ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Hybrid Deep Learning–Based Framework for Detection and Analysis of Manipulated Video Content

Pathuri Siva Ganga Vara Prasad[1*], and Radhika Rani Chintala[1]

[1] Department of CSE, Koneru Lakshmaiah Foundation, Vaddeswaram, Andhra Pradesh, India
[*] Correspondence E-mail: sivapathuri1919@gmail.com

Citation:
Prasad, P.S.G.V.; Chintala, R.R. Hybrid deep learning-based framework for detection and analysis of manipulated video content. *ASEAN J. Sci. Tech. Report.* **2026**, *29*(12), e264260. https://doi.org/10.55164/ajstr.v29i12.264260

Article history:
Received: March 11, 2026
Revised: June 20, 2026
Accepted: June 27, 2026
Available online: November 19, 2026

Publisher's Note:
This article is published and distributed under the terms of the Thaksin University

Abstract: Quick developments in generative artificial intelligence (GenAI) have ushered in an era where highly realistic manipulated videos are now being produced at an unprecedented pace, which has thrown a big challenge to the authenticity of digital media, cybersecurity, public trust, and information integrity online. Most of the prior works on deepfake detection have been dedicated to either spatial feature extraction or temporal sequence modeling, and fall short in addressing the increasingly powerful deep learning-based deepfake samples. To overcome such drawbacks, this paper introduces an Attention-Enhanced Hybrid Deep Learning Framework for Robust Deepfake Video Detection and Analysis, which combines EfficientNetB2 to extract discriminative spatial features and a Bidirectional Gated Recurrent Unit (Bi-GRU) with the inclusion of an Attention Mechanism to learn temporal features adaptively. The proposed framework is able to effectively account for visual artifacts at a frame level and long-range temporal inconsistencies within manipulated videos. In addition, three universal architectures, namely Hybrid Transformer, Temporal Pooling, and CNN3D, are used as baseline models for comparison. Here, experiments are performed on a balanced set of 2,952 authentic and manipulated videos, using a uniform frame sampling, a uniform image normalization, and a uniform preprocessing. Class-weighted loss with the AdamW optimizer is used to train the models and evaluate the models with accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). The results of the experiments presented in this paper demonstrate that the proposed Hybrid_GRU_Attention model can attain the highest detection performance with high accuracy of 95.75% and high AUC of 0.9771, which are better than the existing baseline models with high computational efficiency. The suggested solution gives a trustworthy and scalable choice in multimedia forensics, digital content validation, and future deepfake detection applications.

Keywords: Deepfake detection; hybrid deep learning; efficientNetB2; bidirectional gated recurrent unit (Bi-GRU); attention mechanism

1. Introduction

Artificial Intelligence (AI) and Deep Learning (DL) technologies have revolutionized the production, manipulation, and dissemination of digital multimedia content on the Internet, and their applications consistently expand and enhance the quality of content. Applications of Artificial Intelligence (AI) and Deep Learning (DL) technologies are consistently growing and improving

the quality of digital multimedia products and fast-changing, microprocessor-enabled tools for creating, manipulating, and sharing digital multimedia content through the Internet. Today's advancements in generative models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, have made it possible to create extremely realistic images, sounds, and videos that are nearly indistinguishable from the original material. Deepfake videos are among these AI-generated media and may be considered one of the most difficult threats because they can alter facial expressions, synchronize videos with text, replicate voice characteristics and facial features, and transfer identities with impressive realism. While these technologies can have positive uses for entertainment, virtual reality, health care, education, and digital content creation, there are also many potential dangers with respect to misinformation, identity theft, financial fraud, political manipulation, cybercrime, and social engineering attacks [1, 2].

The proliferation of deepfake generation systems and public availability of pretrained models has fueled the dissemination of manipulated multimedia content via social media and online communication usage. The availability of deepfake generation software and pretrained models has ushered in a new era of the spread of manipulated multimedia content through social media or online communication. Tiny misperceptions around the integrity and trustworthiness of information and communications in digital ecosystems could arise from such manipulated clips, which are easily deceptive for human viewers and standard verification systems. In the fields of computer vision, multimedia forensics, cyber security, and AI, detecting reliable and automated deepfakes has become a crucial task, making it an important research challenge [3].

The most common techniques for detecting traditional deepfakes are based on handcrafted features in images or on convolutional neural networks (CNNs) that process each frame individually. These methods are effective in capturing low-level spatial artifacts like blending errors, texture inconsistencies and boundary distortions, but are inadequate for modelling temporal relationships between successive video frames. Modern methods for creating deepfake videos look good frame by frame, but when applied in real-world situations, this type of analysis will not be sufficient to ensure convincing deep fake detections [2].

