Deepfake Detection Using Deep Learning Methods

Muhammad Zain Haseeb
Dept of Computer Science (student)
Forman Christian College University
Lahore, Pakistan
251739387@formanite.fccollege.edu.pk

Asma Basharat
Supervisor Dept of Computer Science
Forman Christian College University
Lahore, Pakistan
asmabasharat@fccollege.edu.pk

Tayyab Imam
Dept of Computer Science (student)
Forman Christian College Univeristy
Lahore, Pakistan
251741051@formanite.fccollege.edu.pk

***Abstract*—Deep learning is an effective and useful technique that has been widely applied in a variety of fields, including computer vision, machine vision, and natural language processing. Deepfakes uses deep learning technology to manipulate images, videos or audio of a person that humans cannot differentiate them from the real one. In recent years, many studies have been conducted to understand how deepfakes work and many approaches based on deep learning have been introduced to detect deepfakes videos, images or audio.**

This paper presents a comprehensive review of deep learning-based approaches for deepfake video detection techniques. We analyze recent methodologies across benchmark datasets such as DFDC, Celeb-DF , and FaceForensics, emphasizing generalization capabilities. Our study will be beneficial for researchers in this field as it will cover the recent state-of-art methods that discover deepfakes videos in social contents. In addition, it will help comparison with the existing works because of the detailed description of the latest methods and dataset used in this domain.

***Index Terms*—Deepfake Video Detection, LSTM, Hybrid CNN-LSTM, CNN-Attention, ResNext-50, RelU, DFDC dataset, Attention Mechanisms**

I. Introduction

Deepfake technology, fueled by the advancements in Generative Adversarial Networks (GANs) and deep learning techniques, has led to the generation of highly realistic but fabricated digital media, including images, videos, and audio. While deepfake technology originally served creative applications, such as enhancing visual effects and entertainment, its misuse for malicious purposes poses severe threats to privacy, security, and the integrity of information dissemination. Deepfakes have been extensively used to manipulate political speeches, generate fraudulent celebrity videos, and spread misinformation, thereby raising significant ethical, social, and security concerns on a global scale [2]. The accessibility of open-source tools and generative models has contributed to the rapid proliferation of deepfakes, enabling even non-experts to generate convincing fake content with minimal computational resources. This ease of access exacerbates the challenge of combating the spread of misinformation and fraudulent content on digital platforms [1].

Detecting deepfakes presents a complex challenge, primarily due to the continuous evolution of deepfake generation techniques. Traditional detection methods initially focused on identifying visual artifacts, facial asymmetries, or irregularities in head poses. However, as deepfake algorithms have advanced, they have successfully minimized such inconsistencies, making the detection task more difficult [1]. To address this problem, we present a novel deep learning-based technique that successfully distinguishes between AI-generated fake videos (DeepFake Videos) and genuine ones. Understanding the workings of the Generative Adversarial Network (GAN) is pivotal in identifying DeepFake videos. GANs use input videos and images of a target person to replace their faces with those of another person [1]. Deep adversarial neural networks are trained on face photos and target videos to automatically map faces and facial expressions, forming the foundation of DeepFake creation. With appropriate post-processing, these generated videos can achieve a high level of realism. The GAN replaces the input image in each frame by dividing the video into frames and then reconstructing the video using techniques like autoencoders.

Our proposed technique is based on the same underlying process employed by GANs to generate DeepFake videos. We focus on a specific feature of these videos, where face photos synthesized by the DeepFake algorithm undergo an affine warping process to align with the facial features of the source person while maintaining a fixed size [2]. This warping introduces discernible artifacts in the resulting DeepFake video due to resolution discrepancies between the warped face area and the surrounding context. By comparing these created face areas with their surrounding regions, our technique can identify such artifacts. Furthermore, we capture the temporal inconsistencies between frames introduced by GAN during the video reconstruction process by splitting the video into frames, extracting features, and utilizing a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) [2].

