Enhancing Deepfake Detection through Hybrid MobileNet-LSTM Model with Real-Time Image and Video Analysis

Chinna Venkataswamy[1], Pabbathi Jacob Vijaya Kumar[2], Bavanam Yeswanth Kumar Reddy[1], Mangali Suri Babu[1], Neelam Venkatesh[1] and Vuluvala Pavasubralali Reddy[1]
[1]*Department of Computer Science and Design, Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India*
[2]*Department of Computer Science and Engineering (AIML), Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India*

Keywords: Deepfake Detection, MobileNet, LSTM, Hybrid Model, Real-Time Detection, CNN (Convolutional Neural Network), Transfer Learning, TensorFlow Lite, Digital Security, Adversarial Training, GAN (Generative Adversarial Network), Heatmaps, Probability Scores.

Abstract: Deepfake technology is spreading false information and posing a threat to digital security. A Hybrid MobileNetLSTM Model for real-time deepfake detection in videos and images is presented in this project. MobileNet, a lightweight CNN, extracts spatial features, and LSTM records temporal dependencies in video sequences, both of which guarantee high detection accuracy while utilizing minimal computation. Transfer learning is used to train the model on large datasets of real and fake media for better generalization. Heatmaps and probability scores that can be interpreted are provided by the system, which is integrated with OpenCV and TensorFlow. TensorFlow Lite is used to optimize real-time inference, making it possible to use it on mobile and edge devices. On live video feeds, realtime performance is validated, and experimental evaluations demonstrate superior accuracy to traditional CNN-based approaches. This scalable detection system provides support for media forensics, social media verification, and digital security. GAN-based adversarial training, which will bridge the gap between high accuracy and real-time deployment, will be one of the future enhancements. This will make the system more resistant to changing deepfake techniques.

1 INTRODUCTION

Deepfake technology, which is powered by AIdriven generative models, has emerged as a major cause for concern due to its potential misuse in misinformation, identity fraud, and digital security threats. Traditional deepfake detection methods often struggle with real-time processing and accuracy, especially in dynamic video content. This project introduces a cutting-edge Hybrid MobileNet-LSTM Model for enhancing deepfake detection through real-time image and video analysis. MobileNet, a lightweight CNN, effectively extracts spatial features, whereas LSTM networks increase detection precision by capturing temporal dependencies in video sequences.

By employing a hybrid approach, the model achieves a balance between high accuracy and efficient computation, making it suitable for realtime applications. The system is trained on a diverse dataset of genuine and altered media to guarantee robust generalization. When it is integrated with OpenCV and TensorFlow, it provides manipulated region heatmaps and realtime probability scores for improved interpretability.

TensorFlow Lite also makes it possible to deploy on mobile and edge devices, which improves accessibility. Experimental evaluations confirm state-of-the-art performance, outperforming conventional CNN-based methods in deepfake detection. An effective, scalable, and interpretable deepfake detection system for use in media forensics, law enforcement, and social media verification is provided as part of the project to enhance digital security. Future enhancements will also include adversarial training to combat advancing deepfake techniques. Transfer learning methods are used to fine-tune the model, making it able to adapt to changing approaches to deepfake generation.It has become increasingly challenging to distinguish genuine content from manipulated media due to the rapid development of deepfake technology. Security, misinformation, and privacy are gravely endangered

Venkataswamy, C., Kumar, P. J. V., Reddy, B. Y. K., Babu, M. S., Venkatesh, N. and Reddy, V. P.
Enhancing Deepfake Detection through Hybrid MobileNet-LSTM Model with Real-Time Image and Video Analysis.
DOI: 10.5220/0013930400004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 394-399
ISBN: 978-989-758-777-1

by this. Accuracy and real-time processing are frequently issuing for traditional deepfake detection models. To address these issues, this study proposes a hybrid MobileNet-LSTM model for deepfake detection that combines convolutional and sequential processing to enhance feature extraction and temporal analysis.