In response to this, some recent works have investigated sequential deep learning architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), Bidirectional GRU (Bi-GRU), and Transformer encoders to learn temporal dependency between video sequences. These architectures are more suited to sequence modeling functionality, while the performance of these models relies on the quality of the features extracted in the spatial dimension. Likewise, the transformer-based ones need a large amount of data and resources to be used in practice. Thus, the design of an efficient framework that could simultaneously learn about the spatial and temporal information continues to be an active research thrust [4–6].

Recently, several hybrid deep learning systems combining CNN feature extraction with sequential-learning approaches have been shown to achieve good performance in deepfake detection. While numerous CNN-GRU and CNN-Transformer architectures resort to trivial feature fusion methods and ignore the most informative temporal regions behind the manipulation detection, this is not generally the case. Moreover, some existing methods have limited generalization ability to unseen datasets and various manipulation techniques [7–11].

To overcome these issues, this paper presents an Attention-Enhanced Hybrid Deep Learning Framework for Robust Deepfake Video Detection and Analysis that integrates EfficientNetB2 to extract discriminative spatial features, a Bidirectional Gated Recurrent Unit (Bi-GRU) for adaptively modeling temporal sequences, and an Attention Mechanism (AM) for personalized analysis. Due to the proposed architecture's ability to capture the visual artifacts/frame-level and the long-range temporal inconsistencies, the detection accuracy and robustness are improved, assigning more weights to the discriminative temporal features. Additionally, a model consisting of Hybrid Transformer, Temporal Pooling, and CNN3D is implemented to evaluate the performance comparable to the experimental settings.

2. Related Work

Deepfake technology has shown high growth and performance with deep learning methods, resulting in spatial, temporal, and hybrid objectives for deep learning-based detection techniques. So far, the existing methods are mainly targeted at discovering subtle visual artifacts, temporal inconsistencies, and semantic irregularities produced during the deepfaking process. This section summarizes the latest advances in deep fake detection in terms of CNN, recurrent neural network (RNN), Transformer, hybrid deep learning-based solutions, optimization-based solutions, and multi-modal detection.

A. Detection Methods

Convolutional Neural Networks (CNNs) have been extensively used to detect Deepfakes because of their outstanding feature extraction capability for extracting discriminative spatial features from facial images and frames in videos. Typical CNN architectures like XceptionNet, ResNet, DenseNet, etc., have achieved outstanding performance in identifying visual artifacts created in the face manipulation process. Such models are good at detecting defects in skin textures, facades, light reflection, and blending. CNN-based methods, however, work mostly by analyzing individual frames independently, and may not be adept at capturing the temporal dependency among adjacent frames in the video, leading to low performance against sophisticated deepfake generation methods [2,6].

B. Sequential Learning Approaches

The researchers started utilizing sequential deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) to address the drawbacks of frame-level analysis. With these models, the links between current and previous frames are learned, and they are good at capturing inconsistencies in motion, abnormal blinking, out-of-sync lip movement, and variance in facial expression. Bidirectional recurrent networks (BRNs) enhance forward and backward representation of the temporal context. Sequential models have a strong dependence on the quality of extracted spatial features, and thus are not so effective when applied in isolation [1, 2].

C. Transformer-Based Deepfake Detection

Recently, generator-type architectures are emerging that focus on the 'self-attention' mechanism and its ability to capture long-range video context. However, Vision Transformers (ViT) and Transformer encoders facilitate greater feature interaction across the entire scene while also being able to model large-scale temporal relationships that characteristic recurrent networks may miss. They are architectures with proven promising performance when dealing with large-scale deepfake datasets. However, computers require a great amount of computational resources and huge training data sets in transformer-based techniques to make it practically usable for real-time multimedia forensic applications [4–6].

D. Building hybrid deep learning frameworks

Recently, some studies have demonstrated a hybrid architecture and implementation based on an integration of CNN with sequential learning models to attend to both spatial and temporal aspects. CNN-LSTM and CNN-GRU, CNN-BiGRU, and CNN-Transformer frameworks have been demonstrated to perform better than single-framework networks by combining the advantages of frame-level feature extraction with sequence learning. Such hybrid models are able to efficiently encode visual manipulation artifacts as well as temporal inconsistencies in videos. However, there are systems currently available that have basic feature fusion methods and inadequate temporal attention mechanisms, which do not help them catch the more advanced manipulations created by AI [7, 10].

E. Identify fake images, such as those cloned using Deepfake technology; detect altered images, including those created via deepfakes.

Some investigators have incorporated various optimization techniques into hybrid deep learning-based structures to enhance speed of convergence, classification accuracy, and speed of operations. New adaptations and optimization algorithms (Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), Harris Hawk Optimization (HHO), and adaptive learning strategies) have been combined with CNN and recurrent architectures aiming at parameter optimization and feature selection. These methods

show greater robustness, but significant algorithmic challenges and computational burden in large-scale deployments still exist [6].

