A review of deepfake detection techniques and comparative analysis with a ResNeXt + LSTM framework

Jaya Raj[1], Swati Khanve[2], Nitya Khare[3]
[1]M.Tech Scholar, Dept. of Computer Science and Engineering, SIRTE, Bhopal, jayaraj0678@gmail.com, India;
[2]Asst. Prof., Dept. of Computer Science and Engineering, SIRTE, Bhopal, India;
[3]HOD., Dept. of Computer Science and Engineering, SIRTE, Bhopal, India;

Abstract *– Deepfakes are AI-synthesized data particularly images and videos that look so authentic, that they can easily deceive any common man. This study aims to build a hybrid detection system with the help of ResNeXt-LSTM, where the ResNeXt-50 variant is used for extracting spatial features which are fed in an LSTM network that finds temporal inconsistencies across video frames. Conventional neural networks like CNN and LSTM's alone have become incapable for detecting deepfakes obtained from the latest generation techniques. The hybrid approach combines the power of learning from both frame-level artifacts and sequential anomalies. This model is trained on a dataset obtained from combining three different datasets namely FaceForensics++, DFDC and Celeb-DF. Due to the combined strength of ResNeXt-50 and LSTM, and a versatile dataset, the model shows improved accuracy, efficiency and generalization, which makes it suitable for deployment in real-time systems and potentially adaptable to resource-constrained devices.*

Keywords*: Deepfake Detection, ResNeXt, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) Spatiotemporal Analysis, Frame-level feature extraction*

I. Introduction

In a paper published by the U.S. Department of Homeland Security in June 2019, the threats of deepfake technology have been discussed in detail [1]. It explains the origin, development and problems associated with deepfakes. Until the late 90's artificial media was being created using CGI's [2] mainly for the entertainment industry, which could only be created by people who specialized in that field, using only some special tools and software. The creation of artificial media was therefore controlled and restricted to a very few number of people at that time and did not pose any such significant threat of any kind.

Ian Goodfellow, an American computer scientist and engineer introduced a concept of GAN (Generative Adversarial Network) in June 2014, which was a breakthrough in the field of deepfake generation technology. A GAN [3] is a machine learning framework that works on the concept of two neural networks namely generator and discriminator, where the generator is responsible for creating plausible data and the discriminator is responsible for distinguishing real and fake data.

The term deepfake came from a reddit user who started posting non-consensual face-swap pornography featuring celebrities in 2017 under the username deepfakes [4]. This was new and unique to people and therefore it instantly grabbed public attention. Later in January 2018, a desktop application called FakeApp [5] was launched, which was accessible to common people and anybody could use it for creating deepfakes.

Deepfake technology became quite famous in 2018 to the extent where major tech platforms had to make policies to regulate deepfake content on social media platforms [6]. Since then, deepfake technology has been improving at a rapid speed and getting better day by day. This has also led to increase in challenges that deepfakes give rise to, like public distrust, political misinformation, financial scams, identity thefts and many more [7].

These deepfakes were not just limited to images and videos, but also audio deepfakes. Voice cloning software came into the picture in the year 2020 which gave rise to audio deepfake scams [8]. In the recent years, many deepfake generation tools and software were introduced and problem seem to have been increasing ever since. It was important to find a solution to control the increasing frauds happening due to the open availability of deepfake generation tools.

Deepfake detection techniques were taking shape from the 2018 itself and companies like Deeptrace were working on detection methods that focused on the irregular eye-blinking patterns, inconsistent facial features and textures, uneven lip-sync movements, etc. [9].

This paper provides a comparative analysis of the deepfake detection models and discusses the design rationale behind using a ResNeXt-LSTM hybrid approach.

II. Literature Survey

In the recent years, deepfake detection has become a critical research area due to the rapid advancement of generative adversarial networks (GANs) and the increasing accessibility of deepfake creation tools. Numerous detection techniques have been proposed to address this challenge, each employing distinct architecture and learning strategies to identify synthetic or manipulated visual content. The following section provides a consolidated overview of prominent deepfake detection frameworks, their performance and limitations

Early detection models primarily relied on Convolutional Neural Networks (CNNs) to capture spatial inconsistencies in facial regions. Afchar et al. [10] proposed MesoNet (Meso-4 and MesoInception-4), a compact CNN-based model that achieved accuracy between 83%-95% across various datasets. While its shallow design made it computationally efficient, it was less effective on high-resolution or highly compressed videos. Building upon this, Rossler et al. [11] introduced XceptionNet within the FaceForensics++ benchmark, which leveraged depthwise separable convolutions to achieve superior accuracy (~97%). However, it showed reduced generalization under strong compression or unseen manipulations.

