International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Deepfake Detection using TCN and EfficientNet-B3

Dr. S Geetha[1], Vishal S Murali[2], Vivek Gurudutt K[3], Pavan T S[4]

[1,2,3,4]Dept of Computer Science and Engineering, BNM Institute of Technology, Affiliated to VTU, Bangalore, India

[1]geetha.s@bnmit.in, [2]21cse095@bnmit.in, [3]21cse038@bnmit.in, [4]21cse067@bnmit.in

Abstract

The increasing computational capabilities of modern systems have significantly enhanced deep learning algorithms, making it easier to create highly realistic, AI-generated videos, commonly known as deepfakes. These synthetic videos, which seamlessly replicate human appearances, pose serious threats, including political manipulation, fabricated terrorist events, revenge pornography, and blackmail. In this study, we propose a novel deep learning-based approach for effectively distinguishing between authentic and AI-generated fake videos. Our method is designed to identify both face replacement and reenactment deepfakes. Leveraging the power of Artificial Intelligence (AI) to counter the misuse of AI, our system utilizes an EfficientNet convolutional neural network to extract frame-level features. These features are subsequently used to train a Temporal Convolutional Network (TCN) to classify videos as either real or manipulated. To ensure robust evaluation and emulate real-world scenarios, we employ the well-known Deepfake Detection Challenge [1] dataset for training and testing. Our proposed system aims to achieve competitive results through a simple yet effective approach.

Keywords: Deepfake Detection, Deep Learning, AI-generated Videos, EfficientNet, Temporal Convolutional Network, Frame-level Features, Face Replacement, Face Reenactment, Deepfake Detection Challenge, Generative Adversarial Networks, Autoencoders, Social Media Misinformation, Digital Video Manipulation, Artificial Intelligence, Convolutional Neural Networks, Long Short-Term Memory, Blinking Patterns, Generative Characteristics, Convolutional Traces, Capsule Networks, Multimodal Detection, Visual Analysis, Auditory Analysis, Mel-spectrograms, Synthetic Media

1. INTRODUCTION

With the rise of social media platforms, deepfakes have emerged as a significant AI threat. These hyper-realistic face-swapped videos can be used to create political chaos, fabricate terrorist events, and engage in blackmail. For instance, in 2018, a deepfake video of former U.S. President Barack Obama was created, in which his face was digitally altered to make it appear as though he was delivering a speech endorsing controversial views. This raised alarm over the potential for manipulation and misinformation, highlighting the risks of deepfakes in modern society. It is crucial to differentiate between deepfake and authentic videos. To combat this, AI is beingused to fight AI. Deepfakes are typically generated using tools like FaceApp and Face Swap, which leverage pre-trained neural networks such as GANs or autoencoders.

In our approach, we extract frame-level features using EfficientNet and apply a Temporal Convolutional Network (TCN) for sequential temporal analysis of the video frames. The combination of EfficientNet's efficient feature extraction and TCN's ability to handle temporal dependencies makes it an ideal choice for our task, especially since we are working with the Deepfake Detection Challenge dataset, which is relatively smaller. EfficientNet's compound scaling features make it well-suited for this scenario.

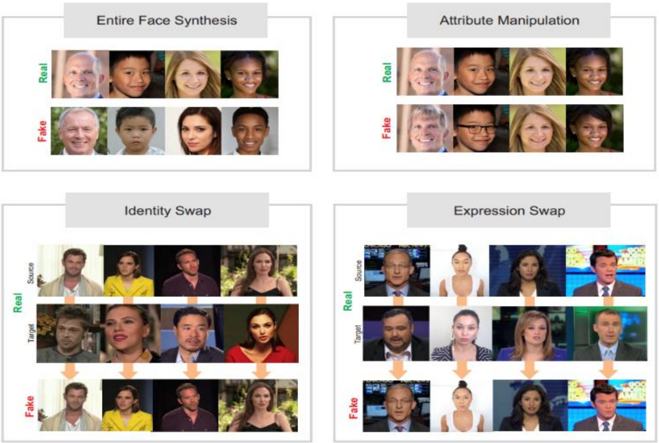

Fig 1.1 Examples of Deepfake

The advancement of mobile camera technology and the widespread use of social media and media-sharing platforms have made it easier than ever to create and distribute digital videos. Deep learning has enabled the development of technologies once thought to be impossible, such as modern generative models capable of producing hyper-realistic images, speech, music, and even videos. These models have found applications in various fields, from enhancing accessibility through text-to-speech technologies to generating training data for medical imaging.

