Vol.no 1, Issue no 2, Sep - Dec, 2025

PUXplore Multidisciplinary Journal of Engineering

Parul® University NAAC GRADE A++

Deepfake Detection Using ResNeXt-50 and LSTM Neural Networks

K. Kumar Soma Sekhar[1], G. Sai Kishore[2], K. Manoj Kumar[3], G.K. Vijitendra[4], Rodda Pavan Kumar[5], Daxa Vekariya[6], Mukesh Patidar[7]

[1,2,3,4,6,7]*Parul Institute of Engineering and technology, Parul University, Vadodara, India.*
[5]*Parul Institute of Engineering and technology (MCA), Parul University, Vadodara, India.*

Corresponding E-mail[5]: roddapavankumar24@gmail.com

***Abstract**: Recent advances in deep learning methods have made it possible to create extremely convincing synthetic media, known by the general public as deepfake videos, that carry enormous dangers such as spreading misinformation, political manipulation, financial scams, and individual blackmail. In this research, we present a system based on deep learning to automate the detection of deepfake videos. In the proposed methodology, we adopt the usage of the ResNeXt-50 architecture to learn the effective frame- level features of the video, which pass through a Long Short-Term Memory (LSTM) network that recognizes the relationships over the whole video. To enhance the generalizability of the model, we adopt a sequential setup with ReLU activation and dropout regularization. To evaluate the model's effectiveness, we utilize benchmark datasets of deepfakes, and our results show that the model of ResNeXt-50 + LSTM is highly effective in distinguishing between real and manipulated videos. The results of this research provide a significant demonstration of the benefit of fusing multi- dimensional data (i.e., spatial and temporal data) to construct a reliable model to identify deepfakes.*

Index Terms—*Deepfake detection, Convolutional Neural Net- works (CNN), ResNeXt-50, Long Short-Term Memory (LSTM), Hybrid deep learning, Temporal feature extraction, Video foren- sics.*

1. Introduction

The exponential growth of artificial intelligence and deep learning has drastically shifted the consumption and pro- duction of digital media. One of the particularly disturbing implications of this technology is its formation into *deep- fakes*, highly realistic computer-simulated videos generated with the help of deep learning algorithms, such as generative adversarial networks (GANs) and convolutional autoencoders. Unlike older photo or video editing technologies that relied on specialized knowledge and equipment, recent open-source software and widely available computing power have enabled nearly anyone to create deepfakes. Widely used platforms such as FaceSwap, DeepFaceLab, and FakeApp have accelerated the democratization of this technology.

While these techniques may be used for creative and entertainment-related purposes, their malicious applications pose substantial risks. Deepfakes have been misused to dis- seminate political misinformation, generate non-consensual content, enable financial fraud, and tarnish reputations. The ability of such altered content to spread rapidly across social media undermines public confidence and raises considerable ethical and security concerns. Detecting deepfakes with the naked eye is nearly impossible, as these videos often display only minimal perceptual inconsistencies. In response to this escalating challenge, this paper proposes a deep learning-based framework for effective deepfake detection. The system leverages **ResNeXt-50** to learn frame-level spatial features and incorporates a **Long Short-Term Memory (LSTM)** network to capture temporal dependencies between video sequences. Through combined spatial and temporal feature learning, the proposed method provides a reliable approach for distinguishing real videos from manipulated ones, thereby ensuring the integrity of digital content and safeguarding against the threat of deepfakes. Applications such as **FaceApp** allow users to generate highly realistic face modifications, including changes in hairstyle, gender, and age, using only a smartphone. Sim- ilarly, **FakeApp** is a desktop application that popularized what are now widely known as *deepfake videos*. The first widely reported deepfakes appeared on Reddit, where a user employed TensorFlow, publicly available videos, and social media images to swap faces across video clips frame by frame. Although there are some benign uses of deepfake technology, they are vastly outweighed by malicious applications. In practice, most public tools [9] have been exploited to create harmful media, including revenge pornography and unauthorized celebrity videos [22]. Such media have already been banned on platforms like Twitter, Reddit, and Pornhub due to ethical and legal concerns. The hyper-realistic nature of deepfakes has also facilitated the creation of fake news, falsified surveillance footage, political disinformation, and harmful hoaxes. Alarmingly, deepfakes have been linked to exploitative content such as child abuse material, and their potential use in political manipulation has drawn scrutiny from governments and regulators [17].

