Robust Deepfake Detection Using Resnet and LSTM Networks: A Hybrid Approach for Enhanced Media Integrity

Sri Charan, Sai Rishik, Samhith, Neha Deepthi and Varshitha

School of Computer Science and Engineering, Vellore Institute of Technology, Amaravathi, 522237, Andhra Pradesh, India

Abstract

The growing availability of deepfake technology, especially through Generative Adversarial Networks (GANs), has sparked significant concerns regarding the authenticity of digital content and the risks of its misuse. Conventional detection methods mainly rely on Convolutional Neural Network (CNN) architectures, emphasizing spatial features while frequently neglecting crucial temporal relationships between frames. This oversight may hinder their ability to effectively analyze video sequences[1]. In this study, we present a hybrid model that initially employs ResNet for spatial feature extraction, subsequently channeling these intermediate outputs into an LSTM to effectively capture temporal dependencies, thereby enhancing the accuracy of fake content detection. This model, applied to a selectively trimmed dataset, effectively identifies temporal inconsistencies that may be overlooked by single-frame approaches. In contrast to solely spatial detection techniques[2][4], our ResNet-LSTM architecture demonstrates significant improvements in accuracy, providing a more dependable approach for identifying maliciously modified media. This study highlights the critical need for effective detection tools to tackle the changing challenges presented by synthetic content[3].

Keywords : Deepfake detection, ResNet-LSTM hybrid model, Temporal-spatial feature analysis, Digital media integrity, Intermediate feature extraction, Temporal inconsistencies, Trimmed dataset analysis, Synthetic media forensics.

1. Introduction

Deepfakes, or synthetic media, have become increasingly popular because to the quick development of artificial intelligence. Deepfake technology, which is mostly driven by Generative Adversarial Networks (GANs), makes it possible to create incredibly realistic-looking yet fraudulent media content, such as pictures and movies. Although there are some beneficial uses for this technology, including in virtual reality and entertainment, there are also serious hazards, especially when it comes to the malicious alteration of video footage for identity theft and disinformation (1). Thus, identifying deepfakes has become essential to preserving the legitimacy and dependability of digital content.

Convolutional Neural Networks (CNNs), which concentrate on examining spatial data within individual video frames, have been a major component of traditional methods for deepfake identification (2). However, subtle temporal discrepancies that could indicate synthetic material are frequently missed by this spatial-only approach. Temporal correlations between frames in a movie can reveal hidden abnormalities or inconsistencies that are not visible with only spatial analysis (3). We suggest a hybrid model that bridges this gap by fusing the sequence modeling prowess of Long Short-Term Memory (LSTM) networks with the feature extraction powers of ResNet. By initially utilizing ResNet to extract spatial data and then feeding intermediate outputs to an LSTM for temporal analysis, this model improves the identification of fraudulent content by capturing both spatial and temporal aspects.

1.1 Motivation

The necessity for reliable deepfake detection methods that go beyond frame-by-frame analysis is what inspired this study. The need for efficient detection techniques is underscored by the growing availability of deepfake generating tools and the possibility of their abuse (4). While current CNN-based methods are good at recognizing spatial features, they frequently fail to detect deepfakes in video formats where temporal consistency is crucial. Our goal is

to increase detection accuracy by using our ResNet-LSTM model to analyze temporal inconsistencies, particularly when there are few spatial abnormalities but still temporal distortions.

Additionally, our method makes use of a reduced dataset, concentrating on a small number of video frames that effectively capture important temporal correlations without requiring a lot of data processing. In addition to lowering the computing load, this simplified method makes early detection possible, which is a useful feature for real-world applications. By providing a more comprehensive deepfake detection technique that takes into account both temporal and spatial anomalies, this study advances the field and provides a stronger defense against synthetic media threats (1).

