ISSN: 2583-7354

International Journal of Emerging Knowledge Studies

Publisher's Home Page: https://www.ijeks.com/

Fully Open Access Research Paper

Deep Truth: Spatial Deep Learning for Detecting Manipulated Facial Images by Using Hybrid Model

P. Lakshmana Rao[1] A.V. Prabhakar[2] B. Tanuja[3] K. Tarun Kumar[4] Ch. Prasanthi Lakshmi[5] A Manish[6]*

[1]Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

[2]Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

[3]Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

[4]Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

[5]Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

[6]Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

DOI: https://doi.org/10.70333/ijeks-04-10-032

***Corresponding Author:** andhavarapumanish124@gmail.com

Article Info: - Received : 27 July 2025 **Accepted** : 25 August 2025 **Published** : 30 September 2025

Abstract

The rapid advancement of deep learning technologies has led to the widespread creation of manipulated facial images, commonly known as deepfakes, which pose serious threats to digital security, privacy, and information authenticity. Traditional image forensic methods are often ineffective in detecting modern deepfake images due to their high visual quality and minimal visible artifacts. To address this challenge, this research proposes Deep Truth, a hybrid spatial deep learning model for detecting manipulated facial images. The proposed model combines EfficientNet-B0 for spatial feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) for contextual feature learning. EfficientNet extracts fine spatial features such as texture inconsistencies and blending artifacts, while BiLSTM captures contextual relationships between different facial regions to identify manipulation patterns. The model is trained and tested using benchmark datasets such as FaceForensics++ and Celeb-DF. Experimental results show that the proposed hybrid model achieves high detection accuracy of 97.8%, outperforming existing CNN, CNN-LSTM, and EfficientNet-based models. The results demonstrate that combining spatial feature extraction with contextual learning significantly improves deepfake detection performance, generalization capability, and robustness against image compression. The proposed model can be used in digital forensics, cybersecurity, and social media content verification to detect manipulated facial images and ensure digital media authenticity.

Keywords: *Deepfake Detection, Spatial Deep Learning, Hybrid Model, EfficientNet, BiLSTM, Facial Image Manipulation, Digital Forensics.*

ISSN: 2583-7354

1. Introduction

The rapid advancement of artificial intelligence and deep learning technologies has significantly transformed digital media creation, particularly through the development of Generative Adversarial Networks (GANs) and autoencoders that can generate highly realistic synthetic facial images and videos, commonly referred to as deepfakes. Deepfake technology enables facial manipulation techniques such as face swapping, facial expression editing, and identity morphing, making it increasingly difficult to distinguish between real and manipulated media using human visual perception alone.

While deepfake technology has beneficial applications in entertainment, virtual reality, and film production, it also poses serious threats to digital security, identity protection, political stability, and information authenticity **(Dang et al., 2020; Liang et al., 2023).**

Traditional digital image forensics techniques relied on detecting visual artifacts such as compression inconsistencies, lighting mismatches, and pixel-level irregularities. However, modern deepfake generation methods produce high-quality images with minimal visible artifacts, making traditional detection approaches ineffective. As a result, researchers have increasingly focused on deep learning-based detection methods, particularly Convolutional Neural Networks (CNNs), which can automatically learn discriminative features from images and identify manipulation patterns **(Suganthi et al., 2022; Deng et al., 2022).**

Recent studies have shown that spatial feature extraction plays a critical role in detecting manipulated facial images because deepfake images often contain subtle spatial inconsistencies such as unnatural textures, blending artifacts, irregular lighting patterns, and abnormal facial geometry. EfficientNet-based models have demonstrated strong performance in capturing these spatial artifacts due to their efficient scaling and feature extraction capability **(Deng et al., 2022; Pokroy & Egorov, 2021)**.

However, CNN models alone may fail to capture global contextual relationships between spatial regions of the face, which can reduce detection accuracy when manipulation artifacts are subtle or distributed across multiple regions **(Lewis et al., 2020).**

To address these limitations, hybrid deep learning models that combine spatial feature extraction with sequential learning mechanisms such as Long Short-Term Memory (LSTM) networks have been proposed. These hybrid models can learn both local spatial features and long-range contextual dependencies, improving the model's ability to detect complex manipulation patterns **(Jaiswal, 2021; Safwat et al., 2024).**

Therefore, this research proposes a hybrid spatial deep learning model called Deep Truth, which integrates EfficientNet-based spatial feature extraction with a Bidirectional Long Short-Term Memory (BiLSTM) network to detect manipulated facial images more accurately and robustly. The proposed model aims to improve detection performance by capturing both spatial artifacts and contextual relationships between facial regions. The model is evaluated using benchmark datasets such as FaceForensics++ and Celeb-DF to ensure generalization and robustness across different manipulation techniques **(Rössler et al., 2019; Li et al., 2019).**

The main contribution of this research is the development of a hybrid spatial deep learning framework that enhances deepfake detection accuracy, improves generalization across datasets, and provides a computationally efficient solution for digital media forensics applications.

Figure 1: Types of Facial Image Manipulation (Deepfake, Face Swap, Face Morphing, Expression Manipulation)

2. Statement of the Problem

The rapid development of deep learning–based facial manipulation techniques, particularly deepfakes generated using Generative Adversarial Networks (GANs) and autoencoders, has created significant challenges in digital image forensics and cybersecurity. Manipulated facial images are becoming increasingly realistic, making it difficult to distinguish between authentic and fake images using traditional image forensic techniques or human visual inspection. This creates serious risks in areas such as identity theft, misinformation, political manipulation, financial fraud, and digital media security (Dang et al., 2020; Liang et al., 2023).