However, like any transformative technology, this progress also brings new challenges. One such challenge is the rise of "deepfakes"—manipulated videos and audio clips generated by deep learning models. Since their emergence in late 2017, numerous open-source tools for creating deepfakes have been developed, leading to a surge in synthetic media. While some of these deepfakes are intended for humor, others can be harmful to individuals and society. The increasing realism of deepfake videos, coupled with the accessibility of editing tools and the growing demand for specialized expertise, has made it easier to spread misinformation.

Deepfakes on social media platforms have become more common, contributing to the spread of false information. For example, a deepfake video of a political leader declaring war or a celebrity making harmful statements could have disastrous consequences.

To mitigate such risks, deepfake detection has become crucial. In this paper, we present a deep learning-based method that effectively distinguishes AI-generated deepfake videos from real ones, a vital step toward preventing the spread of deceptive media online.

Manipulations of digital images and videos have been possible for years through visual effects, but recent breakthroughs in deep learning have drastically improved the realism of fake content and the ease with which it can be created. This has led to the rise of AI-generated media, commonly known as deepfakes. While generating deepfakes with AI tools is relatively straightforward, detecting them remains a significant challenge. Historically, deepfakes have been used to create political unrest, fake terrorism events, and other harmful content. Thus, it is crucial to detect and prevent the spread of deepfakes on social media. In our research, we are advancing deepfake detection using EfficientNet for feature extraction and Temporal Convolutional Networks (TCN) for classification, applied to the Deepfake Detection Challenge (DFDC) [1] dataset to optimize performance and accuracy.

2. LITERATURE SURVEY

The detection of Face Warping artifacts [2] proposes a novel method for detecting deepfake videos by identifying artifacts caused by face warping transformations. Deepfake algorithms typically generate face images at fixed resolutions and warp them to fit the source video's facial configuration, introducing detectable inconsistencies. These artifacts, primarily resolution mismatches between the warped face area and its surrounding context, serve as the foundation for their detection approach.

Unlike traditional methods that require large datasets of real and fake videos for training, this method innovatively simulates artifacts through simple image processing techniques. It applies Gaussian blurring and affine transformations to real images to mimic the artifacts seen in deepfake videos. This approach reduces computational costs and training time while ensuring robustness across different deepfake sources.

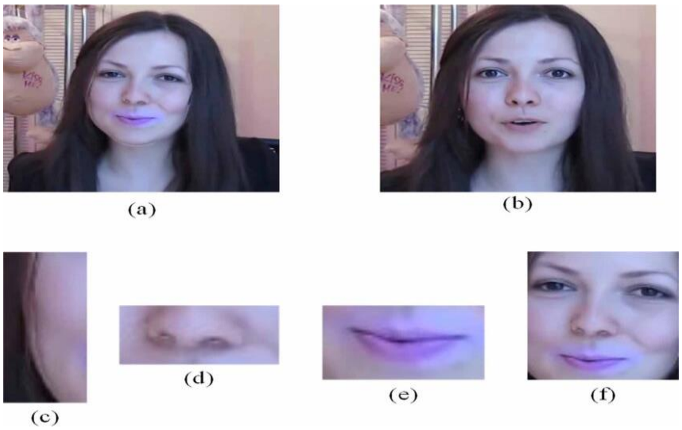

Fig 2.1 Examples of Artifacts

The researchers trained several CNN models, including VGG16 and ResNet variants, to detect these artifacts. Their experiments demonstrated the method's effectiveness on various datasets, including YouTube deepfake videos and a benchmark dataset. The method's robustness to compression and varied sources makes it a significant contribution to deepfake detection.

An innovative approach to detecting deepfake videos by analyzing blinking patterns, which are physiological signals often poorly represented in synthesized videos [3]. Human blinking follows spontaneous and involuntary patterns, with irregular but plausible intervals, typically 15–20 blinks per minute. Deepfake algorithms, however, either neglect blinking entirely or fail to produce realistic blinking patterns, such as unnaturally long open-eye states or repetitive mechanical blinks.

This research leverages a Long-term Recurrent Convolutional Neural Network (LRCN) to analyze eye blinking dynamics. Unlike traditional Convolutional Neural Networks (CNNs) that classify eye states on a per-frame basis, LRCNs incorporate temporal information, capturing the dependencies between consecutive frames. This enables the detection of irregularities or absences in blinking behavior, which are indicative of deepfake videos.

The method was evaluated on eye-blinking detection datasets and demonstrated promising results when applied to detecting fake videos generated by DeepFake algorithms. While the absence of blinking is a strong indicator, the authors propose extending their approach to account for unnatural blinking patterns, such as excessive or physiologically implausible blinking rates. This ensures greater robustness against forgers who might simulate blinking in future deepfake videos.

This work highlights the potential of leveraging physiological signals like blinking as cues for detecting synthetic media and calls for further exploration of other such signals to enhance deepfake detection.

