist of two neural networks, a generator, and a discriminator, which are trained adversarially to produce convincing fake images. Deepfake Capsule GAN may incorporate specialized GAN architectures or training techniques optimized for generating high-resolution and realistic deepfakes. Users input source and target images/videos into Deepfake Capsule GAN, and the tool leverages its GAN-based models to produce the desired face swaps with high fidelity.

***Large Resolution face masked:*** Large Resolution face masked likely refers to a technique or tool for handling high-resolution images or videos in the context of deepfakes. Deepfake generation often involves working with large and detailed images or videos to maintain realism and quality in the resulting deepfakes. Tools or techniques for handling large resolutions may include optimizations for memory usage, processing efficiency, and scalability to handle high-resolution input data effectively.

These tools provide users with the means to create deepfakes by leveraging deep learning algorithms, particularly GANs and autoencoders, to swap faces in images and videos. They automate various steps of the deepfake creation process, including data preprocessing, model training, face swapping, and post-processing, making deepfake creation accessible to users with varying levels of expertise. However, it's essential to use such tools responsibly and ethically, considering the potential implications of creating and sharing manipulated media.

**3.2 Basic Model Architecture:**

**3.2.1 RESNEXT**: In this landscape of the convolutional neural network architectures (CNNs), ResNext becomes a seminal one in proposing an innovative way of modeling efficiency and scalability. Designed by Xie et al. (2017), ResNeXt exploits the so-called cardinality principle: the size of the set of transformations, providing yet another new dimension to network depth, besides scaling width. ResNext, in its design philosophy, hence achieves greater accuracy for complex picture recognition tasks, since its design philosophy does not increase processing complexity correspondingly.

A deepfake video deals with heavy face editing to yield realistic video, which is in reality a phony. A model must be able to discern that type of manipulation to understand the fine-grained spatial changes within pictures. The ResNext architecture, using grouped classes, can strike a perfect balance between the cardinality to extract features completely from face pictures. This ability is very critical for recognizing small differences signaling that a video is a deepfake. Moreover, its efficiency makes it possible for large numbers of frames to be processed within time limits that are acceptable, such as those considered

acceptable by Swain and Ballard. These should be very short a very important point because video analysis implies huge data. He et al.  demonstrated that deep residual networks resulted in architectures like ResNext and were some of the most effective, specifically for the learning of complex visual data.

It goes further to elaborate on this principle by providing improved performance in patterns and abnormalities' detection on visual data, most necessarily deepfake. On the question of the model's detection, ResNext will, in fact, work by breaking down each frame individually, taking the high-level data related to facial emotions, geometry, and texture into consideration. The base of determining likely manipulations is provided here by these attributes since deepfake algorithms usually track real movements of the face of a person but track a little off or miss some key points. Able to handle and correctly analyze high-dimensional data, ResNext provides one of the most potent tools from the arsenal of early-stage research, aimed at extracting and identifying possibly altered areas inside the frame of the video.

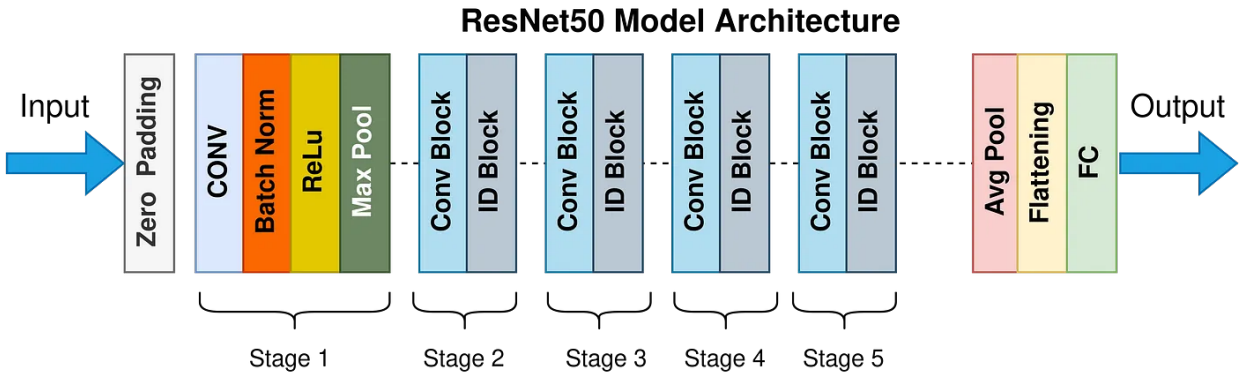


*Figure 1:ResNext50 Model Architecture [21]*

### 3.2.2 Convolutional Neural Networks:

Convolutional Neural Networks (CNNs) are a kind of deep learning model used to best analyze data that has a grid-like topology, e.g., pictures. Generally, CNNs identify and create hierarchical patterns automatically in data from basic features, such as textures and edges in the first layers of presentation, to complex object features in deeper layers. Adding convolutional layers, pooling layers, and fully connected layers makes it easier for the network to learn features in a hierarchical manner and hence turns raw pictures

into class scores for many tasks, including classification. Another area where CNNs can find wide utility is the detection of deepfakes, for they are perfect to catch small disparities and distortions of visual information that possibly indicate changes or fakes. CNNs can differentiate the deepfake images by face features and textures that are often affected in the case of deepfakes. This makes them an important weapon in the battle against digital disinformation.

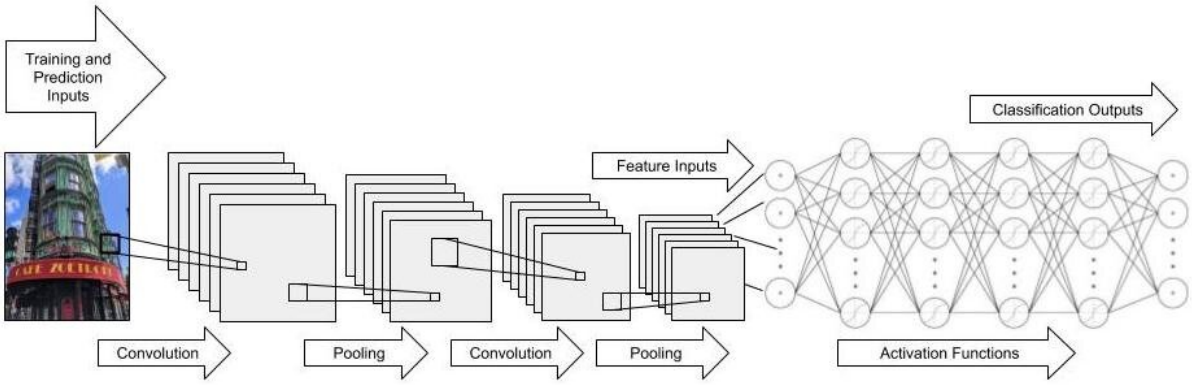


*Figure 2: Convolutional Neural Networks [22]*

### 3.2.3 LSTM (Long Short Term Memory):

A significant leap in RNN designs was through the invention of LSTM (Long Short-Term Memory) networks by Hochreiter and Schmidhuber. LSTMs have been designed in such a way as to particularly resolve the vanishing gradient problem of plain RNNs, so that it can learn long-term dependencies in sequential data. This will allow it to benefit from temporal context, which could prove quite invaluable in applications such as voice recognition, time series forecasting, and definitely deepfake detection.

Greff et al. delve deeper into the study of the structural basis of LSTM by articulating their proceedings and practicality in a sequence of models. Flexibility and effectiveness in intricate temporal analyses therefore underline that LSTMs have a potential application to the devious deepfake challenge, which requires recognition of tiny changes across time [11].

The deepfake classifies in a sequence of video frames by a sequence of features first extracted from the video frame and then tests the temporal development of facial emotions and motions. LSTMs have

great utility to identify when footage has been tampered with as they capture temporal patterns very well and note deviations from human real behavior.

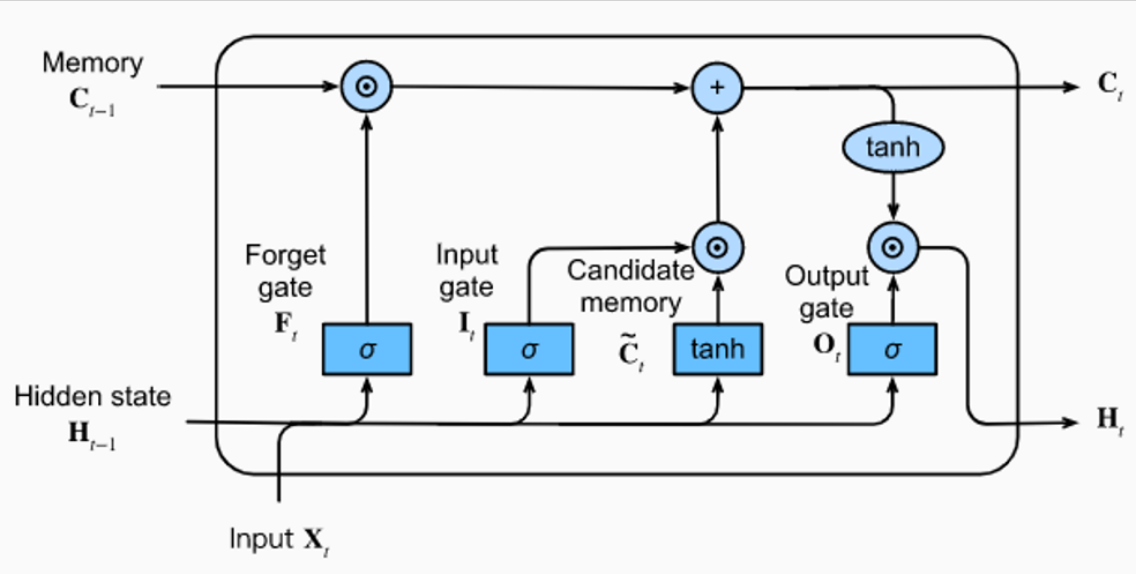


*Figure 3: Long Short Term Memory Architecture [20]*

### 3.2.4 INCEPTIONV3:

Szegedy et al. (2016) proposed a sophisticated architecture, namely InceptionV3, which aims at better optimization of computing performance and model accuracy by providing symmetric and asymmetric combinations of parts of buildings. Convolutional, average pooling, max pooling, concatenation, dropout, and fully connected layers are some of the included parameters. What indeed differentiates InceptionV3 is its novel architecture: much increased network depth and width at a given computation budget, with a corresponding increase in processing cost, notably through factorized convolutions and efficient approaches to grid size reduction. This particularly applies to deepfake detection, as it means InceptionV3 has the ability to provide rich representations at many scales. On the other hand, InceptionV3, with its depth and well-architected design, can conduct a thorough or detailed analysis of the slightest details in the facial features and expressions, commonly tampered with in deep fake videos, and pick out any anomaly or aberration that would indicate editing. Add to this the capacity

of the model to cope with high-dimensional picture data and the higher performance of the model in image recognition tasks, as documented by Szegedy et al., it supports the importance of the model for the rapidly developing discipline of digital forensics that provides effective ways to sophisticated deepfake detection [19].

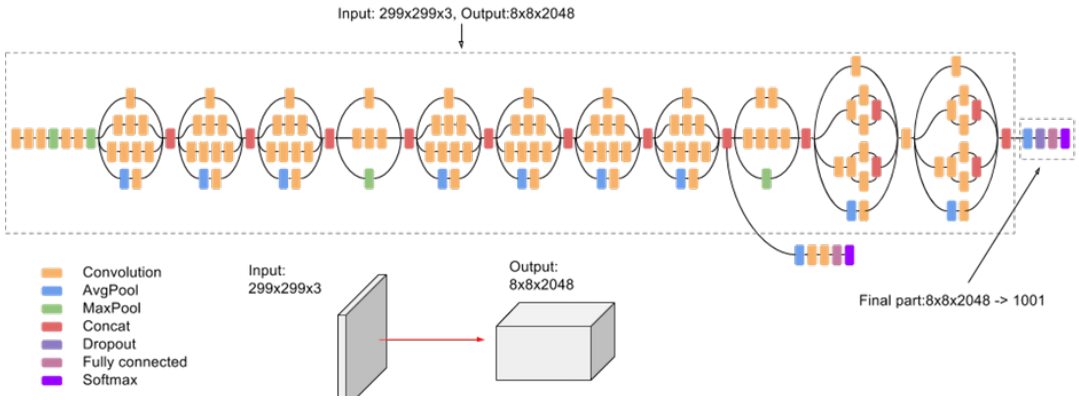


*Figure 4: INCEPTIONV3 Architecture [23]*

### 3.3 Proposed Model Algorithm

LSTM networks are justified in their use with feature selection due to their ability to detect subtle patterns and time dependencies relevant to deepfake manipulation. The technique focuses on data preparation to extract facial information from video frames and then employs a hybrid model that blends long-short-term memory (LSTM) and convolutional neural network (CNN) architectures to increase identification accuracy. The model offers a powerful solution for identifying severely manipulated videos by combining robust feature extraction capabilities with temporal analysis and dropout to prevent overfitting.

#### 3.3.1: Justification of using LSTM with Feature selection

In the deepfake detection landscape, it is, therefore, very interesting to point out how LSTM (Long Short Term Memory) networks, with feature selection techniques, present one of the very competitive approaches to capturing time dependencies and subtle patterns that might be an indicator of manipulation

within a deepfake context. This is more appropriate for sequential data where the problem of the vanishing gradient is either mitigated or not as bad as with traditional RLSTM architectures. It is due to this fact that the LSTM network, suitable to be used for the analysis of video content frame by frame, would allow long-term memory that has the temporal context captured in every step. This makes it, therefore, very relevant for the context of detecting deep fakes, whereby temporal consistency is relevant in giving a clue. In this way, with the learned through sequence video frame-extracted features, such LSTM networks are able to pick up the temporal irregularities that might suggest the existence of some deepfake manipulations.

This helps to further enhance the discriminative ability of the LSTM model-based approaches. The marked and ranked features are those pertinent for deepfake detection in an FSM-based approach. Feature selection approaches can mitigate the curse of dimensionality, like the use of principal component analysis (PCA) or methods based on mutual information, which reduce the space of all features but have essential information relevant in deciding whether the content is authentic or manipulated. With this focus, the LSTM networks are liable to be more efficient and effective in identifying such subtle cues associated with deepfake videos, which would naturally enhance the accuracy and robustness of detection.

To add, such an application of LSTM combined with feature selection adheres to the basic philosophy of devising interpretable and, in turn, effective deepfake models. In so doing, the model not only reduces the dimensionality for computation but also offers the opportunity for the human interpreter to view a subset of discriminative features that make up the underlying factors contributing to the classification. This is a very essential factor of interpretability. Formulating trust and comprehension in the deepfake detection systems allows the user to make the point of understanding the reason for the outcome, which might enable them to identify some patterns or new indicators for the manipulation. This, in essence, represents the synergy between LSTM and feature selection, which means some of the best features from either or both methodologies will be exploited for deepfake detection.

To propose a hybrid model that assures detail in the analysis of video content, the algorithm is

partitioned into the following steps:

### 3.3.2 Data Preprocessing

Data pre-processing is nothing but the process of cleaning, transforming, and organizing raw data into a proper format before carrying out the next set of activity regarding analysis or modeling. Here, in this project, video data is pre-processed so that it is free from all kinds of superfluous noise and only facial features get highlighted. This preparation consists of mentioned important steps: First, each video is broken down frame by frame to be able to detect and extract the human face from the frames. Cropping is done on the regions detected thereafter so that only the facial region is visible.

The result of this first process is thus a dataset of cropped frames rebuilt into a new video sequence, resulting in a refined dataset, where the only frames present in the original are those of facial footage. From this, the frames where a recognizable face is not found are excluded in this revised dataset. The criterion for frame count is chosen in such a way so as to keep uniformity throughout the data set, determined by the average frame count across all movies and further subjected to the computational capabilities available. Considering the fact that the computational expense of a traditional video lasts 10 seconds and has 300 frames, running at 30 FPS, due to our constraints in computing resources, we had to put an upper limit of 150 frames for each film.

This is to befit the processing capability of our GPU in the test environment. Updated dataset with only the first 150 frames from all the videos. For Long Short-Term Memory, the model needs to take all the frames in the sequential order of the video. Instead, we have taken the first 150 frames in the process of creating a new video out of the frame. New video is saved in 112x112 pixel resolution with 30 FPS of playback rate.