MobileNet, a lightweight CNN, effectively extracts spatial features from images and video frames, while LSTM captures temporal dependencies, enhancing detection in dynamic video sequences. Because it is designed to work best with realtime applications, the proposed model can be used on mobile and cloud-based platforms quickly and reliably. This study contributes to the growing demand for robust and scalable deepfake detection methods by providing a lightweight but effective solution for real-world applications in digital forensics and security.

2 LITERATURE REVIEW

Nguyen et al. (2019) created a CNN-based deepfake detection model with high static image accuracy. However, their method was unsuccessful at identifying video sequences with temporal inconsistencies. They emphasized the drawbacks of frame-based detection and advised incorporating temporal analysis for better outcomes. For improved performance, their findings emphasized the need for hybrid models that combine CNNs with sequence-based architectures like LSTM.

Afchar et al. (2018) proposed MesoNet, a lightweight deepfake detection model for detecting subtle face manipulation artifacts using shallow CNN layers. Their model performed well in Realtime scenarios but lacked robustness against high quality deepfakes generated by Abased architectures. They concluded that while lightweight CNNs can offer computational efficiency, incorporating sequential learning methods such as RNNs or LSTMs could significantly enhance detection performance.

Tolosana et al. (2020) provided a comprehensive review of deepfake detection techniques, categorizing methods into pixel-based, frequency-based, and hybrid approaches. Their research demonstrated that standalone CNN-based methods consistently performed worse than hybrid models that make use of both spatial and temporal features. They also talked about how important it is to have a variety of datasets and how many models don't work well when applied to different deepfake datasets.

Chai et al. (2020) introduced an attention-based LSTM model for deepfake video detection, focusing on capturing long-range temporal dependencies. Their approach significantly reduced false positives compared to traditional CNN models. However, the high computational demand of attention mechanisms made real-time implementation challenging. In order to make better trade-offs between speed and accuracy, they suggested optimizing LSTM structures.

Zhao et al. (2021) developed a two-stream CNNLSTM model in which one stream used a CNN to process spatial features and another used LSTMs to analyze temporal variations. Their hybrid method improved the accuracy of detection, especially in videos with subtle facial manipulations. Additionally, they emphasized the significance of frame selection strategies, as the model's efficiency could be compromised by using redundant or irrelevant frames.

Rossler et al. (2019) created the FaceForensics++ dataset, one of the most widely used deepfake benchmark datasets. The gap in generalization between models trained on a single dataset and those trained on multiple datasets was brought to light in their study, which compared multiple deepfake detection models. In order to effectively deal with a variety of deepfake variations, their findings emphasized the necessity of adaptive and hybrid deepfake detection architectures.

Li et al. (2020) proposed a method for frequency-based deepfake detection that looked at how manipulated images were represented in the Fourier domain. Their research showed that many deepfake algorithms leave traces in the frequency spectrum that can be seen. By combining frequency-domain features with deep learning models, they achieved higher accuracy than traditional pixel-based CNN approaches.

Wang et al. (2022) used self-attention mechanisms to capture both local and global dependencies in deepfake videos to implement a transformer-based model for deepfake detection. On highly compressed videos, where CNN-based models struggled, their method performed better. However, the high computational cost of transformers remained a limitation for real-time applications.

Ding et al. (2021) investigated adversarial training to improve the robustness of deepfake detection. In order to strengthen the model's generalization against undetected deepfake manipulations, adversarial examples were used in its training. They found that adversarially trained models performed better against adaptive deepfake generation techniques, making them more resilient in real-world applications.

Jain et al. (2022) proposed an audio and visual cue-based multimodal deepfake detection system. Lip-sync inconsistencies and unnatural facial movements could increase detection accuracy, according to their findings. Their model outperformed vision-only deepfake detection techniques, particularly on low-resolution videos, by combining both modalities.

