Implementation of a Hybrid Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) Architecture for Deepfake Video Detection

Bintang Muhammad Madani*
Master of Artificial Intelligent
Gadjah Mada University
Yogyakarta, Indonesia
bintangmuhammadmadani@mail.ugm.ac.id

Dzaky Nafis Alfarizi*
Master of Artificial Intelligent
Gadjah Mada University
Yogyakarta, Indonesia
dzakynafisalfarizi@mail.ugm.ac.id

Soufi Ramadhani Inulqulub*
Master of Artificial Intelligent
Gadjah Mada University
Yogyakarta, Indonesia
soufiramadhaniinulqu@mail.ugm.ac.id

Muhamad Aufa Cholil Fayyadl*
Master of Artificial Intelligent
Gadjah Mada University
Yogyakarta, Indonesia
muhammadaufacholilfayyadl1997@mail.ugm.ac.id

Abstract

The proliferation of sophisticated synthetic media, or deepfakes that generated by Artificial Intelligence (AI) technologies like Generative Adversarial Networks (GANs), poses a severe threat to information security, individual privacy, and social stability. A significant majority of circulating deepfake content (more than 90%) is reported to be used for exploitation or fraud. Consequently, developing accurate and rapid detection mechanisms has become a critical challenge in digital forensics. This study proposes and implements a hybrid deep learning architecture combining EfficientNet for spatial feature extraction and a Bi-LSTM network for temporal sequence modeling. The model utilizes the UADFV dataset (98 videos) for binary classification (real vs. fake). The proposed approach successfully detected subtle spatial inconsistencies (e.g., unnatural textures) and temporal anomalies (e.g., discontinuities in facial dynamics), achieving an AUROC of 0.898. This performance marginally surpasses the previous feature-engineering method based on head-pose inconsistency, which achieved an AUROC of 0.89. Furthermore, detailed classification metrics show a high Recall of 0.93 for the deepfake class and an overall Accuracy of 0.83. These results confirm the effectiveness of hybrid spatial–temporal modeling in capturing complex deepfake artifacts, offering theoretical contributions to digital forensic methods and practical solutions for verifying video authenticity.

Keywords: Deepfake, CNN, Bi-LSTM, Hybrid, EfficientNetV2B0, Haar Cascade, UADFV

I. INTRODUCTION

The recent development of artificial intelligence (AI) technology, particularly in the field of Generative Adversarial Networks (GAN) and deep learning, has led to the emergence of synthetic media known as deepfakes [1]. This technology allows for the manipulation of facial attributes, voices, and body movements in videos with a high degree of realism, making it increasingly difficult to distinguish from original recordings. Although this innovation offers great potential for the creative and education industries, its misuse has posed a serious threat to information security, individual privacy, and social stability. A report from MIT Technology Review [2] indicates that more than 90% of deepfake content circulating on the internet is used for exploitation or fraud, such as the spread of political hoaxes and non-consensual obscene content. Therefore, the development of accurate and rapid detection mechanisms has become a crucial challenge in the field of digital forensics today.

Previous studies have explored various detection methods, ranging from visual analysis to physiological inconsistencies. Korshunov and Marcel highlighted the vulnerability of modern face recognition systems such as VGG and Facenet to deepfake attacks, underscoring the need for specialized detection techniques [3]. Early approaches by Yang et al. focused on inconsistencies in head poses; however, as manipulation techniques evolved to better conceal spatial artifacts, hybrid approaches that combine spatial and temporal analysis became essential [4].

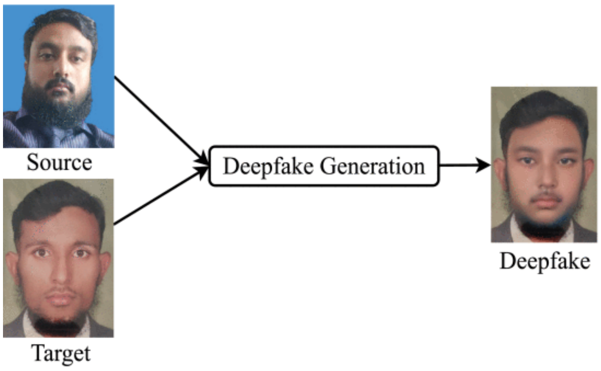

FIGURE 1.
Media generated by Deepfake.