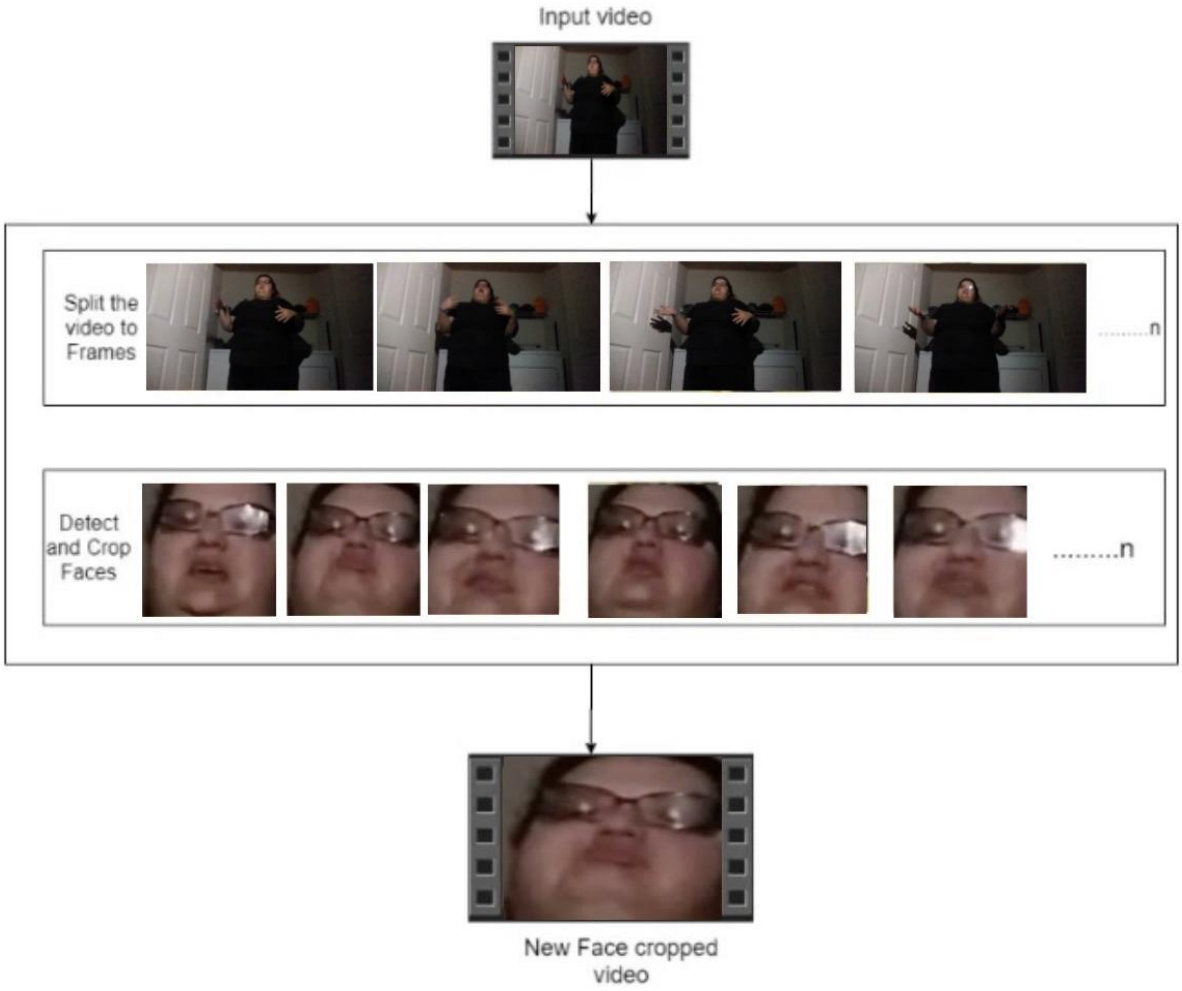


*Figure 5 Preprocessing*

### 3.3.3 Model Creation: Hybrid Proposed ResNext + LSTM

In order to advance the methods of the field for deepfake detection, I propose a novel model that integrates the architecture of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Central to the model is the use of pre-trained ResNext CNN model where resnext50_32x4d variant has been proven to be the most efficient architecture for extracting frame-level information from videos. Taking full advantage of the ResNext architecture, this feature fine-tunes the performance for deeper neural networks and models trained on vast datasets, thereby being able to extract effectively the salient features of concern within a video frame.

After the initial stage of feature extraction with the ResNext CNN, an LSTM network is used to perform sequential analysis on the extracted 2048-dimensional feature vectors. Video frames can be temporally analyzed using the LSTM architecture, which has one layer with 2048 hidden dimensions and 2048 hidden layers. In order to prevent overfitting during model training, dropout is introduced to the LSTM with a probability of 0.4. This allows the model to become more resilient and generalized. With the LSTM model, it would be feasible to compare the two frames at different times, allowing for even smaller alterations that would indicate deepfake manipulation.

The model also includes other layers, such as ReLU activation function layer. These enables it to be capable of learning the relationship between input features and output classification which can identify weather it is a deepfake video or a pristine video. The average pooling layer is adaptive and reshapes output into the necessary format. Because a SoftMax layer measures the degree to which the model is certain of its categorization, it is a useful indicator of the prediction performance of the model. To put it succinctly, the creative model combines the strong feature extraction capabilities of CNN with the temporal analysis capabilities of LSTMs to offer a powerful solution for the identification of deeply fabricated videos.

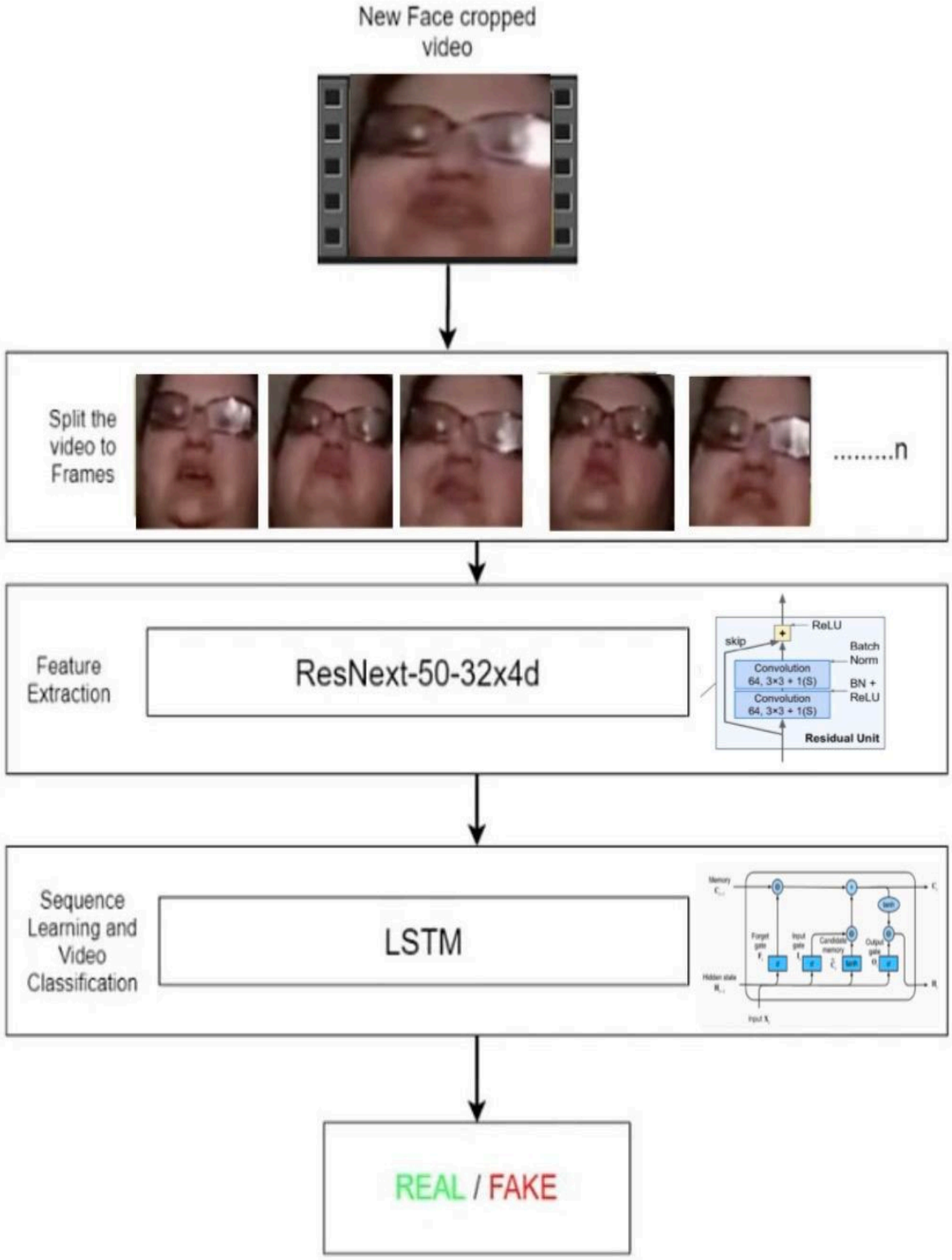


*Figure 6: Model overview*

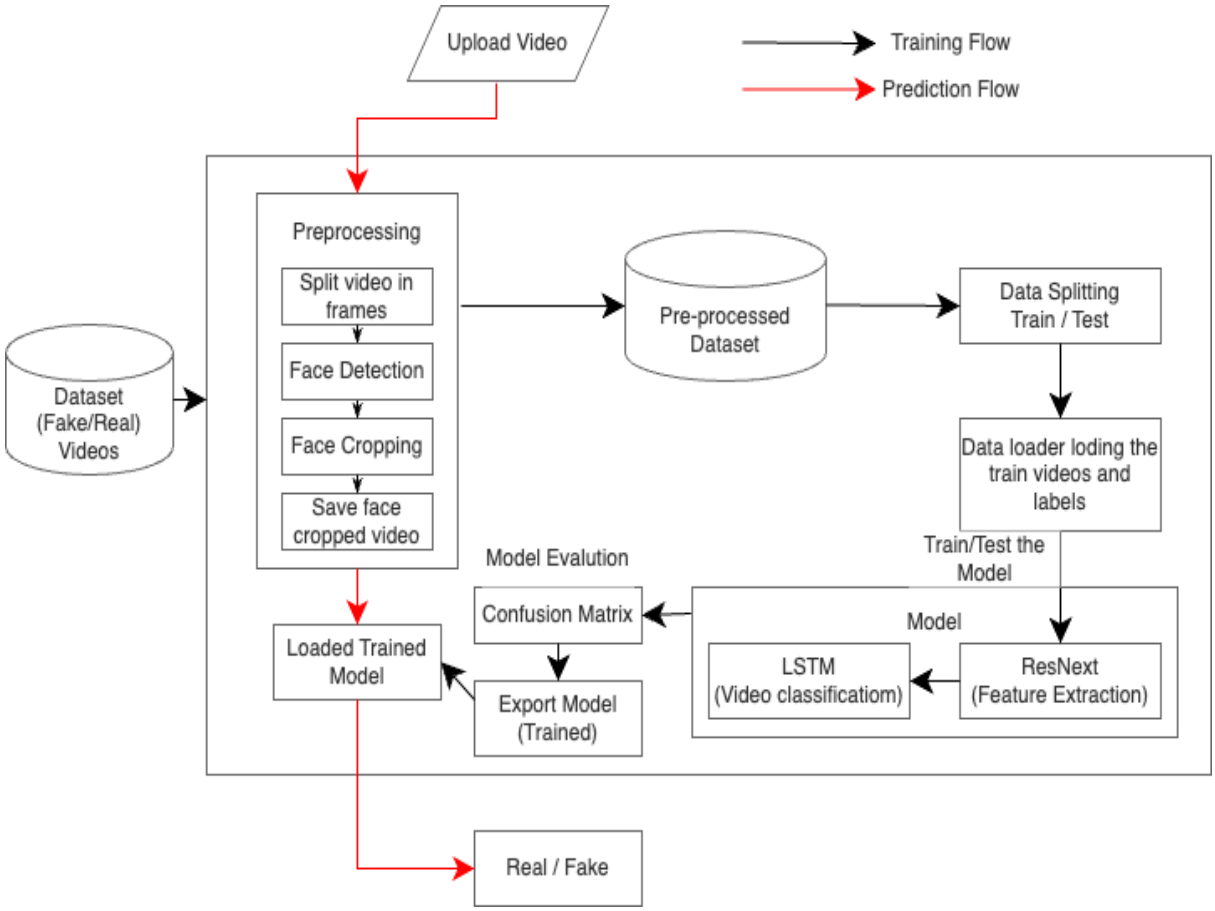


*Figure 7: Model Architecture*

# Chapter 4

## Experiments and Results

This chapter describes the research of deepfake detection in detail and uses a number of deep learning architectures and datasets for the training, validation, and test experiments. These deepfake videos, made with the use of cutting-edge AI technology, manipulate both visual and auditory elements. In this work, three of the most popular datasets FaceForensic++, Celeb-DF, and the Deepfake Detection Challenge (DFDC) are used to train the models by evaluating them using several model architectures. This includes the effective evaluation of models such as Hybrid of ResNext and LSTM), CNN, LSTM, RNN, and InceptionV3 + LSTM against all major performance metrics including accuracy, true positive rate, true negative rate, false positive rate, and false negative rate. This chapter rigorously analyzes and visualizes the hybrid model, reaffirming the efficacy of the hybrid model in attaining very high accuracy levels and balanced detection capability across diverse datasets. Moreover, it will provide enlightenment on the strengths and limitations of individual architectures, which open ways for further optimization and refinements to boost the reliability and robustness of the deepfake detection system in real-world applications.

### 4.1 Data description

G In order to move our fine-tuned model for the best efficiency and resilience in deepfake detection towards real-time prediction, we choose to do so by leveraging the most recent source of real-world deepfake video data. The Deepfake Detection Challenge (DFDC), FaceForensic++, and Celeb-DF videos are where we gathered the dataset. This was a very deliberate strategy to guarantee exposure to a variety of deepfake scenarios and, consequently, the ability to enable them to distinguish between manipulated and genuine video. Additionally, we evenly distributed both actual and false videos inside the dataset to ensure that the file for Python minimizes any potential bias that may arise during the model training process. The purpose of this approach was to guarantee fairness and equity in the training outcomes by

preventing the model from selecting any classes that had a bias of any kind over the others.

We introduced subsequently a model that can equally differentiate between efficient faked content from real content across various scenarios of 50% real and 50% fake video distribution. The DFDC dataset extended valuable insights to cases of deepfake modalities, such as audio-altered videos, but it had to be kept in mind that these were out of the scope of the present study. To facilitate better analysis on the deepfake detection based on videos, we came up with a Python script used in preprocessing the DFDC dataset by filtering all films with a modified audio. This will ensure that our developed model remains poised with the study objective of being able to focus on the video-based deepfake detection process only. After the preprocessing of the DFDC dataset, we, therefore, curated a dataset of 1500 real and 1500 deepfake videos, making sure that it retained a balanced representation of the classes. The approach also includes videos from the FaceForensic++ and Celeb-DF datasets, which cumulatively add another 1000 real and 1000 fake, along with 500 real and 500 fake from Celeb-DF.

This comprehensive strategy of dataset compilation has enriched the model training of ours with learning from diverse scenarios of deepfakes and optimizing the detection capability for real-world applications.

**4.1.1 FaceForensic++:**

This has been bolstered by the rapid development of deep learning algorithms that have contributed to the growth of "deepfake" videos, in which human faces are altered with the intention of deceiving a viewer. This poses a serious challenge in data face identification, for which a reliable system is required in this line of digital forensic work.

This dataset was of critical importance as a benchmark for the deepfake detection system development and evaluation in their paper "FaceForensics++" by Rössler et al. "FaceForensics++ expands significantly on the prior version with the addition of a much larger amount of real and manipulated films, resultingly creating a much larger, multifaceted, and diverse dataset for identification of deep fakes. The

dataset includes four techniques of modification: Deepfakes, Face2Face, FaceSwap, and NeuralTextures, well represented in terms of how to go about doing the changes.

In fact, the diversity of methods used often means that deep-learning models trained on FaceForensics++ fail to generalize and apply to other deep-fake creation methods. This dataset points to the necessity of such data. The FaceForensics++ dataset has now become a new standard for use in digital media forensics, and it has been predominantly used for training deepfakes and detecting altered movies by convolutional neural networks (CNNs) and other deep learning architectures.

Its comprehensive nature, therefore, allows the development of algorithms that can tell authentic from fraudulent content with very high accuracy; clearly pointing in the right direction to urgently develop tools in the fight against the spread of disinformation in defense of digital integrity.

This dataset would enable one to investigate the impact of diversity in video compression rates on the effectiveness of the detection algorithm and, therefore, show some of the robustness that characterizes such approaches when deployed under such real-life situations.

**4.1.2 The Deepfake Detection Challenge:**

The Deepfake Detection Challenge (DFDC) is a large-scale dataset designed within Kaggle, widely known for driving the development and benchmarking of methods in deepfake detection. The dataset was developed in support of the campaign to prevent the spread of deepfake videos. Deepfake is a kind of synthetic media in which a person in an existing picture or video is replaced with the likeness of someone else by means of artificial neural networks. They encourage this DFDC dataset to be exploited by academia and engineers as an important resource in developing automated algorithms for detection of such synthetic creations, given deepfakes huge potential to be used in malicious applications, including disinformation, identity theft, and harassment.

The DFDC data, one of the biggest and most comprehensive freely available labeled datasets for deepfake videos, include both the most extensive sets of diverse deepfake variations gender, race, age, and

more that have ever been released to the public and represent the final DFDC datasets. It has this rich variety that makes it irreplaceable in the cases of building detection systems that work across different demographics. Videos created cover all sorts of generation methods of the deepfake, making the dataset handy in the training and testing of detection models against different types of deepfake methods. Size and Scale: The DFDC dataset consists of hundreds of thousands of motion pictures, partitively divided into two sets - testing and training. The actual number of videos may vary in the case of an update to the dataset.

Each of the videos in the dataset is labeled either "real," meaning the video contains the real footage of the actor, or "fake," indicating the footage in the video has been manipulated to make the real actor's face appear. This results in a binary label, further reducing efforts in binary-model development for deepfake detection.

Diversity: This dataset proposes great diversity in background, lighting conditions, and movements of heads. Such variability contributes to the training of strong detection models which may act effectively in a large number of situations.

### 4.1.3 Celeb-DF

The Celeb-DF dataset represents an extraordinary leap in the development of data creation and benchmarking for detection algorithms of such issues. Celeb-DF has been designed to put forth an authentic benchmark to the hard task of deepfake detection. This illustrates a huge data corpus of deepfake videos that include celebrities made using very advanced video editing and synthesis methods to derive high-quality deepfake material. So, some basic aspects of Celeb-DF, which make it absolutely indispensable for the field, refer to the size of the dataset, its diversity, and the quality of those deepfakes. There are 590 real videos and 5639 DeepFake videos in the Celeb-DF dataset, which is the equivalent of more than two million frames of the video. All videos from the pool are of average length, 13 seconds, and with the default standard frame rate of 30 frames per second.

Its latest version, Celeb-DF (v2), includes a large number of movies in the dataset, hence making

the value of films meaningful for deep-learning-based detection models.

Number of Videos: This dataset contains 5639 deepfake videos and 590 real videos. Which is over 2- million video frames.

Video Quality: One of the main differences in Celeb-DF is that it is based on high-quality video production. Previous datasets had strong artifacts or clear errors observable in deepfake videos, while Celeb-DF zeroes in on producing deepfakes very hard to be distinguished from the original, real footage, thus forming a nice testing bed for detection algorithms.

These video datasets bring in celebrities of all kinds: gender, age, and race. Celeb-DF is varied enough in all attributes so that the developed model should be generally applicable to other sets of the population and should not introduce bias toward some specific attribute.

Annotations and Labels: each video of the Celeb-DF dataset was annotated, labeled carefully with relevant metadata, and given a well-designed title of 'Real' and 'Deepfake'. These ease the training and testing of machine learning models by allowing exact measures of the performance of developed algorithms.

The Celeb-DF dataset is of great advantage to scholars and professionals in studying the detection of deepfake movies. Growing prevalence is given to its everyday life relevance, which demonstrates the growing deepfake material prevalence on social media and other digital platforms with massive threats to information integrity, privacy, and security. The high quality and diversification of deepfake videos in Celeb-DF will help to build an effective and robust way of detection, constraining the risks that would be linked with technology in deepfakes.