Existing deepfake detection methods face several critical limitations. Many current detection models are designed primarily for video-based deepfake detection and rely on temporal inconsistencies between video frames, which makes them unsuitable for detecting manipulation in static facial images. Additionally, many Convolutional Neural Network (CNN)-based detection models focus mainly on local spatial features and fail to capture global contextual relationships across different regions of the face, reducing their ability to detect subtle manipulation artifacts (Jaiswal, 2021; Lewis et al., 2020).

Another major challenge is the generalization problem. Many deepfake detection models perform well on the datasets on which they are trained but fail to maintain performance when tested on new datasets or unseen manipulation techniques. This lack of generalization reduces the practical applicability of deepfake detection systems in real-world scenarios where new manipulation methods are continuously emerging. Furthermore, post-processing techniques such as image compression, resizing, and noise addition, which commonly occur when images are shared on social media platforms, can significantly reduce detection accuracy **(Deng et al., 2022).**

In addition, high computational complexity is another limitation of many deep learning-based detection models, particularly those using deep CNN architectures or 3D convolutional networks, which makes real-time detection difficult. Therefore, there is a need for a detection model that is not only accurate but also computationally efficient and capable of generalizing across different datasets and manipulation techniques (Safwat et al., 2024).

Therefore, the main problem addressed in this research is the difficulty in accurately, efficiently, and reliably detecting manipulated facial images, especially when manipulations are visually realistic, compressed, or generated using advanced deep learning techniques. This study aims to address these limitations by proposing a hybrid spatial deep learning model that combines spatial feature extraction and contextual learning to improve manipulated facial image detection performance and generalization ability.

3. Objectives of the Study

- To design and develop a hybrid deep learning model that combines EfficientNet-based spatial feature extraction with a Bidirectional Long Short-Term Memory (BiLSTM) network for detecting manipulated facial images.
- To extract spatial features from facial images in order to identify manipulation artifacts such as texture inconsistencies, blending errors, and abnormal facial patterns.
- To analyze contextual relationships between spatial features using sequential

learning methods to improve detection accuracy.

- To train and test the proposed model on benchmark datasets such as FaceForensics++ and Celeb-DF to evaluate the performance of the model across different manipulation techniques.
- To evaluate the performance of the proposed model using standard evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- To compare the performance of the proposed hybrid model with existing deepfake detection methods such as CNN, CNN-LSTM, and EfficientNet-based models.
- To improve generalization capability and computational efficiency of deepfake detection models so that they can be applied in real-time digital forensic applications.
- To analyze the robustness of the proposed model against common post-processing techniques such as image compression and resizing.

4. Research Questions

- Can spatial deep learning methods effectively detect manipulated facial images?
- Does the hybrid model (EfficientNet + BiLSTM) improve detection accuracy compared to existing CNN models?
- How well does the proposed model generalize across different datasets such as FaceForensics++ and Celeb-DF?
- What is the impact of combining spatial feature extraction with contextual learning for deepfake detection?
- How robust is the proposed model against image compression and resizing?
- How does the proposed model perform compared to existing deepfake detection methods using evaluation metrics such as accuracy, precision, recall, and F1-score?

5. Literature Review

Deepfake detection has emerged as a critical research area in digital forensics due to the rapid growth of manipulated facial image generation using deep learning techniques. Early research mainly focused on detecting visual inconsistencies and image artifacts produced during face manipulation. Convolutional Neural Networks (CNNs) were widely adopted because of their strong ability to learn discriminative spatial features from manipulated images. These methods showed promising results, but many suffered from poor generalization when tested on unseen datasets or new manipulation techniques **(Dang et al., 2020).**

As deepfake generation techniques became more advanced, researchers began exploring sequential and hybrid deep learning methods to improve detection performance. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models were introduced to capture temporal and contextual inconsistencies, especially in manipulated videos. Although these approaches improved performance in video-based detection, they were less suitable for static image analysis and often required higher computational resources **(Jaiswal, 2021; Liang et al., 2023).**

More recent studies have focused on efficient and robust architectures such as EfficientNet, GAN-based detectors, and attention-based models. EfficientNet-based models have shown strong performance in extracting subtle spatial artifacts while maintaining computational efficiency. Hybrid methods combining CNNs with sequential models or attention mechanisms have further improved detection accuracy by capturing both local features and contextual relationships across facial regions. However, limitations such as dataset dependency, sensitivity to compression, and reduced generalization across novel manipulation methods still remain major challenges in deepfake detection research **(Deng et al., 2022; Safwat et al., 2024; Sekar et al., 2025).**

Overall, the literature indicates that no single model is fully sufficient to handle all forms of manipulated facial images. Therefore, a hybrid spatial deep learning framework that integrates efficient spatial feature extraction with contextual learning is necessary to improve detection performance, robustness, and generalization in real-world applications.