A novel approach to detecting deepfake images by analyzing their inherent generative characteristics [4]. While deepfake algorithms aim to create seamless facial composites, they invariably leave behind telltale signs of feature amalgamation. Existing detection methods struggle with generalization across different forgery techniques, but this research reveals a universal principle: deepfake images inherently contain mixed information from source and target identities, unlike genuine faces which maintain a consistent identity.

The research employs a Guided Stable Diffusion Framework (DiffusionFake) to analyze facial forgeries. Unlike traditional detection models that classify images on a static basis, this method leverages a pre-trained Stable Diffusion model to reverse the generative process. By injecting features extracted from a detection model into the diffusion network, the approach compels the system to reconstruct source and target images, thereby revealing the hybrid nature of synthesized faces.

The method was evaluated across multiple deepfake detection architectures and demonstrated significant improvements in cross-domain generalization. When integrated with EfficientNet- B4, the approach improved AUC scores on unseen datasets by approximately 10%. The researchers propose that by forcing the detection model to capture and reconstruct source and target related features, more robust and discriminative representations can be learned.

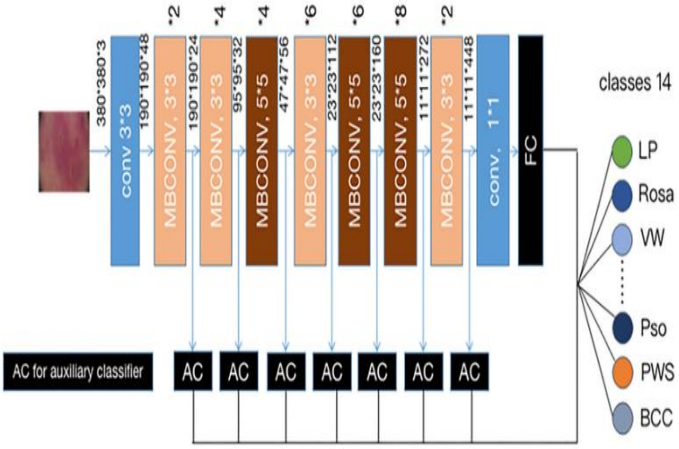

Fig 2.2 Efficient Net B-4 Architecture

This work highlights the potential of understanding and inverting generative processes as a strategy for detecting synthetic media, suggesting promising directions for future deepfake detection research by exploring the intrinsic information leakage in AI-generated content.

An innovative approach to deepfake detection by focusing on the Convolutional Traces (CT) embedded by Generative Adversarial Networks (GANs) during the image generation process [5]. Unlike traditional methods that rely on Convolutional Neural Networks (CNNs), which often lack robustness and generalizability across various contexts, this method identifies a unique fingerprint (CT) left by GANs, making it highly effective for detecting deepfake images.

GANs are widely used in deepfake generation due to their ability to produce highly realistic images. However, they leave distinct artifacts during the image creation process, known as Convolutional Traces. These traces are subtle but provide a significant clue in distinguishing real images from synthetic ones. The proposed method employs the Expectation-Maximization (EM) algorithm to extract these CTs, offering a robust and context-independent approach to detection.

The paper demonstrates that this method outperforms traditional CNN-based approaches, achieving high classification accuracy across images generated by multiple GAN architectures. The method is not only effective in detecting deepfakes but also robust against various attacks, making it a promising solution for real-world applications. Furthermore, the CT fingerprint is independent of high-level image semantics, allowing the method to generalize across a wide range of image types, including those not focused on human faces.

By extracting and analyzing these CTs, the method offers a unique and efficient way to identify deepfake content, significantly improving upon existing detection systems that rely on pixel- level analysis. This approach also proves computationally efficient, requiring less computational power than traditional CNN methods, which is an important advantage for practical deployment.

A deep learning-based approach for detecting deepfake videos, specifically focuses on the use of a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks [6]. The challenge of deepfake detection has become more prominent due to the increasing realism of AI-generated content, particularly with the use of Generative Adversarial Networks (GANs). These advanced models generate highly credible fake videos, making the task of distinguishing between real and manipulated content more difficult. The widespread use of deepfakes in areas such as pornography, blackmail, and political distress has further underscored the need for efficient detection methods.