Verdoliva (2020) conducted a comprehensive survey of forensic methods for detecting deepfakes, highlighting the significance of combining various methods of detection. Their research suggested incorporating CNNs for spatial feature extraction and RNNs/LSTMs for temporal modeling into hybrid deep learning architectures. In addition, they emphasized the necessity of explainable AI methods for increasing trust in deepfake detection systems.

Qi et al. (2021) introduced a hybrid model that is optimized for real-time processing called MobileNet-LSTM. MobileNet was used for light feature extraction, and LSTM was used to look at temporal dependencies in deepfake videos. Their findings demonstrated that hybrid models are suitable for use on mobile devices because they can strike a balance between accuracy, computational efficiency, and real-time applicability.

Chen et al. (2022) used deep learning models and facial landmark analysis to find small distortions in deepfake videos. Their method helped interpret manipulated regions and improved deepfake localization by analyzing facial geometry inconsistencies. Their approach, which worked in conjunction with deep learning models, improved overall detection robustness.

Guera AND Delp (2018) introduced a framework¨ for temporal consistency analysis based on LSTM for the purpose of detecting deepfakes. Their model observed unnatural transitions in deepfake-generated videos by tracking facial movements across video frames.

3 EXISTING SYSTEM

The existing deepfake detection systems primarily rely on Convolutional Neural Networks (CNNs) for feature extraction and classification, but they face significant challenges in detecting manipulated videos due to the lack of temporal consistency analysis. For fake images, traditional CNN-based models like XceptionNet, ResNet, and VGG-16 have demonstrated high accuracy; however, for deepfake videos, these models struggle because they do not capture sequential dependencies between frames. Facial texture analysis, frequency-domain analysis, and anomaly detection are examples of handcrafted feature-based methods that are used in some of the current approaches. However, not all deepfake generation methods benefit from these strategies. Another class of detection model uses Long Short-Term Memory (LSTM) or Recurrent Neural Networks (RNNs) to investigate video temporal inconsistencies. However, standalone LSTM-based methods often lack robust spatial feature extraction, making them less effective for frameby-frame deepfake detection. Although some hybrid models combine CNNs and RNNs to improve detection accuracy, their real-time applicability is limited by their high computational costs. Additionally, large-scale deepfake datasets are used by a lot of deepfake detection models, such as FaceForensics++, Celeb-DF, and DFDC. However, they have a harder time detecting novel deepfakes because these models frequently overfit to particular kinds of synthetic media. Most of the existing deepfake detection frameworks are designed for offline analysis, making them impractical for Realtime detection in live-streamed content or social media applications.

The majority of CNN-based detection systems also face difficulties with adversarial attacks, in which deepfake models are fine-tuned to evade detection mechanisms. Transformer based models and attention mechanisms have recently made progress that has improved detection performance. However, these models and attention mechanisms require a lot of computational power, making them unsuitable for mobile and edge computing environments. Deepfake detection systems primarily use CNN-based models like XceptionNet and ResNet for feature extraction, but they struggle with temporal consistency in videos. Although RNNs and LSTMs aid in the investigation of sequential inconsistencies, they lack robust spatial feature extraction.

Hybrid CNN-RNN models improve accuracy but are computationally expensive, limiting real-time application. Existing models trained on large datasets often overfit to specific deepfake types, reducing generalization. Adversarial attacks further challenge detection mechanisms, and transformer-based models, though promising, require extensive resources. Additionally, trust in automated systems decreases when explain ability is poor. It is necessary to have a robust, scalable hybrid model that strikes a balance between real-time detection, efficiency, and accuracy. Additionally, it is challenging to interpret and validate the decisions made by existing deepfake detection systems due to their lack of explain ability, which reduces trust in automated detection. In conclusion, there is a need for the creation of a hybrid

model that is both more durable and scalable, as the various deepfake detection methods currently in use lack a balanced trade-off between accuracy, computational efficiency, real-time applicability, and generalization across various deepfake techniques.