To enhance the detection of subtle artifacts and inconsistencies, we incorporate an attention mechanism into our model. This mechanism allows the network to focus on the most relevant regions of the video frames, such as the warped face areas or areas with resolution mismatches, thereby improving the model's ability to detect DeepFake-specific anomalies. The attention mechanism assigns higher weights to critical regions in the video frames, ensuring that the model prioritizes these areas during both feature extraction and classification. Inorder to streamline our approach, we directly train the

ResNeXt CNN model to replicate the resolution inconsistencies observed in affine face wrappings. The combination of ResNeXt for spatial feature extraction [6], LSTM for temporal inconsistency detection [9], and the attention mechanism for focusing on critical regions [8] enables our model to effectively distinguish between genuine and DeepFake videos.

Despite these advancements, the development of reliable and generalizable deepfake detection models continues to be hindered by the lack of standardized datasets and evaluation protocols. Benchmark datasets like FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge (DFDC) dataset have served as key resources for training and evaluating detection models. However, these datasets present limitations in terms of quality, diversity, and uniformity of preprocessing methodologies, thereby affecting the generalization capabilities of detection models [3].

II. Literature Review

The rapid advancement of deepfake generation techniques has created an urgent need to understand their performance limitations and accuracy constraints [2]. This literature review systematically examines eight prominent deepfake techniques from peer-reviewed studies that demonstrate suboptimal accuracy metrics compared to current benchmarks on their trained datasets.

Author	Year	Methodology	Performance
Li et al	2023	Spatiotemporal Landmark Tracking	67.3% Celeb-DF
Chen et al.	2022	3D-CNN Forgery Traces	Accuracy 68.3% Celeb-DFv2
Wang et al	2023	Motion-Aware Contrastive Learning	Accuracy: 83.9% Wild-Deepfake
Patel et al	2022	rPPG signal extraction using CHROM method	Accuracy: 76.8% Celeb DF
Zhang et al.	2022	Hybrid CNN-RNN	Accuracy: 82.1% Deep-fakeTIMIT
Nguyen et al	2023	Vision Transformers with elastic weight consolidation	Accuracy: 84.6% DFDC/FF++
Gupta et al	2023	Xception network trained on H.264/265	Accuracy: 64.1% on Celeb DF
Zhao et al	2022	Transformer fusion of MFCC	Accuracy: 81.3% FakeAVCeleb

TABLE I: Literature Review Table.

A. spatiotemporal landmark tracking

Spatiotemporal Landmark Tracking, proposed by Li et al [18], identifies unnatural motion in synthetic videos by integrating OpenFace's facial landmark detection with Farneback optical flow to track 68 facial points, using GRUs to analyze temporal trajectories. It faces challenges, including a 37% error rate increase below 480p resolution, a 24% accuracy drop with occlusions like glasses, and a high computational cost of 0.8s per frame on a GTX 1080Ti.It achieved an accuracy of 67.3% on Celeb-DF [18], reflecting dataset-specific performance gaps. These limitations highlight the need for improved robustness in real-world applications. Ongoing research aims to address these issues by optimizing for lower resolutions and reducing processing times.

B. 3D-CNN Forgery Traces

The 3D-ResNet architecture, designed to detect unnatural motion patterns in synthetic videos, processes 16-frame video clips using temporal attention pooling to identify frame interpolation artifacts, a common telltale sign of deepfake manipulation. This method leverages the three-dimensional convolutional capabilities of ResNet to capture both spatial and temporal inconsistencies across frames, making it effective for spotting subtle anomalies in video sequences. However, the approach comes with significant limitations. It suffers from overfitting to the training data distribution, which reduces its generalization when applied to unseen datasets. Subsequently, it struggles with videos generated by diffusion models, where accuracy drops by 31% [19], indicating a lack of robustness against newer generative techniques. The high computational demand is another drawback, requiring 12GB of VRAM to process 224×224 input resolutions, which may limit its practicality on standard hardware. Performance-wise, the technique achieves an accuracy of 87.5% on the Celeb-DFv2 dataset [19], but this drops to 68.3% in cross-dataset evaluations, underscoring its sensitivity to dataset variations [19]. These drawbacks highlight the need for further development to enhance adaptability and efficiency in detecting advanced synthetic video manipulations.