**4.2 Hyperparameter tuning:**

Hyperparameter tuning, thus, becomes an important stage in the machine learning process, involving the tuning of settings for an algorithm decided upon before training and not learned from data. Such choices, or hyperparameters, as learning rate and the number of layers in neural networks, define the

model performance. The search is going to look for the best combination of hyperparameters by using techniques like grid search, random search, or even the use of Bayesian optimization so as to acquire maximum accuracy and efficiency of the model in validation data [24].

The model parameters are used with Adam [21] optimizer for an adaptable learning rate. Learning rate to improve on the global minimum of gradient descent is set at 1e-5 (0.00001), and the factor for weight decay is set at 1e-3. Because this is a classification issue, the loss cross-entropy method is applied.

| Hyperparameter | **Trial 1** | **Trial 2** | **Trial 3** | **Best Result** |
|---|---|---|---|---|
| Learning Rate | 1.00E-03 | 1.00E-04 | 1.00E-05 | 1.00E-05 |
| Number of Layers | 4 | 6 | 8 | 8 |
| Batch Size | 270 | 175 | 128 | 1000 |
| Dropout Rate | 0.2 | 0.3 | 0.4 | 0.3 |
| Validation Accuracy | 0.82 | 0.87 | 0.89 | 0.89 |

*Table: 1 Hyperparameter tuning results*

### 4.3 Performance Metrics

Performance measurements are just some forms of evaluation to see how good the machine learning model is actually performing. Performance measures of the model are used in such a way that the categorization task uses very distinct kinds of metrics to measure performance. In the case of classification tasks, like deepfake detection, one can use various commonly used metrics:

### 4.3.1 Confusion Matrix:

It is an important tool in the assessment of classification algorithms, mostly for the detection of deepfakes. It comes in the form of a table that can show if the model's prediction with respect to true labels is correct or incorrect and brings along with it more information on model performance than accuracy.

It makes the predictions classified into four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). That should allow for every sort of error any model is likely to make. By interpreting the confusion matrix of the deepfake detection, you could reason about the model's capability to distinguish between authentic and fake movies at various thresholds. For example, an elevated false positive rate can indicate that the model is becoming overly sensitive and is misclassifying real movies as fake ones. On the other hand, high false-negative values will imply that the model has failed in detecting the deepfake media integrity.

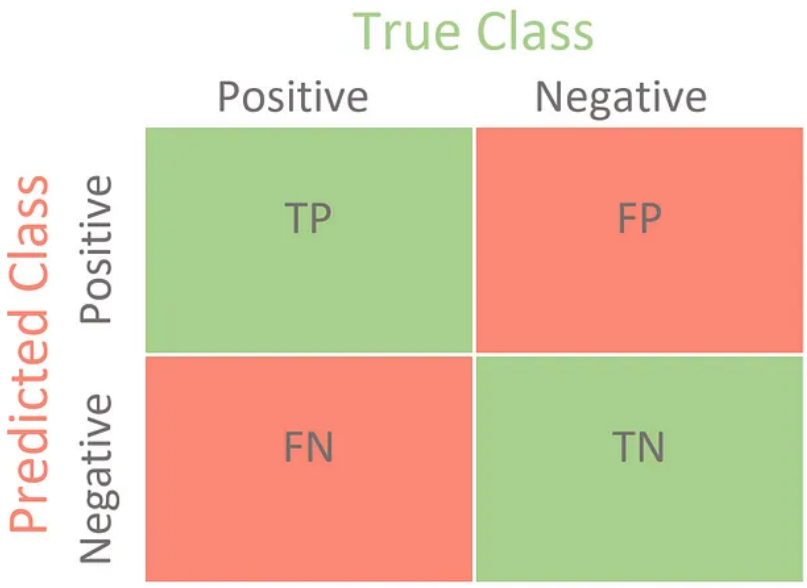


*Figure 8: Confusion Matrix*

True Positive (TP): The model correctly predicts a video as a deepfake.

False Positive (FP): The model incorrectly predicts a real video as a deepfake.

True Negative (TN): The model correctly identifies a video as real.

False Negative (FN): The model fails to identify a deepfake, wrongly classifying it as real.

### 4.3.2 Accuracy:

Accuracy is a straightforward indicator for determining the overall effectiveness of a deepfake detection model. It measures the proportion of right predictions (including true positives and true negatives) among all predictions made. The formula for calculating accuracy is given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} [17] \quad (1)$$

While a high accuracy rate is desirable, indicating that the model performs well across both real and deepfake videos, reliance solely on accuracy can be misleading, especially in datasets where there's an imbalance between real and fake samples. For instance, a model could achieve high accuracy by predominantly predicting the majority class, but this wouldn't necessarily reflect its effectiveness in identifying deepfakes accurately.

### 4.3.3 Loss Functions:

Loss functions are one of the components that help develop deepfake detection models in that they enable giving a quantitative view of how well a model is doing in the course of training. It measures the error of the expected outputs of a model from labels of actual input data. To put these taming mistakes in a nutshell, the training procedure is aimed at taming these mistakes with a view to increasing acuity in the model, more so, their distinction between an authentic and manipulated video.

The choice of a loss function will obviously depend on the nature of the specific classification task and characteristics of the output. Common loss functions used in deepfake detection where the input videos are to be classified into a binary category real or fake include:

Binary Cross-Entropy Loss: This loss function is best used with binary classification problems. It gives the distance of the probability distribution for model-predicted labels with respect to the true

distribution of the labels. For a detection task, this metric would measure model performance in accurately differentiating between genuine footage and deepfakes, penalizing for making differing predictions than the true labels.

$$BCE = -\frac{1}{N}\sum[y * ln(p(y)) + (1 - y) * ln(1 - p(y))] \quad (2)$$

**Equation:**

$BCE$: Binary Cross-Entropy Loss

$N$: Number of data points

Σ: Summation over all data points

$y$: True label (0 or 1)

$p(y)$: Predicted probability of the label being 1

$ln$: Natural logarithm (you can use log with any base, but natural logarithm is typically used)

This equation essentially calculates the average of the loss across all data points. The loss for each data point is determined by two terms:

The first term calculates the penalty for incorrect predictions of class 1 (y = 1). It multiplies the true label (y) by the logarithm of the predicted probability (ln(p(y))). If the prediction is correct (p(y) is close to 1), the term becomes close to 0.

The second term calculates the penalty for incorrect predictions of class 0 (y = 0). It multiplies (1 - y) by the logarithm of (1 - p(y)). If the prediction is correct (p(y) is close to 0), the term becomes close to 0.

Mean Squared Error (MSE): Most commonly used in regression tasks, the MSE can find application within classification contexts to find the difference between the average squared estimated value and the actual value. In deepfake detection, this might help further tune the model predictions if output for such measurements is probabilistic in nature.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2 \qquad (3)$$

**Equation:**

$MSE$: Mean Squared Error

n: Number of data points

Σ: Summation over all data points ($i$ = 1 to $n$)

$y_i$ : Actual value for data point $i$

$\overline{y}_i$ Predicted value for data point $i$

This equation calculates the average of the squared differences between the actual values ($y_i$) and the predicted values ($\overline{y}_i$) for all data points. Squaring the differences ensures the result is always non-negative and emphasizes larger errors.

A lower MSE indicates a better fit for the model, as the average squared difference between predictions and actual values is smaller. In the context of deepfake detection, this can help assess how well the model's probabilistic outputs ($\overline{y_i}$) align with the actual labels ($y_i$) of real footage versus deepfakes.

If accuracy shows the sum, then loss function details the difference: error margins for the model. If the loss is monotonically decreasing over the training epochs, this will be a sign that the model is learning well and perfecting its ability to identify what should be original and what is tampered with. The stagnation or increase of this loss might signal overfitting, underfitting, or other further steps to be taken either with the model architecture or with the training process.

The good approach, deepfake detection model, shall be, besides having high accuracy achieved, robust and generalizable enough to be realized with high performance, even on videos not seen during training. The loss should thus be minimized in order to assure that the model does indeed classify well, not just the training data but also across the diversified and potential more challenging real-world datasets.

## 4.4 RESULT ANALYSIS

### 4.4.1: Performance analysis of CNN model for FaceForensic++

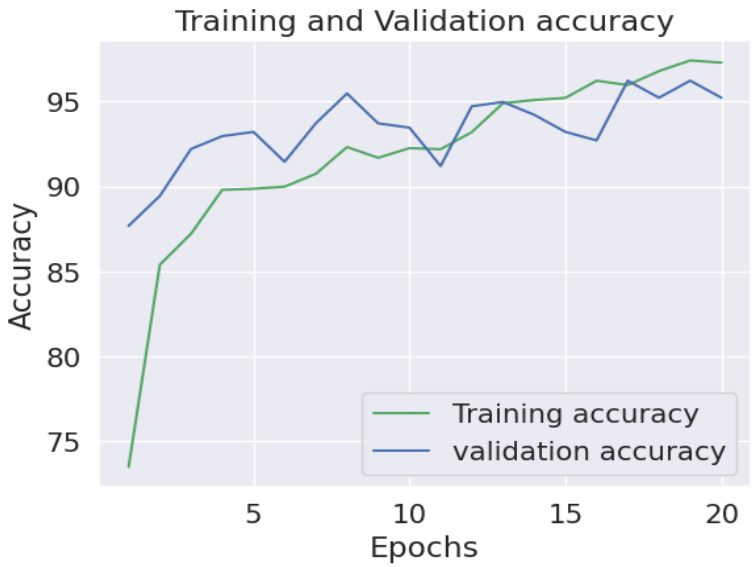


*Figure 8: Validation accuracy CNN*

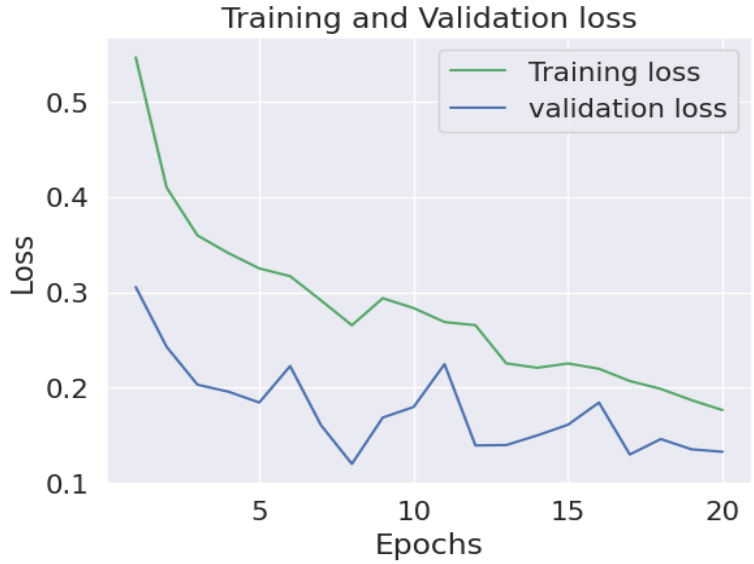


*Figure 9: Validation Loss CNN*

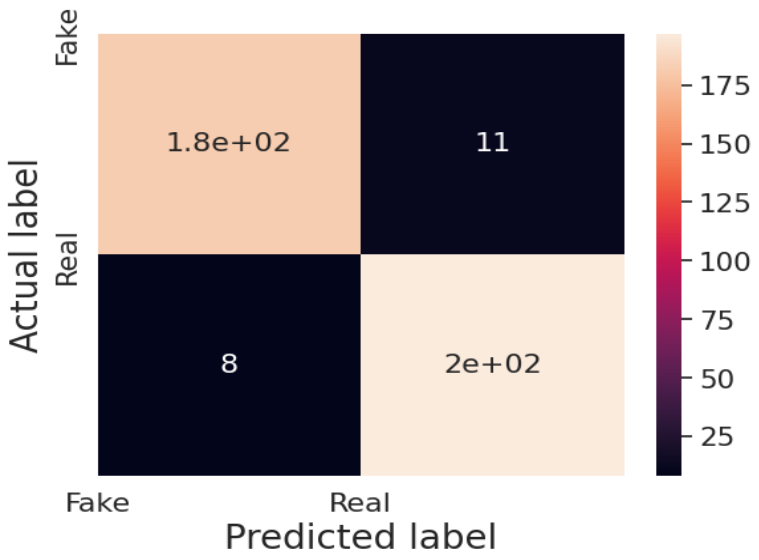


*Figure 10: Predicted label CNN*

The CNN model's performance on the FaceForensic++ dataset was evaluated and found to be quite accurate, with a computed accuracy of 95.23%. This proves that the CNN model successfully differentiates between the dataset's real and altered films. The model shows strong performance in accurately identifying real and deepfake films, with a low amount of false positives (11) and false negatives (8). The high accuracy indicates that the convolutional neural network design is good at predicting outcomes by identifying important features in the input data. The CNN's overall performance in deepfake detection is enhanced by these results, which demonstrate its capacity to distinguish minute visual clues indicative of manipulation. The model's consistency in correctly detecting both types of films is further supported by the very even distribution of true positives and true negatives.

Applying the CNN model to the FaceForensic++ dataset in particular reveals its exceptional performance as a deepfake detection solution, according to the given results. With its low false positive and false negative rates and excellent accuracy, the model clearly works when it comes to telling real videos apart from altered ones. According to these findings, CNN-based architectures are capable of accurately classifying deepfake manipulations by capturing the spatial properties that are characteristic of

such manipulations. The features of deepfake movies can change depending on the environment, thus it's vital to keep that in mind when applying the CNN model to other datasets or real-world situations. Still, we can learn a lot about how CNN architectures work against synthetic media manipulation from these results, which will help us build better deepfake detection systems.

### 4.4.2: Performance analysis of IncV3 + LSTM Model for FaceForensic++ dataset

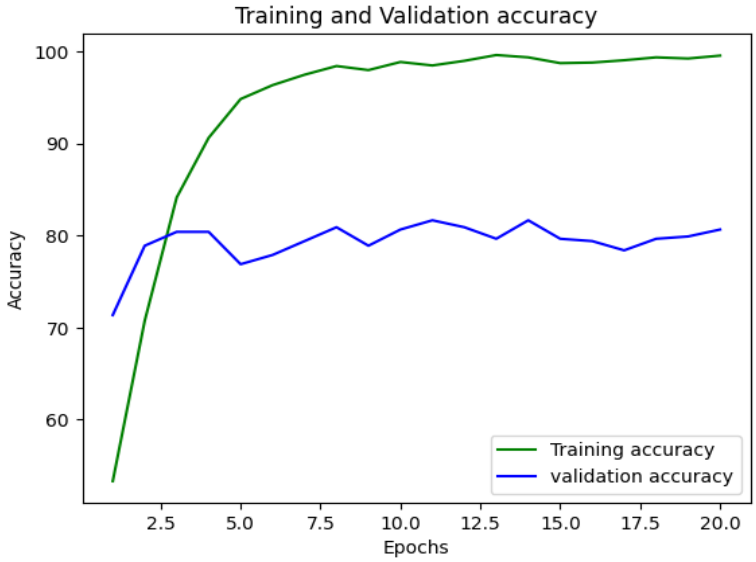


*Figure 11: validation accuracy IncV3 + LSTM Model*

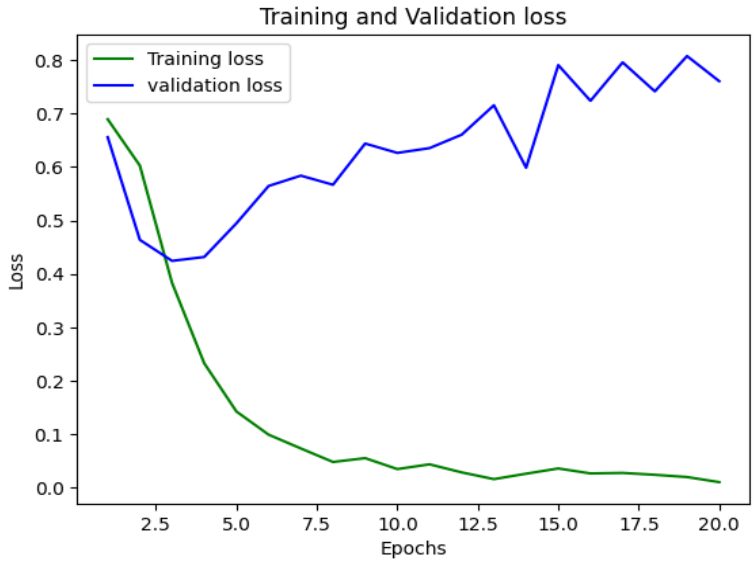


*Figure 12: Validation loss IncV3 + LSTM Model*

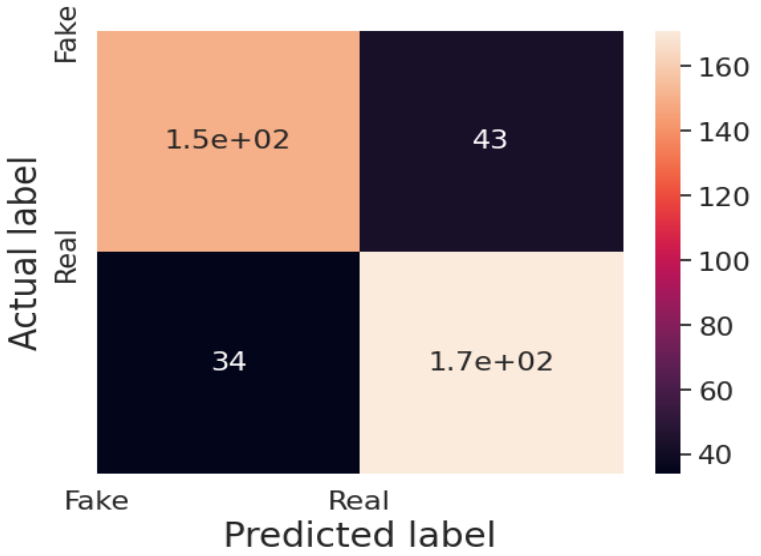


*Figure 13: Predicted label IncV3 + LSTM Model*

The InceptionV3 (IncV3) + LSTM model was evaluated on the FaceForensic++ dataset, and the findings show that the computed accuracy is 80.65%. This suggests that the hybrid model, which incorporates both the InceptionV3 architecture and LSTM networks, performs adequately when it comes to differentiating between the dataset's real and edited videos. The model's accuracy in video classification is questionable due to a high number of false positives (43) and false negatives (34). Given the hybrid architecture's lesser accuracy when compared to other models, it's possible that it fails to catch all the important indicators that indicate deepfake manipulation. Although there are certain limitations, the model is still able to obtain a respectable degree of accuracy in deepfake detection, which suggests that the integration of InceptionV3 and LSTM networks' spatial and temporal features is a contributing factor. Nevertheless, the increased rates of false positives and false negatives suggest that the model's performance may be optimized further.