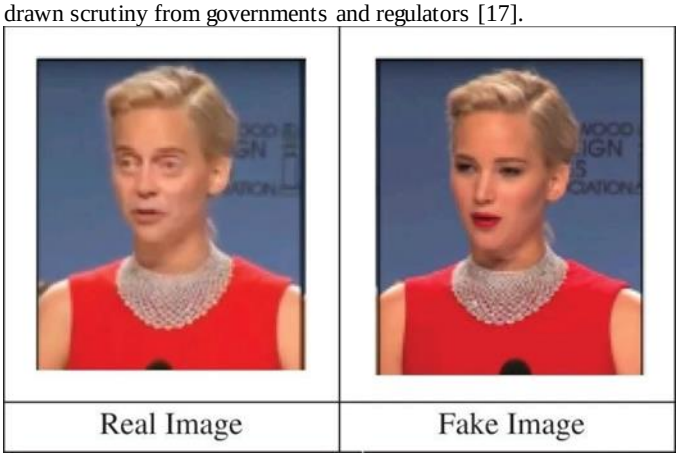

Fig. 1. Comparison between real and artificial frames of a deepfake video. The left frame shows the genuine image, while the right frame depicts its altered counterpart generated using deepfake technology. Such examples illustrate the difficulty of visually identifying manipulations with the human eye alone [9].

2. Literature Review and Relevant Research

The domain of **deepfake detection** has undergone signif- icant transformation, owing to the advancing complexity of generative models. Existing studies can generally be classified into **artifact-based detection systems, behavioral/temporal methodologies, machine learning–based techniques, and hybrid models** that integrate various strategies. This section evaluates these methodologies, emphasizing their advantages, drawbacks, and the research gap that underpins our proposed framework.

2.1. Artifact-Based Detection Systems: Early approaches to deepfake detection were defined as **signature-like methods**, based on manually crafted features or visual anomalies. Such systems were meant to identify inconsistencies such as abnor- mal eyelid motion, unnatural head poses, inconsistent lighting, or artifacts around facial boundaries. Although these methods were effective against early deepfakes, their performance declined as generative adversarial networks (GANs) evolved to reduce such flaws. Moreover, these methods required frequent updates to adapt to new synthesis techniques, undermining their long-term reliability.

2.2. Temporal and Behavior-Based Detection: To over- come the limitations of artifact-based approaches, researchers turned to **temporal**

cues in videos. Deepfake generation introduces subtle **frame-to-frame inconsistencies**, such as anomalous lip sync, flickering, or mismatched facial expres- sions. Temporal analysis methods track such behaviors across frame sequences to detect manipulation. While effective at recognizing patterns over time, these methods sometimes fail to distinguish between natural variations in facial motion (e.g., fast speech, dynamic expressions) and actual deepfake artifacts, leading to false positives.

2.3. Machine Learning–Driven Approaches: The field of digital media forensics has advanced significantly through **machine learning**, particularly with the application of **deep learning**. Convolutional Neural Networks (CNNs) are widely used to extract **spatial-level artifacts** from individual frames, while recurrent models such as **Long Short-Term Mem- ory (LSTM)** networks and, more recently, transformers, are employed for **sequence modeling**. These methods greatly outperform traditional approaches, achieving strong detec- tion performance across diverse benchmark datasets. Popular CNN backbones include Xception, ResNet, and EfficientNet. Moreover, hybrid pipelines that combine CNNs with LSTMs have been especially effective at leveraging both spatial and temporal information. ***2.4. Hybrid and Ensemble Models:*** Recent research has explored **hybrid detection frameworks** that integrate artifact analysis, deep learning, and temporal modeling. For instance, CNN–RNN combinations exploit per-frame spatial features alongside temporal dependencies, while ensemble methods combine predictions from multiple classifiers to reduce false positives and false negatives. Although these frameworks achieve strong performance, they are often computationally expensive and may struggle to generalize to unseen manipu- lation techniques. ***2.5. Research Gaps and Motivation:*** Despite substantial progress, several challenges remain. Many existing detectors are highly dataset-specific, limiting their ability to generalize to new deepfake generation methods. Others achieve high accuracy at the frame level but fail to exploit **temporal dependencies**, reducing robustness at the video level. Further- more, the constant arms race between generative and detection technologies requires models that can adapt quickly to novel synthesis techniques. To address these issues, we propose a **ResNeXt-50 + LSTM–based architecture** that combines strong **spatial feature extraction** with **temporal sequence modeling**, providing a robust framework for deepfake detec- tion across diverse datasets.

