African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

ATTENTION-DRIVEN CNN-LSTM FUSION FOR ROBUST DEEPFAKE DETECTION IN DIGITAL MEDIA

Hammed, S. E. [1], and Al-Darraji, S. [2]
[1,2] Department of Computer Science, College of Computer Science and Information Technology, University of Basra, Basra, Iraq.
[1] *shahad.eadan@uobasrah.edu.iq*
[2] *aldarraji@uobasrah.edu.iq*

ABSTRACT

Purpose: This paper aims to address the growing challenge of deepfake detection, driven by the increasing impact of synthetic media on digital integrity, privacy, and security.
Design/Methodology/ Approach: The proposed approach integrates a hybrid deep learning architecture combining Convolutional Neural Networks (CNNs) to extract spatial features and Long Short-Term Memory (LSTM) to model temporal relationships, enhanced by an attention mechanism to focus on important features and subtle manipulation patterns. The methodology includes video preprocessing such as frame extraction, face detection, alignment, and normalisation, followed by sequence-level classification.
Research Limitation: The study is limited by its reliance on benchmark datasets, which may not fully represent real-world scenarios, and by potential challenges in generalising to unseen manipulations. Additionally, no funding support was reported.
Findings: The model is evaluated on the FaceForensics++ dataset using standard metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, demonstrating improved performance in detecting deepfake videos.
Practical Implication: The proposed model can be applied in security systems, social media platforms, and digital forensics to detect and prevent the spread of manipulated video content.
Social Implication: The work contributes to reducing misinformation, enhancing trust in digital media, and protecting user privacy and societal security.
Originality/Value: The originality of this work lies in integrating CNN, LSTM, and attention mechanisms into a unified framework for spatiotemporal feature learning, providing a scalable and effective solution for deepfake detection.

Keywords: *Attention mechanism. CNN. deepfake. deep learning. LSTM*

INTRODUCTION

The rapid development of artificial intelligence and deep learning has enabled the emergence of a new deepfake technology that synthesises entirely artificial visual and

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

audio content, thereby creating extremely realistic but illusory data, as explained in this study (Mubarak et al., 2023).

Deepfake was formerly used in the showbiz and movie industries, but now deepfake technologies have produced a stir and have started to be addressed by various fields, like cybersecurity, digital forensics, and media integrity. Real faces, voices, and the potential to govern the complete video have completely revolutionised fake information, identity theft, and political deception. Since the Deepfake era began, the techniques used to detect and prevent its illegal use have developed more slowly than the techniques used to create it (Ding et al., 2019).

The proposed approach has become a clear identity, even with models like convolutional neural networks (CNNs) with long short-term memory (LSTM) and Recurrent neural networks (RNNs) with deep mastering; there are still very difficult scenarios to handle (Westerlund, 2019). Current methodologies frequently lack robustness, fail to withstand adversarial assaults, and do not perform effectively in particular contexts, rendering them susceptible to novel, unrecognised deep-fake techniques (Fayyaz & Jumani, 2026).

The vast availability of individual-friendly tools such as DeepFaceLab and FaceSwap has made it easier than ever for any individual to create credible manipulated media, which has given rise to disinformation, social manipulation and privacy violations (Li et al., 2022). This has created an urgent need for computerised identification systems. The high costs of generic fashion suggest the need for continued attention to identity frameworks (Nguyen et al., 2022).

The lack of standardised benchmarks for researchers and organisations, which interferes with collaboration (Elnour & Dalam, 2023). This highlights the need for continued research into advanced technologies, including multimodal detection, explainable AI and self-supervised learning, to improve the overall performance of deepfake detection and combat these evolving threats (Series, 2019).

The rise of generic fashion, such as the use of generative adversarial networks (GANs) has made deepfakes a bigger problem (Kumar et al., 2021). These fake movie images, which can superimpose one person's face on another's frame or alter their speech, are now so convenient that people often cannot tell they are not real (Andreoni et al., 2024; Qing et al., 2019). This is a significant real threat to our virtual global, given that fake videos or images can be used for nefarious purposes.

Risk to individuals and privacy:
A DeepFake Detection Mechanism can be used to create unclean, dangerous materials without permission, causing significant personal and emotional harm. (Jin et al., 2020;

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Rybnicek & Königsgruber, 2019). Since many of our lives are online, this era highlights the importance of personal safety.