4 PROPOSED SYSTEMS

The proposed system aims to improve deepfake detection by incorporating a Hybrid MobileNetLSTM model for real-time image and video analysis. For effective spatial feature extraction, this strategy makes use of the lightweight architecture of MobileNet and the capacity of LSTM networks to capture temporal dependencies in video sequences. By combining these strengths, the system addresses the flaws of conventional CNN-based models, which frequently overlook temporal inconsistencies in deepfake videos. The MobileNet component ensures real-time application-friendly computational efficiency by extracting spatial features from individual frames. The sequences of these features are then analyzed by the LSTM network to find temporal anomalies that point to deepfake manipulations.

The hybrid architecture enables the model to recognize subtle inconsistencies across frames, enhancing its detection accuracy. Utilizing pretrained MobileNet weights to speed up training and boost generalization, transfer learning methods are used to improve performance. In order to seamlessly process video and image inputs, the system integrates frameworks like OpenCV and TensorFlow for real-time processing. TensorFlow Lite makes it easier to deploy on mobile and edge devices, ensuring low-latency inference that is suitable for applications like social media content monitoring and live video streaming. The detection output includes probability scores and visualizations like heatmaps to help users gain trust and comprehend the situation. The hybrid model's efficiency and accuracy have been tested in experiments using benchmark datasets such as FaceForensics++ and DFDC.

The system's ability to function with minimal latency during stress testing on live video feeds further demonstrates its robustness. By providing a scalable and effective detection mechanism suitable for media forensics, digital security, and social media verification, this project contributes to the fight against deepfake misinformation. Future enhancements may be more robust against evolving deepfake generation methods if adversarial training is incorporated. In conclusion, in the field of deepfake detection, the proposed Hybrid MobileNetLSTM model provides a solution that strikes a balance between realtime deployment viability and high detection accuracy. By incorporating a hybrid MobileNet-LSTM model for real-time image and video analysis, the proposed system improves deepfake detection. The spatial features extracted by MobileNet, a lightweight CNN, and the temporal dependencies captured by LSTM in video sequences improve detection accuracy.

5 METHODOLOGY

The proposed method aims to improve deepfake detection by incorporating a hybrid MobileNetLSTM model for Realtime image and video analysis. By utilizing MobileNet's lightweight architecture for efficient spatial feature extraction and LSTM's capacity to capture temporal dependencies, this strategy addresses the shortcomings of conventional CNN based models, which frequently overlook temporal inconsistencies in deepfake videos. To meet MobileNet's input requirements, preprocessing the input videos involves extracting and resizing frames. After that, every frame is processed using the MobileNet model. Using transfer learning, it has already been trained on large image datasets, allowing the model to use learned features and speed up training. Using the extracted MobileNet spatial features, the LSTM network looks at the temporal dynamics between frames to find anomalies that point to deepfake manipulations.

The model is able to identify subtle inconsistencies that might not be apparent in individual frames thanks to this sequential processing. The Face Forensics++ and DFDC datasets are two examples of diverse datasets that the model is trained on to ensure both robustness and generalizability. In order to achieve a balance between detection accuracy and computational efficiency that makes it suitable for real-time applications, the hybrid architecture must be optimized during the training process. Due to its integration with deployment frameworks like OpenCV and TensorFlow, the system handles video and image inputs seamlessly. Low-latency processing is essential for applications like live video streaming and content monitoring on social media platforms. The detection output includes probability scores that indicate the likelihood of manipulation, as well as visualizations like heatmaps that highlight areas of interest to increase user trust and interpretability.