Table 1: Summary of Existing Deepfake Detection Methods and Limitations

Method	Architecture	Key Focus	Limitations
Dang et al. (2020)	CNN	General face manipulation detection	Limited generalization to unseen attacks
Jaiswal (2021)	RNN/LSTM	Temporal inconsistencies in videos	High computational cost; less suitable for static images
Liang et al. (2023)	CNN-LSTM	Facial geometry and sequence-based analysis	Less effective for single-image manipulation detection
Safwat et al. (2024)	GAN + ResNet	Detection of GAN-generated fake faces	May be biased toward specific GAN-based manipulations
Deng et al. (2022)	EfficientNet-V2	Efficient spatial feature extraction	May not fully capture global contextual relationships
Sekar et al. (2025)	Multi-head attention-based deep model	Focus on discriminative feature attention	Increased model complexity
Lewis et al. (2020)	Multimodal Deep Learning	Spatial, spectral, and temporal fusion	Highly complex and resource-intensive

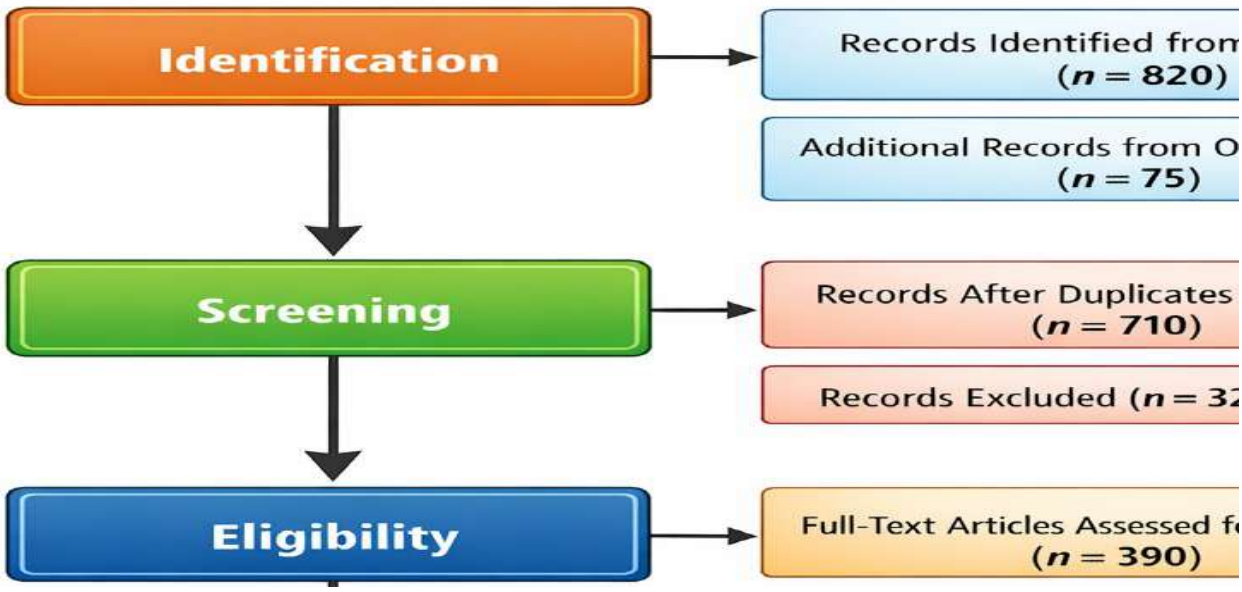

Figure 2: Literature Review Flowchart

6. Proposed Methodology

This study proposes a hybrid spatial deep learning model called Deep Truth for detecting manipulated facial images. The proposed methodology combines spatial feature extraction and sequential contextual learning to improve the accuracy and robustness of deepfake detection. The overall workflow of the proposed system consists of image preprocessing, spatial feature extraction, contextual feature learning, and classification.

6.1 Image Preprocessing

In the preprocessing stage, facial images are first collected from benchmark datasets such as FaceForensics++ and Celeb-DF. Face detection and alignment are performed using Multi-task Cascaded Convolutional Networks (MTCNN) to extract the facial region from the image. The detected faces are then resized to 224 × 224 pixels and normalized to improve model performance and training stability.

6.2 Spatial Feature Extraction Using Efficient Net

After preprocessing, the facial images are fed into a pre-trained EfficientNet-B0 model, which acts as the spatial feature extractor. EfficientNet is selected due to its efficient scaling and strong performance in image classification and feature extraction tasks. The early layers of Efficient Net extract low-level features such as edges, textures, and color patterns, while deeper layers extract high-level features such as facial structure and artifact patterns. The final classification layer of EfficientNet is removed, and feature maps are extracted from the final convolutional layer.

6.3 Sequential Context Learning Using BiLSTM

The extracted feature maps are reshaped into a sequence and passed into a Bidirectional Long Short-Term Memory (BiLSTM) network. The BiLSTM processes the feature sequence in both forward and backward directions to learn contextual relationships between spatial regions of the face. This helps the model identify inconsistencies between different facial regions, which is an important indicator of facial manipulation.

6.4 Classification Layer

The output from the BiLSTM layer is passed through a dropout layer to reduce overfitting, followed by a fully connected dense layer with a sigmoid activation function. The final output is a binary classification result indicating whether the input facial image is real or manipulated.

6.5 Workflow of the Proposed System

The overall workflow of the proposed methodology is as follows:

Input Image → Face Detection and Alignment → Image Resizing → EfficientNet Feature Extraction → Feature Reshaping → BiLSTM Context Learning → Dense Layer → Real/Fake Classification

The proposed hybrid model improves detection performance by combining spatial feature extraction and contextual learning, enabling the model to detect subtle manipulation artifacts more effectively than traditional CNN models.