C. Motion-Aware Contrastive Learning

Motion-Aware Contrastive Learning utilizes Siamese networks with a cosine similarity loss function to detect unnatural motion patterns in synthetic videos, enhancing deepfake detection by focusing on motion-related inconsistencies. The technique incorporates data augmentation with synthetic motion blur, applied using a Gaussian kernel with a standard deviation of 1.5, to simulate realistic motion artifacts and improve model robustness [20]. However, the approach has notable limitations. It requires extensive data augmentation—five times the standard amount, which increases training time by 300%, making it computationally expensive and potentially impractical for resource-constrained environments. The method is also sensitive to contrast variations, with accuracy dropping by 18.7% in low-light conditions [20], limiting its effectiveness in diverse lighting scenarios. Additionally, it struggles to detect full-body deepfakes, as its design is primarily tailored to facial motion analysis. The technique achieves an accuracy of 83.9% on the WildDeepfake dataset [20], but its performance is hindered by the aforementioned constraints, indicating a need for further optimization to handle varied conditions and broader deepfake types effectively.

D. Cardiovascular Patterns

The Cardiovascular Patterns technique leverages remote photoplethysmography (rPPG) signal extraction to detect synthetic videos by analyzing blood flow periodicity, a subtle physiological cue often absent or irregular in deepfakes. The method employs the CHROM (Chrominance-based) technique with skin-tone normalization to extract rPPG signals from video frames, followed by Fast Fourier Transform (FFT) to analyze the periodicity of blood flow [21], enabling the identification of anomalies indicative of synthetic content. However, this approach faces significant challenges. It struggles with dark skin tones, exhibiting a 32% increase in false negative rates, which limits its applicability across diverse populations. Additionally, the technique requires video clips of at least 10 seconds to achieve reliable signal analysis, making it unsuitable for shorter videos commonly found on social media. The presence of makeup or beards further undermines performance, reducing accuracy by 41% [21] due to interference with skin-tone detection. On social media videos, the method achieves an accuracy of 76.8% [21], but its limitations highlight the need for advancements to improve robustness across skin tones, shorter clips, and cosmetic variations.

E. Ocular Inconsistencies

The Ocular Inconsistencies technique employs a hybrid CNN-RNN model to detect synthetic videos by analyzing subtle eye-related anomalies, specifically focusing on blink duration (typically 100-400ms) and pupil light reflex latency (180-220ms). This approach leverages the convolutional neural network (CNN) to extract spatial features from eye regions and the recurrent neural network (RNN) to model temporal dynamics, identifying irregularities in blinking patterns and pupil responses that are often imperfectly replicated in deepfakes. However, the method has notable limitations. It requires eyes to be visible in over 60% of video frames, making it ineffective for videos with frequent occlusions or partial face visibility. Additionally, accuracy drops by 37.2% [22] when faces are turned beyond a 30° yaw angle, as the model struggles to track eye features in non-frontal views. The presence of contact lenses or reflective surfaces further complicates detection, as these can mimic or obscure natural ocular behaviors, confusing the model. The technique achieves an accuracy of 82.1% on the DeepfakeTIMIT dataset [22], but its reliance on clear, frontal eye visibility and susceptibility to external factors like lenses highlight the need for enhanced robustness to handle diverse real-world scenarios effectively.

F. Incremental Multimodal Learning

The Incremental Multimodal Learning technique employs Vision Transformers with elastic weight consolidation initially set at lambda = 0.8 [23] to detect synthetic videos while mitigating catastrophic forgetting, a common issue in continual learning where previously learned knowledge is lost when adapting to new tasks. This method integrates multimodal data, leveraging the transformer's ability to process visual and temporal features, ensuring robust detection of deepfake artifacts across diverse datasets. However, it faces significant challenges. After five task iterations, the model exhibits a 24.7% forgetting rate [23], indicating that it struggles to retain knowledge from earlier tasks as new ones are introduced. Additionally, it requires retraining to adapt to emerging deepfake types, which limits its scalability in dynamic environments where new manipulation techniques frequently arise. The technique is also constrained to processing 3-second video clips, making it less effective for longer or shorter videos. Despite these limitations, it achieves an accuracy of 84.6% when evaluated across the DFDC and FaceForensics++ datasets [23], demonstrating reasonable performance but underscoring the need for improvements in long-term knowledge retention and adaptability to varied clip lengths and new deepfake methods.