The hybrid model outperforms conventional approaches in terms of accuracy and efficiency when

tested experimentally on benchmark datasets. Furthermore, the proposed approach leverages batch normalization and dropout techniques during training to enhance generalization and reduce overfitting, ensuring stable performance across varied deepfake manipulations. The integration of an attention mechanism within the LSTM network further refines the detection process by prioritizing critical temporal features that exhibit inconsistencies. Additionally, the model's resistance to adversarial attacks is enhanced by employing data augmentation techniques like random cropping, rotation, and color jittering. The scalability of the model allows deployment across diverse hardware environments, from high-performance servers to resource-constrained mobile devices. Experimental
evaluations demonstrate significant improvements in precision, recall, and F1-score compared to baseline models, confirming the effectiveness of this hybrid approach. Facial reenactment, expression manipulation, and identity swapping are just a few examples of the deepfake detection scenarios that the system can adapt to. The model's smooth execution on edge devices thanks to TensorFlow Lite makes it useful for real-world applications like fraud prevention and misinformation detection.

6 FUTURE WORK

By combining a Hybrid MobileNet-LSTM model with realtime image and video analysis, the work that will be done in the future to improve deepfake detection includes a number of important directions that will increase the system's robustness, accuracy, and adaptability to changing methods for creating deepfakes. The primary objective is to broaden and diversify the training datasets. FaceForensics++ and DFDC datasets are frequently used by current models; however, incorporating more diverse datasets with a variety of backgrounds, ethnicities, and lighting conditions can boost model generalization and performance in a variety of scenarios. Another important option is to use Generative Adversarial Networks (GANs) for adversarial training. By imitating sophisticated deepfake techniques, the model can learn to recognize subtle manipulations during training. This improves the model's resistance to sophisticated deepfake techniques.

Transformer-based architectures like Vision Transformers (ViTs) could further enhance detection capabilities by capturing global relationships in image data. This would provide an alternative perspective to CNNs' local feature extraction. Using multi-modal analysis to combine audio and visual data is another promising direction. Deepfakes, which frequently exhibit inconsistencies, could be detected more effectively and precisely by a system that analyzes both audio and visual cues. The interpretability of the detection model must also be improved. By creating visual explanations like heatmaps that highlight manipulated areas, user trust can be increased and manual verification processes can be made simpler. For real-time applications, optimizing the model for deployment on edge devices is crucial.

Model pruning and quantization can reduce computational requirements, allowing devices with limited resources to function effectively without sacrificing accuracy. To adapt to emerging deepfake techniques, mechanisms for continuous learning should be established. Using online learning algorithms, the model can adjust its parameters in response to new data, maintaining its effectiveness over time. Working with experts in digital forensics and social media platforms can make the collection of real-world deepfake examples, the enrichment of training datasets, and the assurance that the model addresses practical challenges easier. Finally, conducting comprehensive evaluations against adversarial attacks is imperative. The development of robust countermeasures, which will guarantee the system's reliability in adversarial environments, will be informed by an assessment of the model's vulnerability to techniques designed to evade detection. In conclusion, in order to improve the effectiveness of deepfake detection systems, future research ought to concentrate on the diversification of datasets, the integration of advanced architectures and multi-modal analysis, the enhancement of interpretability, the optimization of deployment at the edge, the implementation of continuous learning, collaboration for realworld data acquisition, and defense against adversarial attacks.

7 CONCLUSIONS

The proliferation of deepfake technology, which leverages advanced artificial intelligence to create hyper-realistic synthetic media, poses significant threats to information authenticity, personal privacy, and societal trust. The creation of a hybrid MobileNet-LSTM model for real-time image and video analysis has emerged as a promising approach to address these issues and improve deepfake detection capabilities. This approach synergistically

combines the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, to effectively analyze both spatial and temporal features inherent in video data. MobileNet, a compact and effective CNN architecture, is skilled at extracting spatial features from individual video frames. It does this by capturing intricate details like facial textures, lighting conditions, and subtle inconsistencies that could indicate manipulation. Its design ensures rapid processing, making it particularly suitable for realtime applications where computational resources and latency are critical considerations.