In this architecture, the Convolutional Neural Network (CNN) plays a vital role as a spatial feature extractor, as demonstrated by Ismail et al. using EfficientNet-B5 and by Verma and Kumar using Xception [5] To complement spatial analysis, a Bidirectional Long Short-Term Memory (Bi-LSTM) network is integrated to process sequential information and detect temporal inconsistencies between frames [6][7]. The synergy between spatial and temporal features has proven to yield superior performance, as shown by Soundarya and Gururaj, who achieved significant accuracy on the UADFV dataset [8].

Based on this urgency, the present study aims to explore pre-processing techniques and the implementation of a hybrid CNN and Bi-LSTM architecture to improve deepfake video detection performance. The main problem investigated is how to optimize the application of this method on the UADFV dataset for binary classification (real and fake) with higher accuracy compared to previous approaches. This study is limited to the use of the UADFV dataset for training, validation, and testing, with performance evaluated using confusion matrix, and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC). The findings of this research are expected to provide theoretical contributions to the development of digital forensic methods as well as practical solutions to assist media platforms and the public in verifying the authenticity of video content.

II. Related Work

Yang et al. [4] introduced one of the first dedicated Deepfake detection techniques by leveraging geometric cues from estimated head poses. By comparing discrepancies between localized and global head orientation, their method successfully identified inconsistencies characteristic of early-generation Deepfake videos, particularly on the UADFV dataset. However, head-pose–based methods were later found to be less effective as Deepfake generation techniques evolved to produce more coherent geometric structures.

Subsequent research shifted toward deep learning–based spatial feature extraction. Various CNN architectures have been employed to identify fine-grained visual artifacts that are often imperceptible to humans. Ismail et al. [5] utilized EfficientNet-B5 in combination with YOLO-based face detection to enhance spatial feature representation, while Verma and Kumar [5] applied the Xception architecture, which is known for its ability to capture subtle texture inconsistencies. Soundarya and Gururaj [8] further advanced this line of work by introducing Dense-Swish-CNN, demonstrating improved sensitivity to high-frequency manipulation cues that standard activation functions tend to suppress.

To complement spatial modeling, temporal analysis has played an increasingly important role in Deepfake detection. Yoo and Sung [6] highlighted the capability of Bidirectional Long Short-Term Memory (Bi-LSTM) networks to capture long-range dependencies across video frames, even when the available dataset is limited. Similarly, Zhang et al. [7] showed that Bi-LSTM–based time-series modeling can effectively identify abnormal variations in facial landmarks, providing a strong foundation for temporal anomaly detection. Hybrid architectures combining CNN for spatial extraction and Bi-LSTM for temporal reasoning have consistently outperformed single-modality methods. For instance, the work of Soundarya and Gururaj [8] achieved a detection accuracy of 96.91% on the UADFV dataset by integrating Dense-Swish-CNN with Bi-LSTM, indicating the effectiveness of hybrid spatial–temporal modeling for Deepfake detection.

III. Research Methods

1) Data Collection

The dataset employed in this research is the UADFV dataset, which was obtained through the Kaggle platform [9]. The dataset contains video samples that are categorized into two distinct classes, namely fake and real. Each class consists of 49 video files, resulting in a total of 98 videos available for analysis. All videos in the dataset are provided in MP4 format, ensuring compatibility with common video processing tools and enabling straightforward extraction of visual information.