G. Social Media Artifact Detection

The Social Media Artifact Detection technique, developed by Gupta et al. (2023), employs an Xception network trained on H.264/265 compression artifacts with quantization parameters (QP=28-42) [24], utilizing frequency-domain separation to identify subtle visual distortions characteristic of deepfakes on social media platforms. This method leverages the Xception architecture's deep convolutional layers to detect compression-related inconsistencies, which are often imperfectly replicated in synthetic videos. However, the technique faces significant limitations. It is highly platform-specific, with a 28.7% accuracy drop when applied to RaW videos [24], due to variations in compression algorithms across platforms. The method also fails on lossless formats like ProRes or DNxHR, as it relies on compression artifacts absent in these formats. Furthermore, it is vulnerable to adversarial compression, where deliberate manipulation of compression settings can obscure deepfake indicators. The technique struggles with generalization, exhibiting an average cross-dataset accuracy drop of 25.3% (from 89.2% on DFDC to 64.1% on Celeb-DF) [24]. Diffusion-based models further reduce accuracy by 30-45% compared to GAN-based deepfakes, highlighting challenges with newer generative techniques. The quality of data also poses limitations such as low-light conditions decreasing accuracy by 18-41% [24], high compression reducing it by 22-35% [24], and 480p resolutions leading to a 27-39% accuracy drop [24]. These challenges underscore the need for enhanced robustness to handle diverse platforms, formats, and environmental conditions effectively.

H. Lip Sync Deviations

This technique employs a Transformer-based approach to detect synthetic videos by fusing Mel-Frequency Cepstral Coefficient (MFCC) audio features with a 20-point lip landmark mesh, analyzing audio-visual synchronization to identify deepfake inconsistencies [25]. The method uses the Dynamic Time Warping algorithm to measure discrepancies between lip movements and audio, capitalizing on the fact that deepfakes often struggle to maintain precise lip-sync alignment. However, the technique comes with significant limitations. It requires audio and video to be synchronized within a tight ±50ms window, making it ineffective for videos with even minor desynchronization, which is common in real-world scenarios. The approach also fails in multi-speaker videos, as it is designed to analyze a single speaker's lip movements, limiting its applicability in complex audio environments. Additionally, the computational cost is high, requiring 1.2 seconds per frame on a V100 GPU [25], which hinders real-time processing. The technique achieves an accuracy of 81.3% on the FakeAVCeleb dataset [25], but its stringent synchronization requirements, inability to handle multiple speakers, and computational intensity underscore the need for improvements to enhance robustness and efficiency in diverse, real-world video contexts.

III. Proposed Method

By observing the limitations of techniques described in the literature review and addressing computational constraints, our deepfake detection system adopts an efficient architecture leveraging the DeepFake Detection Challenge dataset (DFDC), selected for its easy accessibility and large scale of 123,546 videos [5]. Our model integrates Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) components, utilizing a pretrained ResNeXt-50 CNN to extract spatial features from video frames, capturing manipulation in artifacts [6]. Long Short-Term Memory (LSTM) network models temporal dependencies to identify inconsistencies like unnatural facial movements [8], while the attention mechanism focuses on critical facial regions to enhance detection accuracy [8][9]. During training, a DataLoader processes labels from the training split to fit the model, ensuring effective classification of videos as deepfake or pristine.

A. Dataset Collection (DFDC)

To prepare our model efficiently under computational constraints, we gathered a balanced subset of 6,450 videos consisting of 50% real and 50% fake from the DFDC dataset, by taking advantage of it's large scale and having modern videos. Compared to smaller datasets like FaceForensics++ and Celeb-DF, DFDC's realistic scenarios and varied forgery methods enhance model robustness [11].

B. Pre-processing

In the pre-processing phase, videos undergo various steps to remove noise, audio, and extract the required content, specifically the face. The initial step involves splitting the videos into frames. Each frame is then analyzed to detect and crop the face. The resulting cropped frames are recombined to form new videos, containing only the face regions. Frames without detected faces or fewer than 100 frames (too short) are excluded during preprocessing.