MobileNet's ability to adapt to the particular nuances of deepfake detection by making use of pretrained weights obtained through transfer learning improves both its accuracy and its generalizability to a variety of datasets. Complementing this, LSTM networks are designed to capture temporal dependencies by analyzing sequences of data over time. LSTMs process the sequential MobileNet frames in the context of video analysis, allowing for the detection of temporal inconsistencies and unnatural transitions that are frequently found in deepfake videos. The hybrid model is able to thoroughly evaluate the authenticity of video content thanks to this dualstage processing, which begins with the extraction of spatial features and continues with the analysis of temporal patterns.

The efficacy of this hybrid strategy has been demonstrated by empirical research. For instance, research integrating CNN and LSTM architectures achieved a precision of 98.21 percentage on opensource datasets such as the DeepFake Detection Challenge (DFDC) and Ciplab datasets, indicating a limited false-positive rate and robust detection capabilities. Another study leveraging optical flow features within a hybrid CNNLSTM framework reported an accuracy of 66.26 percentage on the DFDC dataset, further validating the model's effectiveness in discerning deepfake content. The adaptability of the hybrid MobileNet-LSTM model to diverse datasets underscores its robustness and potential for widespread application. By training on a variety of authentic and manipulated media, the model learns to generalize across different scenarios, enhancing its resilience against various deepfake generation techniques. In an environment where deepfake techniques are constantly evolving, posing new challenges to detection systems, this adaptability is essential. The model's lightweight architecture makes it easier to use in practice to perform realtime inference, which is a crucial requirement for applications like live video streaming, social media monitoring, and digital forensics.

REFERENCES

Ali, F., Hussain, A. (2023). Countering Deepfake Threats Using Generative Adversarial Training. Transactions on Cybersecurity, 9(4), 1123-1140.

Brown, C., Jones, M. (2023). A Hybrid Deep Learning Framework for Deepfake Video Detection. Neural Processing Letters, 58(2), 409428.

Chen, Z. et al. (2023). Towards Real-Time Deepfake Identification Using Mobile-Friendly Neural Networks. Sensors, 23(12), 1567.

Dai, Y., Xu, B. (2024). Multi-Modal Deepfake Detection Using AudioVisual Analysis. IEEE Transactions on Multimedia, 26(3), 451-467.

Harrison, Olivia, and James Wilson." A deep learning framework for identifying manipulated videos using MobileNet and LSTM." Neural Networks and Applications 21, no. 2 (2024): 6678.

Huang, X., Li, C. (2023). Explainable AI forDeepfake Detection: Heatmap and SaliencyBased Interpretability. International Journal of Artificial Intelligence Research, 51(5), 223-242.

Kumar, Sanjay, and Li Wei." Mobile Net-LSTM based approach for detecting synthetic media in video footage." Pattern Recognition and Machine Learning 17, no. 3 (2024): 56-72.

Lee, S., Kim, J. (2024). Lightweight Deepfake Detection Using Knowledge Distillation in CNN-LSTM Models. Neural Networks Journal, 157, 65-78.

Nguyen, T. et al. (2023). Deepfake Detection Using Optical Flow Analysis and Recurrent Neural Networks. IEEE Transactions on Information Forensics and Security, 18, 32713285.

Patel, K., Rao, B. (2023). Transfer Learning for Deepfake Video Detection: A Hybrid CNN-RNN Approach. ACM Transactions on Multimedia Computing, 29(5).

Ramirez, J., Torres, L. (2024). CNN-LSTMFusion for Real-Time Deepfake Detection in Social Media Content. Journal of Image Processing, 45(7), 349-366.

Roy, S., Kaur, P. (2023). Evaluating the Role of LSTM in Detecting Deepfake Sequences. Journal of Computational Intelligence, 14(1), 2239.

Wang, H., Lin, J. (2024). Robust Deepfake Detection Through Hybrid CNN-LSTM Models. Journal of Machine Learning Research, 23, 153175.

Zhang, Wei, and Laura Garcia." Real-time analysis of deepfake content using temporal CNN-LSTM networks." Computer Vision and AI Security 14, no. 4 (2024): 89-101.