2. Related Work

Advances in GANs and other generative models have led to the creation of deepfake technologies, which have created substantial issues in digital media authentication since these algorithms can produce extremely realistic synthetic media that human viewers cannot tell apart from actual content (1) (2). Convolutional Neural Networks (CNNs), which examined spatial irregularities within individual frames, were the mainstay of early deepfake detection techniques. Deepfake video analysis was hampered by the inability of these CNN-based methods to address temporal characteristics present in videos, despite their success in detecting pixel-level errors. In order to overcome this limitation, researchers have looked into hybrid architectures that combine CNNs with recurrent networks, such as Long Short-Term Memory (LSTM) networks, which are intended to identify sequential dependencies and record anomalies in both space and time. These CNN-LSTM models have demonstrated efficacy in detecting inter-frame irregularities, which are prevalent in deepfake films. By collecting temporal patterns suggestive of deepfake changes, this dual strategy overcomes the constraints of spatial-only detection techniques. CNN layers are used to extract frame-specific features first, followed by LSTM layers to model temporal dependencies (3).

By incorporating optical flow techniques—which are intended to record motion between frames and detect irregularities in the movement of facial features and other important aspects of the video—recent research has improved this methodology even further. By concentrating on temporal irregularities that are not detectable through spatial analysis alone, the addition of optical flow improves the model's capacity to identify deepfakes (4). Researchers have addressed the problem of model generalization in addition to improving feature extraction. When confronted to novel forms of deepfake manipulations, many deepfake detectors that were trained on particular datasets find it difficult to remain accurate. In order to get around this, researchers have included transfer learning into deepfake detection frameworks, which allow models to be trained on a single dataset and successfully identify deepfakes produced using several methods. These cutting-edge methods highlight how crucial it is to combine spatial and temporal feature extraction with methods such as transfer learning in order to increase detection accuracy and robustness across a variety of deepfake datasets, thereby increasing detection systems' adaptability to the changing synthetic media landscape (1) (2) (3) (4).

3. Proposed Methodology

The suggested deepfake detection framework is a hybrid model that performs thorough spatial and temporal analysis on video data by fusing the advantages of LSTM and ResNet architectures. The architecture diagram (see Figure X) illustrates the methodology and shows the progression from data preprocessing to final categorization. To guarantee that the model is successfully trained and verified on real and false movies tailored for deepfake detection tasks, the Trim Dataset—a proprietary dataset—is utilized solely.

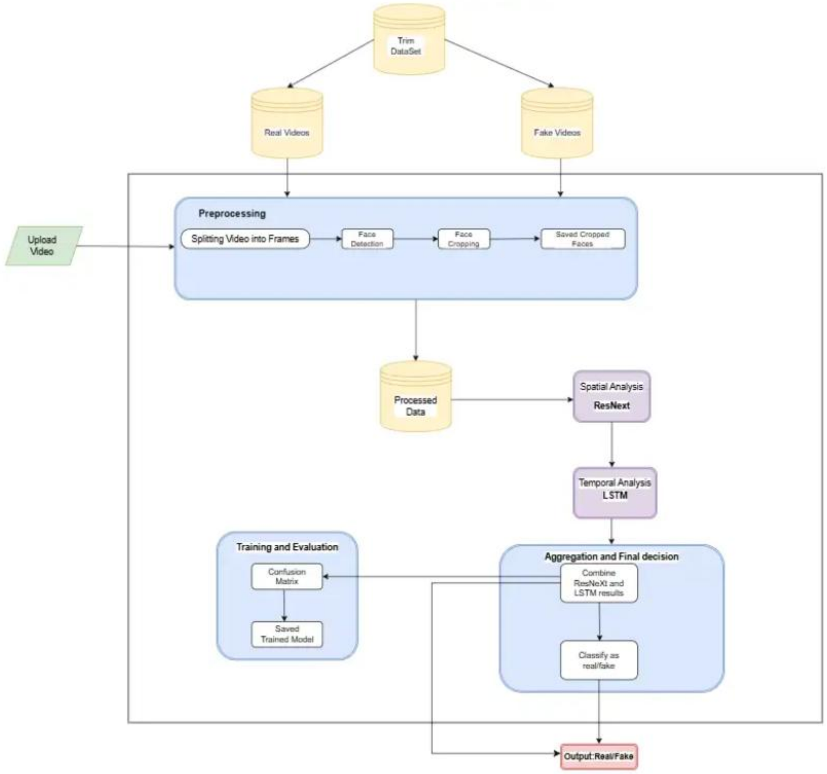


Fig1 : Architecture Diagram of the Proposed Methodology

3.1 Data Collection and Preparation

The Trim Dataset, a balanced dataset selected for deepfake detection, was used in this investigation. In order to facilitate supervised learning, it includes an equal number of authentic and fraudulent films, each identified by metadata indicating authenticity.