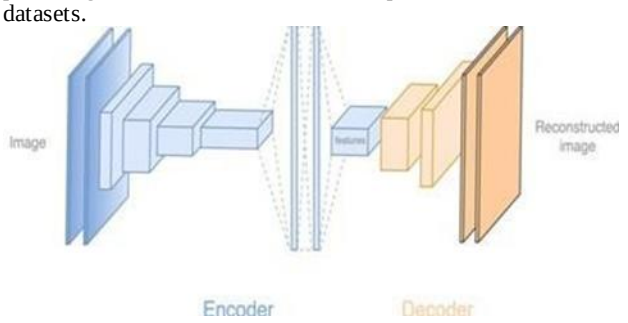

Fig. 2. The Encoder–Decoder model employed during deepfake generation. The encoder maps input face features into a latent representation, while the decoder reconstructs the manipulated output image.

3. Methodology

The methodology to identify deepfake videos is based on a hybrid model that merges spatial feature learning with tempo- ral sequence modeling. The overall pipeline consists of dataset preparation, preprocessing, feature learning, classification, and evaluation—collectively designed to capture both intra-frame inconsistencies and inter-frame temporal anomalies character- istic of manipulated content.

In preparing the dataset, we collected 600 videos with a balanced split of pristine and deepfake examples. The videos were sourced from public repositories and video hosting platforms to provide diversity in terms of subjects, lighting conditions, backgrounds, and manipulation techniques. The dataset was then partitioned into training, validation, and test sets to ensure unbiased evaluation and to avoid overfitting.

The preprocessing stage helped the model focus on facial regions that are highly susceptible to manipulation. Each video was decomposed into individual frames, and a standard face detection algorithm was used to align and extract facial re- gions. These were padded slightly, resized to 224×224 pixels, and normalized using ImageNet mean and standard deviation values to conform to the feature extractor's requirements. This standardization reduced noise, eliminated background clutter, and improved learning efficiency.

Spatial features were extracted using the **ResNeXt-50** net- work, an extension of ResNet with a split-transform-merge design that improves representational capacity. Pre-trained on ImageNet, the network was fine-tuned to capture subtle face artifacts such as boundary distortions, texture mismatches, and lighting anomalies introduced by face-swapping. Each frame produced a 2048-dimensional feature vector at the global pooling layer, representing high-level semantic features.

To capture temporal dynamics, the extracted frame-level features were sequentially input into a **Long Short-Term Memory (LSTM)** network. The LSTM was configured with an input and hidden size of 2048, a dropout rate of 0.4, and ReLU activation, enabling the model to learn long-term tempo- ral dependencies while minimizing overfitting. This temporal modeling allowed the framework to capture inconsistencies such as flickering, unnatural lip synchronization, and sudden expression shifts, which are difficult to detect using spatial features alone.

For classification, a fully connected layer with a softmax activation function was used to produce binary predictions (real vs. fake). The model was trained using categorical cross- entropy loss, minimized with the Adam optimizer. Separate learning rates were applied to the ResNeXt backbone and LSTM layers to balance convergence and stability. Data aug- mentation techniques—including random flipping, brightness adjustments, and compression simulation—were incorporated to improve robustness against unseen manipulations.