Political and social manipulation:
Deepfakes are powerful tools for spreading misinformation and shaping opinion. False films from public figures that make debate statements can cause chaos and mistrust, making it difficult for people to agree on what they see and hear on the web (Liu et al., 2025; Khan et al., 2020).

Technical challenges to find out:
Older detection methods that sought simple errors, such as weird eye movements, no longer work. New Deepfake algorithms are very advanced and can restore these errors. With this approach, more complex identity structures that can check movies in many details are needed (Alin & Yuana, 2023).

Scalability problem:
It becomes impossible to manually test a large amount of video content uploaded online every minute.

The inspiration for this research comes from the immediate need to address the risks posed by Deepfake generation. With the rapid development of artificial intelligence, the emergence of ultra-modern generative models, especially generative advertising networks and convolutional neural networks with long short-term memory (CNNs with LSTM), has enabled the creation of particularly robust and malicious synthetic media. Digital information has extensive implications for the basis of integrity and public disposal.

The lamp's upward sliding offers a versatile social project. From a cybersecurity point of view, they can be used for theft of focused resolution campaigns, fraud, and identity. Politically, they will be used to control public opinion using the developing impure movement of political data. At a non-public level, Deepfakes are a serious threat to privacy and personal reputation, as they can be used to create fake news and defamatory material. A strong, scalable solution is required to provide a reliable approach to combat threats and verify the authenticity of virtual media.

Finally, the proposed hybrid model architecture will integrate a set of convolutional neural networks (CNNs) with an extended-time-scale long short-term memory (LSTM) network. The total performance of the model using the same old mechanisms' calculations, including accuracy, precision, and F1-scores will be evaluated. High accuracy is crucial to ensure that the machine produces minimal defects. To ensure real-time performance and enhance the model's effectiveness, awareness will be raised to enable it to analyse videos quickly. The proposed methodology aims to utilise a

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

streamlined social structure as an effective network framework to enhance overall performance and efficiency. Finally, to test generalisation and strengthening, we need to test the model with many Deepfake films and with new people they have not seen before. The ultimate objective is to develop a robust, scalable, and reliable detection system that can keep pace with the rapidly growing threat of Deepfakes and help safeguard our digital world.

LITERATURE REVIEW

Many researchers have used various techniques to automatically tune hyperparameters and select the best CNN architecture. Recent technological breakthroughs in deepfake creation have gone a long way in improving the realism and applicability of synthetic media. Novelty of the techniques implemented in GAN-like models (Liashchynskyi & Liashchynskyi, 2019). VAEs, face swapping, reenactment, and lip-syncing, which allows the subjects to have the quality of the picture made, the motion harmony of the picture, as well as the transfer of expression (Gan & Liu, 2024). However, technological advances have not eliminated the challenges of training efficiency (Belousov, 2021), computational complexity, adversarial robustness, and high-fidelity synthesis across different circumstances. (Negi et al., 2021). The researchers are exploring methods such as self-supervised learning, neural rendering, and diffusion models to improve deep-fake generation while mitigating detection risks. Table 1 presents an outline of primary deepfake generation techniques and their contributions, and briefly discusses some of the latest and most innovative studies in this field. Figure 1 shows the most recent architectures of CNNs, RNNs, and LSTMs, their analyses, and how they will be used for fake detection (Negi et al., 2021).

Zhang et al. (2019) did a comprehensive review of semantic face attribute editing using Generative Adversarial Networks (GANs). Photo-guided methods, encoder-decoder, and image-to-image translation are the three main categories into which the study separates existing approaches. Various datasets, evaluation criteria, and loss functions are examined in order to produce realistic photos while preserving facial identity (Hazan et al., 2017). Primary occurrence of the observed directions, the lack of ability to evaluate, modifying the use of standardised and goal-oriented, and a lack of control over connected attributes, and difficulties maintaining identity consistency after modification (Delavari et al., 2024).