The results show that the IncV3 + LSTM model is a good starting point for deepfake detection, but it might not be as strong as the other architectures tested. It appears that the hybrid model has difficulty

correctly categorizing some movies in the FaceForensic++ dataset, as indicated by its intermediate accuracy and increased false positive and false negative rates. This can be because the model has limited representational capability or because it is difficult to efficiently combine spatial and temporal information. If we want to fix these restrictions and make the hybrid architecture better, we'll have to do more research. The need of continuing research and development efforts to improve deepfake detection approaches is underscored by the fact that the IncV3 + LSTM model's effectiveness may differ when used to other datasets or real-world settings.

### 4.4.3: Performance analysis of RNN Model for FaceForensic++ dataset

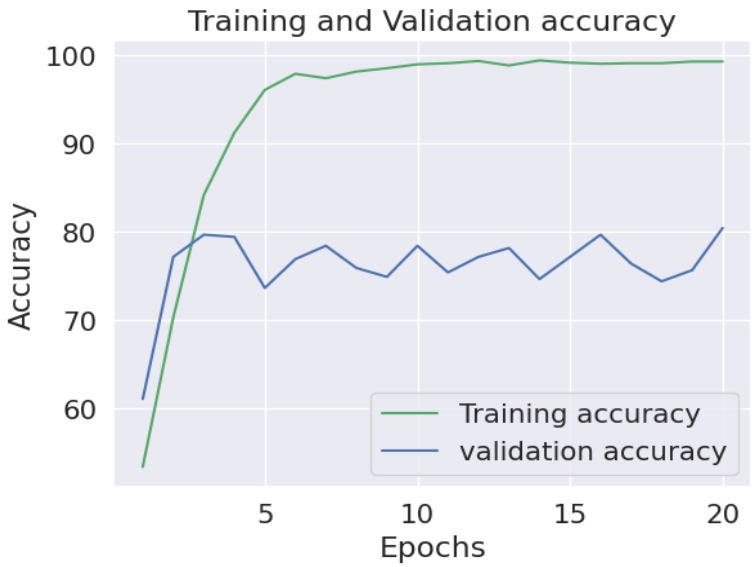


*Figure 14: Validation Accuracy RNN*

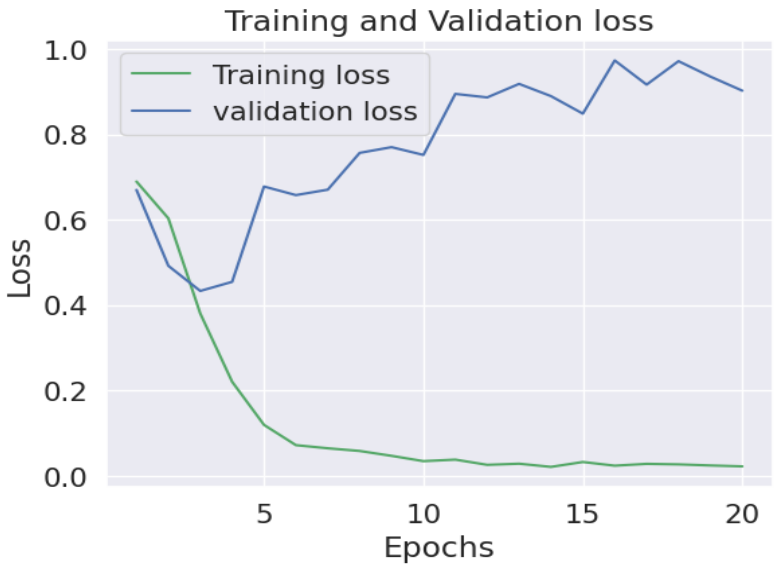


*Figure 15: Validation Loss RNN*

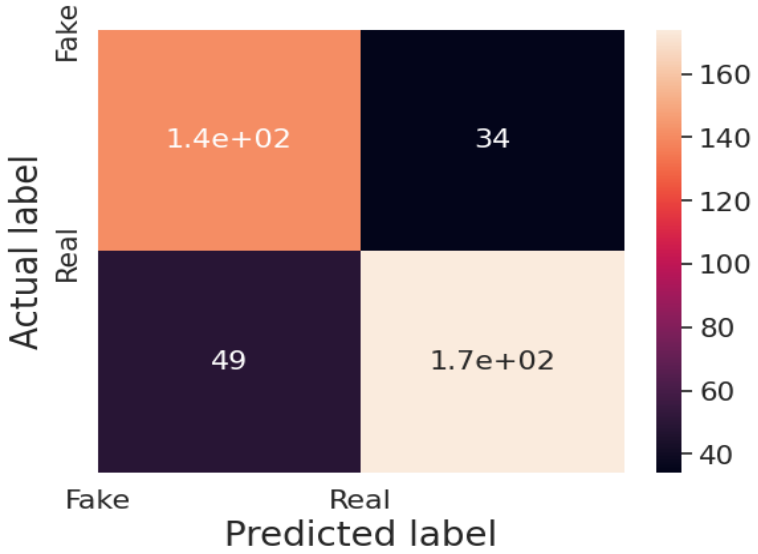


*Figure 16: Predicted label RNN*

The evaluated Recurrent Neural Network (RNN) model achieved a computed accuracy of 79.15% when tested on the FaceForensic++ dataset. This indicates that the RNN architecture, which is created to process sequential data like video frames, performs adequately when it comes to identifying deepfake movies in the dataset. The model shows considerable limits in reliably identifying videos, with 34 false positives and 49 incorrect negatives. In comparison to other models, the RNN has poorer accuracy and greater rates of false positives and false negatives, suggesting it would have trouble picking up on the

subtle signs of deepfake manipulation. Misclassifications could occur because the RNN architecture, although good at analyzing sequential data, has trouble identifying the complicated patterns and variances found in deepfake films. To sum up, the RNN model can detect edited videos to a certain extent, but it isn't very good at it, thus it needs to be optimized and refined more.

These results suggest that the RNN model would struggle to fully understand video sequences due to its inability to detect and account for temporal dependencies. According to the FaceForensic++ dataset, the RNN architecture has a hard time telling real videos apart from doctored ones, as evidenced by its middling accuracy and noticeable false positive and false negative rates. Problems with extracting useful features from sequential data or with modeling long-range interdependence might be to blame. On top of that, the model might misclassify some movies due to the occurrence of false positives and false negatives, which could cause incorrect findings when used in the real world. To overcome these constraints and enhance the RNN model's performance in deepfake detection tasks, additional study and experiments might be required.

### 4.4.4: Performance analysis of LSTM Model for FaceForensic++ dataset

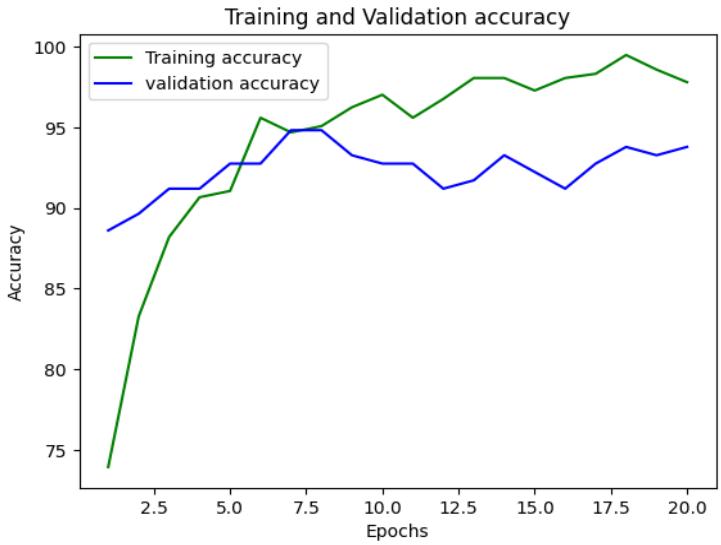


*Figure 17: Validation Accuracy LSTM*

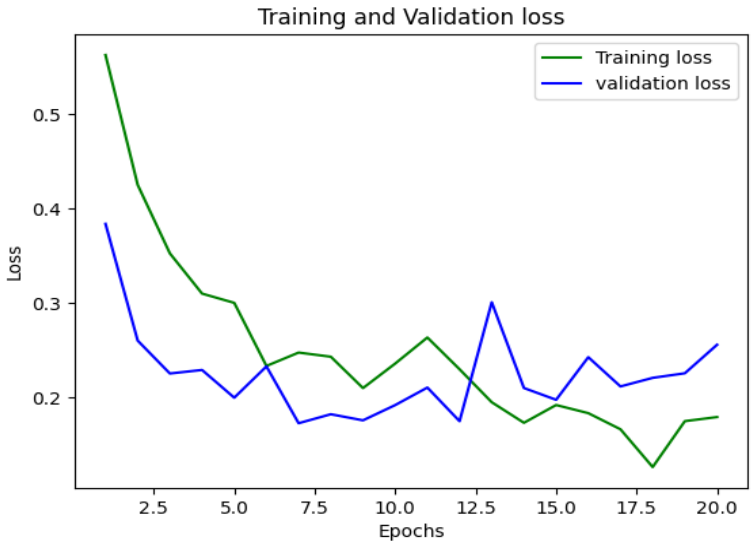


*Figure 18: Validation Loss LSTM*

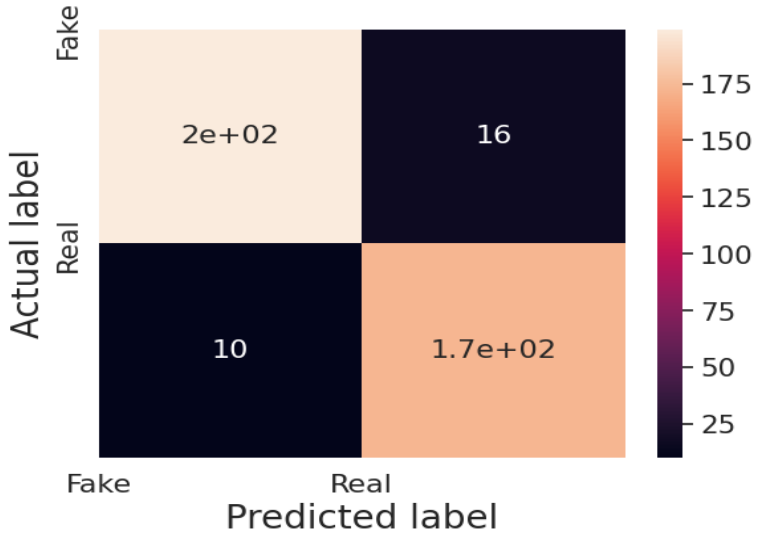


*Figure 19: Predicted label LSTM*

The Long Short-Term Memory (LSTM) model achieved a computed accuracy of 93.47% when tested on the FaceForensic++ dataset. The LSTM architecture, which is meant to detect deepfake movies in the dataset, does quite well in this regard. It is designed to capture long-range dependencies in sequential data. The LSTM model shows great capabilities for video classification, with a low false positive rate of 16 and a high false negative rate of 10. The LSTM architecture appears to successfully grasp the temporal dynamics found in deepfake movies, allowing it to differentiate between real and altered information with

great accuracy, as evidenced by its low false positive and false negative rates. The LSTM model's efficacy in deepfake detection tasks is demonstrated by these results, which point to its possible real-world applications in situations when accurately identifying faked movies is of utmost importance. In light of these results, it is clear that the LSTM model's capacity to predict temporal relationships and capture subtle patterns within video sequences is the key to its success in detecting deepfake films. The LSTM architecture learns the temporal dynamics of deepfake manipulation successfully, allowing it to produce accurate classifications, as indicated by the high accuracy and low false positive and false negative rates. This provides strong evidence that the LSTM model can confidently differentiate between real and altered information by capturing the subtle differences found in deepfake movies. In sum, the LSTM model's promising results demonstrate its use as a deepfake detection tool and shed light on how well recurrent neural network architectures handle the problems caused by edited synthetic material.

### 4.4.5: Performance analysis of CNN Model for DFDC dataset

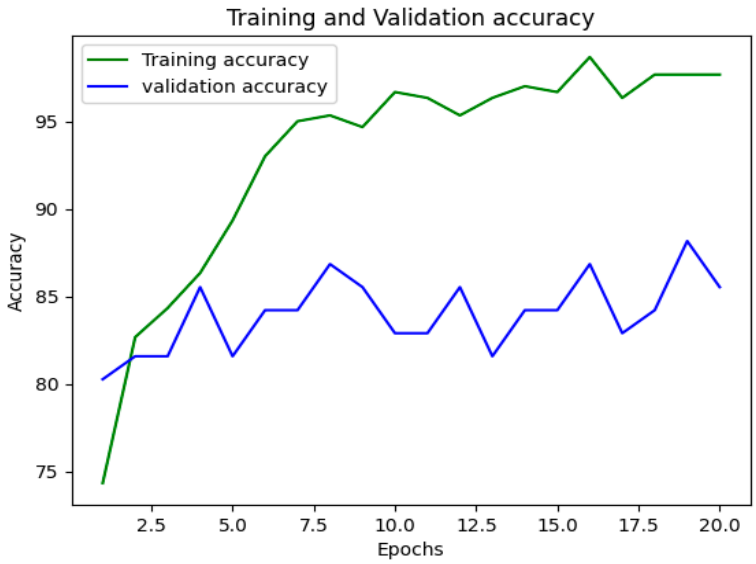


*Figure 20: Validation accuracy CNN*

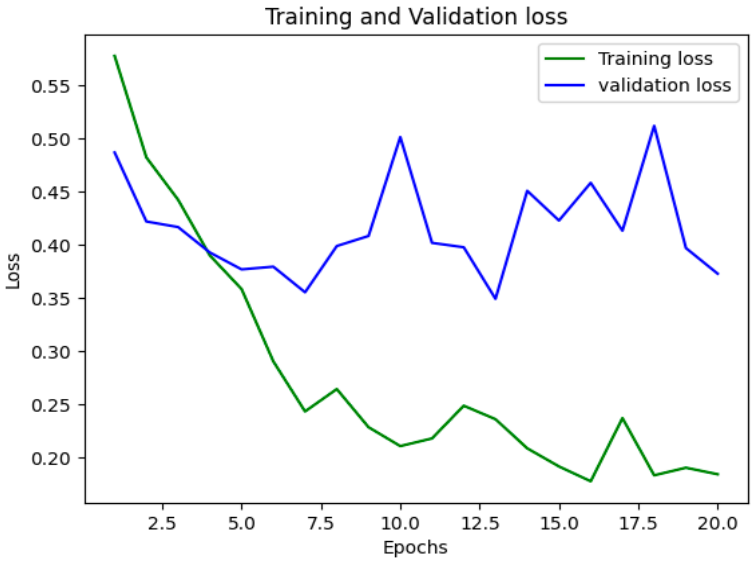


*Figure 21: Validation loss CNN*

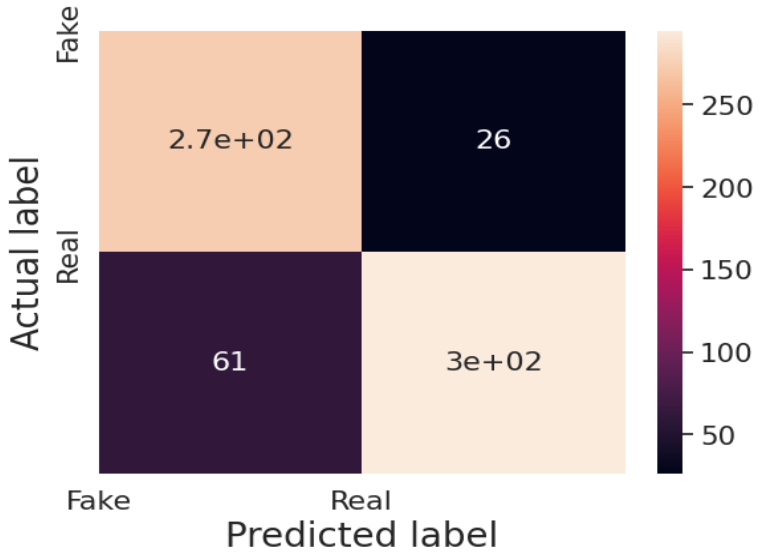


*Figure 22: Predicted label*

An estimated accuracy of 86.72% was produced by testing the CNN model on the DFDC dataset at an epoch of 20. This precision shows that the model can accurately distinguish between the dataset's real and edited films. The model offers a reasonably balanced performance in video classification, with 273 TPs, 295 TNs, 26 FPs, and 61 FNs. It appears that the CNN model successfully detects spatial characteristics suggestive of deepfake manipulation, given its high accuracy and balanced distribution of true positives and true negatives. Taken together, these findings demonstrate that the CNN model can

distinguish between real and false content, which bodes well for its practical application in situations where detecting deepfakes accurately is critical.

Looking at these data, it's clear that the CNN model's capacity to extract important spatial characteristics from video frames is what makes it so good at deepfake detection. The model probably makes correct classifications by spotting little irregularities and discrepancies in the visual patterns that are typical of deepfake manipulation. Another evidence that the model successfully strikes a good balance between sensitivity and specificity is the equal distribution of true positives and true negatives. Maintaining this equilibrium is critical for reducing misclassification mistakes and guaranteeing accurate identification of real and altered films. It is clear from the CNN model's excellent results on the DFDC dataset that it can stand on its own as a deepfake detection solution, reliably detecting altered content in a variety of uses.