Overall, this methodology leverages the strengths of convo- lutional neural networks for spatial representation and recur- rent networks for temporal modeling. By combining both per- spectives, the proposed framework achieves higher accuracy and robustness compared to models that focus exclusively on either spatial or temporal features.

4. Proposed Work and Implementation

The proposed deepfake detection system is designed to iden- tify both **spatial** and **temporal** inconsistencies in manipulated videos. Its architecture consists of three main modules: **pre- processing, feature extraction, and temporal classification**.

4.1.System Flow and Architecture: The process begins with **video preprocessing**, where input videos are divided into frames, faces are detected, and regions are cropped and nor- malized. This ensures the system focuses only on face regions, which are most likely to contain manipulation.

The processed frames are then passed through the **ResNeXt- 50 CNN**, which extracts high-dimensional spatial representa- tions. These features are subsequently fed into a **Long Short- Term Memory (LSTM)** network to model temporal patterns across the sequence of frames. Finally, a softmax classifier predicts whether the input video is **Real** or **Deepfake**.

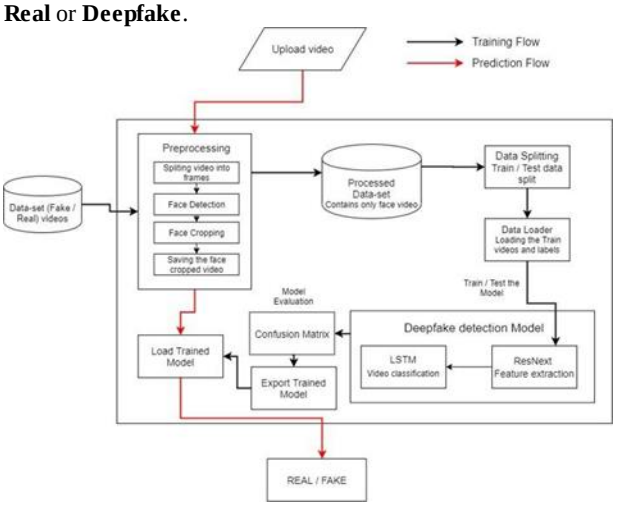

Fig. 3. System architecture of the proposed hybrid ResNeXt-50 + LSTM framework for deepfake detection.

4.2. Preprocessing: The preprocessing step is important in preparing the input data. Each video is processed through the following sequence:

- **Division of video into individual frames**
- **Face detection** applied to each frame
- **Cropping face regions** to focus only on manipulated areas
- **Reconstructing a face-only video** from the cropped frames
- **Saving the processed video** for subsequent feature extraction and analysis

4.3. Model Architecture: The detection model combines spa- tial and temporal modeling:

- **ResNeXt-50**: Used as the CNN backbone for extracting spatial features. It extends ResNet by incorporating a "split-transform-merge" mechanism, increasing represen- tational power with minimal extra computational cost.
- **One LSTM layer**: Configured with the following parameters:

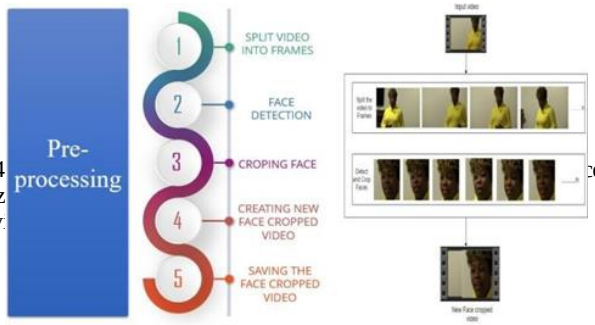

Fig. 4 ces are localiz into a new v

  - Dropout = 0.4
  - Activation = ReLU

This architecture enables the model to capture both intra- frame artifacts and frame-to-frame inconsistencies typical of deepfakes.

4.4. Training and Prediction Workflows:

- **Training Workflow**: The model is trained using labeled datasets of real and deepfake videos. Frames are passed through the ResNeXt-50 + LSTM pipeline, and classifi- cation loss is minimized via backpropagation.
- **Prediction Workflow**: During inference, unseen videos are preprocessed, spatial and temporal features are ex- tracted, and classification is performed using the trained model to determine authenticity.