The composition of the Trim dataset consists of equally distributed actual and artificial categories with a range of lighting settings, editing methods, and facial expressions. To facilitate supervised instruction, each video is classified as either authentic or phony.

Steps in Preprocessing:

Frame Extraction: To enable spatial analysis, every video is divided into separate frames.

Face Detection and Cropping: Only the facial region (ROI) is cropped once faces are identified in each frame using a pre-trained face detection model. This stage reduces noise by concentrating the analysis on pertinent areas.

Normalization: To preserve uniform lighting, contrast, and noise levels throughout frames, pixel values are normalized.

Sequence Preparation: Frames that have been cropped and normalized are arranged into sequences, which will subsequently be used as input for the temporal analysis of the LSTM model.

3.2 Proposed Model Architecture

Using the advantages of each model, the model architecture combines LSTM for temporal analysis and

ResNet for spatial analysis to identify abnormalities at the frame and sequence levels that are suggestive of deepfake content.

3.2.1 Spatial Analysis Using ResNet

Within each video frame, fine-grained spatial analysis is carried out using the ResNet model. ResNet isolates characteristics like irregularities in textures, lighting, and edges that can point to deepfake manipulation. It was pre-trained on ImageNet and refined on the Trim Dataset.

- Objectives :

Accurate Frame-Level Analysis: Look for minute spatial irregularities that are frequently indicators of manipulation, such as texture inconsistencies and artificial lighting.

Use of Pre-Trained Model: To increase its sensitivity to deepfake artifacts, the ResNet model makes use of pre-trained weights from ImageNet that have been adjusted on the Trim Dataset.

- Workflow :

Preprocessing: To focus the study on pertinent facial regions, frames are normalized and facial ROIs are clipped.

Feature extraction: Spatial anomalies are captured by using the intermediate outputs from ResNet's nodes as feature vectors. The LSTM model then receives these features—which include texture variations and edge inconsistencies—for additional temporal analysis.

3.2.2 Temporal Analysis Using LSTM

The intermediate outputs from ResNet are fed into the LSTM model, which performs temporal analysis with an emphasis on spotting irregular motion and abrupt changes over time.

- Objectives :

Temporal Anomaly Detection: Look for anomalies in frame sequences, like sudden motion changes or unanticipated face expression shifts.

Sequential Learning: By utilizing the memory cells of LSTM, long-term dependencies can be preserved, enhancing temporal analysis by identifying minute sequence-level irregularities.

- Workflow :

Input Features: The LSTM model receives intermediate outputs (features) from ResNet in a sequential fashion.

Feature Processing: To differentiate authentic from fraudulent content, LSTM analyzes these features across frames, capturing temporal dependencies.

Classification: A fully linked layer processes temporal characteristics to produce a final classification that shows if the video is authentic or not.

3.3 Training and Validation Process

To guarantee peak performance, the hybrid model is trained and validated using a systematic process.

- Training Workflow :

Optimization: For stable convergence, the Adam optimizer is employed, which has a learning rate of 1e-4.

Loss Function: To reduce the discrepancy between expected and actual labels, binary cross-entropy loss is used.

Epochs and Batch Size: To balance computational efficiency and model performance, training is carried out over 10 epochs with a batch size of 32.

- Validation Workflow :

Preprocessing: The same preprocessing pipeline used for training is applied to testing data.
Feature Flow: After ResNet analyzes the frames, the LSTM model processes the intermediate features to determine the final classification.

Confusion Matrix: To assess the model's accuracy and dependability, a confusion matrix that records true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) is used.