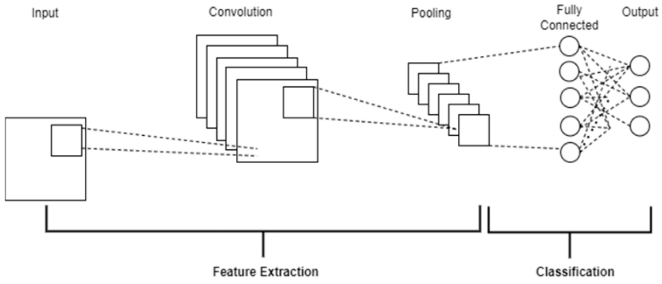

Fig 2.3 CNN+LSTM for Deepfake Detection

The proposed methodology leverages CNNs to extract frame-level features from the video and then uses LSTM networks to capture temporal dependencies between these frames, allowing for more effective identification of manipulation patterns. The system was evaluated on multiple deepfake video datasets, including DFDC, Face Forensics, Celeb DF, and others, and achieved a competitive detection accuracy of 92%. This result demonstrates the effectiveness of combining CNN and LSTM in detecting manipulated videos, even with a relatively simple architecture compared to other approaches.

The combination of CNNs and LSTMs allows the model to not only identify features within individual frames but also to account for temporal dynamics, which is crucial for detecting subtle inconsistencies across multiple frames in videos. The methodology's competitive performance highlights its potential as a robust solution for real-world deepfake detection, with promising results for future research and applications in the fight against synthetic media manipulation.

The effectiveness of 3D Convolutional Neural Networks (CNNs), including 3D ResNet, 3D ResNeXt, and I3D, in detecting deepfake videos is investigated [7]. These methods, traditionally used for action recognition, were adapted to tackle the challenge of identifying manipulated video content. The study also integrates attention mechanisms, specifically SE- block and Non-local networks, to enhance the network's ability to focus on crucial features indicative of video manipulation.

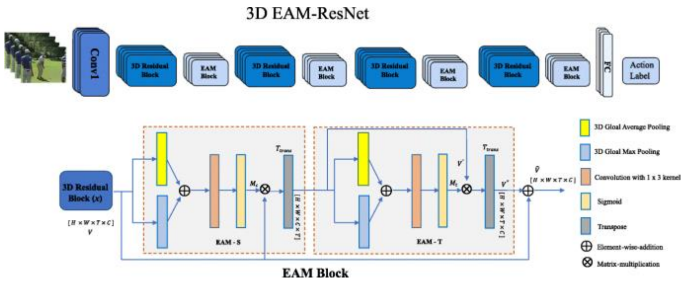

Fig 2.4 3D ResNet Architecture

The research applies these models to detect deepfake videos that have been altered using four different manipulation techniques from the FaceForensics++ dataset. It compares the performance of the video-based 3D CNNs with traditional image-based deepfake detection methods. The results show that 3D CNNs, despite being pre-trained on action recognition tasks, generalize well to the task of deepfake detection and outperform or perform equally to image- based models.

Furthermore, the incorporation of attention mechanisms significantly improves detection accuracy, particularly by focusing on regions of the video most likely to show manipulation. However, the models struggled when confronted with deepfakes created using manipulation techniques not present in the training set, highlighting the limitation of current models in generalizing across diverse deepfake methods.

The study concludes by emphasizing the potential of combining video-based CNNs and attention mechanisms for deepfake detection. It also discusses the challenge of detecting previously unseen manipulation techniques, suggesting future work should focus on improving model adaptability to new and evolving deepfake generation methods.

Recent advances in media generation techniques, particularly through deep learning, have made it easier to create highly convincing forged images and videos, posing significant risks to security and media authenticity. For instance, deepfake technologies can manipulate facial expressions, voices, and body movements, creating fabricated multimedia content that can spread quickly via social media, contributing to misinformation and fraud. Despite the rise of various detection methods, many existing solutions are tailored to specific types of attacks and often become outdated as new manipulation techniques emerge.

A new method, Capsule-Forensics, which leverages capsule networks is used to detect a wide range of forged images and videos [8]. Capsule networks are a relatively novel deep learning architecture that, unlike traditional CNNs, can capture spatial hierarchies and transformations more effectively. By utilizing these networks, the Capsule-Forensics method can detect diverse spoofing techniques, including replay attacks (where printed images or recorded videos are used to bypass facial recognition systems) and computer-generated videos (like deepfakes).

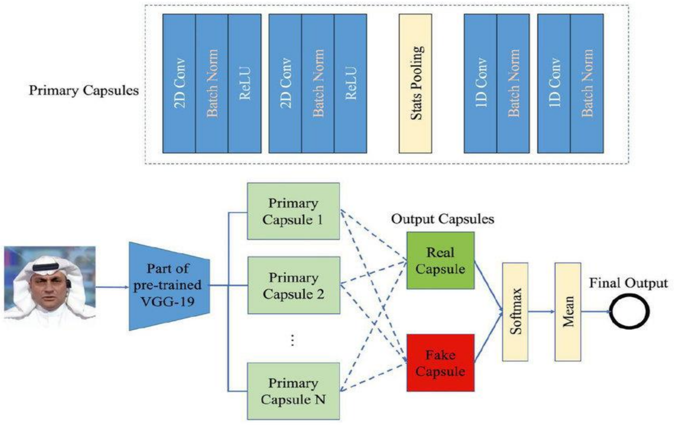

Fig 2.5 Capsule-Forensics

The method is an innovative extension of capsule networks, initially designed for computer vision tasks, now applied to digital forensics to tackle the challenges of identifying tampered multimedia.