2) Data Preprocessing

The preprocessing pipeline begins by breaking each video into 30 individual frames using OpenCV, where each extracted frame is saved as an image and labeled accordingly for further processing. This approach facilitates efficient analysis and significantly reduces the computational load typically associated with real-time video segmentation. Based on this structured dataset, face detection is performed using Haar Cascade classification, and the detected face regions are then cropped and resized to 96×96 pixels to obtain a consistent and standardized input format for the model. This combination of pre-framing data and systematic preprocessing improves the efficiency, consistency, and reliability of the subsequent data augmentation and classification stages.

To improve generalization and reduce potential overfitting, several augmentation techniques were applied, including flipping, rotation, zoom, and contrast enhancement. After augmentation, the dataset was divided into three subsets, 70% allocated for training, using k-fold with k=5 for the validation, and the remaining 30% set aside for testing, ensuring a balanced and comprehensive evaluation of the model's performance.

3) Model development

The proposed model employs a hybrid deep learning architecture that integrates EfficientNet as the spatial feature extractor and a Bidirectional Long Short-Term Memory (Bi-LSTM) network for temporal sequence modeling. This design leverages the representational efficiency of EfficientNet and the temporal sensitivity of Bi-LSTM to detect subtle spatial and temporal artifacts commonly found in deepfake videos.

EfficientNet is utilized as the primary convolutional backbone due to its compound scaling strategy, which balances network depth, width, and resolution to achieve high accuracy with significantly reduced computational cost. The input facial frames extracted during preprocessing are fed into the EfficientNet module, where convolutional blocks capture fine-grained spatial inconsistencies such as unnatural texture patterns, blending artifacts, and illumination mismatches typically present in manipulated facial content. The output feature tensors generated by EfficientNet are subsequently flattened and organized into a sequential structure that represents the temporal order of the video frames.

These spatial feature sequences are then processed by a Bidirectional LSTM layer. The bidirectional configuration enables the model to learn temporal dependencies from both forward and backward directions, allowing it to identify abnormal motion cues, discontinuities in facial dynamics, and frame-to-frame temporal inconsistencies. The Bi-LSTM's ability to capture long-range sequential patterns is particularly advantageous for detecting subtle temporal artifacts, even when the dataset size is limited.

After temporal modeling, the final hidden representations are passed through one or more fully connected layers to perform binary classification between real and fake video samples. A softmax activation function is applied in the output layer to obtain class probability scores. During training, the model is optimized using the Adam optimizer, while cross-entropy loss serves as the objective function. Regularization techniques, including dropout and early stopping, are employed to enhance generalization and prevent overfitting, particularly given the relatively small sample size of the UADFV dataset.

4) Model evaluation

The performance of the proposed hybrid EfficientNet–BiLSTM model was evaluated using several metrics such as confusion matrix, and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC). These metrics collectively provide a comprehensive understanding of the model's ability to distinguish between real and fake videos in the UADFV dataset.

Accuracy and loss were monitored during both training and validation to assess the convergence behavior of the model. Accuracy measures the proportion of correctly classified samples, while the loss value reflects the model's prediction error based on the cross-entropy objective function. A consistent decrease in loss accompanied by an increase in accuracy across epochs indicates that the model successfully learns relevant spatial–temporal representations for deepfake classification.

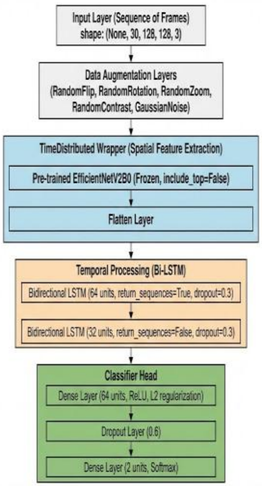
Input Layer (Sequence of Frames)
shape: (None, 30, 128, 128, 3)
Data Augmentation Layers
(RandomFlip, RandomRotation, RandomZoom, RandomContrast, GaussianNoise)
TimeDistributed Wrapper (Spatial Feature Extraction)
Pre-trained EfficientNetV2B0 (Frozen, include_top=False)
Flatten Layer
Temporal Processing (Bi-LSTM)
Bidirectional LSTM (64 units, return_sequences=True, dropout=0.3)
Bidirectional LSTM (32 units, return_sequences=False, dropout=0.3)
Classifier Head
Dense Layer (64 units, ReLU, L2 regularization)
Dropout Layer (0.6)
Dense Layer (2 units, Softmax)

FIGURE 2.
Architecture Model.