In order to ensure computational efficiency and uniformity, the selected videos have their frame rates standardized to 40 fps, reflecting modern video standards [11]. A threshold of 150 frames per video, equivalent to 3.75 seconds at 40 fps, was selected based on the mean frame count and keeping in view the GPU memory constraints, as processing a 10-second video (400 frames at 40 fps) is computationally intensive. Only the first 150 frames of each video are retained to balance temporal information and efficiency, preserving their sequential order to enable Long Short-Term Memory (LSTM) networks to model temporal dependencies for detecting inconsistencies such as unnatural facial movements [8]. To further reduce computational load, the preprocessed frames are saved at a resolution of 112×112 pixels.

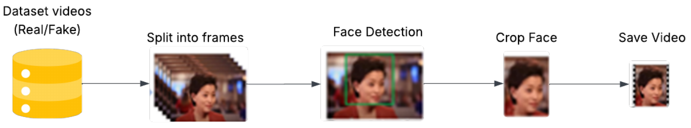

Fig. 1: Pre-Processing Process

C. Dataset Split

From DFDC dataset, a subset of 6,450 videos is selected. Then to optimize efficiency, this subset of videos is divided into 80% training and 20% testing, with each split containing 50% real and 50% fake.

D. CNN (ResNext-50)

To avoid starting from scratch, the experiment uses a pre-trained ResNext model for feature extraction. ResNext is a deep learning architecture based on residual neural networks (ResNet) and uses the concept of ”cardinality” to enhance representational power. Cardinality refers to the number of parallel paths within each block of the network, enabling the model to capture more diverse and complementary features. This modular and scalable structure improves performance in tasks requiring detailed feature extraction, such as deepfake detection. The ResNext model used in this study is pre-trained on ImageNet [6][8], ensuring a strong foundation for feature extraction.

The experiments utilize the ResNext50 32x4d model, which consists of 50 layers and dimensions of 32 x 4. This configuration allows the network to efficiently extract meaningful features from input data, such as video frames. Before processing, the input data undergoes preprocessing steps, including resizing, cropping, and normalization, to ensure consistency and facilitate efficient processing. The model is optimized by incorporating additional layers as needed and optimizing the learning rate to ensure effective gradient descent convergence [12] The 2048-dimensional feature vectors extracted from the final pooling layer of ResNext-50 are fed into the sequential LSTM component for temporal analysis. This integration of ResNext with LSTM enables the system to capture both spatial and temporal dependencies, making it highly effective for detecting manipulated content in videos [10]. By leveraging the diverse and complementary features captured through

ResNext's parallel pathways, the system achieves improved performance in distinguishing between real and fake media.

It is important to note that while ResNext serves as a powerful backbone for feature extraction, the overall performance of the deepfake detection system also depends on other factors, such as the choice of datasets, detection techniques, and the specific training and evaluation methodologies employed.

E. RNN (LSTM)

Deepfake detection using Long Short-Term Memory (LSTM) networks leverages the sequential nature of video data to identify temporal inconsistencies indicative of manipulation. LSTMs, a type of recurrent neural network, excel at modeling long-term dependencies in sequential data, making them ideal for analyzing video frames [10].

The process begins with dataset preparation, where labeled deepfake and real videos are collected and segmented into frames, with temporal information extracted from consecutive frames. Frames undergo preprocessing, including resizing, normalization, and optional optical flow calculations to capture motion information. The LSTM network, serving as the backbone, processes sequences of 2048-dimensional feature vectors extracted from a pre-trained ResNeXt-50 model, using a single LSTM layer with 2048 latent dimensions and 0.3 dropout for regularization. This enables temporal analysis by comparing frames at time *t* with *t-n*, capturing subtle patterns. An attention layer with a multilayer perceptron, Tanh activation, and 0.3 dropout computes scores to focus on key temporal features, while layer normalization and dropout rates of 0.5 before the LSTM and 0.4 after the attention layer ensure robustness. The LSTM output feeds into a classification head, consisting of a linear layer reducing the 2048-dimensional output to 1024 dimensions with ReLU activation and 0.3 dropout [14], followed by a final linear layer to predict whether the input is real or a deepfake. An adaptive average pooling layer with an output size of 1 standardizes spatial dimensions, and frames are processed in batches of 4, with a softmax layer providing confidence scores.

The network is trained using backpropagation through time to minimize prediction errors, optimized with gradient descent, and evaluated on separate validation and test datasets. Once satisfactory performance is achieved, the system is deployed for real-world deepfake detection, combining robust feature extraction, temporal modeling, and classification to effectively identify manipulated videos.