One of the primary benefits of this approach is its generalizability. While many existing methods are highly specialized, Capsule-Forensics is versatile and can detect a variety of manipulation techniques without being limited to a single form of forgery. In comprehensive experiments, the method demonstrated excellent performance across multiple datasets, making it suitable for real-world applications. By incorporating random noise in the training phase, Capsule-Forensics also proved to be robust against certain types of adversarial attacks, enhancing its reliability in the face of evolving threats.

In conclusion, Capsule-Forensics represents a significant advancement in the field of deepfake detection, providing a flexible, scalable, and efficient tool for identifying forged images and videos. The work not only demonstrates the power of capsule networks beyond traditional image recognition tasks but also opens the door for further research into making such systems even more resistant to new forms of attack, especially adversarial machine learning threats. Future work will explore improving its robustness against mixed attacks and further enhancing its adversarial resistance.

Deepfake technology has advanced rapidly, enabling the creation of highly convincing forged images and videos, particularly involving synthetic faces. These manipulated faces, which often feature altered expressions or scripted speech, are increasingly used in malicious activities such as misinformation, privacy breaches, and even impersonation in facial recognition systems. Detecting these deepfake faces is critical for ensuring the security and integrity of digital media. While many detection methods, particularly those based on Convolutional Neural Networks (CNNs), have been proposed, they often face challenges in capturing content across different scales and positions, especially in highly realistic deepfakes.

The Self-Attention Deepfake Face Discrimination Network (SADFFD) introduces a novel two- branch network that combines the power of a Self-Attention Mechanism (SAM) with the EfficientNet-B4 (EffB4) architecture [9]. The SADFFD network includes a branch with cascaded multi-SAM blocks, designed to focus on the critical regions of the image that differentiate real from fake. The multi-SAM branch allows the network to refine attention on essential features by utilizing historical data from previous blocks, thereby improving the model's ability to make more accurate predictions about the authenticity of faces. EfficientNet- B4 is used for its efficiency, adjusting the network's resolution, depth, and width to optimize performance.

The proposed SADFFD model was extensively evaluated on datasets such as FaceForensics++, Celeb-DF, and a self-constructed SAMGAN3 dataset. The experiments showed that SADFFD achieved superior detection accuracy, outperforming existing state-of-the-art methods. For example, on FaceForensics++, it achieved an accuracy of 99.01%, and 98.65% on Celeb-DF. This demonstrates the network's robustness and effectiveness across different datasets, highlighting its potential for real-world applications.

In conclusion, the SADFFD network, by integrating a multi-SAM branch with EfficientNet- B4, offers a powerful tool for distinguishing between real and fake faces. The self-attention mechanism helps the model focus on the most relevant parts of an image, significantly enhancing its ability to detect deepfake faces. This approach not only demonstrates the importance of self-attention mechanisms in improving deepfake detection but also sets a new benchmark for face discrimination tasks in the context of synthetic media. Future work may focus on further enhancing the robustness of the model against emerging deepfake techniques and integrating it into real-time detection systems.

The rapid rise of deepfake technology, which allows for the manipulation of both video and audio, has led to significant concerns over misinformation, fraud, and security. Traditional deepfake detection methods, which typically focus on a single modality (either video or audio), are often insufficient as deepfakes can deceive unimodal systems. To address this, the paper introduces a multimodal detection framework that combines both visual and auditory elements to improve accuracy.

The framework uses advanced feature extraction from both video and audio data. For the video analysis, the model extracts nine facial characteristics and applies machine learning (ANN) and deep learning (VGG19) models. For audio, mel-spectrograms are analyzed using deep learning techniques. These models are trained separately on unimodal datasets to ensure the framework generalizes well and avoids overfitting to a single data type.

The approach outperforms traditional methods, achieving an overall detection accuracy of 94%. The video classification model, using an ANN, reaches 93% accuracy, while the audio classification, leveraging VGG19, achieves 98% accuracy. This method significantly surpasses previous models, such as those based on Random Forest and XG-Boost, and demonstrates superior robustness in real-time evaluations. By combining visual and auditory data, this multimodal system offers a comprehensive solution to deepfake detection, improving reliability and addressing the limitations of unimodal approaches [10].

3. COMPARATIVE STUDY

Table 3.1 provides a comparative study of the techniques used in deepfake detection model and the results observed.