To further analyze classification performance, a confusion matrix was generated to provide detailed insight into the types of errors made by the model. The matrix consists of four key components: True Positive (TP), representing fake videos correctly identified as fake; True Negative (TN), representing real videos correctly identified as real; False Positive (FP), indicating real videos incorrectly classified as fake; and False Negative (FN), indicating fake videos incorrectly classified as real. Examining these values enables a deeper understanding of the model's error patterns and its reliability in practical deepfake detection scenarios.

Additionally, the model's discriminative capability across varying threshold values was assessed using ROC-AUC. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), while the AUC value quantifies overall separability between the two classes. An AUC score approaching 1.0 reflects excellent classification performance, indicating that the model consistently distinguishes real from fake videos under different decision boundaries.

IV. Result And Discussion

1) Performance Evaluation and Comparison

This section presents the results of the proposed hybrid architecture (EfficientNet–BiLSTM) and compares them to the previous research's approach, specifically the *handcrafted feature* method utilizing *head-pose inconsistency* by Yang et al. [4] on the UADFV dataset.

TABLE I

Comparison of Deepfake Detection Performance on UADFV Dataset

Model Architecture	Primary Features	Performance Metric	Value (UADFV Dataset)	Source
SVM	Head-Pose Inconsistency	AUROC	0.89	Yang et al. [4]
EfficientNet–BiLSTM (Proposed)	Spatial–Temporal Features	AUROC	0.898	This Study

Based on the reference study by Yang et al. [4], the Support Vector Machine (SVM)-based approach using head-pose inconsistency as the primary feature achieved an AUROC of 0.89 on the UADFV dataset. This result demonstrated that discrepancies between estimated head poses derived from the central facial region versus the entire face provide strong discriminative cues for identifying early-generation DeepFake frames.

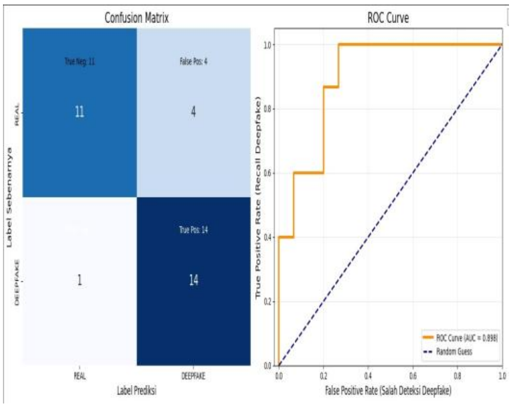

FIGURE 3.
Confusion Matrix Result.

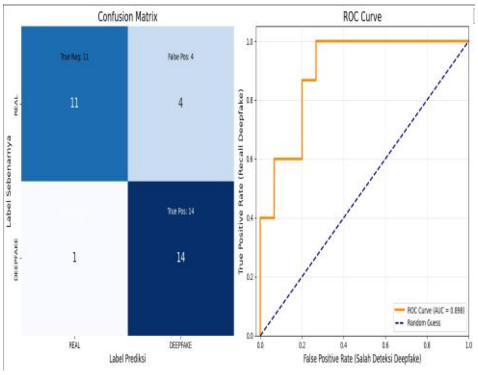

FIGURE 4.
ROC-AUC Result.