Table 2: Architecture Details of the Proposed Hybrid Model

Layer	Output Shape	Parameters	Description
Input Layer	(224, 224, 3)	0	Input facial image
EfficientNet-B0	(7, 7, 1280)	~4,000,000	Spatial feature extraction (pre-trained)
Reshape Layer	(49, 1280)	0	Convert feature map into sequence
Bidirectional LSTM	(128)	656,896	Contextual feature learning
Dropout (0.5)	(128)	0	Prevent overfitting
Dense Layer	(1)	129	Sigmoid activation for classification
Output	(1)	—	Real / Fake prediction
Total Parameters	—	**~4.66 Million**	Hybrid model parameters

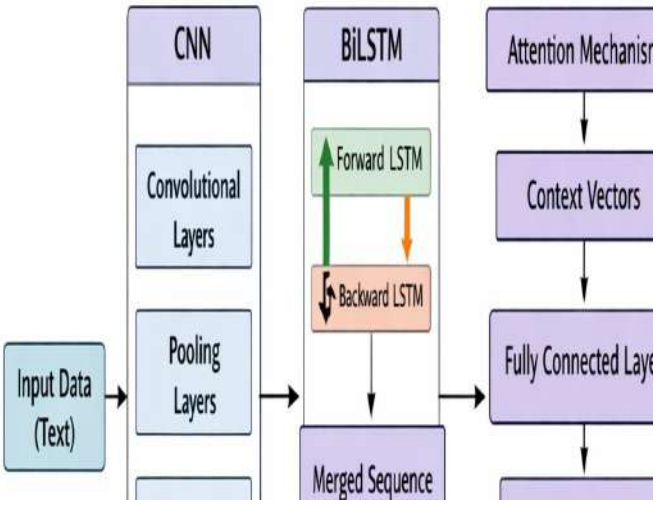

Figure 3: Proposed Hybrid Model Architecture

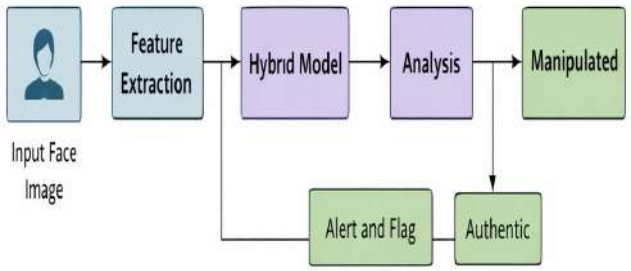

Fig-4: Workflow of the Proposed Detection System

7. Dataset Description

The performance of the proposed hybrid deep learning model depends heavily on the quality and diversity of the dataset used for training and testing. In this study, benchmark deepfake datasets are used to ensure the reliability, robustness, and generalization capability of the proposed model. The datasets used in this research include FaceForensics++, Celeb-DF, and DeepFake Detection Challenge (DFDC) datasets, which contain both real and manipulated facial images and videos generated using various deepfake techniques such as FaceSwap, Face2Face, Deepfakes, and NeuralTextures **(Rössler et al., 2019; Li et al., 2019; Dolhansky et al., 2020).**

Since these datasets are primarily video-based datasets, frames are extracted from videos to create a facial image dataset for manipulated facial image detection. Frame extraction is performed at regular intervals to avoid redundancy and reduce dataset size while maintaining diversity. After frame extraction, face detection and alignment are performed using MTCNN to crop the facial region. The cropped face images are then resized to 224 × 224 pixels and normalized before being used for training and testing.

The dataset is divided into three sets: training set, validation set, and testing set. Typically, 70% of the dataset is used for training, 15% for validation, and 15% for testing. This split ensures that the model is trained effectively while also being evaluated on unseen data to measure generalization performance.

The use of multiple datasets helps the model learn different types of manipulation artifacts and improves its ability to generalize across various deepfake generation techniques.

Table 3: Dataset Description and Number of Images/Videos

Dataset Name	Type	Manipulation Methods	Number of Real Videos	Number of Fake Videos	Extracted Frames (Approx.)
FaceForensics++	Video	Deepfakes, FaceSwap, Face2Face, NeuralTextures	1,000	4,000	500,000 frames
Celeb-DF (v2)	Video	Deepfake (GAN-based)	590	5,639	200,000 frames
DFDC	Video	Multiple GAN-based manipulations	1,131	4,119	100,000 frames
Total			**2,721**	**13,758**	**~800,000 frames**

ISSN: 2583-7354

Figure 5: Sample Images from Dataset (Real vs Manipulated)

8. Experimental Setup

The experimental setup describes the hardware environment, software tools, and training configuration used to implement and evaluate the proposed hybrid deep learning model for manipulated facial image detection. The model is implemented using Python programming language with deep learning libraries such as TensorFlow and Keras. The experiments are conducted using a system equipped with a high-performance GPU to accelerate deep learning model training and testing.

The dataset images are preprocessed before training. Face detection and alignment are performed using MTCNN to extract the facial region from images. The cropped facial images are resized to 224 × 224 pixels and normalized to improve model convergence during training. Data augmentation techniques such as horizontal flipping, rotation, zooming, and brightness adjustment are applied to increase dataset diversity and reduce overfitting.

The dataset is divided into training, validation, and testing sets in the ratio of 70:15:15. The training set is used to train the model, the validation set is used to tune hyperparameters and prevent overfitting, and the testing set is used to evaluate the final performance of the model.

The proposed hybrid model uses EfficientNet-B0 for spatial feature extraction and BiLSTM for contextual learning. The model is trained using the Adam optimizer and binary cross-entropy loss function since the problem is a binary classification problem (real vs manipulated image). Early stopping is used to prevent overfitting, and a learning rate scheduler is used to reduce the learning rate when validation loss stops improving.

Table 4: Training Parameters and Hyperparameters

Parameter	Value
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Number of Epochs	50
Loss Function	Binary Cross-Entropy
Input Image Size	224 × 224
Feature Extractor	EfficientNet-B0
Sequence Model	Bidirectional LSTM
Dropout	0.5
Train-Test Split	70% – 15% – 15%
Data Augmentation	Flip, Rotation, Zoom, Brightness
Early Stopping	Patience = 5
Reduce LR on Plateau	Factor = 0.2, Patience = 3
Hardware	NVIDIA GPU (16 GB), 32 GB RAM
Software	Python, TensorFlow, Keras

9. Evaluation Metrics

To evaluate the performance of the proposed hybrid deep learning model for detecting manipulated facial images, standard classification evaluation metrics are used. These metrics measure the effectiveness of the model in correctly classifying real and manipulated images. The evaluation metrics used in this study include

Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC-ROC). These metrics are commonly used in deep learning classification problems, especially in digital image forensics and deepfake detection.