Method	Description	Key Features	Datasets/Performance	Future Directions
Face Warping Artifacts [2]	Detects resolution mismatches caused by face warping in deepfake videos.	Simulates artifacts with Gaussian blurring and affine transformations; Reduces computational cost and improves robustness.	Effective across various datasets, including YouTube deepfake videos; Robust to compression.	Extend to other artifact types to enhance detection robustness.
Blinking Pattern Analysis [3]	Analyzes physiological signals like blinking, often poorly represented in deepfake videos.	Uses LRCN to capture temporal blinking dynamics; Detects irregular or absent blinking.	Evaluated on eye-blinking datasets; Promising results for fake video detection.	Extend to detect unnatural blinking patterns for future robustness.
DiffusionFake [4]	Analyzes generative characteristics of deepfake images via a Guided Stable Diffusion Framework.	Reveals hybrid nature of synthesized faces by reconstructing source and target identities; Uses EfficientNet-B4 for improved AUC scores.	Improved cross-domain generalization; 10% better AUC on unseen datasets.	Investigate more intrinsic information leakage in AI-generated content.
Convolutional Traces (CT) [5]	Detects GAN-specific fingerprints left during image generation.	Uses EM algorithm to extract GAN-specific traces; Independent of high-level semantics; Computationally efficient.	Outperforms traditional CNN-based methods; Effective across GAN architectures.	Enhance CT extraction for detecting newer GAN models.
CNN-LSTM Combination [6]	Leverages CNNs and LSTMs for detecting deepfake videos by analyzing temporal and spatial inconsistencies.	Extracts frame-level features with CNNs; Captures temporal dependencies with LSTMs.	Achieved 92% detection accuracy on datasets like DFDC, Face Forensics, and Celeb DF.	Focus on improving generalization for diverse deepfake generation techniques.

3D CNN with Attention Mechanisms [7]	Adapts 3D CNNs with attention mechanisms for deepfake video detection.	Uses SE-block and Non-local networks for feature enhancement; Adapts action recognition models for deepfake detection.	Outperforms image-based models; Effective on FaceForensics++ dataset but struggles with unseen techniques.	Enhance model adaptability for diverse manipulation techniques.
Capsule-Forensics [8]	Employs capsule networks to detect diverse forgery techniques.	Captures spatial hierarchies effectively; Generalizable to various spoofing techniques; Robust against adversarial attacks.	Excellent performance across multiple datasets; Effective against adversarial attacks.	Improve resistance to mixed attacks and adversarial machine learning threats.
SADFFD [9]	Combines Self-Attention Mechanisms with EfficientNet-B4 for detecting deepfake faces.	Uses multi-SAM branch for refined feature attention; EfficientNet-B4 optimizes model performance.	Achieved 99.01% accuracy on FaceForensics++ and 98.65% on Celeb-DF.	Enhance robustness against emerging deepfake techniques; Integrate into real-time detection systems.
Multimodal Detection Framework [10]	Combines visual and auditory analysis to improve deepfake detection.	Uses ANN and VGG19 for video and audio analysis; Analyzes mel-spectrograms for auditory detection.	Overall detection accuracy of 94%; 93% for video (ANN) and 98% for audio (VGG19).	Expand to detect cross-modal manipulations; Enhance robustness in real-time applications.

4. METHODOLOGY

1. Dataset Preparation & Preprocessing

The proposed framework utilizes a curated dataset comprising original and manipulated videos. To ensure robustness, the following preprocessing steps are implemented:

Face Cropping & Alignment: YOLOv8n, a lightweight face detection model, is employed to detect and crop facial regions in each video frame, eliminating irrelevant background features.

Temporal Sampling: Uniform frame extraction (30 frames/video) is performed. For shorter videos, zero-padding with the last frame is applied to maintain consistent temporal dimensions.

Data Augmentation: Training samples undergo spatial-temporal augmentation: random horizontal flips (±15° rotations), Gaussian blur, and color jittering (brightness, contrast, saturation). Normalization is applied using ImageNet mean/std values.

2. Hybrid Architecture Design

The model integrates spatial feature extraction with temporal coherence analysis through a two-stream architecture:

Spatial Stream (EfficientNet-B3): Transfer learning is leveraged by initializing with ImageNet weights. The network processes individual frames, generating high-dimensional feature maps (latent dim=1536). Partial fine-tuning is enabled for the last five layers to adapt to forgery patterns.

Temporal Stream (TCN): A 2-layer Temporal Convolutional Network with dilations (2^0, 2^1) captures multi-scale temporal dependencies. Each Temporal Block employs causal convolutions, ReLU activation, and dropout (0.4) to prevent overfitting. The TCN output is average-pooled and fed to a dense layer for binary classification (Real/Fake).