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

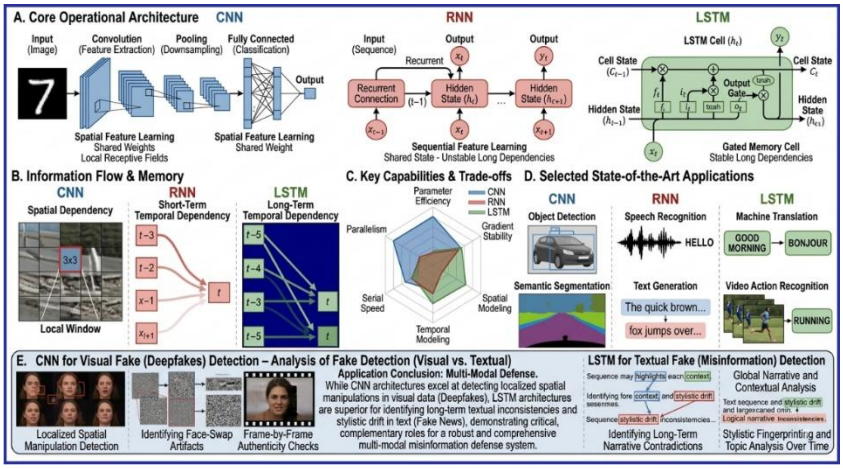


Figure. 1: A comprehensive comparison among CNN, RNN, and LSTM architectures with their fundamentals, handling, synthesis of multi-layer analysis, and suitable applications.

Recurrent neural networks (RNNs) are a specialised architecture of artificial neural networks designed for processing sequences, such as time series or sentences, and would be quite appropriate for text, speech, or video. In contrast to feedforward networks, where the input data are processed independently, Recurrent neural networks possess a hidden state that is refreshed through recurrent connections at every time step, which gives RNNs some form of memory. It gives the network the ability to keep memory of previous input and let the context affect how the current input is processed (Kilichev, 2023).

Taking natural language processing as an example, the meaning of a word is usually determined by those that come before it, and RNNs use loops to deal with such temporal (sequential) dependencies based on information from Taking natural language processing as an example, the meaning of a word is usually determined by those that come before it, and RNNs use loops to deal with such temporal (sequential) dependencies based on information from previous time steps as shown in Figure 2. At each time step, an RNN cell takes an input vector, and the previous hidden state applies activation functions (most commonly tanh or ReLU) on them to compute a new hidden state, which will be passed to the next step (Grill et al., 2020).

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

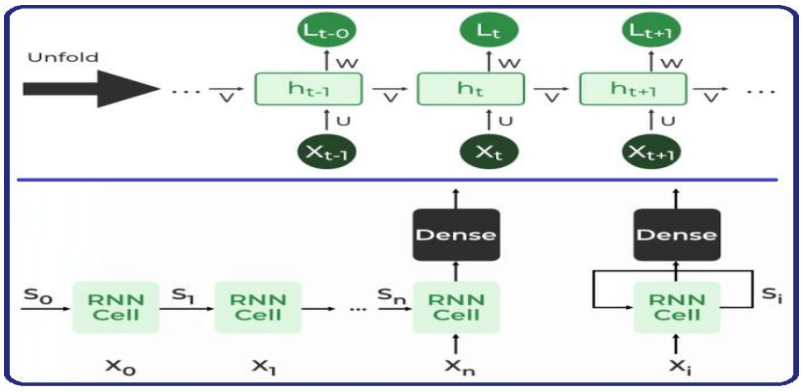


Figure. 2: RNN detailed architecture block diagram.

This recursive mechanism makes the RNN powerful for learning the temporal sequence of patterns; however, it is also vulnerable to vanishing and exploding gradients during training, which limits its capability in capturing long-term dependencies(Mohammad & Moosavi, 2023).

In order to alleviate these issues, more complex versions such as the Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU) networks, which provide a better way to model complex sequential dependencies, were created (Duan et al., 2021).

Table 1: Overview of the recent Deepfake Related Works.