Feature-based approaches using VGG16/VGG19 have also been investigated for transfer learning [12], demonstrating around 85-92% accuracy. Despite solid results, these models tend to overfit and incur higher computational costs. To address spatial hierarchy modeling, Capsule Networks (CapsNet) were utilized by Nguyen et al. [13], yielding approximately 93% accuracy. Although capsule routing improved spatial feature capture, the model faced scalability issues on larger datasets.

To capture temporal dynamics in videos, researchers integrated motion cues with CNNs. Sabir et. al. [14] proposed a Two-Stream CNN framework combining RGB and optical flow inputs, achieving 90-94% accuracy. Similarly, Guera and Delp [15] combined CNNs with Recurrent Neural Networks (RNNs) such as LSTMs to exploit temporal dependencies, improving performances up to 96% on video datasets. Nonetheless, these methods are sensitive to frame alignment errors and demand significant processing power.

More recently, hybrid and attention-based architecture have been proposed. Zhao et al. [16] employed EfficientNet combined with attention mechanisms, achieving 95-98% accuracy while focusing on discriminative facial regions. Transformer-based architectures like Vision Transformers (ViT) [17] demonstrated exceptional performance (96-99%) by modeling global relationships among image patches, although they require extensive data and computational resources for effective training.

In comparison, the proposed ResNeXt + LSTM hybrid model combines the parameter-efficient grouped convolution structure of ResNeXt for spatial feature extraction with the temporal modeling capability of LSTM networks. This design achieves accuracy levels up to 99% on benchmark datasets while maintaining lower parameter counts than deeper CNNs. However, due to sequential computational power during training and inference. Thus, the ResNeXt + LSTM framework represents an optimal trade-off between efficiency and detection accuracy in spatiotemporal deepfake analysis.

All these machine learning and deep learning based detection techniques were discovered since 2018 that worked on different principles and trained on different datasets. They are summarized in the table below:

Table 1 Performance comparison of existing deepfake detection models

Model	Frame work	Accuracy	Reference Paper	Limitations
MesoNet / Meso-4	CNN-based shallow architecture	~83-95% depending on datasets Trained on FaceForensics++	Afchar et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, 2018	Struggles with high resolution deepfakes, sensitive to compression; limited temporal analysis
Xception Net	Deep CNN with depthwise separable convolutions	~97% Trained on FaceForensics++ (raw)	Rossler et al., “FaceForensics”, 2019	Accuracy drops significantly on compressed/lower-quality videos; lacks motion modeling
VGG16/ VGG19 Feature-Based	Transfer learning with	~85-92% depending on training	Nataraj et al., “Detecting GAN	Overfits easily; high computati

Classifiers	pre-trained CNNs	dataset Trained on Celeb-A	generated Fake Images”, 2019	onal cost; less robust to unseen manipulation types
Capsule Network	Capsule layer capture spatial feature hierarchies	~93% on certain deepfake datasets Trained on UADFV	Nguyen et al., “Capsule-Forensics”, 2019	Struggles with large-scale datasets; training instability
Two-Stream CNN (RGB + Motion Stream)	CNN + Optical Flow Motion Detection	~90-94% depending on data Trained on Deepfake TIMIT	Sabir et al., “Recurrent Convolutional Strategies for Face Manipulation Detection”, 2019	Motion estimation is computationally heavy; performance decreases with frame rate variation
LSTM / GRU-Based Temporal Models	CNN Feature Extractor + LSTM Temporal Classifier	~92-96% on video datasets Trained on DFDC	Guera and Delp, “Deepfake Video Detection Using Recurrent Neural Networks”, 2018	Only captures temporal patterns; reliant on CNN quality; struggles when face alignment fails
EfficientNet + Attention Mechanisms	Efficient CNN + Attention Layers	~95-98% in controlled benchmarks Trained on Celeb-DF	Zhao et al., “Learning to Detect Deepfakes with Attention”, 2021	High compute cost; performance drops with real-world noise and compression
Vision Transformers (ViT)-Based Methods	Transformer architecture for global patch relationships	~96-99% on recent datasets Trained on Celeb-DF and DFDC	Dosovitskiy et al., “An Image is Worth 16×16 Words”, 2020 (applied in later DF detection research)	Requires large training datasets; lower performance on low-light or occluded faces
ResNeXt + LSTM (Proposed Hybrid Model)	ResNeXt CNN for Spatial Features + LSTM for Temporal Dynamics	Typically ~80-96% depending on dataset Trained on Celeb-DF, FaceForensics++ and DFDC	My Thesis and Research	The computational cost of training this model a bit high, but once trained it could be deployed on low-end devices