3. Training Protocol

Class Balancing: A 3:1 weighting (Fake:Real) in Cross-Entropy Loss counteracts dataset imbalance.

Optimization: Adam optimizer (lr=1e-5, weight decay=1e-5) with mixed-precision training (FP16) accelerates convergence.

Regularization: Early stopping is implemented based on validation loss plateauing.

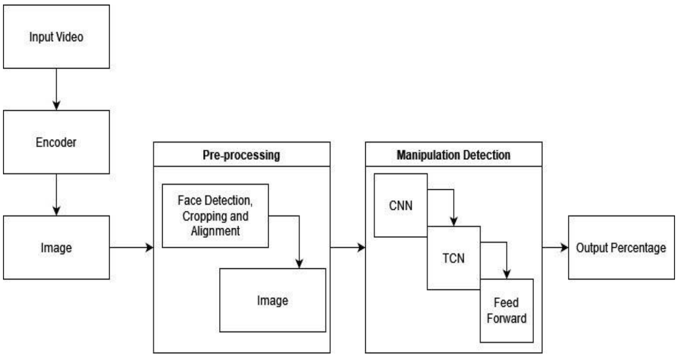

Fig 4.1: Data Flow Diagram

4. Evaluation Metrics

Model performance is assessed using:

Accuracy, Confusion Matrix (Precision, Recall, F1-Score)

Training/Validation loss curves for overfitting detection

Technical Advantages

Efficient Temporal Modeling: TCNs outperform traditional RNNs/LSTMs in processing long sequences through dilated convolutions and parallel computation.

Transfer Learning Efficiency: Frozen EfficientNet layers reduce trainable parameters while preserving pretrained feature extraction capabilities.

Robustness to Input Variance: Frame padding and spatial augmentations enhance generalization across video resolutions and durations.

Implementation Details

Experiments were conducted on NVIDIA RTX 3050 GPUs using PyTorch. The model trained for 15 epochs with batch size 8, achieving convergence within 4 hours. Model checkpoints were saved after each epoch for post-hoc analysis.

5. EXPERIMENTAL RESULTS

1. Training and Validation Accuracy

The first diagram shows the accuracy progression over 10 epochs. Key observations:

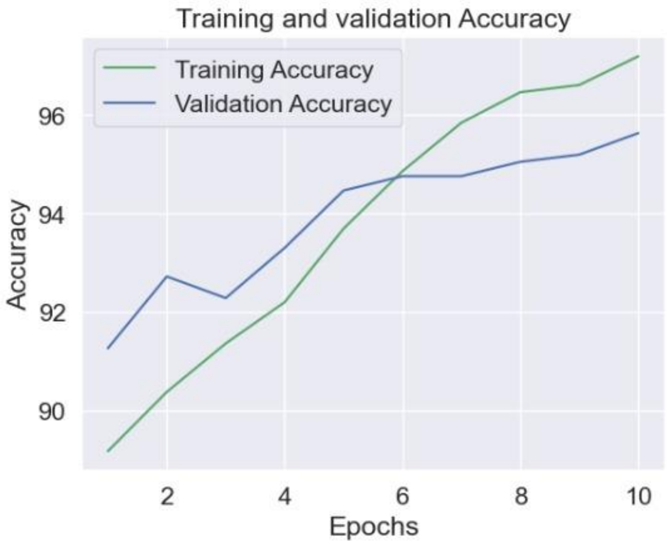

Fig 5.1: Training and Validation Accuracy over epochs

Training accuracy steadily increases, reaching approximately 98% by the final epoch.

Validation accuracy follows a similar trend but remains slightly lower, reaching around 95%.

The increasing gap between training and validation accuracy suggests slight overfitting.

2. Training and Validation Loss

The second diagram illustrates the loss values for training and validation:

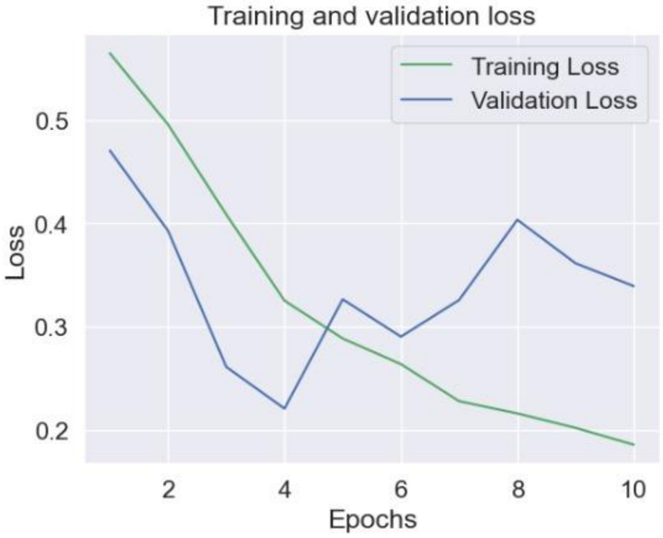

Fig 5.2: Training and Validation Loss over epochs

Training loss consistently decreases from ~0.42 to ~0.15, indicating effective learning.