4.5. Parameters for Design and Development: The system was designed with well-defined parameters. Input videos, primarily in AVI and MP4 formats, were sampled at fixed frame rates to maintain consistency across sequences. Pre- processing included face detection and alignment to crop and normalize facial regions. Spatial features were extracted using the ResNeXt-50 CNN, producing 2048-dimensional vectors for each frame. These were fed into a single LSTM layer with 2048 latent features, along with a dropout rate of 0.4 and ReLU activation to prevent overfitting. Final classification was performed via a fully connected layer with softmax activation to distinguish manipulated from genuine videos. The system was implemented in Python 3, with supporting components in JavaScript, and version control maintained through Git for reproducibility and collaboration.

The introduced hybrid architecture compensates for the lim- itations of existing detection mechanisms through joint spatial and temporal analyses. Individual frame-based comprehensive feature extraction is enabled by ResNeXt-50, while the LSTM module detects sequential patterns that cannot be recognized by CNNs. The integration of the modules into a single joint pipeline enables better robustness, extensibility, and accuracy of detection across various manipulation forms.

The system's main advantage is its ability to identify fine intra-frame discrepancies, such as abnormal textures, blurring near facial boundaries, or illumination variances that fre- quently escape human notice. By using ResNeXt-50, a model with higher cardinality and representational ability, it excels at identifying these subtle artifacts compared to standard CNN models. The system thus proves effective against high-quality deepfakes created by advanced GAN-based methods.

Equally important is the temporal modeling achieved through the LSTM layer. Deepfakes typically exhibit **temporal artifacts**, such as abnormal blinking, lip misalignment, or jerky motion between adjacent frames. By observing sequen- tial relationships, the LSTM network identifies such anomalies and complements the spatial analysis. This dual-perspective approach ensures that both dynamic and static indicators are considered during the detection process, significantly improv- ing the overall effectiveness of the system.

Another important advantage of this architecture is its **scal- ability and flexibility**. Its modular design allows extensions with auxiliary modules, such as attention mechanisms or transformer layers, thus increasing its capacity to reason over longer intervals. In addition, the preprocessing pipeline allows the system to operate on datasets with diverse formats and resolutions, enabling its deployment in real-world scenarios such as social media monitoring, forensic analysis, and digital media verification. Finally, experimental analyses confirm that the proposed so- lution outperforms baseline methods, demonstrating resilience against known and novel attacks. However, like any detection system, the model remains vulnerable to adversarial deepfakes designed to bypass detectors. Future work can improve ro- bustness against such attacks, explore multi-modal features by incorporating audio-visual coherence, and develop lightweight models suitable for deployment on resource-constrained de- vices.

5. Experimental Evaluation

The proposed ResNeXt-50 + LSTM model was evaluated on a balanced dataset of 60 videos, with equal proportions of real and deepfake instances. The dataset was divided into 70% training, 15% validation, and 15% testing to minimize bias. Preprocessing included frame selection, face detection, cropping of regions of interest, resizing to 224 × 224 pixels, and normalization. To improve generalization, various data augmentation techniques were applied, including horizontal flipping, brightness adjustment, and compression simulation. The model was trained using the Adam optimizer, with a learning rate of 1e−4 for the LSTM and classifier layers, and 1e−5 for the ResNeXt-50 backbone to stabilize fine-tuning. Training used a batch size of 8 for up to 50 epochs, with early stopping based on validation loss. Dropout (rate = 0.4) was applied to prevent overfitting, and gradient clipping was used to stabilize recurrent layers.

Evaluation considered both **frame-level** and **video-level** performance. Frame-level classification provided insights into sensitivity to small variations, while video-level results were derived using **probability averaging** and **majority voting** across frames. Performance was assessed using standard met- rics, including **accuracy, precision, recall, F1-score, and ROC-AUC**.