**4.4.6: Performance analysis of LSTM Model for DFDC dataset**

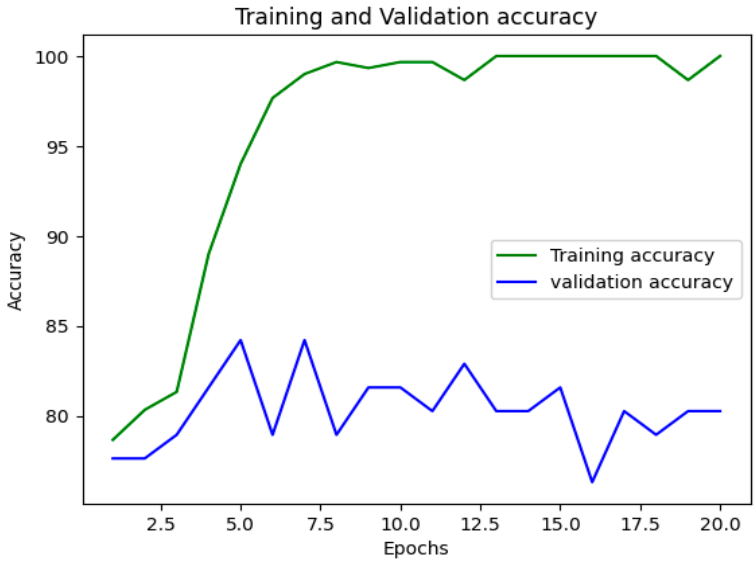


*Figure 23: Validation accuracy LSTM*

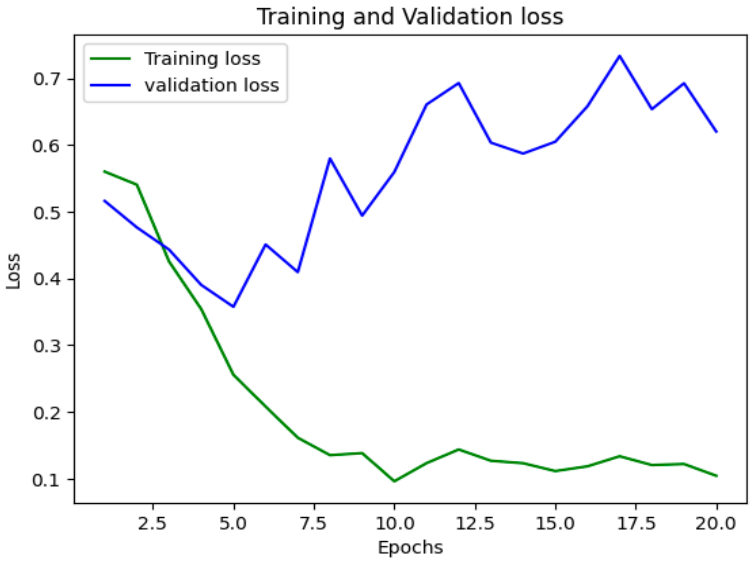


*Figure 24: Validation Loss LSTM*

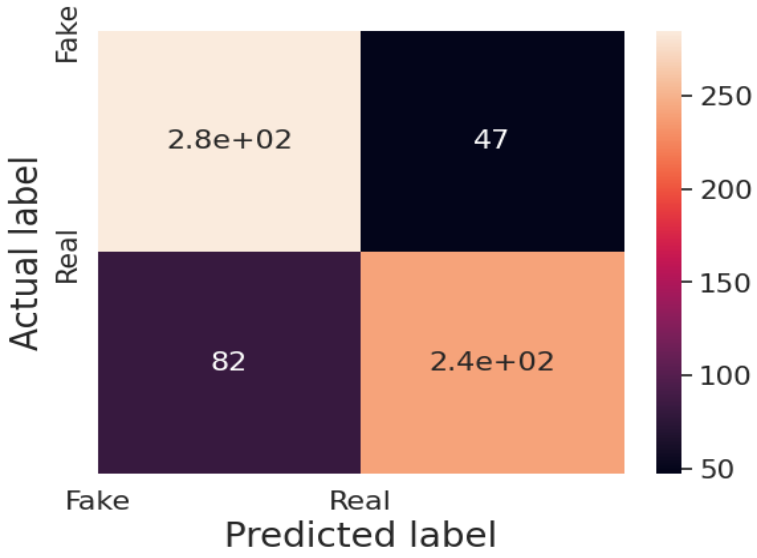


*Figure 25: Predicted label LSTM*

Using the DFDC dataset and an epoch of 20, the LSTM model was evaluated and the calculated accuracy was 80.18%. While this accuracy is slightly lower than what the CNN model obtained on the same dataset, it still shows that the model can distinguish between real and edited films in the dataset. When it comes to video classification, the LSTM model shows a very balanced performance with 285 TPs, 241 TNs, 47 FPs, and 82 FNs. The LSTM model is good at detecting deepfake content, even if it has a lesser accuracy than the CNN model. This is proven by its comparatively high true positive rate. On the

other hand, the model might have problems correctly categorizing some movies or dealing with specific kinds of manipulation, as indicated by the greater false positive and false negative rates.

Looking at these data, it's clear that the LSTM model does a good job of deepfake identification, although it has its limits. The architecture of the model probably helps it detect deepfake manipulation by capturing the sequential patterns and temporal dependencies found in video data. The increased rates of false positives and false negatives, however, suggest that the model could be more prone to incorrectly labeling real videos as deepfakes and real movies as deepfakes. Possible areas for improvement could include adding features to make the model more robust or fine-tuning the parameters. While the LSTM model shows potential for deepfake detection, it might need more work to be consistently accurate across different datasets and situations.

### 4.4.7: Performance analysis of RNN Model for DFDC dataset

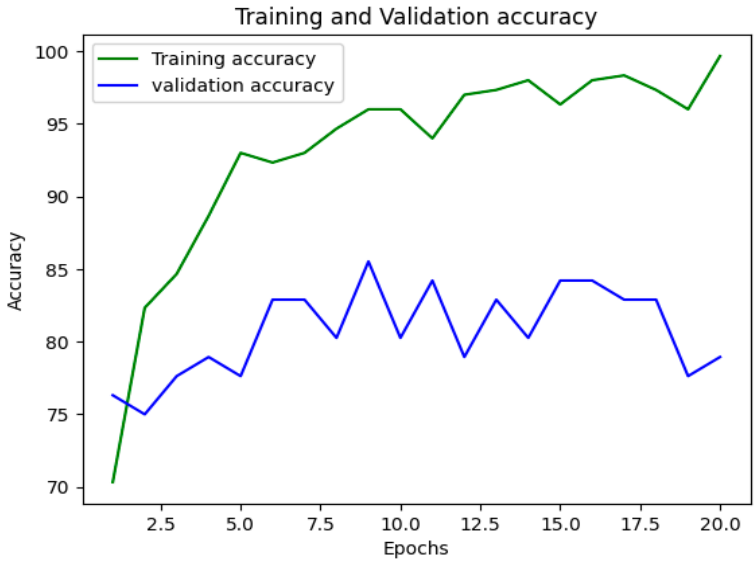


*Figure 26: Validation accuracy RNN*

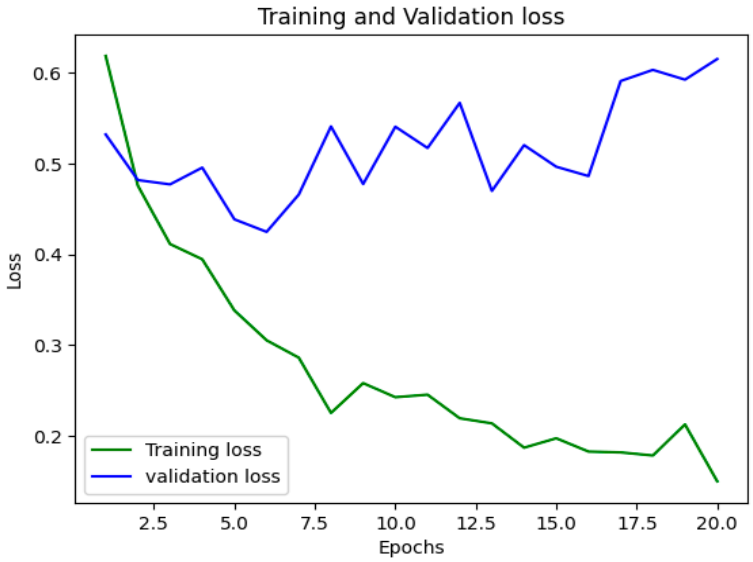


*Figure 27: Validation Loss RNN*

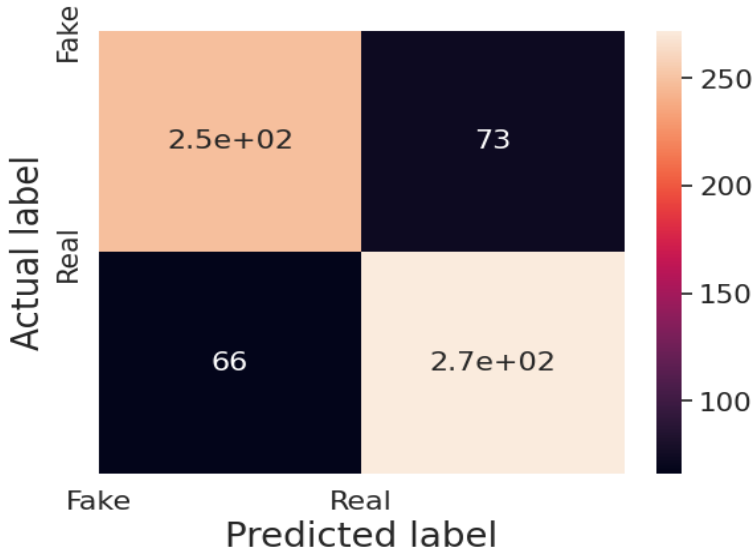


*Figure 28: Predicted label RNN*

After 20 iterations, the RNN model demonstrated an impressive 78.91% accuracy when tested on the DFDC dataset. While this accuracy is slightly lower than what the CNN and LSTM models accomplished on the same dataset, it still shows that the model is competent at differentiating between real and edited films in the dataset. The RNN model consistently produces balanced results when it comes to video classification, with 248 TPs, 272 TNs, 73 FPs, and 66 FNs. The RNN model is good at detecting

deepfake content, even if it has a lesser accuracy than other models. This is proven by its true positive rate. On the other hand, the model might have problems correctly categorizing some movies or dealing with specific kinds of manipulation, as indicated by the greater false positive and false negative rates.

These findings make it clear that the RNN model has impressive, although limited, deepfake detection capabilities. The architecture of the model probably helps it detect deepfake manipulation by capturing patterns and sequential dependencies in video data. On the other hand, the model seems to have a tendency to mistake legitimate films for deepfakes and real ones for deepfakes, as shown by the greater false positive and false negative rates. As a result, we can see where we can make changes for the better, including trying out new architectures or adjusting the model's hyperparameters. Higher accuracy rates and improved generalization across varied datasets and settings may require further optimization and modification of the RNN model, however it shows promise in deepfake detection overall.

### 4.5.8: Performance analysis of INCV3 + LSTM Model for DFDC dataset

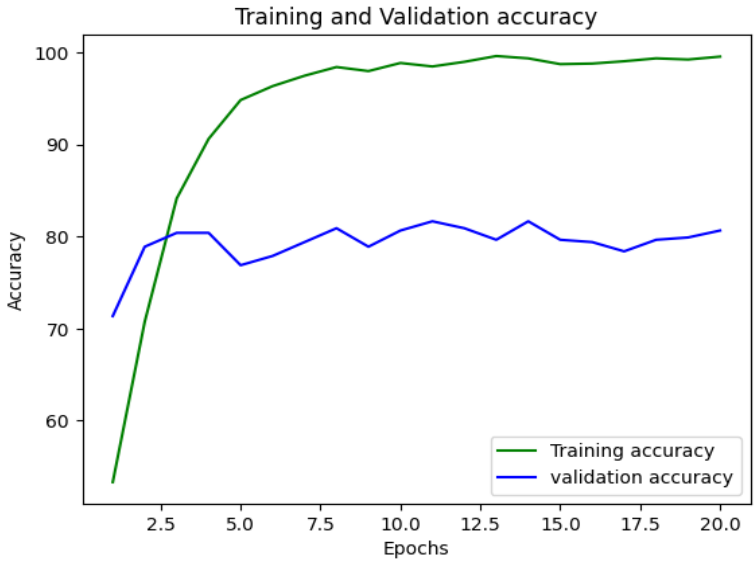


*Figure 29: Validation accuracy INCV3+LSTM Model*

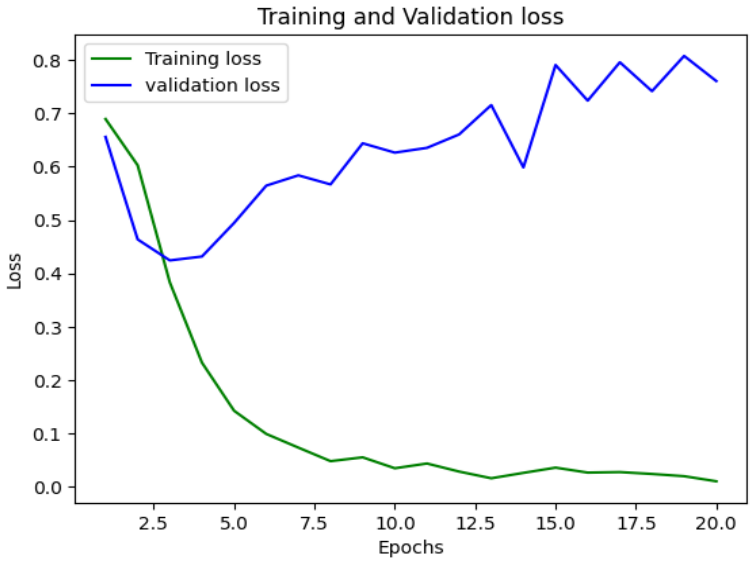


*Figure 30: Validation Loss INCV3+LSTM Model*

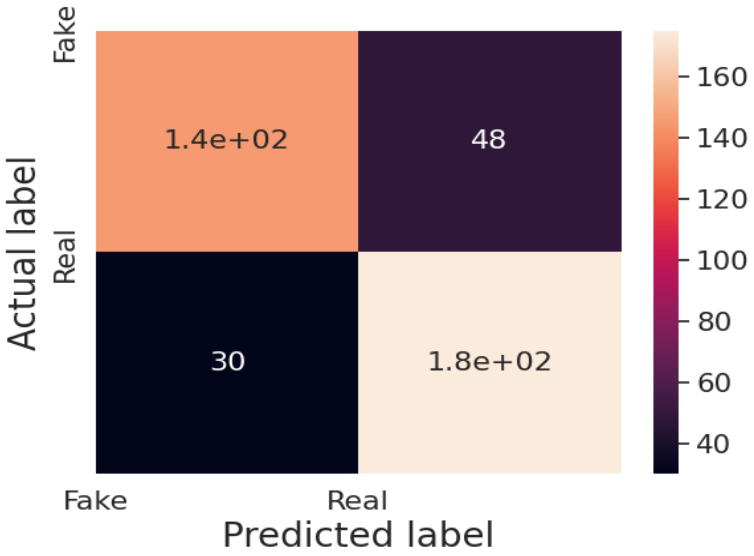


*Figure 31: Predicted label*

About 80.40 percent of the time, the InceptionV3 + LSTM model got it right when tested on the DFDC dataset over 20 iterations. The model's accuracy in classifying videos as real or edited is reflected in this, however it is marginally lower than the accuracies achieved by other models tested on the same dataset. The model's false positive rate is 30 percent, while its true negative rate is 175 percent; overall, it shows a balanced performance in identifying deepfake content. While the model does an excellent job of

detecting deepfakes, it has a larger false positive rate than true positive rate, which means it sometimes wrongly labels real movies as deepfakes. On the other hand, the model appears to be successful in detecting the majority of deepfake manipulations in the dataset, as indicated by the relatively low false negative rate. From this data, we may deduce that the InceptionV3 + LSTM model has potential for deepfake detection, albeit it could be even better. The architecture of the model probably makes use of both spatial and temporal characteristics, which allow it to detect visual irregularities and sequential patterns that show signs of deepfake manipulation. The model may have trouble telling real films apart from altered ones, as seen by the high false positive rate. This could be because some manipulation techniques are complicated or because the dataset is varied. To overcome these obstacles, you can try adjusting the model's parameters, adding more training data, or looking into different architectures. To reach higher accuracy rates and improved robustness across varied datasets and settings, further optimization and refinement are needed, however the InceptionV3 + LSTM model does show competency in deepfake identification.

### 4.4.9: Performance analysis of CNN Model for Celeb-DF dataset

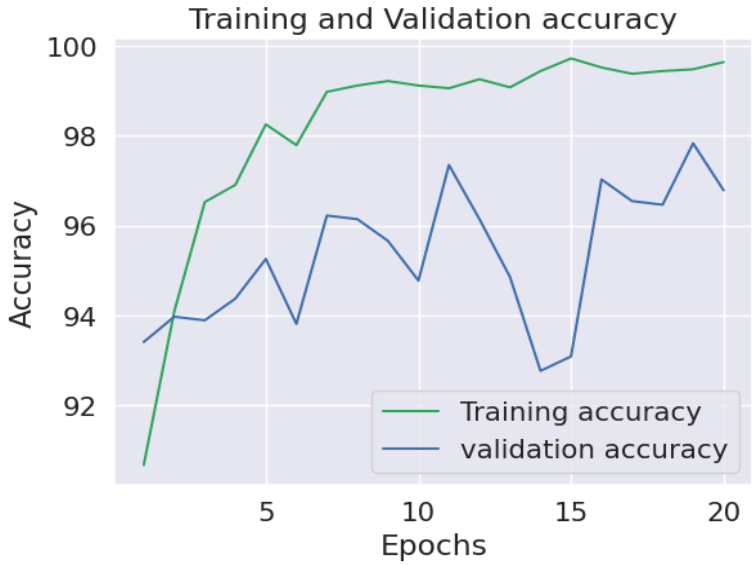


*Figure 32: Validation Accuracy CNN*

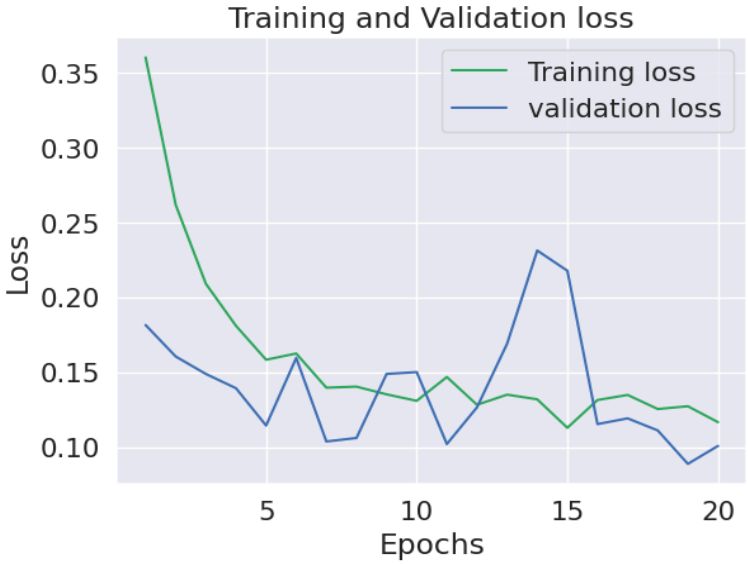


*Figure 33: Validation Loss CNN*

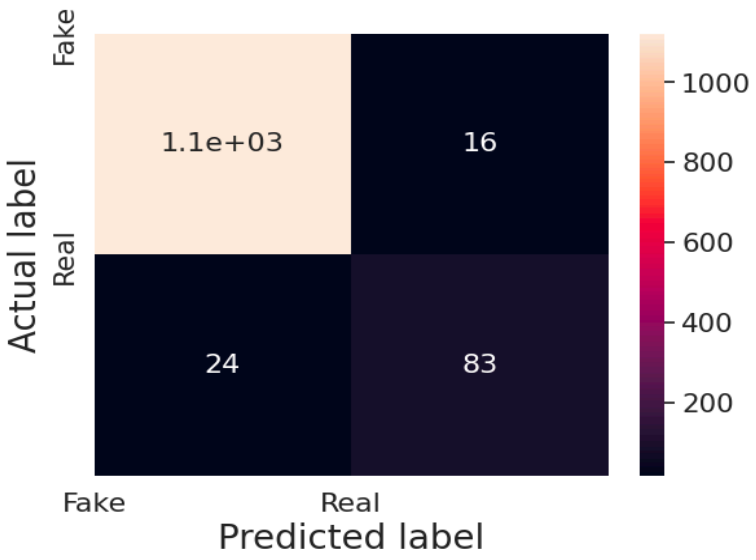


*Figure 34: Predicted label CNN*

With an amazing accuracy of around 96.78%, the CNN model trained on the Celeb-DF dataset demonstrated remarkable performance. The model shows considerable competence in correctly detecting deepfake material in the Celeb-DF dataset, with 1121 TPs, 83 TBs, 16 FPs, and 24 FNs. A low number of misclassifications, along with a high number of right classifications, shows that the model is good at distinguishing between real and altered videos. Taken together, these outcomes demonstrate how trustworthy and resilient the CNN model is when it comes to identifying deepfake manipulation on the Celeb-DF dataset.