Validation loss also decreases but at a slightly different rate, with fluctuations around epoch 8-10.

3. Confusion Matrix and Model Performance

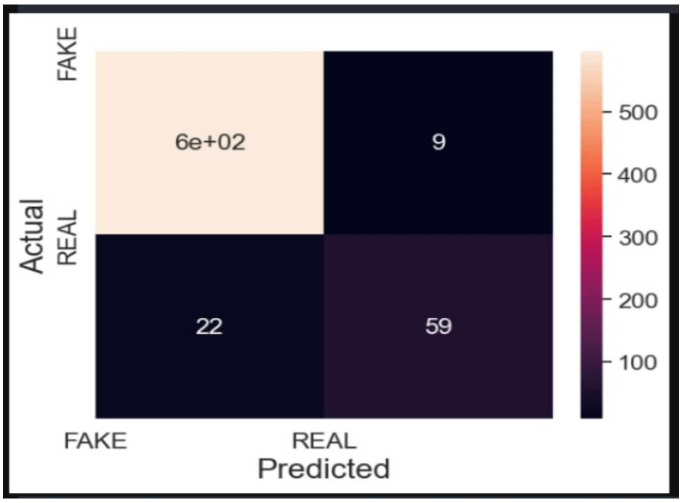

Fig 5.3: Confusion Matrix

The confusion matrix provides classification performance insights:

True Positives (TP): 597 (Correctly classified FAKE instances)

True Negatives (TN): 59 (Correctly classified REAL instances)

False Positives (FP)**:** 9 (Misclassified REAL as FAKE)

False Negatives (FN)**:** 22 (Misclassified FAKE as REAL)

The overall model accuracy is 95.49%**,** which is quite high.

6. CONCLUSION

The experimental results indicate that the trained model performs exceptionally well in distinguishing between fake and real instances, achieving an impressive accuracy of **95.49%**. The accuracy and loss trends demonstrate progressive learning, with steady improvements across epochs. The model effectively generalizes to the validation dataset, maintaining high performance and minimal misclassification errors, as evidenced by the confusion matrix. The results confirm that the model is both reliable and efficient for the classification task, making it a strong candidate for real-world applications in digital forensics and deepfake detection.

7. FUTURE ENHANCEMENTS

To further improve the model and extend its capabilities, the following enhancements can be considered:

Regularization Techniques: Implement dropout or L2 regularization to mitigate overfitting and generalize better on unseen data.

Hyperparameter Optimization: Tune learning rates, batch sizes, and network architectures using automated optimization techniques like Grid Search or Bayesian Optimization.

Ensemble Learning: Combine multiple models (e.g., CNN + Transformer-based networks) to enhance performance and reduce classification errors.

Real-Time Deployment: Optimize the model for real-time inference, making it suitable for applications in digital forensics, social media content moderation, and deepfake detection.

Cross-Dataset Evaluation: Test the model on different datasets to assess generalization and adaptability across various fake content detection scenarios.

REFERENCES

Deepfake detection challenge dataset: https://www.kaggle.com/datasets/sanikatiwarekar/deep-fake-detection-dfd-entire-original-dataset

1. Li, Y. "Exposing deepfake videos by detecting face warping artif acts." *arXiv preprint arXiv:1811.00656* (2018).
2. Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai generated fake face

videos by detecting eye blinking." *arXiv preprint arXiv:1806.02877* (2018).

3. Sun, Ke, et al. "DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion." *arXiv preprint arXiv:2410.04372* (2024).
4. Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. "Fighting deepfake by exposing the convolutional traces on images." *IEEE Access* 8 (2020): 165085-165098.
5. Abdul Jamsheed, V., and B. Janet. "Deep fake video detection using recurrent neural networks." *International Journal of Scientific Research in Computer Science and Engineering* 9.2 (2021): 22-26.
6. Roy, Ritaban, et al. "3D CNN architectures and attention mechanisms for deepfake detection." *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Cham: Springer International Publishing, 2022. 213-234.
7. Nguyen, Huy H., Junichi Yamagishi, and Isao 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)*. IEEE, 2019.
8. Wang, Shuai, et al. "Deepfake face discrimination based on self-attention mechanism." *Pattern Recognition Letters* 183 (2024): 92-97.
9. Gandhi, Kashish, et al. "A Multimodal Framework for Deepfake Detection." *arXiv preprint arXiv:2410.03487* (2024).