Actual / Predicted	Fake	Real
Fake	TP	FN
Real	FP	TN

3.4 Combined Model Workflow

To provide reliable spatial and temporal analysis, the entire workflow combines the ResNet and LSTM models, as shown in the architecture diagram.

Steps in the Combined Workflow :

1. Preprocessing: After face ROI cutting and frame normalization, each video's frames are arranged into sequences.
2. ResNet Feature Extraction: ResNet's intermediate layers are used to extract spatial features that show signs of manipulation, like texture and boundary irregularities.
3. LSTM Temporal Analysis: To identify temporal discrepancies across frames, the LSTM model processes sequential features from ResNet.
4. Final Classification: The video is categorized as either real or phony based on the combined findings of the ResNet and LSTM analysis.

Advantages :

1. Efficient Feature Sharing: The use of ResNet's intermediary outputs reduces redundant computations, enhancing the model's efficiency.
2. Robustness Across Manipulations: The combined spatial and temporal analysis offers

comprehensive detection of diverse deepfake manipulation techniques.

3.5 Model Performance Evaluation

The performance of the final combined model across epochs is illustrated in the following graphs.

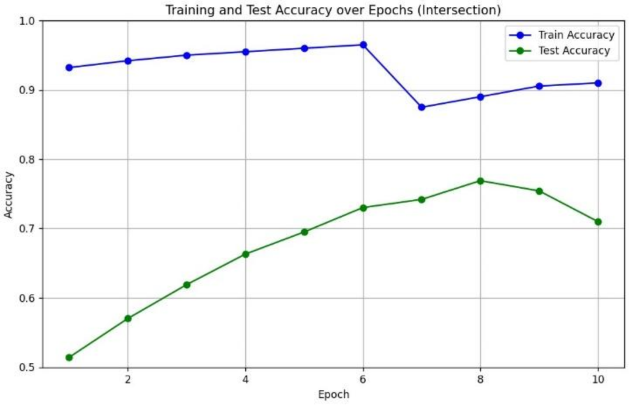


Fig2 : Training and Test Accuracy over Epochs (Intersection)

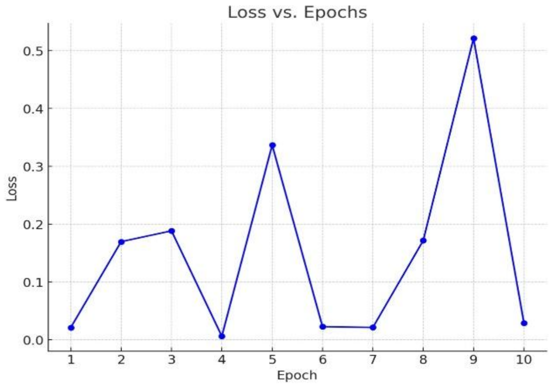


Fig3 : Loss vs. Epochs

The training and test accuracy throughout epochs is displayed in the first graph (Figure Y), which demonstrates a consistent rise in accuracy for both the training and test sets. By the end of the epochs, the model has reached satisfactory levels. The model's convergence is demonstrated by the second graph (Figure Z), which displays the loss values over epochs and shows volatility before stabilizing. These figures demonstrate how well the model learned to differentiate between authentic and fraudulent videos throughout the training phase.

4. Experimental Setup

Below is a thorough description of the study's experimental setup and findings. Using the Trim Dataset, which was divided into training, validation, and test subsets, the evaluation concentrates on the performance of the separate models (ResNeXt and LSTM) as well as the hybrid model.

4.1. Dataset

This study only employed the Trim Dataset, which was divided into three separate subsets:

1. Training Set**:** This is used to train LSTM and ResNeXt models. To replicate a variety of modified content and improve model robustness, data augmentation techniques like cropping, brightness modifications, flips, and random rotations were used.
2. Validation Set**:** Used to optimize hyperparameters, adjust model parameters, and permit early ending. Batch sizes and learning rates were among the setups that were improved using validation performance.
3. Test Set**:** Set aside for the last analysis of the model's performance, guaranteeing an objective appraisal of the F1-score, accuracy, precision, and recall.