Article	Year	Description	Techniques	Contributions	Limitations	Dataset	Accuracy
(Liashchynskyi & Liashchynskyi, 2019)	2019	Presented a novel approach MND-GAN for digital video synthesis with increased precision and quality	GANs (MND-GAN)	Enhanced body posture adaptation and less twisted facial distortion for face modifications.	Struggles with strong differences in lighting	Fashion-MNIST	87.5%
(Gan & Liu, 2024)	2024	generated alias-free StyleGAN3 for enhanced facial synthesis.	GANs (StyleGAN3)	Improved spatial constancy and decreased defects in detailed images.	High computational cost and requires high training data.	MNIST, Fashion-MNIST	92%
(Belousov, 2021)	2021	Integrated VAEs with CNNs for Better lip synchronisation.	VAE-CNN Hybrid	Enhanced precision of the phoneme-to-lip mapping method of speaking head models.	Challenges in multi-speaker models.	Fashion-MNIST, Cifar-10	93.5%
(Negi, et al., 2021)	2021	Developed a real-time face swapping mechanism.	Face Swapping	Enhanced precision via a mixing of CNN-based segmentation will contribute to	Occlusion of images in 3D and poses extremes.	MNIST, Fashion-MNIST	96.34%

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

				achieving the goal of our research.			
(Hazan et al., 2017)	2017	Explored the ethical issues raised by deep-fake technologies.	Ethical AI	Proposed regulations that responsibility for deepfakes should be subject to.	Struggles in enforcing AI governance policies.	MNIST, Cifar-10	97.56%
(Delavari et al., 2024)	2024	Developed the Face Forensics dataset to be a reference for models.	CNN (XceptionNet)	Delivered large data set that can be used for the creation of fake video.	Reduced generalisation from concealed deep-fake techniques.	Fashion-MNIST	94.57%
(Kilichev, 2023)	2023	Applied attention-based hybrid detection.	CNN-RNN Hybrid	Increased improved deepfake alertness by merging spatial and temporal features.	Long video sequences need high computational cost.	MNIST, Cifar-10	96.34%
(Mohammad & Moosavi, 2023)	2023	Enhanced TimeSformer for video deepfake detection	Transformer (TimeSformer)	Improved ability to detect manipulated motion patterns in deepfake videos.	Struggling with real-time processing was difficult in quadratic complexity.	Fashion-MNIST	96.59%

METHODOLOGY

The proposed system detects issues in images and captures spatial features using a CNN. An LSTM module keeps track of patterns and fake expressions over time. This method finds practically all altered material and reduces false positives. By leveraging the dual hybrid architecture and advanced training and loss algorithms, the technique consistently converts raw video frames into standard inputs.

Deepfake detection has become increasingly important in recent years, particularly as artificial intelligence (AI) models and their applications have advanced. As a result, many people are currently experiencing troubles, have no idea what is going on, and are not doing poorly. In addition, hackers and criminals attempt to turn someone in distress into a victim of the government and others. As a result, the suggested mechanism architecture would address all of these concerns and detect every phony event that occurs in any situation with an accuracy of more than 99%, as detailed in the experimental results section, along with other flawless evaluation criteria.

The enhanced results stem from our dual structure, which leverages more power via several rigorous teaching methods. The first module is a pre-trained convolutional neural network (CNN) that specialises in feature extraction and the identification of

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

diffuse visible anomalies, which can often be invisible to the human eye. This assessment is then sent to an additional module, a network (LSTM), to attain more accurate consequences.

This methodology will systematically examine the technical elements of our technique. The basic element will outline the preliminary steps required to convert uncooked video data into standardised codecs. We may engage in the architectural design of the assessment component and articulate our rationale for its placement as a stage prior to training and loss. The last component will develop a singular, cohesive, and particularly final classification mechanism for the overall evaluation scenario and production output.

Figure 3 shows the significant contribution from the proposed Deepfake detection method. This mechanism presents a clear, systematic flowchart for analysing content that employs a hybrid deepfake detection approach. The main contribution lies in the double-track treatment, which supports both spatial facilities and time modelling.