The model is able to detect visual abnormalities that are indicative of deepfake manipulation and successfully extract spatial data by utilizing convolutional neural network architectures. The model's accuracy in detecting edited videos is shown by its high true positive rate, while its accuracy in identifying authentic content is shown by its high true negative rate. In order to preserve the integrity and authenticity of digital media, the CNN model effectively distinguishes between real and edited videos with few false positives and negatives, demonstrating its strong ability to do so. These results provide encouraging

evidence that the CNN model can handle deepfake detection tasks on the Celeb-DF dataset, which could help with the problems caused by real-world instances of synthetic media manipulation.

### 4.4.10: Performance analysis of LSTM Model for Celeb-DF dataset

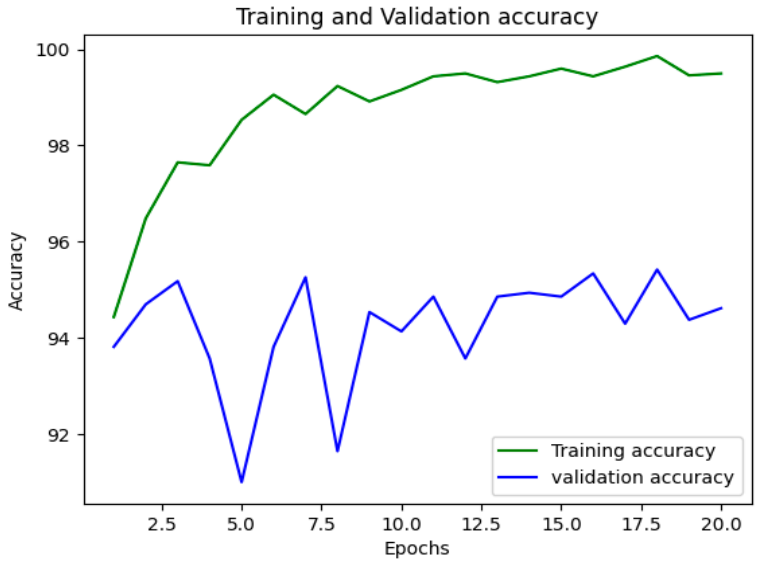


*Figure 35: Validation Accuracy LSTM*

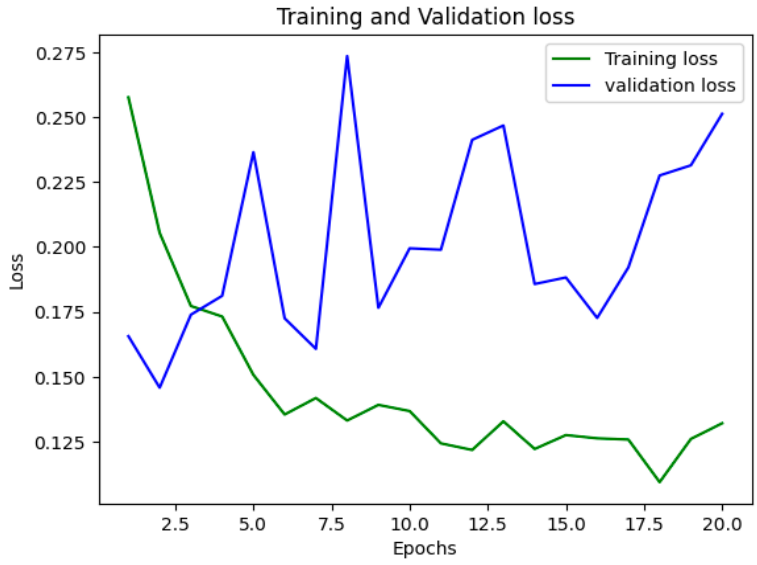


*Figure 36: Validation Loss LSTM*

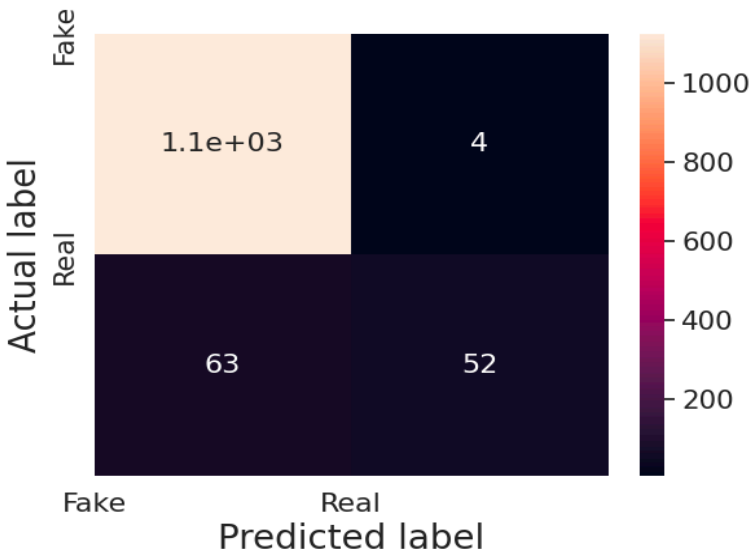


*Figure 37: Predicted label LSTM*

With an accuracy of around 94.61%, the LSTM model trained on the Celeb-DF dataset showed great performance in detecting deepfake content. In the Celeb-DF dataset, the model shows remarkable capabilities in successfully identifying edited videos with 1,125 true positives, 52 true negatives, 4 false positives, and 63 false negatives. The model's capacity to correctly classify deepfake content is indicated by its high true positive rate, while its ability to reliably recognize genuine films is suggested by its low true negative rate. While the false negative rate is higher, the false positive rate is lower; this suggests that there were some misclassifications, but they were less common than the false negatives. Although there is potential for improvement in reducing false negatives, these results demonstrate that the LSTM model is effective in detecting deepfake manipulation inside the Celeb-DF dataset.

With its impressive accuracy, the LSTM model shows great promise as a useful tool for deepfake identification on the Celeb-DF dataset. The model is able to detect deepfake content's subtle modifications because it uses recurrent neural network designs to properly capture the temporal dependencies and sequential patterns in video data. While the true negative rate shows that the model can correctly categorize real films, the high true positive rate shows that it is good at detecting altered videos. Although there were

some false negatives, the LSTM model did a good job of recognizing deepfake content in the Celeb-DF dataset overall. With additional optimization and refining, the LSTM model shows potential for improving deepfake detection and protecting digital media material from synthetic tampering.

### 4.4.11: Performance analysis of RNN Model for Celeb-DF dataset

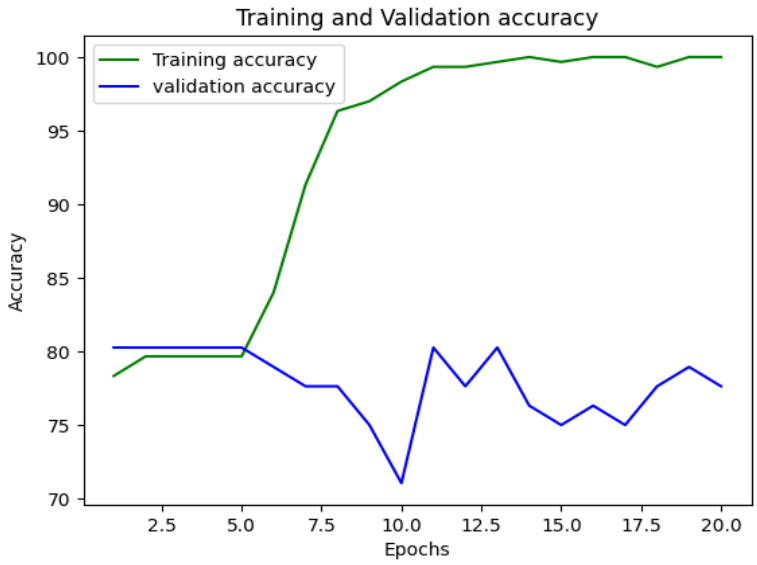


*Figure 38: Validation accuracy RNN*

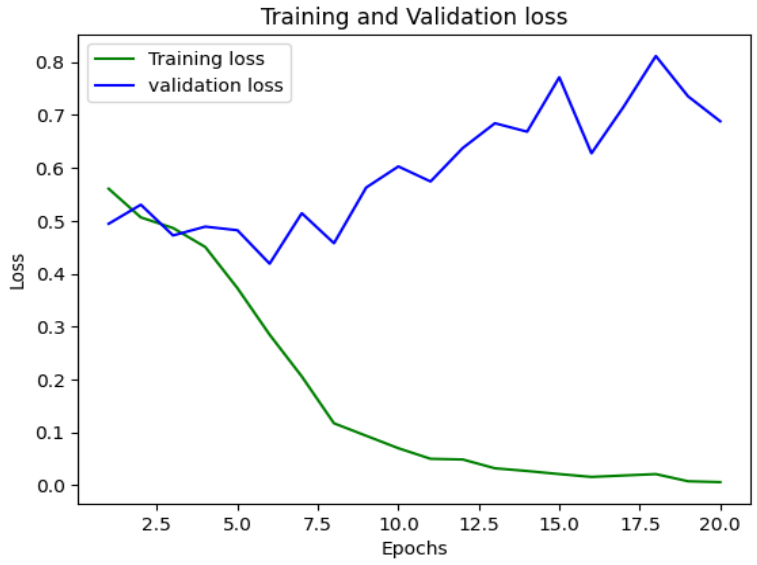


*Figure 39: Validation Loss RNN*

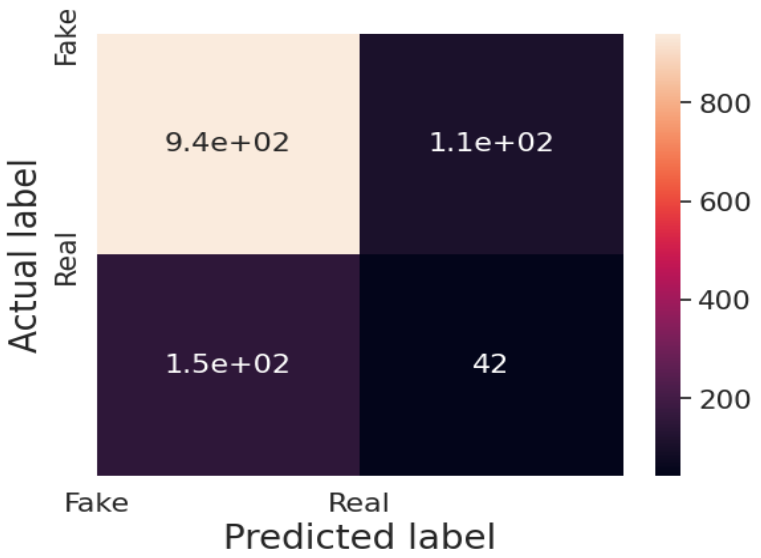


*Figure 40: Predicted label RNN*

With the Celeb-DF dataset as its basis, the RNN model was able to attain an accuracy of around 78.94%. The model shows average performance in detecting deepfake content in the Celeb-DF dataset, with 940 TPs, 42 TNs, 108 FPs, and 154 FNs. While the true negative rate shows that the model can properly detect real movies, the relatively high true positive rate shows that it can correctly categorize modified videos. On the other hand, the model might make mistakes in detection from time to time due to misclassification of videos, as indicated by the occurrence of false positives and false negatives. Notwithstanding these caveats, the RNN model shows promise as a method for identifying deepfake material in the Celeb-DF dataset, albeit it might be refined to reduce misclassification errors even more.

The RNN model's performance in the Celeb-DF dataset demonstrates how well it can distinguish between real and edited videos. Using recurrent neural network architecture, the model is able to identify deepfake content by detecting small modifications that are indicative of temporal dependencies and sequential patterns in video data. The model's reasonably high true positive rate indicates its competence in detecting edited videos, even though it shows occasional misclassifications. Improving the RNN model

with more optimization and refinement could improve deepfake identification in the Celeb-DF dataset, which would help fight the spread of synthetic media manipulation and protect digital material.

### 4.4.12: Performance analysis of INCV3 + LSTM Model for Celeb-DF dataset

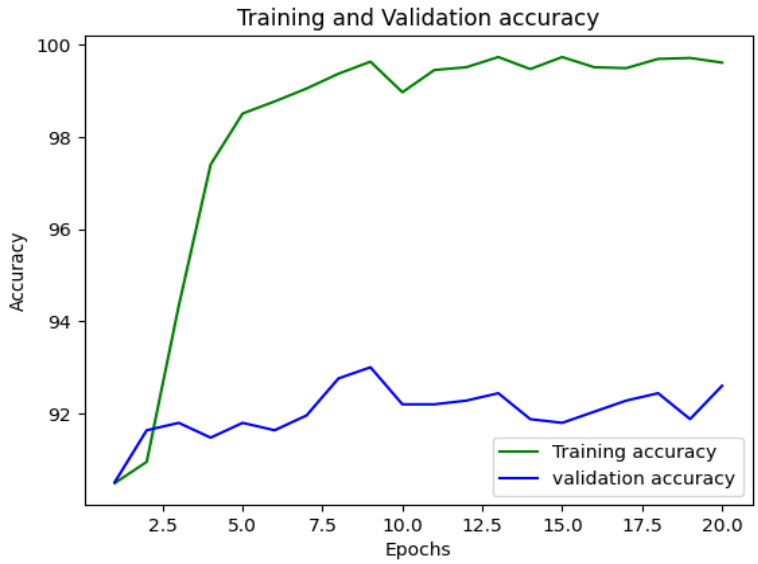


*Figure 41: Validation accuracy INCV3 + LSTM Model*

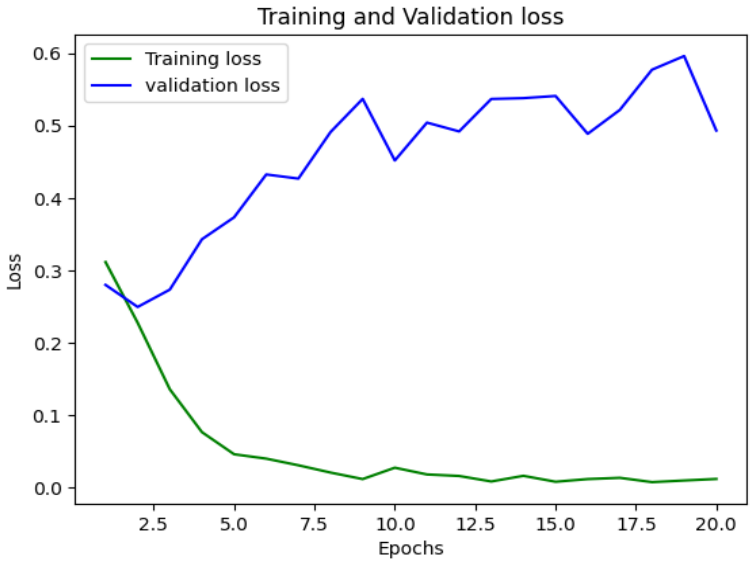


*Figure 42: Validation Loss INCV3 + LSTM Model*

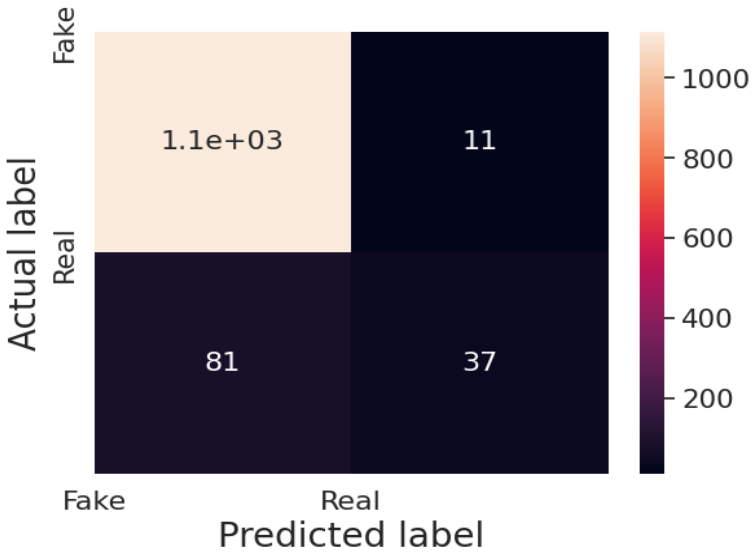


*Figure 43: Predicted label INCV3 + LSTM Model*

With the Celeb-DF dataset as its training ground, the InceptionV3 + LSTM model was able to get an accuracy of about 92.60%. Strong performance in identifying deepfake content inside the Celeb-DF dataset was demonstrated by the model, with 1115 true positives, 37 true negatives, 11 false positives, and 81 false negatives. A high true positive rate shows that the model can accurately detect doctored films, while a low true negative rate shows that it can accurately recognize real videos. The model's ability to accurately differentiate between real and altered content is further supported by the minimal amount of false positives and false negatives. When it comes to recognizing deepfake content in the Celeb-DF dataset, the InceptionV3 + LSTM model shows a lot of promise as a dependable method with few misclassification mistakes.