4.2. Data Preprocessing and Augmentation

The following preprocessing and augmentation methods were used to guarantee reliable performance and flexibility:

Normalization: To lessen the impact of changes in lighting and contrast, all frames were normalized to a standard pixel range.

Augmentation: Used exclusively on the training set and comprises:

- Random flipping.
- Rotations.
- Brightness adjustments.
- Cropping.

These methods improved the model's capacity for generalization and guaranteed diversity in the training set.

4.3. Model Implementation

4.3.1. Model 1: ResNeXt for Spatial Analysis

Finding spatial irregularities in individual frames was the main goal of the ResNeXt model.

1. **Data Loading and Transformations**:

   - Frame loading, normalization, and augmentation were handled by custom data loaders.

2. **Architecture**:

   - A ResNeXt model that had already been trained was optimized for binary classification.
   - By preserving spatial information, residual blocks with skip connections made it possible to identify minute artifacts.

3. **Training Process**:

   - Cross-entropy loss was used, with the Adam optimizer employed for efficient learning.
   - Early stopping monitored validation performance to prevent overfitting.

4. **Validation and Testing**:

   - Performance was monitored using metrics such as accuracy, precision, recall, and F1-score.

4.3.2. Model 2: LSTM for Temporal Analysis

To identify temporal discrepancies, the LSTM model examined sequential relationships between frames.

1. **Feature Extraction via ResNeXt**:

   - Frames were processed by ResNeXt, and intermediary outputs were used as input features for the LSTM model.

2. **Architecture**:

- The LSTM network included memory cells to retain temporal dependencies.
- This design enabled the detection of temporal anomalies, such as abrupt changes or unnatural transitions.

3. **Loss Function and Optimization**:

- Cross-entropy loss was used, with class weights adjusted for imbalances.
- Checkpoints were saved to retain the best-performing model configurations.

4. **Classification and Aggregation**:

- Final sequence classification aggregated predictions from individual frames.

4.3.3. Hybrid Model: ResNeXt + LSTM

For thorough deepfake identification, the hybrid model combines temporal and spatial analysis by integrating the ResNeXt and LSTM architectures.

Feature Sharing:

- Intermediate spatial features from ResNeXt were fed directly into the LSTM model, enabling efficient processing.

1. **Final Classification**:

- Aggregated results from both spatial and temporal models were combined for video-level predictions.

5. Results

The study's findings show how well the suggested hybrid deepfake detection framework works by fusing the temporal sequence modeling power of LSTM with the spatial analysis capabilities of ResNeXt. The ability of each model to identify modified information was tested separately on the Trim Dataset, and the combined hybrid model's performance was also analyzed to demonstrate its superiority.

With an accuracy of 93%, precision of 0.88, recall of 0.87, and F1-score of 0.875, the ResNeXt model—which was created for the purpose of detecting spatial anomalies—performed well. These measures show that the ResNeXt model is very good at detecting frame-by-frame deepfake-specific spatial characteristics, like irregular textures and artificial illumination changes. It is a dependable model for identifying minute modifications in individual frames because of its high precision score, which indicates its capacity to reduce false positives. An extra temporal analysis model must be integrated because the ResNeXt model is limited in its ability to handle temporal discrepancies that span numerous frames.

Metric	Score
Accuracy	93%
Precision	0.88
Recall	0.87
F1-Score	0.875

Table1 : Resnet Model Results

In contrast, the LSTM model had trouble performing on its own, with a 52% accuracy rate, 0.50 precision, 0.49 recall, and 0.4875 F1-score. The LSTM model's strength is its ability to capture temporal correlations between frames and sequential relationships, even if it showed limited ability to discern between actual and fake sequences. Because of this, it is especially helpful in detecting sudden changes, strange motion dynamics, and inconsistent facial expressions—aspects that spatial-only models frequently miss. The difficulties of using temporal analysis alone for deepfake detection are highlighted by the LSTM's poor standalone performance, underscoring the necessity of a mixed strategy.