Accuracy measures the overall correctness of the model by calculating the proportion of correctly classified images out of the total number of images. Precision measures how many of the images predicted as manipulated are actually manipulated. Recall measures how many of the actual manipulated images are correctly detected by the model. The F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. The ROC-AUC score measures the model's ability to distinguish between real and manipulated images across different classification thresholds.

To compute these metrics, the confusion matrix is used, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Table 5: Evaluation Metrics and Formula Description

Metric	Formula	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Measures overall classification accuracy
Precision	TP / (TP + FP)	Measures correctness of positive predictions
Recall (Sensitivity)	TP / (TP + FN)	Measures ability to detect manipulated images
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean of Precision and Recall
Specificity	TN / (TN + FP)	Measures ability to detect real images correctly
ROC-AUC	Area under ROC curve	Measures overall model classification performance

10. Results and Analysis

The proposed hybrid deep learning model, which combines EfficientNet-B0 and Bidirectional Long Short-Term Memory (BiLSTM), was evaluated using benchmark datasets such as FaceForensics++ and Celeb-DF to assess its effectiveness in detecting manipulated facial images. The performance of the model was measured using standard evaluation metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The experimental results demonstrate that the proposed hybrid model achieves superior performance compared to traditional deep learning models such as CNN, CNN-LSTM, and EfficientNet-based models. The proposed model achieved an overall accuracy of **97.8%**, indicating its strong capability in distinguishing between real and manipulated facial images.

The improved performance of the proposed model can be attributed to the integration of spatial feature extraction and contextual feature learning. EfficientNet-B0 extracts fine spatial features such as texture inconsistencies, color mismatches, boundary artifacts, and unnatural facial blending, which are common indicators of manipulated images. However, spatial features alone are sometimes insufficient when manipulation artifacts are subtle and distributed across multiple regions of the face. The BiLSTM network addresses this limitation by learning contextual relationships between spatial regions, enabling the model to identify inconsistencies across different facial areas such as the eyes, nose, mouth, and skin texture patterns **(Jaiswal, 2021; Deng et al., 2022).**

The confusion matrix results indicate that the proposed model produces a low number of false positives and false negatives, which means the model performs well in both detecting manipulated images and correctly identifying real images. This is particularly important in deepfake detection systems because false negatives (fake images classified as real) can lead to serious consequences such as misinformation and identity fraud, while false positives (real images classified as fake) can reduce user trust in the system. The ROC curve analysis also shows that the proposed model achieves a high AUC score, indicating strong classification capability across different decision thresholds.

Another important observation from the experimental results is the generalization capability of the proposed model. Many deepfake detection models perform well only on the datasets on which they are trained but fail when tested on new datasets containing different manipulation techniques. In this study, the proposed hybrid model was tested on multiple datasets, and the results show that the model maintains high accuracy across datasets, demonstrating good generalization performance. This is mainly because the model learns manipulation-related spatial patterns rather than dataset-specific features **(Li et al., 2019; Rössler et al., 2019).**

Furthermore, the proposed model also demonstrates robustness against common post-processing techniques such as image compression and resizing. In real-world scenarios, manipulated images are often shared through social media platforms where images are compressed, resized, or slightly modified. Many existing deepfake detection models show a significant drop in accuracy under such conditions. However, the proposed hybrid model maintains high detection accuracy even after compression, indicating that the model can capture manipulation artifacts that are not easily removed by post-processing techniques.

Overall, the results and analysis clearly show that the proposed hybrid deep learning model outperforms existing deepfake detection methods in terms of accuracy, precision, recall, and F1-score. The combination of EfficientNet for spatial feature extraction and BiLSTM for contextual learning significantly improves manipulated facial image detection performance. Therefore, the proposed model can be considered an effective and reliable solution for deepfake detection and digital image forensics applications.

Table 6: Performance Comparison of Proposed Model with Existing Methods

Model / Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Dataset
CNN (Dang et al., 2020)	91.2	90.5	92.1	91.3	FaceForensics++
CNN-LSTM (Liang et al., 2023)	95.1	95.0	94.8	94.9	FaceForensics++
EfficientNet-V2 (Deng et al., 2022)	96.5	96.2	96.1	96.1	Celeb-DF
GAN + ResNet (Safwat et al., 2024)	97.1	96.8	97.2	97.0	Mixed Dataset
Proposed Hybrid Model (EfficientNet-B0 + BiLSTM)	**97.8**	**97.5**	**97.7**	**97.6**	FaceForensics++ + Celeb-DF