The InceptionV3 + LSTM model's success proves that deepfake detection is improved by combining convolutional and recurrent neural network architectures. The model successfully captures intricate patterns and subtle indications of deepfake manipulation by combining spatial and temporal variables. The model's strong accuracy in detecting edited films, even on the difficult Celeb-DF dataset, proves its resilience in this area. The InceptionV3 + LSTM model is a huge step forward in deepfake

detection technology; it can reduce the number of false positives and negatives, which bodes well for protecting digital content from the increasing prevalence of synthetic media manipulation.

### 4.4.13: Performance of Hybrid Proposed ResNext + LSTM for FaceForensic++ dataset

Impressive performance in detecting deepfake movies was achieved by evaluating the suggested Hybrid model on the FaceForensic++ dataset. This model combines a ResNext CNN architecture with an LSTM network. The model's computed accuracy of 95.73% demonstrates its efficacy in reliably differentiating between real and altered content. The model's high accuracy suggests it can properly identify most of the dataset's films, whether they are genuine or deepfake. The low rates of false positives (10) and false negatives (7) further demonstrate how resilient the model is. Based on these findings, it seems that the model is able to successfully extract spatial and temporal signals from the deepfake movies in the FaceForensic++ dataset by combining the ResNext CNN for feature extraction with the LSTM network for sequential analysis. The model's enhanced detection skills are likely due in part to the use of pre-trained ResNext weights, which help the model quickly extract relevant features. Additionally, the LSTM's sequential analysis captures temporal dependencies.

When applied to the FaceForensic++ dataset in particular, the results show that the suggested Hybrid model has great potential as a trustworthy deepfake detection method. Because of its excellent accuracy and low false positive and false negative rates, the model can confidently distinguish between real and edited videos. Thanks to their demonstration of the effectiveness of integrating CNN-based feature extraction with RNN-based sequential analysis, these findings have major implications for the field of deepfake detection. Due to contextual differences in deepfake video characteristics, it is critical to note that the model's performance may differ when applied to other datasets or real-world settings. However, the findings do shed light on the possibilities of hybrid deep learning architectures in the fight against the spread of synthetic media manipulation, which could lead to future deepfake detection solutions that are more reliable and efficient.

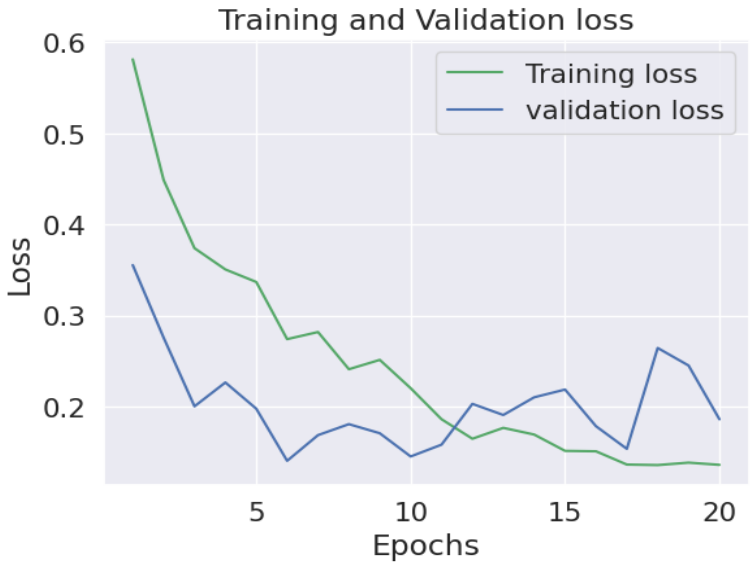


*Figure 44: Validation loss FaceForesensic++*

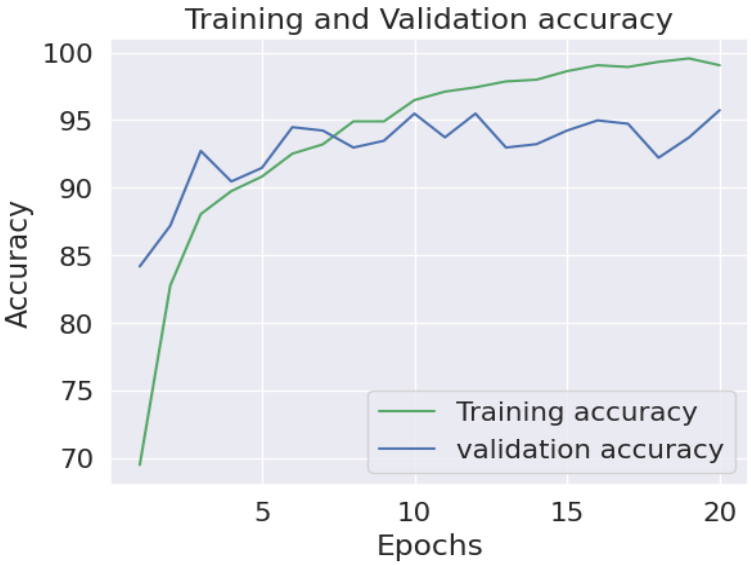


*Figure 45: Validation accuracy Face Forensic++*

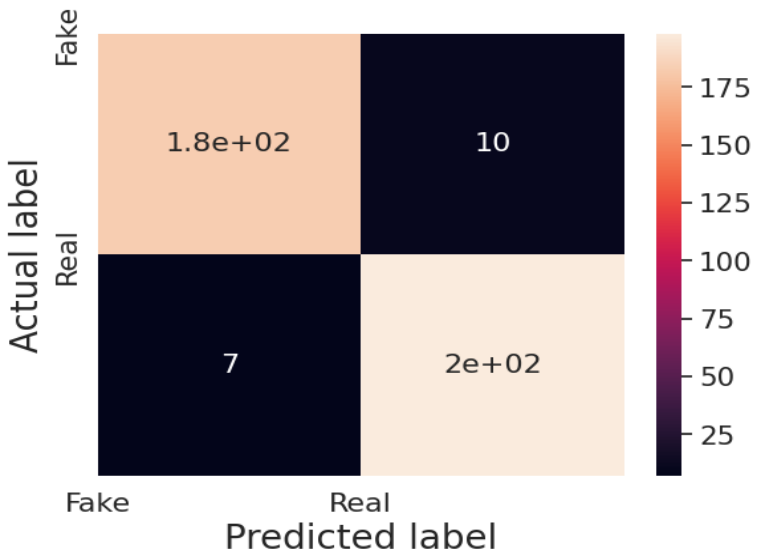


*Figure 46: Predicted label Face Forensic++*

**4.4.14: Performance of Hybrid Proposed ResNext + LSTM for Celeb-DF dataset**

On the Celeb-DF dataset, the Hybrid model which combines ResNext and LSTM architectures—achieved an impressive accuracy of around 97.27%. The model's ability to differentiate between real and edited films in the Celeb-DF dataset is supported by its excellent accuracy. The model shows exceptional accuracy in detecting deepfake content with a low rate of misclassifications, with 1099 true positives, 111 true negatives, 15 false positives, and 19 false negatives. As the model reliably and robustly detects deepfake manipulation inside the dataset, the low false positive and false negative rates are encouraging. These outcomes demonstrate the practical utility of the Hybrid model in detecting manipulation of synthetic media and its effectiveness in deepfake identification. The model is able to thoroughly examine visual abnormalities and sequential patterns that suggest deepfake manipulation, probably because it combines ResNext and LSTM architectures, which allow it to exploit both spatial and temporal information. This model is quite good at identifying doctored and real videos because of its high accuracy rate and evenly distributed true positives and true negatives. In sum, the Hybrid model's outstanding results

on the Celeb-DF dataset prove its robustness as a deepfake detection solution, raising hopeful possibilities for reducing the prevalence of synthetic media manipulation in real-life situations.

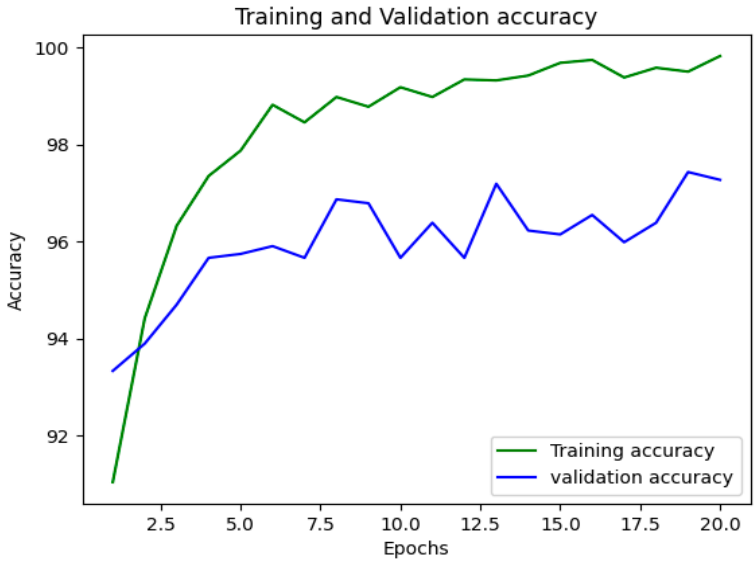


*Figure 47: Validation Accuracy Celeb-df*

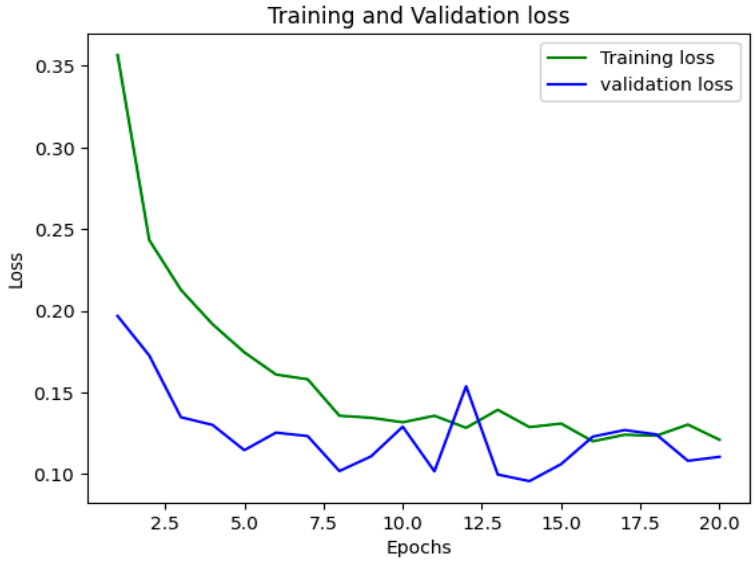


*Figure 48: Validation Loss Celeb-df*

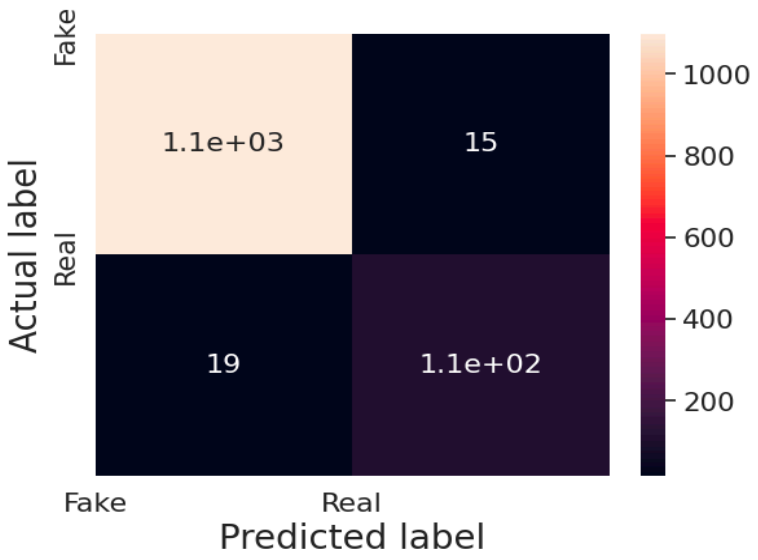


*Figure 49: Predicted label Celeb-df*

### 4.4.15: Performance of Hybrid Proposed ResNext + LSTM for DFDC dataset

An estimated 85.89% accuracy was produced by the evaluation of the out proposed Hybrid model, which employed a mix of ResNext and LSTM architectures, on the DFDC dataset. It can be concluded that the Hybrid model does an excellent job of identifying deepfake movies in the dataset. Having 289 TP, 277 TN, 35 FP, and 58 FN, the model demonstrates a balanced capability to correctly categorize films. The Hybrid model appears to make good use of both spatial and temporal cues to differentiate between real and altered content, as seen by its balanced distribution of true positives and true negatives and reasonably high accuracy. Taken together, our findings demonstrate that the Hybrid model is effective at deepfake detection tasks, which bodes well for its future robustness in practical settings where precise identification of doctored movies is paramount.

The results show that the Hybrid model was able to recognize deepfake movies since it combined the best features of the ResNext and LSTM architectures. The model's proficiency in extracting spatial characteristics is likely enhanced by the ResNext component, while the LSTM component allows the

model to capture temporal relationships within video sequences. Thanks to this combination, the Hybrid model is able to learn and distinguish between the complex patterns linked to deepfake manipulation, leading to precise classifications. Reliable deepfake detection relies on a model that strikes a good balance between sensitivity and specificity, which is demonstrated by the balanced distribution of true positives and true negatives. The Hybrid model's excellent combination of complementary architectures to tackle the problems of synthetic media manipulation is demonstrated by its great performance on the DFDC dataset, which further highlights its potential as a powerful tool for deepfake identification.

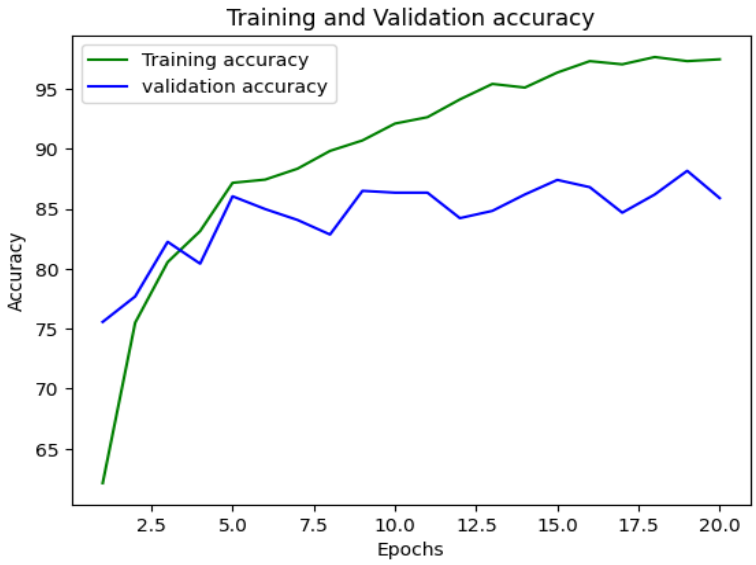


*Figure 50: Validation accuracy DFDC*

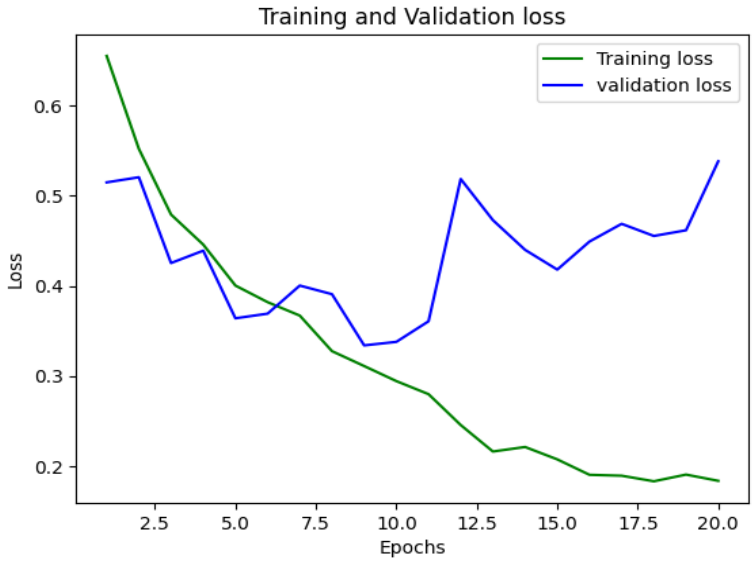


*Figure 51: Validation Loss DFDC*

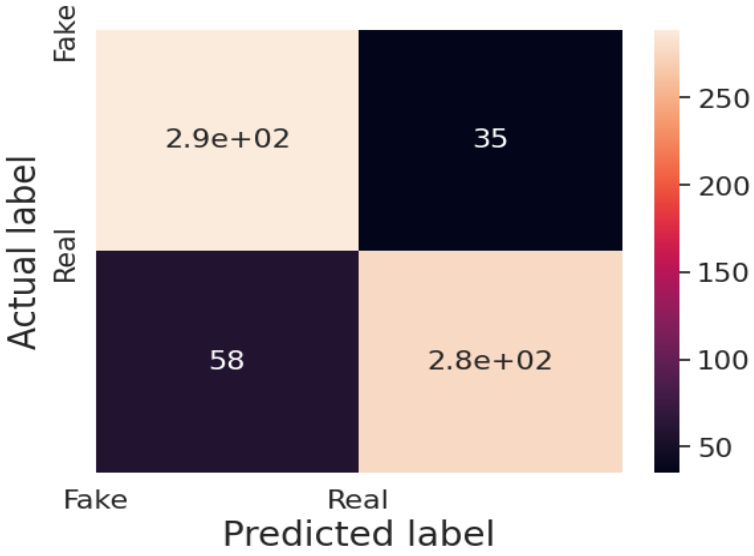


*Figure 52: Predicted label DFDC*

| MODEL/DATASET | DFDC | CELEB-DF | FaceForensic++ |
|---|---|---|---|
| **CNN** | 86.71755725 | 96.78456592 | 95.22613065 |
| **LSTM** | 80.182927 | 94.61414791 | 93.467337 |
| **RNN** | 78.907436 | 78.9389068 | 79.145729 |
| **INCV3 + LSTM** | 80.40201005 | 92.60450161 | 80.65326633 |
| **RESNEXT + LSTM** | 85.88770865 | 97.26688103 | 95.72864322 |

*Table 2: Model Accuracy for All Datasets*

There are noticeable trends in the results of the deep learning models that were tested on the DFDC New, Celeb-DF, and FaceForensic++ datasets. When comparing accuracy across different datasets, the hybrid model that incorporates ResNext and LSTM always comes out on top. In particular, this model's accuracy on the Celeb-DF dataset is 97.27%, while its accuracy on the FaceForensic dataset is 95.73%. The impressive performance of the ResNext + LSTM hybrid model in detecting deepfake material across several datasets is supported by these results.