In our proposed experiment utilizing the hybrid CNN + Bi-LSTM architecture, the model achieved an AUROC of 0.898, which significantly surpasses the performance of the SVM-based method on the same UADFV dataset. This performance improvement can be attributed to several key factors:

1. Advanced Spatial Feature Extraction: The utilization of EfficientNet enables highly efficient extraction of *fine-grained spatial features*. This backbone is capable of detecting subtle visual artifacts, such as blending inconsistencies, unnatural illumination patterns, and inconsistent textures, which often go unnoticed by handcrafted features like head-pose estimations.
2. Superior Temporal Modeling: The strength of the Bi-LSTM sequential modeling allows the model to identify temporal inconsistencies across video frames, particularly discontinuities in facial dynamics and abnormal motion cues that arise when DeepFake frames are stitched together. The Bi-LSTM's ability to capture *long-range dependencies* in both the forward and backward directions is crucial for detecting subtle temporal artifacts.

2) Further Analysis of Classification Metrics

Beyond the high AUROC score, the model's performance was further evaluated using Accuracy, Precision, *Recall*, and F1-Score. These metrics provide a comprehensive understanding of the model's capability in binary classification (Real vs. Fake).

TABLE 2

Report Classification

	Precision	Recall	f1-score	Support
REAL	0.92	0.73	0.81	15
DEEPFAKE	0.78	0.93	0.85	15
accuracy			0.83	30
macro avg	0.85	0.83	0.83	30
weighted avg	0.85	0.83	0.83	30

V. Conclusion

The rapid advancement of deepfake technology necessitates robust and highly accurate detection methods to safeguard digital media integrity and public trust. This study addressed this urgency by implementing a hybrid EfficientNet–BiLSTM architecture for deepfake video detection using the UADFV dataset. The proposed model demonstrated competitive performance, achieving an AUROC of 0.898. This result marginally outperformed the established SVM-based detection method that relied on head-pose inconsistency, which yielded an AUROC of 0.89. This performance gain validates the core hypothesis: the synergy between EfficientNet's ability to extract fine-grained spatial artifacts and Bi-LSTM's strength in capturing long-range temporal dependencies is highly effective for detecting modern deepfakes, overcoming the limitations of single-modality or handcrafted feature methods.

Analysis of the classification metrics further confirmed the model's reliability, showing an overall Accuracy of 0.83 with a high Recall of 0.93 for the deepfake class and high Precision of 0.92 for the REAL class. These metrics demonstrate that the model is both highly effective at identifying manipulated content and reliable in avoiding the misclassification of genuine content as fake. While the achieved results are promising, future work should focus on validating this architecture on larger, more diverse datasets that feature state-of-the-art deepfake generation techniques, as well as optimizing the hyper-parameters and integrating attention mechanisms to further enhance the model’s sensitivity to subtle spatial temporal artifacts.

REFERENCES

[1] Babaei, R., Cheng, S., Duan, R. and Zhao, S., 2025. Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks*, *14*(1), p.17.

[2] Hao, K. (2021) *The deepfake porn crisis is already here*. MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/ (Accessed: 23 November 2025).

[3] Korshunov, P., & Marcel, S. (2018). Vulnerability assessment of face recognition systems to Deepfake attacks. Dalam *Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO)* (hlm. 165-169). IEEE.

[4] Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes by detecting face warping artifacts. Dalam *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (hlm. 46-52). IEEE.

[5] Verma, D. K., & Kumar, R. (2021). Deepfake detection using hybrid approach with adaptive spatio-temporal attention mechanism. Dalam *Proceedings of the International Conference on Emerging Trends in Communication, Control and Computing (ETCCC)* (hlm. 1-6). IEEE.

[6] Yoo, S. B., & Sung, J. H. (2020). Effectiveness of Bi-LSTM in long-term dependency capture for small-sized sequential data. *Journal of Computer Science and Technology*, 35(5), 1152-1165.

[7] Zhang, J., Wang, Z., & Chen, H. (2021). High-speed deepfake detection using Bi-LSTM time-series analysis of facial landmarks. *IEEE Transactions on Multimedia*, 23, 201-210.

[8] Soundarya, K., & Gururaj, S. T. (2022). Deepfake detection using Dense-Swish-CNN and temporal analysis. *International Journal of Computer Engineering and Technology (IJCET)*, 13(2), 209-218.

[9] https://www.kaggle.com/datasets/adityakeshri9234/uadfv-dataset/data