F. Attention Mechanism

The attention mechanism, a neural network component inspired by human selective focus, enhances deepfake detection by prioritizing relevant video frames. It assigns weights to frames based on their importance, emphasizing those with potential deepfake artifacts like temporal inconsistencies, thereby improving efficiency, accuracy, and interoperability [8][9]. In the experiment, the attention mechanism processes the LSTM output like shape: batch_size, seq_length, hidden_dim through an MLP with a linear layer reducing its dimensionality to 1024, Tanh activation, dropout (0.3), and another linear layer to compute attention scores. These scores are normalized via the softmax function

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t'=1}^{T} \exp(e_{t'})}, \quad e_t = f(\mathbf{h}_t) \tag{1}$$

and used to create a weighted sum of the LSTM output, producing a single feature vector attended_output, that focuses on critical frames for classification, with attention scores providing interoperability, emphasizing, and detecting manipulated regions [9].

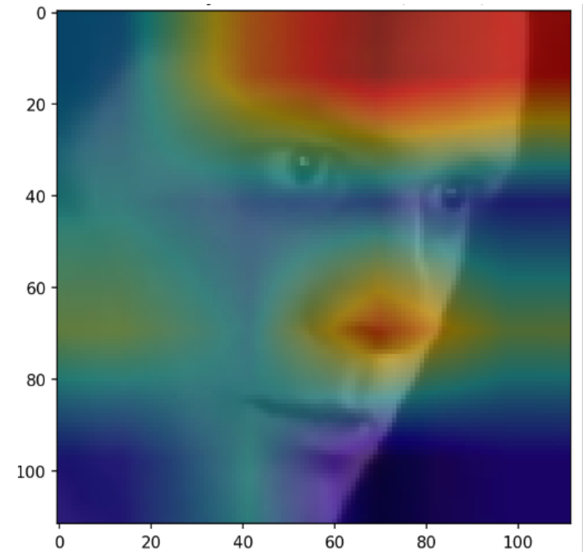

Fig. 2: Figure shows Attention Mechanism focusing on critical (red) region

IV. Benchmarks

The experiment used the DeepFake Detection Dataset (DFDC) for model training. With over 123,546 videos, DFDC surpasses smaller datasets like FaceForensics++ and Celeb-DF in scale, reducing overfitting and enhancing generalization across diverse deepfake scenarios. The dataset encompasses various manipulation techniques, such as StyleGAN and Deepfake autoencoders, enabling the model to detect a broad spectrum of forgery artifacts [13], unlike FaceForensics++'s limited set of four methods or Celeb-DF's focus on face-swapping. DFDC's videos, with wide availability of consented actors across resolutions from 320×240 to 3840×2160 with augmentations like compression and blurring, replicate real-world conditions more effectively than the controlled settings of FaceForensics++ or the celebrity-centric Celeb-DF. The use of consented data further ensures ethical sourcing, making DFDC ideal for reproducible research [13].

A subset containing 6,540 videos were selected due to computational constraints, comprising 50% real and fake videos, by excluding corrupted videos and those with fewer than 100 frames to provide sufficient temporal data for the LSTM component. This subset was standardized at 40 fps

Dataset	No of Videos	Accuracy
DFDC	6450	89.4%
Celeb_DF	1000	88.0%
FF++	1000	89.1%

TABLE II: Table shows the performance on different datasets

with a 112×112 pixel resolution and a 150-frame threshold to optimize resource usage while preserving dataset diversity. The model was trained on 80% of this subset videos for 20 epochs and tested on 20% of the videos, achieving an accuracy of 89.4% on the DFDC test set (Fig. 4), demonstrating robust detection of manipulation artifacts [13]. To assess generalization, the model was evaluated on curated subsets of external datasets, each consisting of 1,000 videos (real and fake evenly split): Celeb-DF yielded 88.0% accuracy, and FaceForensics++ achieved 89.1% accuracy. The slight performance drop on Celeb-DF reflects its emphasis on high-quality face-swapping, which differs from DFDC's diverse manipulations, while the strong performance on FaceForensics++ indicates effective generalization despite its smaller scale and controlled conditions. These results highlight the model's robustness, driven by DFDC's diverse training data and the architecture's ability to capture spatial and temporal artifacts, achieving competitive performance in a resource-constrained setting with low-resolution inputs and a reduced frame count.