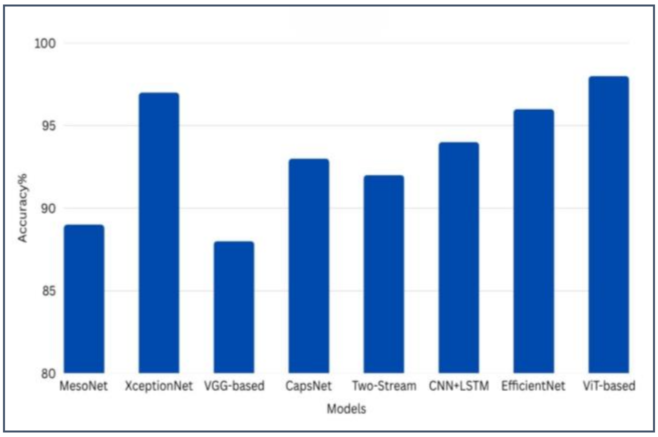

Figure 1 Accuracy comparison across different detection models

All these models are fine and work great when trained on a single or known dataset, but one major problem faced by all these models is that their accuracy drops drastically when it is fed with a combination of multiple and unknown datasets. These models are unfit for real-time detection and cannot be deployed on low-end devices as their performance falls drastically when given real-time data and they work only on devices that have a GPU support.

For the fulfillment of this research gap, frame-level detection [18] combining ResNeXt-50 and LSTM is proposed. Frame-level detection offers real-time robustness and a small model size which makes it fit for devices with a lower range of configuration thus making it available for all types of devices. This will also increase the accuracy of the model as it leverages the benefits of both CNN and RNN being able to detect spatial and temporal defects in a video [19].

III. Proposed Methodology

The proposed approach utilizes a hybrid deep learning architecture that combines the strengths of spatial feature extraction and temporal sequence modeling for deepfake detection. The framework employs ResNeXt as the primary feature extractor to learn discriminative spatial representations from individual video frames. ResNeXt is selected due to its grouped convolution mechanism, which enables strong representational capacity while maintaining a comparatively smaller number of parameters, allowing the model to remain efficient in terms of memory and scalability.

Following spatial feature extraction [20], the sequential frame-level embeddings are passed into an LSTM (Long Short-Term Memory) network to capture temporal dependencies across video frames. This component enables the model to detect subtle inconsistencies in facial expressions, head motion, and temporal coherence that may not be apparent in single-frame analysis. By integrating spatial and temporal information, the hybrid architecture targets both texture-level artifacts and motion irregularities commonly present in manipulated videos.