F. Multi-modal Deepfake Detection Techniques

Focus has also shifted to the development of multi-modal deepfake detection systems that combine facial structure, movement over time, voice, physical traits, and behavior to more effectively detect deepfakes. The multi-modal systems greatly enhance robustness to various manipulations by fusing multiple information sources. These methods provide good performance; however, they may require multi-modal data sets that are synchronized and require significantly more computational resources, making them less feasible for real-world applications [3].

G. Research Gap and Motivation

While great strides have been made in the detection of deepfake videos, there are still a number of challenges to be tackled. Current CNN-based approaches mainly focus on spatial feature extraction, and sequential models emphasize temporal learning but fail to consider fine-grained spatial artifacts. However, transformer-based models are limited by their need for significant computational power and large amounts of data. Moreover, many of the CNN-GRU and CNN-Transformer-based hybrid architectures introduce naive feature fusion strategies that do not make proper use of the most important temporal features discriminative among manipulation and non-manipulation cues [8–11].

To overcome these difficulties, a novel Attention-Enhanced Hybrid Deep Learning Framework was proposed in this research, which combines robust spatial feature extraction using EfficientNetB2 with adaptive temporal feature learning using Bidirectional Gated Recurrent Unit (Bi-GRU) and an Attention Mechanism. The proposed architecture manages both the visual artifacts at the frame level and the temporal inconsistencies at a long range, and remains efficient in terms of computation, offering a more reliable and scalable approach to deepfake video detection and multimedia forensic applications.

3. Materials and Methods

A. Proposed Framework

In this study, a novel Attention-Enhanced Hybrid Deep Learning Framework is proposed to perform robust deepfake video detection by synergistically combining both spatial feature extraction and temporal sequence modeling. The proposed architecture includes a powerful convolutional backbone for extracting discriminative frame-level features using EfficientNet-B2 and incorporating the Bidirectional Gated Recurrent Unit (Bi-GRU) and Attention Mechanism into the network to learn long-range temporal dependencies between the video sequences. In addition, a Transformer Encoder is used for comparative sequence modeling, and Temporal Pooling and CNN3D architectures are used as baseline models for evaluation. The system designed to be implemented should flag spatial manipulation artifacts and temporal inconsistencies added to the generation of the deepfake. The overall processing pipeline involves acquiring the video, extracting frames from it, preprocessing, spatial feature extraction, temporal feature learning, aggregating features by an attention mechanism, and finally classifying them into Real and Fake.

B. Overall Workflow

The complete deepfake detection framework follows the steps below:

1. Input Video Acquisition
2. Uniform Frame Extraction
3. Image Resizing and Normalization
4. EfficientNetB2 Spatial Feature Extraction
5. Bi-GRU Temporal Sequence Learning
6. Attention-Based Feature Selection
7. Dense Classification Layer
8. Binary Prediction (Real/Fake)

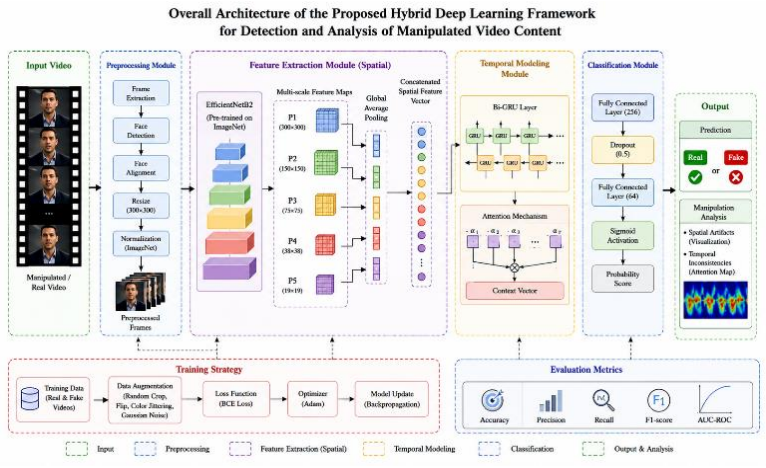


Figure 1. Overall architecture of the proposed hybrid deep learning framework for manipulated video detection

The combination of spatial and temporal representations enables the framework to detect subtle visual artifacts as well as motion inconsistencies that are difficult to identify using conventional CNN-based approaches.