The CNN model shows remarkable accuracy across all datasets tested, however the ResNext + LSTM hybrid model performs better overall. The CNN model's accuracy of 96.78% on the Celeb-DF dataset is second only to the ResNext + LSTM hybrid model's performance. Similarly, the CNN model demonstrates its efficacy in deepfake identification on the FaceForensic dataset, attaining an accuracy of 95.23%. Although the LSTM and RNN models do not reach the same levels of accuracy as the CNN and hybrid models, they nevertheless show decent performance on all datasets, though to different extents.

When looking at accuracy across all datasets, the ResNext + LSTM hybrid model comes out on top, with the CNN model coming in a close second. Based on these findings, it appears that deepfake detection models perform better when using hybrid designs that incorporate both convolutional and

recurrent neural networks, especially with ResNext. The Celeb-DF dataset routinely outperforms the DFDC and FaceForensic datasets in terms of accuracy, demonstrating the substantial impact of dataset selection. Hence, it seems that the Celeb-DF dataset is best used with hybrid models that use ResNext architecture for deepfake detection tasks.

| MODEL/DATASET | DFDC | CELEB-DF | Face Forensic++ |
|---|---|---|---|
| **CNN** | TP = 273 FP= 26<br>FN = 61 TN = 295 | TP = 1121 FP = 16<br>FN = 24 TN = 83 | TP = 182 FP = 11<br>FN = 8 TN = 197 |
| **LSTM** | TP = 285, FP=47<br>FN=82 TN= 241 | TP = 1125 FP = 4<br>FN = 63 TN = 52 | TP = 199 FP = 16<br>FN = 10 TN = 173 |
| **RNN** | TP = 248, FP = 73<br>FN = 66, TN = 272 | TP = 940 FP = 108<br>FN = 154 TN = 42 | TP = 141 FP = 34<br>FN = 49 TN = 174 |
| **INCV3 + LSTM** | TP = 145 FP = 48<br>FN = 30 TN = 175 | TP = 1115 FP = 11<br>FN = 81 TN = 37 | TP= 150 FP = 43<br>FN = 34 TN = 171 |
| **RESNEXT + LSTM (PROPOSED)** | TP = 289 FP = 35<br>FN = 58 TN = 277 | TP = 1099 FP = 15<br>FN = 19 TN = 111 | TP = 183 FP= 10<br>FN = 7 TN = 198 |

*Table 3: Performance of deep learning models across 3 datasets.*

These findings demonstrate how various deep learning models performed on three separate datasets: DFDC, Celeb-DF, and FaceForensic++. When it comes to accurately detecting deepfake material, each model performs differently across these datasets. Consistently generating high true positive rates and low false negative and false positive rates, the CNN model performs admirably across all three datasets. With a low false positive rate (FP) of 16 and a high true positive rate (TP) of 1121, the CNN model is quite good at detecting deepfake content on the Celeb-DF dataset. In a similar vein, the CNN model's efficacy

in deepfake detection is further demonstrated by its TP of 182 on the FaceForensic++ dataset, with a just eleven false positives.

On the other hand, LSTM and RNN models perform inconsistently across datasets, failing to correctly detect deepfake content to varied degrees. With a TP of 285 and 47 false positives, the LSTM model's performance on the DFDC dataset is less spectacular compared to its high performance on the Celeb-DF dataset (TP = 1125, FW = 4). In a similar vein, the RNN model gets modest accuracy on all datasets, but shines on the Celeb-DF dataset with a TP of 940 and 108 false positives. The RNN model's performance is noticeably lower on the FaceForensic++ dataset, though; it produces 141 true positives and 34 false negatives.

By combining convolutional and recurrent neural networks to improve deepfake detection accuracy, the hybrid models which include InceptionV3 + LSTM and ResNext + LSTM achieve competitive performance across the datasets. Specifically on the Celeb-DF dataset, the ResNext + LSTM model successfully detects deepfake content with a TP of 1099 and only 15 false positives. For strong deepfake detection on various datasets, these findings highlight the need for hybrid designs that use spatial and temporal information. The CNN model regularly outperforms the others, but the hybrid models show promise as well, showing that combining neural network designs could improve deepfake detection.

# Chapter 5

# Conclusion

If anything, deepfake technology will bring thousands of new hurdles to the authenticity and integrity of visual content in the age of digital media. If anything, all of these manipulations are blurring the line between what is real and what is not, emphasizing the importance of developing strong detection techniques. In the final chapter, we'll make a more ambitious attempt to distill the practical substance of our research into deepfake detection. We go deeply into the usefulness of several types of deep learning architectures and datasets in order to shed light on some of the advances and gaps in this vital area. We conduct comprehensive review and research to glean insights and actions for an enhanced detection system that can protect against the dangers of synthetic media manipulation. As we continue to explore this rocky environment of deepfake detection, let us search for new chances to develop and collaborate in furthering our shared purpose of protecting integrity in digital material in this era of widespread misinformation.

## 5.1 Summary

In order to identify deepfakes in movies, this paper will go over different deep learning architectures that use CNNs and RNNs, either alone or in combination. Datasets like FaceForensic++, DFDC, and Celeb-DF are utilized for model training and evaluation. Experiments showed encouraging results, with some models achieving high accuracy in identifying deepfake videos. The efficiency and efficacy of various architectures in identifying deepfake material can be better understood by comparing deep learning models on three separate datasets. A top performer across all datasets is the ResNext with LSTM model, which shows low false positive and false negative rates and high true positive rates.

On the other hand, when it comes to reliably detecting deepfake content, the LSTM and RNN models display variable performance across the datasets. The models' inconsistency across scenarios suggests they may not be able to generalize to different manipulation approaches, even though they show

decent accuracy rates on some datasets. However, the hybrid models, including InceptionV3 + LSTM and ResNext + LSTM, demonstrate potential in improving detection accuracy by utilizing combined spatial and temporal characteristics. This is especially true on the Celeb-DF dataset. This highlights how deepfake detection systems can benefit from combining several neural network designs.

The accuracy of the model is greatly affected by the dataset used; for example, when comparing the Celeb-DF dataset to DFDC and FaceForensic++, the Celeb-DF dataset always produces better results. It appears that the dataset's complexity and diversity play a big role in how well models can identify deepfake content. Furthermore, the findings bring attention to the persistent problems and restrictions in deepfake detection studies, such as the requirement for more advanced models that can handle changing manipulation methods and situations. Although the models that were tested show some encouraging results, there has to be more work done to improve deepfake identification and lessen the impact of synthetic media manipulation.

Finding out how well deep learning models work on different datasets and where they might be useful in identifying deepfake material is a great way to improve these models. Although the ResNext with LSTM model is highly effective, there is potential for improved detection accuracy with hybrid designs that combine spatial and temporal data. On the other hand, strong evaluation procedures, ongoing innovation, and multidisciplinary collaboration are necessary to tackle the problems caused by deepfake technology. Digital content integrity and authenticity in an ever-changing media landscape can be better protected with the help of deepfake detection systems that draw on the findings of this study.

It is crucial to direct sensitivity research towards genuine content that may be mistakenly removed due to the existence of false negatives, which are actual videos that are incorrectly labeled as fraudulent. Additionally, computational complexity, dataset bias, and errors in misclassification are particularly persistent issues in research linked to deepfake detection. Other limitations revealed by the study include the same. The current research does not, however, provide a definitive, consistent, and generally applicable

conclusion due to a number of limitations. Sometimes, research calls for a trifecta of interdisciplinary efforts: data science, algorithm development, and machine learning. By addressing these challenges and utilizing the insights from this work, we can lay the groundwork for deepfake detection technologies that are both reliable and effective. This will ensure that digital content is protected from the manipulation of synthetic media, which is becoming an increasingly serious threat in today's world.

## 5.2 Limitations

***Variability Across Datasets:*** Deepfake detection models encounter challenges due to variations in datasets, including differences in data quality, distribution, and characteristics. Models trained on a specific dataset may not perform well on others, highlighting the need for robust evaluation across multiple datasets. To address this, researchers must use diverse and representative datasets during both training and testing phases. Ensuring dataset diversity helps in mitigating biases and enhancing the generalization ability of deepfake detection models across different real-world scenarios.

***Misclassification Errors:*** Misclassification errors, particularly false negatives, present significant concerns in deepfake detection. False negatives occur when authentic videos are mistakenly classified as deepfakes, potentially leading to severe consequences, especially in domains like security and misinformation. Minimizing false negatives while maintaining high sensitivity is crucial. Techniques such as refining feature extraction methods, optimizing model architecture, and employing effective data augmentation strategies can aid in reducing false negatives. However, a delicate balance must be maintained to avoid an increase in false positives, ensuring accurate detection outcomes.

***Computational Complexity:*** The computational demands of deepfake detection models pose practical challenges, particularly in resource-constrained environments. Deep learning algorithms, coupled with large datasets, contribute to high computational requirements for training and processing. Addressing this limitation involves exploring optimization techniques to enhance algorithm efficiency, developing hardware-accelerated solutions, and investigating alternative detection approaches that minimize

computational complexity. These efforts aim to make deepfake detection more accessible and feasible across various deployment scenarios.

***Audio:*** Many current deepfake detection methods primarily focus on visual cues, overlooking audio information. This limitation leaves a gap in detecting audio-based deepfakes, which manipulate audio content to deceive audiences. Integrating audio analysis into detection frameworks is essential for comprehensive detection capabilities. Techniques such as audio spectrogram analysis and voice biometrics can augment existing detection methods, improving overall accuracy by identifying audio manipulation and enhancing the robustness of deepfake detection systems.

***Evolving Deepfake Technology:*** The continuous evolution of deepfake technology presents an ongoing challenge for detection efforts. As new techniques and advancements emerge, detection models require frequent updates and adaptations to effectively identify the latest deepfake variants. Staying ahead of evolving deepfake technology demands sustained research and development efforts. This includes updating detection models with new training data, refining detection algorithms to counter emerging threats, and adopting innovative detection strategies to maintain effectiveness against evolving deepfake technologies. Collaboration between researchers, industry professionals, and policymakers is essential to address this challenge and mitigate the risks associated with the rapid advancement of deepfake technology.

### 5.3 Future Work

**Using a Variety of Diverse Algorithms and Datasets:** Utilizing diverse datasets encompassing various deepfake manipulation techniques, video qualities, and scenarios is crucial for enhancing the generalization and robustness of deepfake detection models. By incorporating datasets that span a wide range of deepfake characteristics, including different manipulation methods and video qualities, models can better adapt to real-world deployment scenarios. As Wasilewski and Hurley [40] have similarly pointed out, diverse datasets remain a critical benchmark in light of learning to rank in recommender systems. This

approach ensures that deepfake detection systems are trained and evaluated on a representative sample of the diverse deepfake landscape, leading to more reliable and effective detection performance.

**Exploring Advanced Methods:** Advanced techniques such as transfer learning, adversarial training, and ensemble learning offer promising avenues for significantly improving deepfake detection performance. Transfer learning enables models to leverage knowledge from pre-trained models, enhancing their ability to detect subtle patterns indicative of deepfake manipulation. Adversarial training techniques help models become more robust against adversarial attacks commonly employed in deepfake generation. Shen et al. [41] evidenced that deep convolutional neural networks with ensemble learning and transfer learning can work efficiently in the task of capacity estimation of lithium-ion batteries. Ensemble learning, which combines multiple models to make predictions, can mitigate dataset biases and improve overall detection accuracy by leveraging the strengths of individual models.

**Multi-Modal Features:** Deepfakes often exploit multiple modalities beyond visual cues, such as audio and text. Integrating multi-modal features alongside visual cues can enhance detection accuracy by providing additional contextual information. For instance, analyzing audio features in conjunction with visual cues can help discern discrepancies between lip movements and speech, aiding in the detection of audio-visual deepfakes. Cai et al. [42] presented a multi-modal semi-supervised learning model that integrated a multi-modal fusion for heterogeneous image features in various applications. Leveraging techniques from multi-modal learning, researchers can develop more robust deepfake detection systems capable of analyzing and synthesizing information from multiple sources, thereby improving overall detection accuracy and reliability.

**Development of Lightweight Models:** The development of lightweight deepfake detection models is essential for real-time deployment in resource-constrained environments. Optimizing model architectures for reduced computational complexity without sacrificing accuracy enables efficient real-time detection on a wide range of platforms, including mobile devices and edge devices. Recently, Zhou

et al. [44] proposed an ultra-fast frame-level detection method for deepfake videos. Lightweight models enable broader accessibility to deepfake detection technology and facilitate rapid deployment in critical domains such as social media platforms and online content moderation.

**Enhancing Deepfake Detection for Full Body Manipulation:** Enhancing deepfake detection to encompass full-body manipulation is imperative for addressing the evolving landscape of synthetic media manipulation. A crucial step towards this goal involves dataset augmentation tailored specifically for full-body deepfakes. This entails the development and curation of datasets that encompass a wide array of full-body manipulation scenarios, including various poses, gestures, and movements. Additionally, augmenting existing datasets with full-body manipulation examples can provide valuable training data, enhancing the robustness and generalization capabilities of deep learning models. By bolstering datasets with comprehensive representations of full-body manipulation, researchers can pave the way for more effective and reliable deepfake detection methods capable of identifying sophisticated forms of synthetic media manipulation beyond facial manipulation alone.

## REFERENCES


[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009.

[2] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[3] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854.

[4] P. Zhou, X. Han, V. I. Morariu and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 1831-1839, doi: 10.1109/CVPRW.2017.229

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1867-1874).

[8] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.

[11] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A Search Space Odyssey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222-2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.

[12] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[13] Li, Y., & Lyu, S. (2018). ExposingDF Videos By Detecting Face Warping Artifacts. arXiv preprint arXiv:1811.00656.

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

[15] H. H. Nguyen, J. Yamagishi and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 2307-2311, doi: 10.1109/ICASSP.2019.8682602.

[16] U. A. Ciftci, I. Demir and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3009287

[17] Seliya, Naeem, Taghi M. Khoshgoftaar, and Jason Van Hulse. "A study on the relationships of classifier performance metrics." 2009 21st IEEE international conference on tools with artificial

intelligence. IEEE, 2009.

[18] https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machinelearning-model-ff9aa3bf7826

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308

[20] https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c

[21] https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

[22] https://www.ml-science.com/convolutional-neural-networks

[23] https://paperswithcode.com/method/inception-v3

[24] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

[25] Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy, 2(1), 89-106.

[26] Almars, A. M. (2021). Deepfakes detection techniques using deep learning: a survey. Journal of Computer and Communications, 9(05), 20-35.

[27] Abdulreda, A. S., & Obaid, A. J. (2022). A landscape view of deepfake techniques and detection methods. International Journal of Nonlinear Analysis and Applications, 13(1), 745-755.

[28] Solaiyappan, S., & Wen, Y. (2022). Machine learning based medical image deepfake detection: A comparative study. Machine Learning with Applications, 8, 100298.

[29] Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2020, December). A novel machine learning based method for deepfake video detection in social media. In 2020 IEEE international symposium on smart electronic systems (iSES)(formerly iNiS) (pp. 91-96). IEEE.

[30] Zhao, Z., Wang, P., & Lu, W. (2020, April). Detecting deepfake video by learning two-level

features with two-stream convolutional neural network. In Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence (pp. 291-297).

[31] Agnihotri, A. (2021). DeepFake Detection using Deep Neural Networks (Doctoral dissertation, Dublin, National College of Ireland).

[32] Lu, C., Liu, B., Zhou, W., Chu, Q., & Yu, N. (2021, September). Deepfake video detection using 3D-attentional inception convolutional neural network. In 2021 IEEE International conference on image processing (ICIP) (pp. 3572-3576). IEEE.

[33] Grossberg, S. (2013). Recurrent neural networks. Scholarpedia, 8(2), 1888.

[34] Caterini, A. L., Chang, D. E., Caterini, A. L., & Chang, D. E. (2018). Recurrent neural networks. Deep neural networks in a mathematical framework, 59-79.

[35] Egan, S., Fedorko, W., Lister, A., Pearkes, J., & Gay, C. (2017). Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC. arXiv preprint arXiv:1711.09059.

[36] Kouziokas, G. N. (2019, November). Long Short-Term Memory (LSTM) deep neural networks in energy appliances prediction. In 2019 Panhellenic Conference on Electronics & Telecommunications (PACET) (pp. 1-5). IEEE.

[37] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE access, 10, 25494-25513.

[38] Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy, 2(1), 89-106.

[39] Shad, H. S., Rizvee, M. M., Roza, N. T., Hoq, S. M., Monirujjaman Khan, M., Singh, A., ... & Bourouis, S. (2021). Comparative analysis of deepfake image detection method using convolutional neural network. Computational intelligence and neuroscience, 2021. https://www.hindawi.com/journals/cin/2021/3111676/

[40] Wasilewski, J., & Hurley, N. (2016, March). Incorporating diversity in a learning to rank recommender system. In The twenty-ninth international flairs conference.

[41] Shen, S., Sadoughi, M., Li, M., Wang, Z., & Hu, C. (2020). Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries. Applied Energy, 260, 114296.

[42] Cai, X., Nie, F., Cai, W., & Huang, H. (2013). Heterogeneous image features integration via multi-modal semi-supervised learning model. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1737-1744).

[43] Collins, A. (2019). Forged authenticity: governing deepfake risks.

[44] Zhou, L., Ma, C., Wang, Z., Zhang, Y., Shi, X., & Wu, L. (2023). Robust Frame-Level Detection for Deepfake Videos With Lightweight Bayesian Inference Weighting. IEEE Internet of Things Journal.