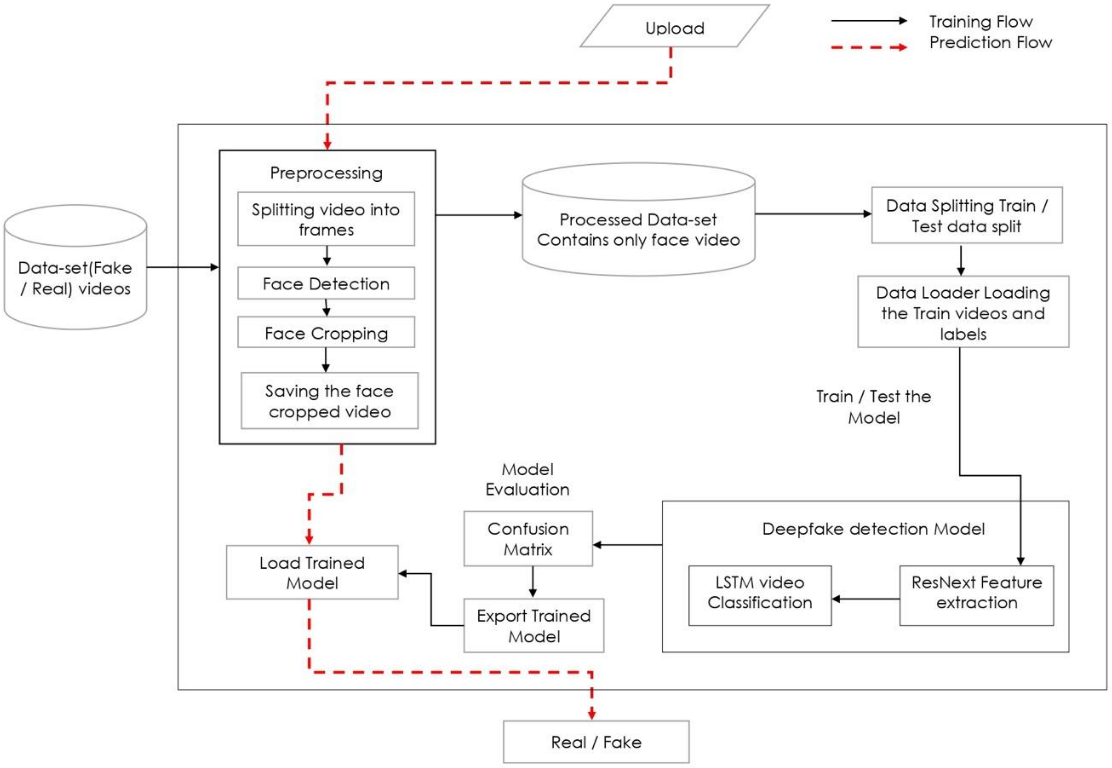
Upload
Training Flow
Prediction Flow
Preprocessing
Splitting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Data-set(Fake / Real) videos
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM video Classification
ResNext Feature extraction
Load Trained Model
Export Trained Model
Real / Fake

Figure 2 Proposed ResNeXt + LSTM model framework

The system processes input video by first sampling frames at uniform intervals and normalizing them for consistent face alignment. The output classification layer then assigns a probability of authenticity, allowing the model to differentiate between real and forged content. This combined spatial-temporal design provides improved detection performance compared to the performance compared to purely spatial CNN approaches, while remaining less computationally intensive than full 3D convolutional or transformer-based alternatives. Overall, the methodology aims to offers a balanced framework that enhanced deepfake detection accuracy while maintaining practical deployability.

IV. Conclusion

This review examined the evolution of deepfake detection techniques, highlighting the shift from traditional frame-based CNN models to more advanced spatiotemporal and transformer-driven approaches. While early models demonstrate strong performance under controlled conditions, they often struggled to generalize across varying video qualities, compression levels, and unseen manipulation techniques. Incorporating temporal analysis has proven increasingly important, as inconsistencies across frames are among the most reliable indicators of synthetic media.

In this context, the hybrid ResNeXt + LSTM framework offers a balanced solution by combining efficient spatial feature extraction with temporal sequence modeling. Although it can be computationally demanding during inference, it remains more parameter-efficient than many high-capacity models and provides improved detection accuracy in real-world video scenarios. The approach underscores the value of integrating spatial and temporal cues for robust deepfake detection.

Overall, the field continues to face challenges related to cross-dataset generalization, real-time deployment, and resilience against rapidly evolving generation techniques. Future research should focus on compression-aware training, lightweight temporal aggregation strategies, and adaptive learning across diverse datasets to ensure reliable and scalable deepfake detection in practice applications.

V. Future Scope

Future research in deepfake detection should prioritize improving generalization across diverse datasets and manipulation techniques. Current models often achieve higher accuracy on benchmark datasets, but fail when exposed to real-world variations such as compression, artifacts, low lighting, occlusion or new synthesis methods, not seen during training. Developing domain-adaptive and compression-aware training strategies, as well as incorporating large scale diverse datasets can enhance their robustness and reduce over fitting to specific artefact patterns. Cross-dataset evaluation should become a standard practice to ensure practical effectiveness.