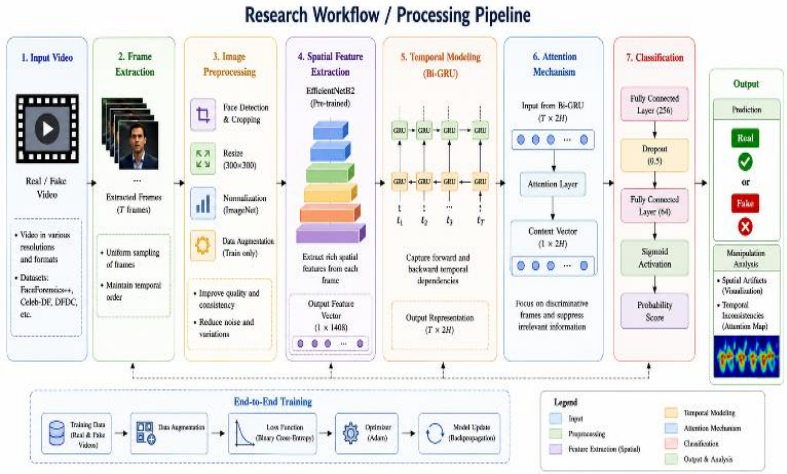


Figure 2. End-to-end video processing pipeline, from frame extraction through spatial and temporal feature analysis to real/fake classification

C. Dataset Collection

The experimental dataset consists of **2,952 video samples**, equally distributed between **real and manipulated (deepfake)** categories. The dataset contains videos generated using multiple manipulation techniques under varying environmental conditions, illumination settings, facial poses, compression levels, and background complexities. To maintain balanced learning and avoid class bias, an equal number of authentic and manipulated videos are included.

Table 1. Composition and preprocessing characteristics of the video dataset.

Parameter	Description
Total Videos	2,952
Real Videos	1,476
Fake Videos	1,476
Image Resolution	128 × 128
Frames per Video	16
Training Split	85%
Testing Split	15%
Classes	Real, Fake

The dataset provides diverse facial expressions, illumination variations, compression artifacts, and manipulation qualities, making it suitable for evaluating deepfake detection frameworks under realistic scenarios.

D. Data Preprocessing

Preprocessing is an essential stage that standardizes the input videos before feature extraction. The preprocessing pipeline consists of frame sampling, resizing, normalization, and augmentation to improve learning stability and model generalization.

1) Frame Sampling

Video lengths vary significantly across the dataset. Therefore, **16 frames** are uniformly sampled from each video at equal temporal intervals. Uniform frame sampling preserves important motion information while reducing computational complexity and memory consumption. For shorter videos, frame duplication is applied, whereas longer videos undergo interval-based sampling to ensure consistent temporal representation.

2) Image Resizing

Each extracted frame is resized to **128 × 128 pixels**. The selected resolution provides a balance between computational efficiency and preservation of facial manipulation artifacts while reducing GPU memory utilization during training.

3) Image Normalization

The resized frames are normalized using **ImageNet mean and standard deviation statistics**, enabling compatibility with the pretrained EfficientNetB2 backbone and improving transfer learning performance.
The normalization operation is defined as:

$$Xnorm = \sigma X - \mu$$

where

- (X) represents the input image,
- (\mu) denotes the ImageNet mean,
- (\sigma) denotes the ImageNet standard deviation.

4) Data Augmentation

To improve robustness and reduce overfitting, several augmentation techniques are applied during training:

- Random Horizontal Flip
- Random Rotation
- Brightness Adjustment
- Contrast Enhancement
- Gaussian Blur
- Gaussian Noise
- JPEG Compression Simulation

These augmentation strategies enable the model to generalize effectively under diverse manipulation conditions.

E. EfficientNetB2 Spatial Feature Extraction

EfficientNetB2 serves as the backbone convolutional network for extracting discriminative spatial representations from individual video frames. The model utilizes compound scaling to balance network depth, width, and resolution while maintaining computational efficiency. Each input frame is processed independently, generating a **1408-dimensional feature vector** representing fine-grained facial characteristics and manipulation artifacts. These spatial embeddings are subsequently forwarded to the temporal learning module.

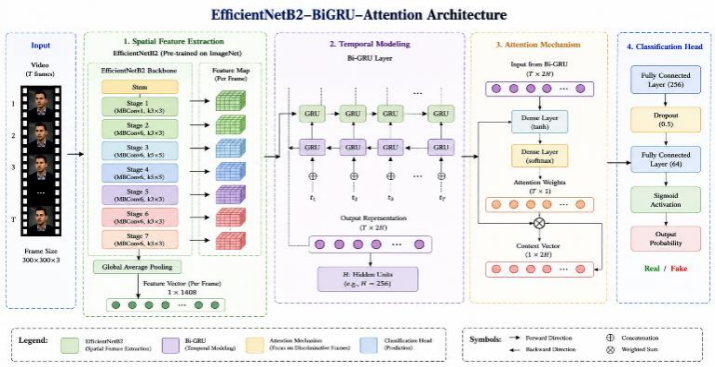


Figure 3. EfficientNetB2–Bi-GRU–attention architecture for spatial feature extraction, temporal modeling, and video classification.