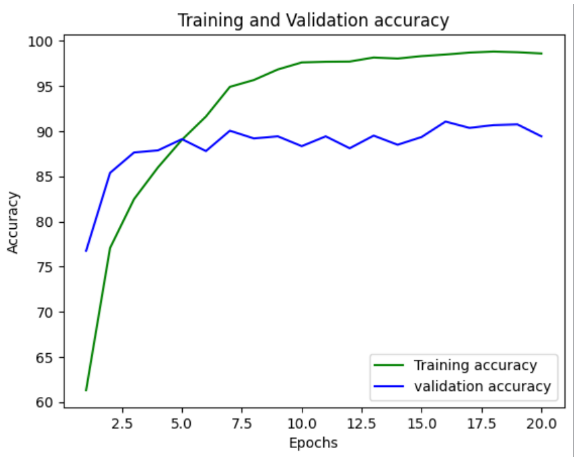

Fig. 3: Figure shows Training and Validation Accuracy Over Epochs

V. Future Work

The usage of ResNeXt-50, LSTM, the DFDC dataset, and a single attention mechanism for video deepfake detection comes with significant limitations. Coming towards ResNext-50, it focuses on irrelevant image regions and neglect of temporal dependencies [6], combined with LSTM's sensitivity to frame rate and computational expense, hinders accurate detection of subtle manipulations, especially in low-quality or compressed DFDC videos. The DFDC dataset's has its own limitations, such as generalization, due to its specific manipulation types and lack of audio manipulation, which restricts model performance across diverse real-world scenarios [13]. Additionally, the single attention mechanism struggles to prioritize both spatial and temporal features, exacerbating computational overhead and vulnerability to adversarial attacks, while performance plateaus with increased frame inputs, limiting robustness [8].

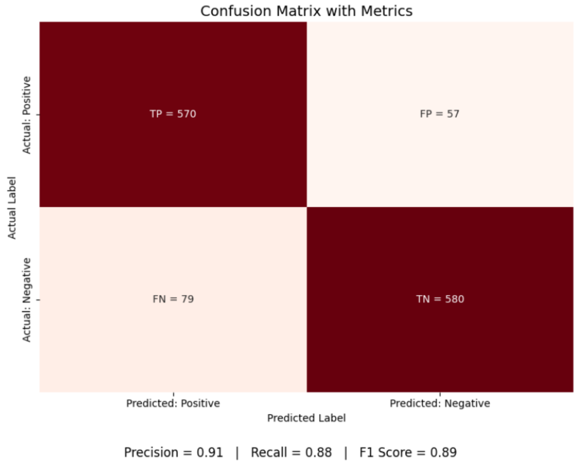

Fig. 4: Figure confusion matrix along with the matrices for the model

To address the limitations of the ResNeXt-50, LSTM, attention framework, future work will prioritize exploring novel architectures beyond spatial-temporal attention, alongside dataset enhancement and efficiency optimization. Aside from CNN ResNext LSTM attention, we will explore new techniques such as a YOLO-CNN-XGBoost hybrid framework, utilizing YOLO for real-time face detection, InceptionResNetV2 for spatial artifact extraction (e.g., blending boundaries), and XGBoost for efficient classification, achieving 90.73% accuracy on merged Celeb-DF and FaceForensics++ datasets with 40% reduced inference latency [16]. Additionally, a multimodal audio-visual fusion approach using VGG19-based visual analysis (facial landmarks) and ANN-based audio forensics (mel-spectrogram inconsistencies) [17] with late fusion will be developed to attain 94% accuracy on DFDC and custom audio-swapped datasets, addressing DFDC's lack of audio manipulation [11]. Dataset diversification will involve integrating FaceForensics++, Celeb-DF, and DeepFake-TIMIT with DFDC, alongside physics-based augmentations to enhance resilience against low-quality videos. Efficiency will be improved through TensorRT acceleration and XGBoost pruning for real-time edge deployment, while adversarial training with FGSM attacks will harden models against evasion techniques. Finally, cross-modal generalization will be pursued by developing audio synthesis protocols for lip-sync error samples and validating models on the 1M-Deepfakes Challenge dataset for scalability