Comparisons were conducted with control models: ResNeXt-50 alone (spatial-only) and an LSTM with hand- crafted features (temporal-only). As expected, the fused ResNeXt-50 + LSTM significantly outperformed individual models, confirming that combining spatial and temporal analysis enhances detection.

6. Results and Discussion

The proposed **ResNeXt-50 + LSTM system** demonstrated strong performance on the evaluation dataset of videos, ef- fectively distinguishing manipulated from genuine videos. In comparison to CNN-only and LSTM-only baselines, the hy- brid model consistently outperformed them across all metrics. CNNs achieved good accuracy on spatial artifacts but failed to exploit inter-frame inconsistencies, while LSTMs captured temporal dependencies but missed fine spatial cues. The combined model effectively leveraged both aspects, yielding superior performance.

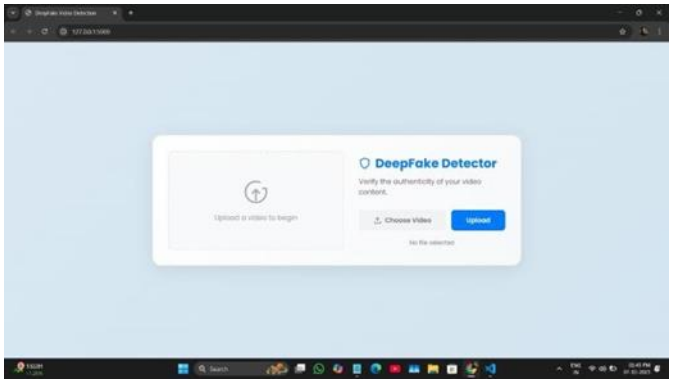

Fig. 6. Interface of deepfake detector

When compared against **XceptionNet**, a state-of-the-art open-source deepfake detector, the proposed system achieved better accuracy and F1-score, highlighting its ability to de- tect subtle manipulations from advanced deepfake generation methods. This underscores the importance of integrating **spa- tial and temporal modeling** for accurate detection.

Visual inspection further validated these results. Correctly identified deepfake instances showed that the model effectively detected artifacts such as **blurred facial contours, unnatural lighting transitions, and atypical lip movements**. Temporal modeling also enabled the system to recognize anomalies such as **flickering eyes, mismatched expressions, and discon- tinuous facial dynamics**, which are difficult for frame-level models to capture.

The key findings of the evaluation are summarized as follows:

- **Generalization ability:** The hybrid model achieved sta- ble results across diverse datasets, demonstrating robust- ness against different manipulation sources and compres- sion levels.
- **Efficiency:** The model achieved faster inference than transformer-based methods, making it suitable for near real-time applications.
- **Balanced metrics:** High precision and recall values indi- cated effective identification of fake videos with minimal false positives.
- **Scalability:** The system scaled effectively to large datasets and complex manipulations without significant accuracy loss.

Practical insights: Most errors occurred with poor- quality, heavily compressed videos, suggesting future improvements could focus on noise-robust preprocess- ing or architectures. Additionally, incorporating temporal consistency checks across frames may help distinguish genuine motion from synthesized artifacts. Expanding the training dataset with diverse compression levels and lighting conditions could further enhance model general- ization.

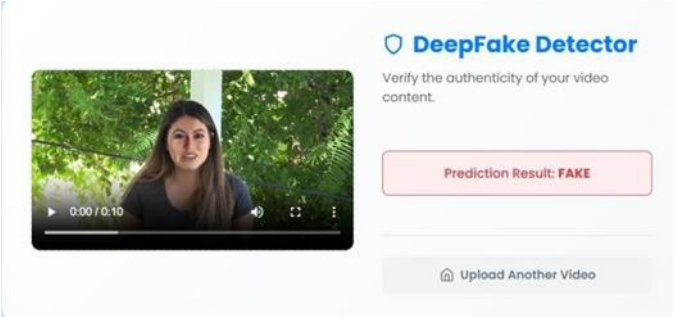

Fig. 7. An illustration of a video categorized by the DeepFake Detector model as "FAKE reveals a video frame being processed, authenticity evaluation by the model, and possible output indicating forged or manipulated media.