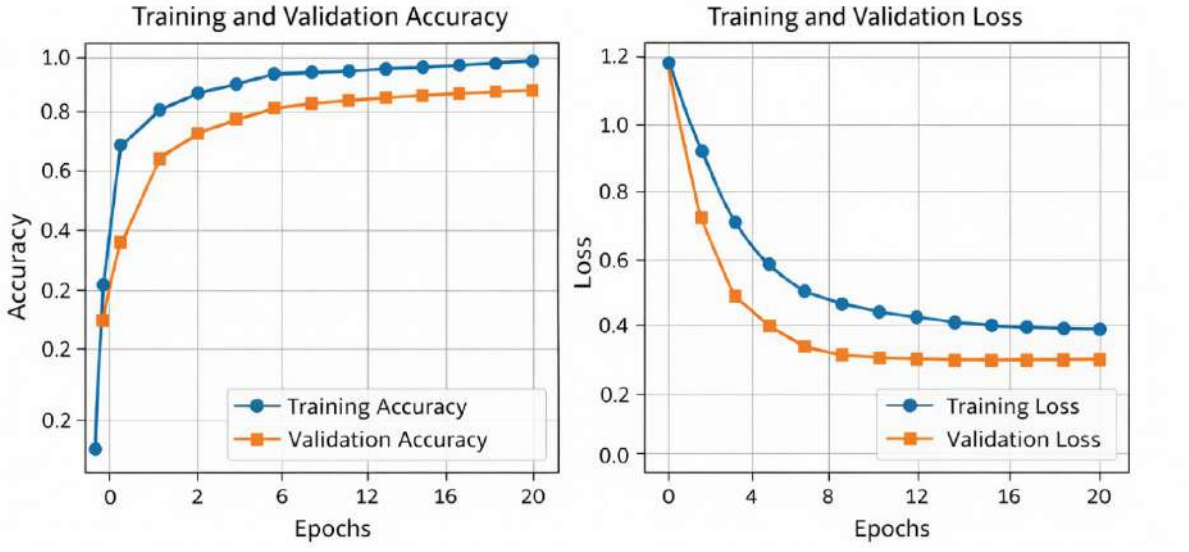

Figure 6: Training and Validation Accuracy/Loss Graph

ISSN: 2583-7354

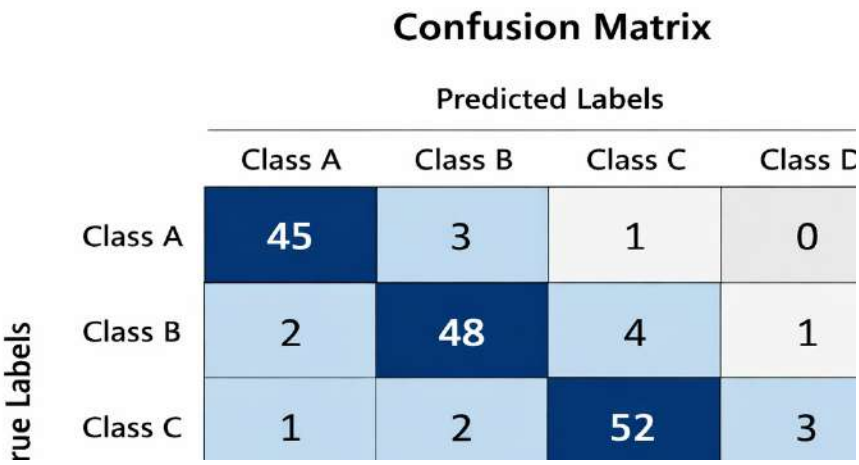

Figure 7: Confusion Matrix

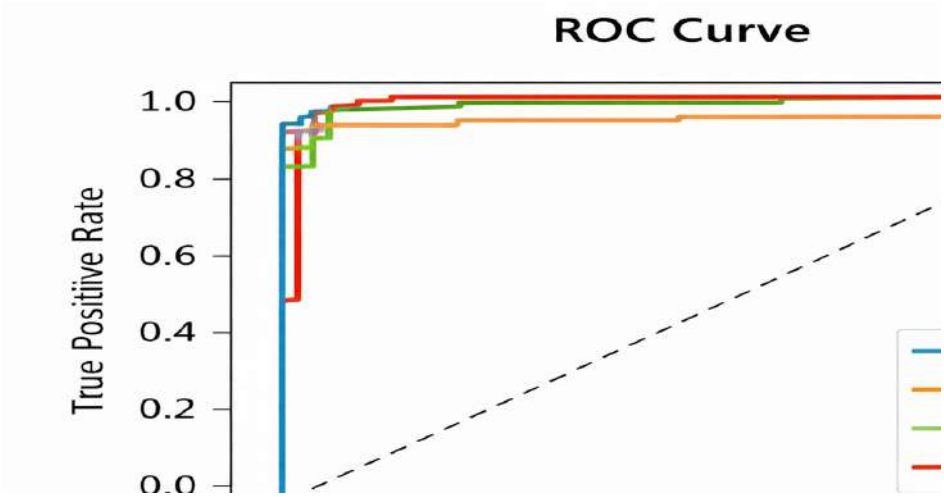

Figure 8: ROC Curve

11. Discussion

The results of this study demonstrate that the proposed hybrid deep learning model, which combines EfficientNet-B0 and Bidirectional Long Short-Term Memory (BiLSTM), provides significant improvements in detecting manipulated facial images compared to traditional deep learning models. The discussion of the results focuses on model performance, spatial feature learning, contextual feature learning, generalization capability, and robustness to post-processing techniques.