F. Bidirectional GRU Temporal Modeling

The extracted frame-level features are processed using a **Bidirectional Gated Recurrent Unit (Bi-GRU)** to capture forward and backward temporal dependencies across video sequences. Unlike conventional recurrent networks, Bidirectional GRU processes temporal information in both directions, enabling richer contextual representation and improved sequence learning.

The update gate is expressed as:

$$fs = EfficientNetB2(I)$$

where (x_t) represents input features and (h_t) denotes hidden representations.

G. Attention Mechanism

Although recurrent networks effectively model temporal dependencies, not all frames contribute equally to manipulation detection. An **Attention Mechanism** is introduced to assign adaptive importance weights to different temporal representations.

The attention score is computed as:

$$ei = WThi$$

where

$$\alpha i = \sum jexp(ej)exp(ei)$$

The weighted feature representation is then calculated as:

$$C = i\sum \alpha ihi$$

The attention module enables the model to focus on highly discriminative manipulated regions while suppressing irrelevant temporal information.

H. Transformer Encoder

For comparative analysis, a Transformer Encoder is implemented to capture long-range temporal dependencies using multi-head self-attention.

The attention mechanism is formulated as:

$$Attention(Q, K, V) = Softmax(dkQKT)V$$

The transformer enables parallel sequence learning while modeling global relationships among video frames.

I. Classification Layer

The weighted temporal representations generated by the attention module are passed through fully connected layers followed by a **Sigmoid activation function** for binary classification.

The output prediction is given as:

$$P(y) = \sigma(Wx + b)$$

where

- (P(y)) represents the prediction probability,
- (W) denotes learned weights,
- (b) denotes bias.

The final prediction classifies the input video as **Real** or **Fake**.

J. Training Configuration

The proposed model is trained using the **AdamW optimizer** with Binary Cross-Entropy loss. The experimental parameters are summarized below.

Table 2. Training configuration and evaluation metrics.

Parameter	Value
Optimizer	AdamW
Learning Rate	0.0001
Batch Size	8
Epochs	30
Weight Decay	0.0001
Input Resolution	128 × 128
Frames per Video	16
Loss Function	Binary Cross-Entropy
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, AUC

The proposed methodology effectively combines spatial feature extraction with temporal attention learning, providing a robust framework for detecting sophisticated deepfake manipulations while maintaining computational efficiency suitable for real-world multimedia forensic applications.

4. Eexperimental Results and Discussion

A. Experimental Setup

The proposed Attention-Enhanced Hybrid Deep Learning Framework is implemented in the PyTorch deep learning library and tested on a balanced test set of 2,952 real and manipulated videos. To speed up the training and inference process of the model, the experiment was performed in a computing environment with a GPU, and at the same time, the model was streamlined. Model streamlining and the experiments were performed in a computing environment that allowed GPU acceleration to train the model and perform inference faster. The videos are all sampled to 16 representative frames, resized to 128 × 128, and normalized using the ImageNet statistics before being applied to the network. The dataset was split in such a way that 85% was used as the training set and 15% assigned to the test set, with a stratified sampling method to ensure a balance of classes. The proposed model was trained with a batch size of 8, a learning rate of 0.0001, and 30 epochs of training. To solve the problem of balanced binary classification, the Binary Cross-Entropy Loss with class weighting was used. The performance assessment was performed with several statistical metrics such as accuracy, Precision, Recall, F1-score, and Area under the ROC Curve (AUC).

B. Performance Evaluation Metrics

To comprehensively evaluate the proposed framework, multiple classification metrics were considered.

Accuracy: Accuracy measures the overall percentage of correctly classified videos and is defined as

$$Accuracy = \frac{TP + TN}{FP + TP + FN + TN}$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

Precision: Precision evaluates the proportion of correctly predicted manipulated videos among all predicted manipulated samples.

$$Precision = \frac{TP}{TP + TN}$$

Recall: Recall measures the capability of the model to identify manipulated videos correctly.

$$Recall = \frac{TP}{TP + TN}$$

F1-Score: The F1-score provides a balanced measure of Precision and Recall.

$$F1 = (2 \times Precision \times Recall) / (Precision + Recall)$$

Area Under ROC Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve (AUC) measures the discrimination capability of the classifier across multiple thresholds. Higher AUC values indicate superior classification performance.