Another important direction involves reducing computational complexity while maintaining strong detection accuracy. Many high performing models including hybrid, spatial temporal frameworks and transformer-based architecture is require significant computational resources, which limits their deployment in real-time or resource constrained environments. Future models may explore, lightweight, temporal attention we can resume affective frame sampling strategies and model compression techniques such as pruning and quantization advancements in hardware, neural architecture search (NAS) may also help in designing optimize models for mobile or edge devices.

Finally, with deepfake generation methods advancing rapidly, future direction system should shift towards forgery, explanation and interpretability rather than binary classification alone. Visual explanations, anomaly heatmaps, and uncertainty scoring can provide users an analyst with insight into why clip is considered manipulated. In addition, integrating detection frameworks into special media, social media platforms, authentication pipelines and legal forensic systems will be critical to combating misuse. Collaboration across technical researchers, policymakers and media organizations will be necessary to ensure responsible and effective deployment of detection technology.

Acknowledgment

I would like to acknowledge my gratitude towards my guide Prof. Swati Khanve for helping me understand the process of research and for her constant support during this experiment. I would also like to thank my HOD Prof. Nitya Khare for regularly giving me the motivation to keep going and all the faculties of the CSE department that helped during my M.tech endeavour.

References

[1] U.S. Department of Homeland Security (2022) "Increasing Threats of Deepfake Identities" https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

[2] S. Das (2023) "The evolution of visual effects in cinema: A journey from practical effects to CGI" https://www.researchgate.net/publication/375989472_The_Evolution_Of_Visual_Effects_In_Cinema_A_Journey_From_Practical_Effects_To_Cgi

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio (2014) "Generative Adversarial Networks" 10.48550/arXiv.1406.2661

[4] M. Westerlund (2019) "The emergence of deepfake technology: A review" 10.22215/timreview/1282

[5] B. U. Mahmud, A. Sharmin (2021) "Deep insights of Deepfake technology: A review" https://arxiv.org/abs/2105.00192

[6] D. Ojha, S. Malhotra (2023) "Exploratory note on deepfakes and policy considerations" https://www.dsci.in/files/content/knowledge-centre/2023/DSCI%20Exploratory%20Note%20on%20Deepfakes_0.pdf

[7] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, H. Malik (2023) "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward" 10.1007/s10489-022-03766-z

[8] B. Zhang, H. Cui, V. Nguyen, M. Whitty (2025) "Audio deepfake detection: What has been achieved and what lies ahead" 10.3390/s25071989

[9] M. S. Rana, M. N. Nobi, B. Murali, A. H. Sung (2022) "Deepfake Detection: A Systematic Literature Review" 10.1109/ACCESS.2022.3154404

[10] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen (2018) "MesoNet: A Compact Facial Video Forgery Detection Network" 10.1109/WIFS.2018.8630761

[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner (2019) "FaceForensics++: Learning to Detect Manipulated Facial Images" 10.1109/ICCV.2019.00009

[12] "VGG-19 Convolutional Neural Network" (n.d.) https://www.sciencedirect.com/topics/computer-science/vgg-19-convolutional-neural-network

[13] H. H. Nguyen, J. Yamagishi, I. Echizen (2019) "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos" 10.1109/ICASSP.2019.8682602

[14] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan (2019) "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos" https://arxiv.org/abs/1905.00582

[15] D. Güera, E. J. Delp (2018) "Deepfake Video Detection Using Recurrent Neural Networks" 10.1109/AVSS.2018.8639163

[16] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu (2021) "Multi-attentional Deepfake Detection" 10.1109/CVPR46437.2021.00220

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby (2020) "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" 10.48550/arXiv.2010.11929

[18] M. Naveenkumar, A. Vadivel (2015) "OpenCV for Computer Vision Applications" https://www.researchgate.net/publication/301590571_OpenCV_for_Computer_Vision_Applications

[19] N. Zhang, J. Luo, W. Gao (2020) "Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks" 10.1109/9239720

[20] R. C. Staudemeyer, E. R. Morris (2019) "Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks" 10.48550/arXiv.1909.09586