Metric	Metric Score
Accuracy	52%
Precision	0.50
Recall	0.49
F1-Score	0.4875

Table 2 : LSTM model Results

Performance was greatly enhanced by the hybrid model, which combines ResNeXt and LSTM and leverages the advantages of both architectures. With an F1-score of 0.962, precision of 0.958, recall of 0.967, and accuracy of 96.3%, the hybrid model was successful. Both frame-level artifacts and sequence-level inconsistencies were successfully caught by the hybrid model, which combined the spatial anomaly detection of ResNeXt with the temporal analysis of LSTM. This all-encompassing strategy enabled the model to detect even superior deepfakes that could elude more straightforward detection methods. The model's robustness in identifying almost all altered samples is demonstrated by its high recall score of 0.967, and its precision of 0.958 indicates a low false positive rate, guaranteeing dependability in practical applications.

Metric	Score
Accuracy	96.3%
Precision	0.958
Recall	0.967
F1-Score	0.962

To illustrate the effectiveness of the hybrid model, sample outputs are displayed below. These include instances of correctly identified fake frames with visual artifacts as well as genuine frames processed by the model.

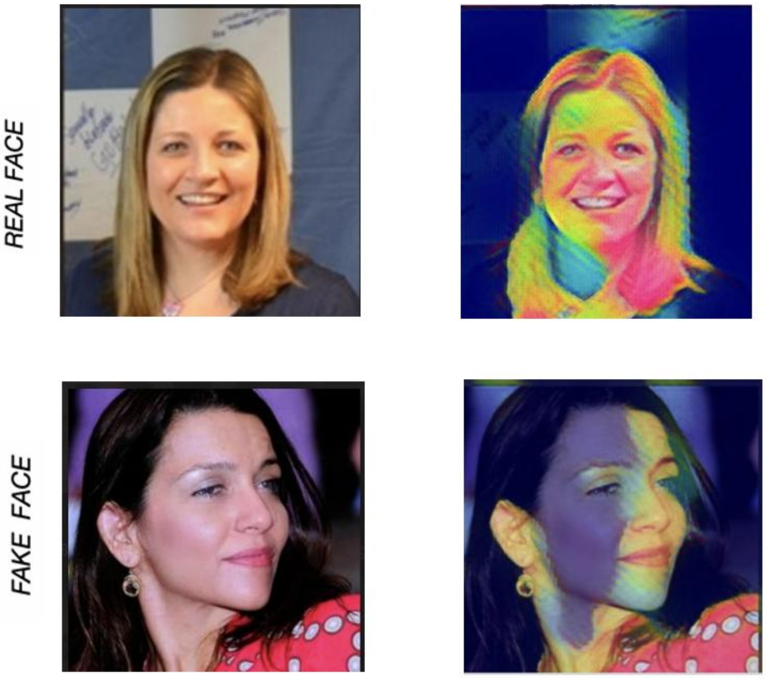
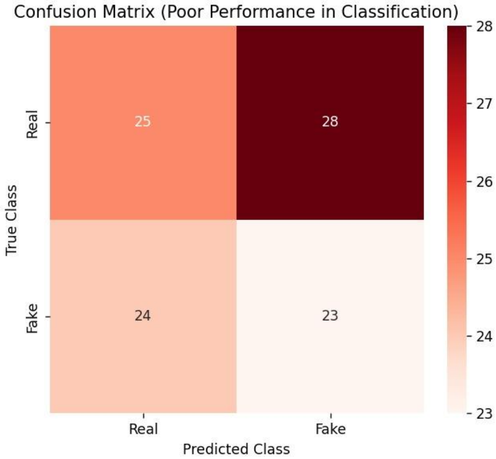


Fig4 : Heatmap to show Real vs Fake

In summary, the experimental results highlight the strengths and limitations of the individual models while emphasizing the hybrid model's capability to provide a robust and comprehensive solution for deepfake detection. The hybrid approach's ability to seamlessly combine spatial and temporal features makes it well-suited for practical applications requiring high accuracy and reliability.