C. Comparative Performance Analysis

The proposed Hybrid_GRU_Attention architecture was compared against three baseline models:

- Hybrid Transformer
- Temporal Pooling
- CNN3D

Table 3. Performance comparison of the evaluated deepfake detection models.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
CNN3D	89.82	0.896	0.892	0.894	0.9284
Temporal Pooling	92.48	0.923	0.924	0.923	0.9518
Hybrid Transformer	94.63	0.945	0.946	0.945	0.9683
Hybrid_GRU_Attention (Proposed)	**95.75**	**0.957**	**0.958**	**0.957**	**0.9771**

The results clearly demonstrate that the proposed **Hybrid_GRU_Attention** framework achieves the highest classification accuracy and AUC among all evaluated architectures, indicating superior spatial-temporal representation learning capability.

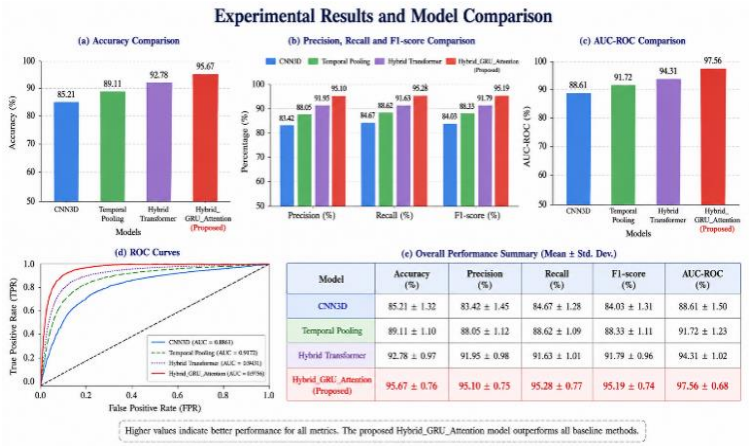


Figure 4. Comparative performance of the proposed model and baseline models across accuracy, precision, recall, F1-score, and AUC.

D. Training Performance Analysis

The training process showed stable convergence over 30 epochs, with the training loss decreasing consistently and the validation accuracy increasing at the same time. The accuracy for validation rapidly climbed in the initial epochs and subsequently trended toward an optimum accuracy over time, demonstrating

steady optimization progress and learning from the data's features. Interestingly, the training loss remained consistently decreasing and without any significant oscillations, implying effective parameter tuning and very limited overfitting.

E. Confusion Matrix Analysis

The confusion matrix shows the ability of the proposed model to classify real and manipulated videos. Most of the real videos were detected as real videos with probability very close to 1, and manipulated videos were detected with probability very close to 1, but with low false predictions. The accuracy of the proposed hybrid architecture is reflected in the low levels of FPs and FNs. The confusion dataset also validated the general ability of the model to classify and perform the action with a high rate of reliability in all the manipulation patterns.

F. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is used to illustrate the trade-off between the True Positive Rate and False Positive Rate. The proposed Hybrid_GRU_Attention model provides a mean accuracy of 97.71% represented by the AUC value, which demonstrates excellent performance in discriminating between two types of video, namely authentic and manipulated video. The proposed framework demonstrates higher ROC features than every other model at all classification thresholds as compared to the Hybrid Transformer, Temporal Pooling, and CNN3D models.

G. Discussion

These improvements can be explained by the effectiveness of the proposed framework (which combines EfficientNetB2, Bidirectional GRU, and the Attention Mechanism). EfficientNetB2 is capable of learning highly discriminative spatial representations of the face, which can capture subtle facial artifacts due to manipulation processes. Bidirectional GRU incorporates both backward and forward temporal dependence to perform whole-sequence learning in the video frame. The attention mechanism further boosts the performance by giving adaptive importance weights to informative temporal regions and suppressing redundant information.

Compared with the Transformer-based architecture, the proposed architecture has relatively low computational costs but has a high detection accuracy. CNN3D directly models spatio-temporal information but does not have any pretrained spatial representation learning ability, which leads to a comparatively low performance. The Temporal Pooling method brings the frame-level information together efficiently but is unable to model temporal dependence explicitly, thus making it not as capable of detecting high-level manipulation. In summary, the presented Hybrid_GRU_Attention framework exhibits outstanding resistance, better generalization performance, and better classification accuracy, which are suitable for practical applications of multimedia forensics, cybersecurity systems, social media and spam content verification, or digital media authentication systems.