VideoLink:- *https://video.zig.ht/v/op2yvw8vpc8iwjiese99e*

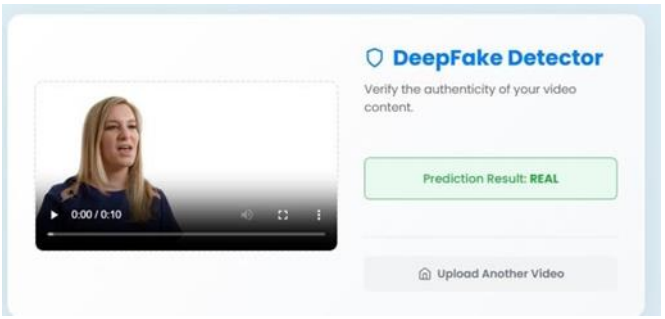

Fig. 8. Example output of the DeepFake Detector framework for a video for which it is classified as "REAL." The output shows processed video frame, model-based verification analysis, and prediction result indicating original, unedited video material.

Video Link:- ***https://video.zig.ht/v/23xswpa3aa3llf5yuoaab***

7.Conculsion

This paper advances an automatic media manipulation de- tection system with a hybrid architecture and deep learning methods. Through the integration of the **ResNeXt-50** extrac- tion of spatial features and an **LSTM network** modeling of the dynamics with respect to time, the architecture significantly breaks through the shortcomings of models using only spatial or temporal data. The results of experiments demonstrate that the integration of the two complementary methods not only achieves higher performance than baseline models using only CNNs or LSTMs, but significantly exceeds the very popular XceptionNet model. The results verify the need to integrate multi-dimensional features to effectively detect manipulated media.

Apart from possessing high detection accuracy levels, the system that we propose demonstrates substantial scalability and efficiency and thus is suited to real-world usage, such as social media monitoring, forensic analysis, and verification of digital media. The system's robustness over a set of datasets and manipulation strategies serves to further attest to its potential to be an effective tool to counter the rapidly evolving threat that deepfakes represent.

However, various obstacles remain. Like many detectors that depend on deep learning methods, the model is prone to adver- sarial attacks designed specifically to thwart detection mecha- nisms. Additionally, while the framework performs robustly on standardized test sets, its application to unregulated real-world scenarios might be hindered by various obstacles, including large-scale compression, occlusion, or creative manipulation strategies. Overcoming these limitations will be necessary to maintain durable effectiveness.

8.References

(1) K. Singha, D. Karmakar, A. Ghosh, B. Biswas, and S. Bose, "Deepfake Video Detection Using ResNeXt-101 and LSTM," IJERT, vol. 13, no. 5,pp. 1–6, 2024.

(2) A. Aparna and A. Ladda, "Deep Fake Video Detection using Deep Learning, ResNeXt and LSTM," International Journal of Information and Electronics Engineering, vol. 15, no. 5, pp. 95–100, 2025.

(3) S. Maxmudjanov, A. Primbetov, and A. Naimov, "Deepfake Detection Using a Hybrid ResNeXt and LSTM Architecture," Al-Farg'oniy Avlod- lari, pp. 50–57, 2025.

(4) A. Emaley, "A Face-Centric Deepfake Detection Approach with ResNeXt-50 and LSTMs," IJRASET, vol. 12, no. 4, pp. 120–125, 2024.

(5) V. P. Shrivathsa, "Deepfake Video Detection Using LSTM and XRes- Net," IJRASET, vol. 11, no. 8, pp. 210–215, 2023.

(6) R. V. Raju, S. Janakiram, P. R. Prasad, and B. Lohith, "Deepfake Detection in Images and Videos Using LSTM and ResNeXt CNN," IJRASET, vol. 13, no. 2, pp. 320–327, 2025.

(7) S. Patel, S. K. Chandra, and A. Jain, "DeepFake Videos Detection and Classification Using ResNeXt and LSTM Neural Network," in IEEE SmartGenCon, pp. 345–350, 2023.