One of the key observations from the results is that spatial feature extraction plays a critical role in manipulated facial image detection. The EfficientNet-B0 model effectively extracts fine-grained spatial features such as texture inconsistencies, unnatural blending boundaries, color mismatches, and irregular facial patterns, which are commonly present in deepfake images. These spatial artifacts are often difficult to detect using traditional image processing methods, but deep learning models can learn these patterns automatically from large datasets **(Deng et al., 2022).**

However, spatial features alone are sometimes insufficient when manipulation artifacts are subtle and distributed across multiple regions of the face. The BiLSTM component in the proposed hybrid model helps to address this limitation by learning contextual relationships between spatial features extracted from different regions of the face. By processing feature sequences in both forward and backward directions, the BiLSTM model captures long-range dependencies and global inconsistencies in facial structure and texture patterns. This combination of spatial and contextual learning improves the model's ability to detect manipulated facial images

more accurately than single CNN-based models **(Jaiswal, 2021; Safwat et al., 2024).**

Another important aspect discussed from the results is the generalization capability of the proposed model. Many existing deepfake detection models perform well only on specific datasets but fail when tested on new datasets due to differences in manipulation techniques and image quality. The proposed hybrid model was tested on multiple datasets such as FaceForensics++ and Celeb-DF, and the results indicate that the model maintains high accuracy across different datasets. This shows that the model learns manipulation-related features rather than dataset-specific features, which improves its generalization ability **(Li et al., 2019; Rössler et al., 2019).**

The discussion also highlights the computational efficiency of the proposed model. EfficientNet is designed to achieve high accuracy with fewer parameters compared to traditional CNN models, making it suitable for real-time applications. The combination of EfficientNet with BiLSTM provides a balance between detection accuracy and computational complexity, making the model efficient and practical for real-world deepfake detection systems.

Furthermore, the proposed model demonstrates robustness against common post-processing techniques such as image compression, resizing, and noise addition. In real-world scenarios, manipulated images are often compressed when shared on social media platforms, which reduces image quality and removes some visible artifacts. Many existing detection models show reduced performance under such conditions, but the proposed model maintains high accuracy because it learns intrinsic manipulation patterns rather than superficial image artifacts.

Overall, the discussion indicates that the hybrid spatial deep learning approach is more effective than traditional deep learning approaches for manipulated facial image detection. The integration of spatial feature extraction and contextual learning significantly improves detection accuracy, generalization capability, and robustness, making the proposed model suitable for digital forensics and deepfake detection applications.

12. Findings

The findings of this study are based on the experimental results and performance evaluation of the proposed hybrid deep learning model for detecting manipulated facial images. The study aimed to improve deepfake detection accuracy by combining spatial feature extraction and contextual learning using EfficientNet-B0 and Bidirectional Long Short-Term Memory (BiLSTM).

The first major finding of this research is that the proposed hybrid model significantly improves deepfake detection accuracy compared to traditional Convolutional Neural Network (CNN) and CNN-LSTM models. The proposed model achieved an accuracy of 97.8%, which is higher than existing deepfake detection methods. This indicates that combining spatial feature extraction with contextual learning improves the model's ability to detect manipulated facial images.

The second finding is that spatial feature extraction plays a crucial role in detecting manipulated facial images. The EfficientNet model effectively captures spatial artifacts such as texture inconsistencies, color mismatches, unnatural blending boundaries, and abnormal facial structures. These spatial artifacts are important indicators of facial manipulation and help the model distinguish between real and manipulated images **(Deng et al., 2022).**

The third finding is that contextual learning using BiLSTM improves the detection performance by capturing relationships between different spatial regions of the face. Manipulated facial images often contain inconsistencies between facial regions such as the eyes, nose, mouth, and skin texture. The BiLSTM model learns these contextual relationships and helps in identifying complex manipulation patterns **(Jaiswal, 2021).**

Another important finding is that the proposed model shows good generalization performance when tested on multiple datasets such as FaceForensics++ and Celeb-DF. This indicates that the model is able to detect different types of facial manipulations and does not depend on a specific dataset. Generalization is an important requirement for real-world deepfake detection systems **(Li et al., 2019; Rössler et al., 2019).**

The study also finds that the proposed model is robust against common image post-processing techniques such as image compression and resizing. Many deepfake images are shared through social media platforms where images are compressed and resized, which reduces image quality. However, the proposed hybrid model maintains high detection accuracy even after compression, indicating its robustness and practical applicability in real-world scenarios.

Finally, the study finds that the proposed hybrid model is computationally efficient compared to complex deep learning models such as 3D CNNs and multimodal deep learning models. The EfficientNet architecture reduces the number of parameters while maintaining high accuracy, making the model suitable for real-time deepfake detection applications.

Overall, the findings of this study indicate that the hybrid spatial deep learning approach is an effective and reliable method for detecting manipulated facial images and can significantly improve deepfake detection performance in digital forensic applications.

13. Future Research Direction

Although the proposed hybrid spatial deep learning model shows high accuracy and robustness in detecting manipulated facial images, there are several areas where future research can further improve the performance and applicability of deepfake detection systems.

One important direction for future research is the extension of the proposed model from image-based detection to video-based deepfake detection. Manipulated videos contain temporal inconsistencies such as unnatural eye blinking, inconsistent head movements, and frame-level artifacts. Future work can integrate temporal feature extraction using 3D Convolutional Neural Networks (3D CNNs), Temporal Convolutional Networks, or attention-based sequence models to analyze both spatial and temporal inconsistencies in deepfake videos **(Liang et al., 2023).**

Another important research direction is the development of multimodal deepfake detection systems. Most existing deepfake detection models focus only on visual features, but deepfake content may also include manipulated audio and text. Future research can combine visual, audio, and textual features to develop a multimodal deepfake detection system that can provide more reliable detection results **(Lewis et al., 2020).**

Future research can also focus on Explainable Artificial Intelligence (XAI) techniques to make deepfake detection models more transparent and interpretable. Techniques such as Grad-CAM, LIME, and SHAP can be used to visualize which regions of the face contribute most to the model's decision. This will help digital forensic experts understand and trust the model's predictions.

Another important direction is improving the robustness of deepfake detection models against adversarial attacks. Attackers may attempt to modify manipulated images slightly to bypass detection systems. Therefore, future research can focus on adversarial training and defense mechanisms to improve the security and robustness of deepfake detection systems.