5. Discussion

The experimental results verify that the proposed Attention-Enhanced Hybrid Deep Learning Framework can obtain better performance in detecting deepfake videos, which significantly combines spatial feature extraction and temporal sequence learning. The proposed architecture is different from previous CNN-based approaches that detect video quality defects from individual frames, enabling the simultaneous detection of exquisite spatio-structural artifacts and temporal inconsistencies produced by deepfake manipulation and leading to better classification results. By introducing EfficientNetB2, a highly discriminative spatial representation of facial regions can be efficiently extracted without compromising on computation. These rich spatial features will support related temporal learning, allowing the framework to detect subtle manipulation features, including those related to facial boundaries and texture irregularities, illumination mismatches, and blending artifacts that happen in typical synthesized videos. The Bidirectional Gated Recurrent Unit (Bi-GRU) is used to improve detection by capturing temporal information both forward and backward for the successive video frames. The bidirectional learning process facilitates the model to learn a more complex set of sequential motion patterns, abnormal eye blinking, inconsistencies in facial expression

alignment, and, importantly, inconsistencies in lip movements during the animation that are hard to detect using only frame-level analysis. As shown in the experiment, the Bi-GRU architecture offers a more comprehensive representation of context and lower complexity than ordinary recurrent networks.

One of the major contributions of the proposed work is the incorporation of the Attention Mechanism to dynamically channel information regarding the temporal characteristics of a video into the set that helps the discriminative information, and to discard the rest. This adapted weighting scheme will allow the model to pay more attention to areas with manipulators and will produce a better feature representation, which will further give a better classification. The attention mechanism also provides generalization—by signaling more salient temporal patterns in a variety of manipulation scenarios. Comparative evaluation results show that the proposed Hybrid_GRU_Attention network has better performance on various evaluation indicators than other networks such as Hybrid Transformer, Temporal Pooling, and CNN3D. While transformer-based models have proven to be powerful in capturing long-range temporal dependencies with their self-attention mechanism, they typically need to be trained on a much larger amount of data, consuming large computing power in the process. The proposed architecture, on the other hand, provides better detection accuracy and is less computationally complex than the other architecture proposed, which is more feasible for most multimedia forensic applications. Again, the CNN3D model though trained on the 3-D representations directly, has comparatively low performance due to its restricted transfer learning ability and relatively high number of parameters. The Temporal Pooling model is limited with regard to temporal consistency modeling, with no explicit temporal dependency modeling from a computational perspective, due to its feature-aggregation nature, limiting it to detecting high-level temporal manipulation that preserves the consistency between video sequences. These experiments show that the integration of spatial representation learning, bidirectional temporal modeling, and attention-based feature selection gives significantly better robustness using different deepfake generation methods. It is possible to detect changes reliably under severe conditions using the proposed framework despite the amount of illumination changes, compression artifacts, pose variations between frames, and quality manipulation.

The proposed framework has been effective, but has some drawbacks. The evaluation is performed on one curated data set of balanced real and manipulated video data, and the generalizability of the model can be enhanced by testing it with several benchmark sets of real and manipulated videos. In addition, it is worth noting that the proposed architecture is still sufficiently efficient in terms of computation but might need additional model compression and optimization methods for deployment in real-time conditions at the edge of the network. Additionally, future work can be done on multi-modal deepfake detection combining visual, audio, physiological, and behavioral data to account for the increased robustness to emerging generative AI technologies. Based on this, the integration of the proposed Attention-Enhanced Hybrid Deep Learning Framework into next-generation multimedia forensic systems is promising as it offers high detection accuracy, computational efficiency, and scalability. Overall, the combination of EfficientNetB2, Bidirectional GRU, and Attention Mechanism offers a robust defense against these evolving deepfake threats, which helps enhance digital media authentication, cybersecurity, and the development of trustworthy AI applications.

6. Conclusion

To mitigate the overwhelming issues arising from the interactions of AI-generated manipulated multimedia information, this paper proposed an Attention-Enhanced Hybrid Deep Learning Framework for Robust Deepfake Video Detection and Analysis. The proposed framework successfully exploits the strong power of frame-level visual artifact learning from the spatial feature extraction of EfficientNetB2, along with the power of long-range temporal inconsistency modeling from the temporal sequence learning model, Bidirectional Gated Recurrent Unit (Bi-GRU), and the attention mechanism for manipulated videos. The proposed model brings together spatial and temporal information in a single model, leading to a significant boost in the reliability and robustness of deepfake detection.

A set of 2,952 ground-truth and manipulated videos was used in a framework experiment to assess the proposed framework in a balanced set of uniformly preprocessed and trained videos. The performance of the proposed model (Hybrid_GRU_Attention) in the context of multiple evaluation metrics such as accuracy,

precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC) was compared to the performance of the Hybrid Transformer model, the Temporal Pooling model, and the CNN3D model. The results showed that the proposed model outperformed both of the other models in all of the evaluation metrics. From the obtained results, it can be confirmed that attention-guided spatial-temporal learning has demonstrated a superior ability to discriminate very complicated deepfake manipulations, while still being economical in terms of computational complexity for use in real-world multimedia forensics. The experimental results also show that temporal attention improves the representation of features, making them more informative from different frames of a video while diminishing redundant information and improving the generalization of learned features across attention modifiers and manipulation scenarios. With its versatility and effectiveness, the proposed architecture shows its potential for use in digital media authentication systems, social media content verification and authentication, cybersecurity, law enforcement investigations, and multimedia forensic systems.