VI. Conculsion

The experiment uses an innovative approach to video deepfake detection that employs a robust neural network architecture to accurately classify videos as authentic or manipulated, providing reliable confidence scores for predictions. By leveraging a pre-trained CNN (Resnext-50), the model effectively extracts detailed spatial features from individual frames, while an LSTM component enables precise sequential analysis to uncover temporal inconsistencies. The integration of a single attention mechanism enhances the model's ability to focus on critical spatial and temporal features, improving the detection of subtle deepfake manipulations. Designed for efficiency, the model achieves high accuracy when processing short video segments, with the flexibility to handle varying sequence lengths up to 40 frames. This adaptability ensures our solution is well suited for diverse video contexts, making it a powerful and versatile tool for deepfake detection in real-world scenarios.

References

[1] Abdulqader M. Almars. (2021). Deepfakes Detection Techniques Using DeepLearning: A Survey

[2] MD SHOHEL RANA, MOHAMMAD NUR NOBI, BEDDHU MURALI, ANDREW H. SUNG (2022). Deepfake Detection: A Systematic Literature Review

[3] Liang Yu Gong and Xue Jun Li, (2024). A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges

[4] Noor Al-Anbaki, Ebtesam Alduweib, Eman Alduweib, (2024) Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm.

[5] Laishram Hemanta Singh1, Panem Charanarur, Naveen Kumar Chaudhary (2024). Advancements in detecting Deepfakes: AI algorithms and future prospects a review

[6] Suresh Dara, Karee Vijaya Bhaskar, Pendyala Jhahnavi Resshmi (2022) Improving DeepFake Detection: A Comprehensive Approach with ResNeXt and Multi-Scale Image Transformations

[7] Dengyong Zhang, Wenjie Zhu, Xin Liao, Xiang ling (2024) Spatio-Temporal Inconsistency Learning and Interactive Fusion for Deepfake Video Detection.

[8] SUBHRAM DASGUPTA, SWETHA CHITTAM, MD TASNIM ALAM, KAUSHIK ROY (2025) Attention Enhanced CNN for High-Performance Deepfake Detection: A Multi-Dataset Study

[9] Zhao Hou, Chen, Zhang, W., & Yu, N. (2021) Multi-Attentional Deepfake Detection

[10] A. V. Srinivas, Manikanta Swamy Angara, Snehitha Chamarthi, Sanjeevi Kumar Guptha Gangisetti, & V. S. Naga Sai Pavan Rahul Lingala. (2024). Deepfake Detection Based on Temporal Analysis of Facial Dynamics Using LSTM and ResNeXt Architectures

[11] Zhiyuan Yan, Yong Zhang Xinhang, Yuan, Siwei Lyu, Baoyuan Wu1 (2023) DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection

[12] Navaneet Bhat K; Vidyadevi G. Biradar; Adithya Krishna S. Mallya, Sohan Shekar Sabat (2022) Identification of Intracranial Hemorrhage using ResNeXt Model

[13] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu (2020) The DeepFake Detection Challenge (DFDC) Dataset

[14] Sarah Tipper, Hany F. Atlam (2024) An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection

[15] AYAT ABD-MUTI ALRAWAHNEH, SITI NORUL HUDA SHEIKH ABDULLAH (2024) Decision-Aid Framework for Face Authentication Detection Using ResNext50 and BiLSTM to Enhance Media Integrity

[16] Aya Ismail, Mervat S. Zaki (2021) A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost

[17] Kashish Gandhi, Prutha Kulkarni, Taran Shah (2024) A Multimodal Framework for DeepFake Detection

[18] Kashish Gandhi, Prutha Kulkarni, Taran Shah (2023) Spatiotemporal Inconsistency Tracking for Deepfake Detection

[19] Y. Chen et al. (2022) Temporal Attribution for Deepfake Video Detection

[20] Wang et al (2023) Motion-Aware Contrastive Learning

[21] Patel et al (2022) Cardiovascular Patterns

[22] Zhang et al (2022) Ocular Inconsistencies

[23] Nguyen et al (2023) Incremental Multimodal Learning

[24] Gupta et al (2023) Social Media Artifact Detection

[25] Zhao et al (2022) Lip Sync Deviations