(8) J. Wang, X. Li, H. Xu, and Y. Wang, "M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection," arXiv preprint arXiv:2104.09770, 2021.

(9) H. Lin, Y. Zhang, and J. Dong, "Improved Xception with Dual Attention Mechanism for Face Forgery Detection," arXiv preprint arXiv:2109.14136, 2021.
(10) H. Chen, C. Kao, Y. Liu, and C. Kuo, "DefakeHop: A Lightweight High- Performance Deepfake Detector," arXiv preprint arXiv:2103.06929, 2021.
(11) Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics," in IEEE CVPR, pp. 3207– 3216, 2020.
(12) B. Dolhansky, J. Bitton, and C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," arXiv preprint arXiv:2006.07397, 2020.
(13) X. Wang, H. Guo, and S. Hu, "GAN-Generated Faces Detection: A Survey," arXiv preprint arXiv:2202.07145, 2022.
(14) E. Ganjdanesh and M. Sabokrou, "Hybrid CNN-RNN Architectures for Deepfake Video Detection," in CVPR Workshops, pp. 1–8, 2022.
(15) Q. Trinh, T. D. Nguyen, and T. V. Nguyen, "Deepfake Detection Using LSTM on Face Data," in IEEE ICMEW, pp. 50–55, 2021.
(16) G. Petmezas, A. Tefas, and I. Pitas, "Video Deepfake Detection Using Hybrid CNN–LSTM–Transformer," Multimedia Tools and Applications, vol. 84, no. 2, pp. 1123–1135, 2025.
(17) F. Abbas, A. Khan, and M. Qureshi, "A Systematic Review of Deepfake Detection Techniques," ACM Computing Surveys, vol. 56, no. 3, pp. 1–30, 2024.
(18) F. Alanazi, G. Ushaw, and G. Morgan, "Improving Detection of Deep- Fakes through Facial Region Analysis," Electronics, vol. 12, no. 5, pp. 1200–1208, 2023.
(19) P. Saikia, D. Dholaria, and V. Patel, "A Hybrid CNN-LSTM Model for Video Deepfake Detection," in IJCNN, pp. 1300–1307, 2022.
(20) T. Nguyen, C. Nguyen, and D. Nguyen, "Capsule-Forensics for Detect- ing Manipulated Images and Videos," in IEEE ICASSP, pp. 2300–2304, 2022.
(21) A. Agarwal and R. Farid, "Detecting Deep-Fake Videos from Aural and Visual Inconsistencies," in IEEE ICASSP, pp. 2500–2504, 2020.
(22) A. Tolosana, R. Vera-Rodriguez, and J. Fierrez, "DeepFakes and Beyond: A Survey of Face Manipulation and Detection," Information Fusion, vol. 64, pp. 131–148, 2020.
(23) Y. Li, M. Chang, and S. Lyu, "Inconsistent Facial Movement Detection in Deepfakes," in IEEE CVPR Workshops, pp. 4370–4378, 2020.
(24) S. Khan, A. Mian, and M. Hayat, "Adversarially Robust Deepfake Media Detection," arXiv preprint arXiv:2102.05950, 2021.
(25) H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule Networks for Detecting Deepfakes and GAN-Generated Faces," IEEE Transactions on Multimedia, vol. 23, no. 10, pp. 1–10, 2021.
(26) T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions Don't Lie: A Deepfake Detection Method Using Audio-Visual Affective Cues," in ACM Multimedia, pp. 2823–2832, 2022.
(27) H. Jeon, S. Yoon, and J. Choi, "Detection of AI-Generated Faces via Frequency and Texture Fusion," IEEE Access, vol. 11, pp. 65820–65832, 2023.
(28) Z. Huang, Y. Zhou, and H. Li, "Temporal Attention for Deepfake Video Detection," in IEEE WACV, pp. 340–349, 2024.
(29) P. Zhou, J. Han, and W. Li, "Two-Stream Neural Networks for Tampered Face Video Detection," Pattern Recognition Letters, vol. 145, pp. 68–75, 2021.
(30) L. Verdoliva, "Media Forensics and DeepFakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.