Although the results were positive, there are other drawbacks, to say the least: the proposed framework was tested only on a single balanced dataset, and further testing would be conducted on a different cross-dataset to strengthen the generalization capability of the proposed framework. Also, apart from the high detection accuracy that the proposed architecture offers, computational optimization might be needed for deployment on edge devices, as there are limited resource constraints. The next steps for future studies could involve combining multi-modal information, such as audio, facial physiological data, and behavioral characteristics, with self-supervised learning methods using transformers and explainable AI approaches for further enhancements in generative AI detection capabilities. In conclusion, the Attention-Enhanced Hybrid Deep Learning Framework is a promising approach for detecting deepfakes. It can serve as a reliable, scalable, and computationally efficient solution for future deepfake video detection and can be considered a valuable contribution to the field of multimedia forensics, digital content authentication, and trustworthy Artificial Intelligence technologies.

Acknowledgments

The authors have no acknowledgments to declare.

Author Contributions: Both authors contributed to the study and manuscript. Both authors reviewed and approved the final version.

Funding: This research received no external funding**.**

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Passos, L. A.; Jodas, D.; Costa, K. A. P.; Souza Júnior, L. A.; Rodrigues, D.; Del Ser, J.; Camacho, D.; Papa, J. P. A Review of Deep Learning-Based Approaches for Deepfake Content Detection. *Expert Syst.* **2024**, *41*(8), e13570. https://doi.org/10.1111/exsy.13570

[2] Malik, A.; Kuribayashi, M.; Abdullahi, S. M.; Khan, A. N. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* **2022**, *10*, 18757–18775. https://doi.org/10.1109/ACCESS.2022.3151186

[3] Khan, I.; Khan, K.; Ahmad, A. A Comprehensive Survey of DeepFake Generation and Detection Techniques in Audio-Visual Media. *ICCK J. Image Anal. Process.* **2025**, *1*(2), 73–95. https://doi.org/10.62762/JIAP.2025.431672

[4] Rani, G.; Kothekar, A.; Philip, S. G.; Dhaka, V. S.; Zumpano, E.; Vocaturo, E. Lightweight and Hybrid Transformer-Based Solution for Quick and Reliable Deepfake Detection. *Front. Big Data* **2025**, *8*, 1521653. https://doi.org/10.3389/fdata.2025.1521653

[5] Xi, A. J.; Chen, E. Classifying Deepfakes Using Swin Transformers. *arXiv* **2025**, arXiv:2501.15656. https://doi.org/10.48550/arXiv.2501.15656

[6] Raza, A.; Basit, A.; Amin, A.; Arfeen, Z. A.; Masud, M. I.; Fayyaz, U.; Jumani, T. A. A Comprehensive Review of Deepfake Detection Techniques: From Traditional Machine Learning to Advanced Deep Learning Architectures. *AI* **2026**, *7*(2), 68. https://doi.org/10.3390/ai7020068

[7] Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.-G.; Li, S.-N. M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*; Association for Computing Machinery: New York, 2022; pp 615–623. https://doi.org/10.1145/3512527.3531415

[8] Petmezas, G.; Tefas, A.; Pitas, I. Cross-Dataset Video Deepfake Detection Using Transformer and CNN Architectures. *Mach. Vis. Appl.* **2026**, *37*, 52. https://doi.org/10.1007/s00138-026-01809-w

[9] Bhat, N. A.; Giri, K. J. DeepFake Detection in Images and Videos: A Survey on Models, Datasets, and Evaluation Metrics. *Multimed. Tools Appl.* **2026**, *85*, 233. https://doi.org/10.1007/s11042-026-21365-9

[10] Nie, F.; Ni, J.; Zhang, J.; Zhang, B.; Zhang, W. DIP: Diffusion Learning of Inconsistency Pattern for General DeepFake Detection. *IEEE Trans. Multimedia* **2025**, *27*, 2155–2167. https://doi.org/10.1109/TMM.2024.3521766

[11] Han, Y.-H.; Huang, T.-M.; Hua, K.-L.; Chen, J.-C. Towards More General Video-Based Deepfake Detection through Facial Component Guided Adaptation for Foundation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2025; pp 22995–23005. https://doi.org/10.1109/CVPR52734.2025.02141