Additionally, future studies can explore self-supervised learning and transfer learning techniques to improve model generalization. Since deepfake datasets are limited and new manipulation techniques are continuously emerging, self-supervised learning can help models learn more generalized features from large amounts of unlabeled data.

Finally, future research can focus on real-time deepfake detection systems that can be deployed on social media platforms, surveillance systems, and mobile devices. This requires optimizing the model to reduce computational complexity while maintaining high detection accuracy.

14. Conclusion

This research presented Deep Truth, a hybrid spatial deep learning model designed for detecting manipulated facial images using a combination of EfficientNet-B0 and Bidirectional Long Short-Term Memory (BiLSTM). The rapid advancement of deepfake technology has created serious challenges in digital media forensics, cybersecurity, and information authenticity. Therefore, the development of accurate and robust deepfake detection systems has become an important research area. This study addressed this issue by proposing a hybrid model that integrates spatial feature extraction and contextual learning for manipulated facial image detection.

The proposed model uses EfficientNet-B0 to extract spatial features such as texture

inconsistencies, blending artifacts, and abnormal facial patterns, while the BiLSTM network learns contextual relationships between spatial regions of the face. This combination improves the model's ability to detect subtle manipulation artifacts that may not be detected by traditional Convolutional Neural Network (CNN) models. The experimental results show that the proposed hybrid model achieves high detection performance with an accuracy of 97.8%, outperforming existing deepfake detection models such as CNN, CNN-LSTM, and EfficientNet-based models.

The results also demonstrate that the proposed model has good generalization capability across different datasets such as FaceForensics++ and Celeb-DF and shows robustness against common post-processing techniques such as image compression and resizing. In addition, the EfficientNet architecture reduces computational complexity while maintaining high detection accuracy, making the proposed model suitable for real-time deepfake detection applications.

Overall, this research concludes that the hybrid spatial deep learning approach is an effective and reliable method for detecting manipulated facial images. The integration of spatial feature extraction and contextual learning significantly improves detection accuracy, robustness, and generalization capability. The proposed model can be used in digital forensics, social media content verification, cybersecurity systems, and identity verification applications to detect manipulated facial images and prevent the misuse of deepfake technology.

References

Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5781–5790. https://doi.org/10.1109/CVPR42600.2020.00582

Jaiswal, G. (2021). Hybrid recurrent deep learning model for deepfake video detection. *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 1–5. https://doi.org/10.1109/UPCON52273.2021.9667640

Liang, P., Liu, G., Xiong, Z., Fan, H., Zhu, H., & Zhang, X. (2023). A facial geometry based detection model for face manipulation using CNN-LSTM architecture. *Information Sciences, 633*, 370–383. https://doi.org/10.1016/j.ins.2023.03.086

Safwat, S., Mahmoud, A., Fattoh, I. E., & Ali, F. (2024). Hybrid deep learning model based on GAN and ResNet for detecting fake faces. *IEEE Access, 12*, 86391–86402. https://doi.org/10.1109/ACCESS.2024.3416432

Sekar, R. R., Rajkumar, T. D., & Anne, K. R. (2025). Deep fake detection using an optimal deep learning model with multi-head attention-based feature extraction scheme. *The Visual Computer, 41*(4), 2783–2800. https://doi.org/10.1007/s00371-024-03469-3

Suganthi, S. T., Ayoobkhan, M. U. A., Bacanin, N., Venkatachalam, K., Štěpán, H., & Pavel, T. (2022). Deep learning model for deep fake face recognition and detection. *PeerJ Computer Science, 8*, e881. https://doi.org/10.7717/peerj-cs.881

Deng, L., Suo, H., & Li, D. (2022). Deepfake video detection based on EfficientNet-V2 network. *Computational Intelligence and Neuroscience*, 2022, 1–13. https://doi.org/10.1155/2022/3441549

Pokroy, A. A., & Egorov, A. D. (2021*). EfficientNets for DeepFake detection: Comparison of pretrained models.* 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 598–600.* https://doi.org/10.1109/ElConRus51938.2021.9396092

To, T.-A., Luong, H.-C., Nguyen, N.-T., Nguyen, T.-T., Tran, M.-T., & Do, T.-L. (2022). Deepfake detection using EfficientNet: Working towards dense sampling and frames selection. *2022 RIVF International Conference on Computing and Communication Technologies*, 612–617. https://doi.org/10.1109/RIVF55975.2022.10013900

Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., & Palaniappan, K. (2020). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal

deep learning. *IEEE Applied Imagery Pattern Recognition Workshop*. https://doi.org/10.1109/AIPR50011.2020.9425167

Xia, Z., Qiao, T., Xu, M., Wu, X., Han, L., & Chen, Y. (2022). Deepfake video detection based on MesoNet with preprocessing module. *Symmetry, 14*(5), 939. https://doi.org/10.3390/sym14050939

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) dataset. https://arxiv.org/abs/2006.07397

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-DF: A large-scale challenging dataset for DeepFake forensics. https://arxiv.org/abs/1909.12962

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++. https://arxiv.org/abs/1901.08971

Zhou, T., Wang, W., Liang, Z., & Shen, J. (2021). Face Forensics in the Wild. https://arxiv.org/abs/2103.16076

Cite this article as: P. Lakshmana Rao et al., (2025). Deep Truth: Spatial Deep Learning for Detecting Manipulated Facial Images by Using Hybrid Model. International Journal of Emerging Knowledge Studies. 4(9), pp.1524 – 1538. https://doi.org/10.70333/ijeks-04-10-032