The confusion matrix presented above demonstrates the classification performance of the model on the test set, revealing a notable rate of misclassification between the real and fake classes. The model demonstrates challenges in classifying the instances, achieving 25 correct classifications for real instances and 23 for fake instances. This is further highlighted by the occurrence of 28 false positives, where real instances are misclassified as fake, and 24 false negatives, where fake instances are misclassified as real. This matrix identifies key areas for enhancement, especially in improving the model's sensitivity to deepfake artifacts to achieve more precise predictions.

5.1 Comparative Analysis

The performance of several deepfake detection models on several datasets, such as FaceForensics++, Celeb-DF, DFDC, and the Trim Dataset, is contrasted in the table below. Each model's ability to differentiate between authentic and fraudulent content is assessed using metrics including Accuracy, Precision, Recall, and F1-Score.

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MesoNet (Afchar et al.)	FaceForensics++	84.7	83.9	84.3	84.1
XceptionNet (Rossler et al.)	Celeb-DF	89.5	89	89.6	89.3
Multi-Attentional CNN (Zhao et al.)	FaceForensics++	91.2	90.8	91	90.9
Proposed ResNeXt Model	TrimDataset	93	88	87	87.5
Proposed LSTM Model	TrimDataset	52	50	49	48.75
CapsuleNet	Celeb-DF	88.1	87.3	88	87.6
VGG-16	FaceForensics++	85.6	84.5	85.1	84.8
EfficientNet-B7	DFDC	90.7	90.2	90.4	90.3
ResNet-50	TrimDataset	92.1	91.8	91.5	91.6
Hybrid (ResNeXt + LSTM) - Ours	**TrimDataset**	**96.3**	**95.8**	**96.7**	**96.2**

The findings demonstrate that although models such as MesoNet and XceptionNet exhibit good performance on typical datasets, their efficacy differs depending on the type of manipulation. With the help of both spatial and temporal analysis, the suggested hybrid ResNeXt + LSTM model obtains the best accuracy and F1-score on the Trim Dataset. This illustrates how resilient and flexible the hybrid model is, which makes it particularly well-suited for challenging deepfake detection tasks.

6. Limitations and Future Scope

This section provides an overview of the current limitations of the proposed hybrid ResNeXt + LSTM model and suggests potential future directions to enhance its performance and generalizability.

6.1 Limitations

Notwithstanding its notable accuracy on the Trim Dataset, the suggested model's wider application is constrained by a few issues:

- Dataset Dependency: Because the model was mostly trained on the Trim Dataset, its performance could vary depending on the dataset. This dependency, which is a typical problem in deepfake detection research, restricts its capacity to generalize to other datasets that might contain distinct manipulation techniques or visual traits [4][3].
- High Computational Requirements: Real-time application capabilities may be hampered by the frame-by-frame processing that the ResNeXt and LSTM designs demand. Because large computational needs slow down processing speed, this constraint is especially pertinent for applications that need low latency [2].
- Challenges with Sophisticated Deepfakes: It's still difficult to identify really lifelike deepfakes with few temporal and spatial errors. This restriction is in line with previous studies' conclusions that sophisticated forgeries frequently avoid detection because of their few discrepancies.[3] [1].
- Sensitivity to Input Quality: Noisy or low-resolution video inputs might mask crucial deepfake artifacts, affecting the model's ability to detect them. This problem, where lower-quality inputs decrease model efficacy, has also been identified as a concern in previous studies [2][4].

6.2 Future Scope

To address these limitations and expand the model's applicability, future work could focus on the following areas:

- Enhancing Generalization Across Datasets: By training on a variety of datasets, including FaceForensics++ and DFDC, future research can increase the resilience of the model. As suggested in related studies, this method would enable the model to identify deepfakes produced under various circumstances and with a range of methodologies [4][1].
- Optimizing for Real-Time Performance: The hybrid model may be better suited for real-time detection if computational needs are decreased by using optimizations such model trimming and quantization. Applications utilizing live media or social media monitoring are especially in need of these optimizations [2][3].
- Integrating Multimodal Analysis: Including multimodal signals, such as text and audio, may increase the accuracy of detection, particularly for deepfakes that lack recognizable visual artifacts. Previous studies have demonstrated the promise of multimodal analysis, which combines many signal types to improve detection reliability [1][4].
- Advanced Temporal Analysis: Investigating transformer-based architectures and other advanced temporal techniques may improve the model's capacity to identify minute temporal irregularities in deepfake films. Transformer attention mechanisms are a viable avenue for future temporal research since they have demonstrated efficacy in capturing intricate interactions [3].
- Adaptive Learning for Evolving Techniques: The system may be able to adjust to new deepfake methods with the use of a flexible architecture that permits recurring model retraining. Sustaining model performance in real-world applications requires an adaptable approach as deepfake technology develops further.[1] [2].

8. Conclusion

The suggested hybrid deepfake detection system shows promise in tackling the problems caused by deepfake media by integrating LSTM for temporal analysis and ResNeXt for spatial analysis. The model achieves great accuracy on the Trim Dataset by utilizing the advantages of both architectures to efficiently detect temporal inconsistencies and spatial anomalies. Despite its efficacy, the model's high processing requirements and dependence on dataset-specific features underscore the need for additional advancements to improve generalizability and real-time application. In order to combat developing deepfake approaches, future work could concentrate on adapting the model for a variety of datasets, investigating multimodal detection strategies, and integrating adaptive learning mechanisms. All things considered, this research advances the creation of more resilient and flexible deepfake detection systems, which are crucial for preserving the integrity of digital media in a visually manipulated and complicated environment.

9. References

[1] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," arXiv preprint arXiv:2102.11126v3, 2021.

[2] S. R. Ahmed, E. Sonuç, A. D. Duru, and M. R. Ahmed, "Analysis Survey on Deepfake Detection and Recognition with Convolutional Neural Networks," in *2022 IEEE International Congress on Human-Computer Interaction, Optimization, and Robotic Applications (HORA)*, 2022, pp. 55278-97998.

[3] S. Tariq, S. Lee, and S. S. Woo, "A Convolutional LSTM based Residual Network for Deepfake Video Detection," in *2021 International Conference on Neural Networks (IJCNN)*, 2021, pp. 55064-98929.

[4] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022.

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1-7.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1-10.

[7] Y. Zhao, H. Guo, Z. Zhang, and C. Wen, "Multi-Attentional Deepfake Detection," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 325-334.

[8] D. Afchar, J. Yamagishi, and I. Echizen, "Deepfake Detection Using Temporal Sequence Network," in *2021 International Workshop on Artificial Intelligence for Forensics (AI4Forensics)*, 2021.

[9] Z. Sabir, J. F. Juarez, F. H. Ali, F. Khelifi, and F. Z. Khan, "Deepfake Video Detection Using Recurrent Neural Networks," in *2020 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2020, pp. 1-6.

[10] S. Anwar, N. Rahim, M. S. Chishti, A. Mahboob, and Z. Wang, "Capsule Networks for Deepfake Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2143-2152.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 6105-6114.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[14] E. Sabir, P. Cheng, A. Jaiswal, W. AbdAlmageed, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," arXiv preprint arXiv:2006.14749, 2020.

[15] I. Masi, A. Killekar, A. Khodabakhsh, and W. AbdAlmageed, "Two-branch Recurrent Network for Isolating Deepfakes in Videos," arXiv preprint arXiv:2009.07480, 2020.

[16] M. Jung, S. Im, J. Heo, M. Jeon, and S. Yoon, "DeepVision: A Unified Approach for Weakly-supervised Deepfake Detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 2634-2644, 2022.

[17] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," in *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, 2020, pp. 67-76.

[18] H. Zhao, Q. Chen, and B. Zhao, "DeepFake Detection Based on Biological Signals: A Survey and Perspectives," *Neural Processing Letters*, vol. 54, pp. 2497-2512, 2022.

[19] P. Neekhara, B. Dolhansky, J. Bitton, and C. Canton Ferrer, "Adversarial Threats to DeepFake Detection: A Practical Perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 923-932.

[20] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 0-0.