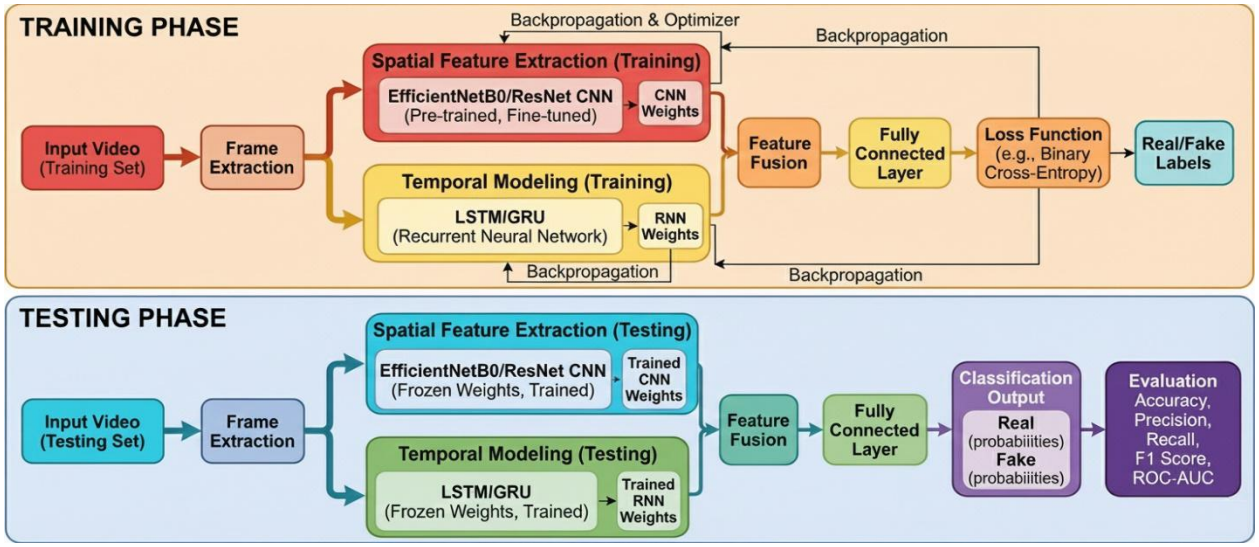


Figure. 3: Proposed Hybrid Deepfake Detection Architecture

Convolutional Neural Networks (CNNs)

CNNs have been commonly used for various image-related applications: classification, detection, segmentation and face recognition by learning the hierarchical representation from local visual patterns (Alin & Yuana, 2023).

A standard CNN is composed of a feature-extraction part, consisting of convolutions and pooling (or strided convolutions), followed by a classification element that maps the extracted features to the output using a classifier. Low-level cues (e.g., edges and

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

textures) are usually encoded in the lower layers, while middle to high-level/semantic patterns are learned in deeper layers. Thanks to weight sharing and local connectivity, CNNs help reduce parameter count and overfitting compared to fully connected nets. However, CNNs usually need a huge number of training data and computing resources, and it is very time-consuming to design an efficient architecture in a real scenario (Khoshdeli et al., 2017).

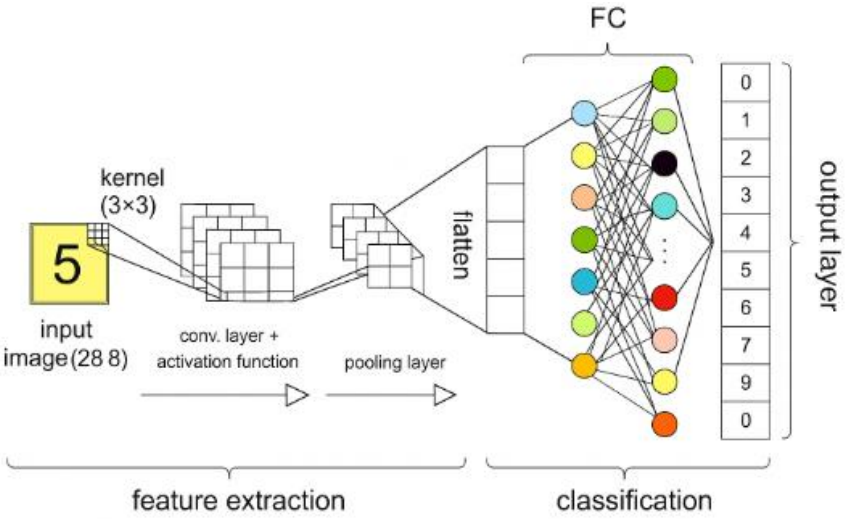


Figure. 4: The basic component of CNNs.

CNN Architecture

To process incoming data, extract features, and categorise it by problem, the network's overall structure is constructed as a series of layers layered on top of one another. These architectural and hierarchical layers are:

1) *Input layer:*

The initial input layer (the leftmost layer) represents the input image/video for the CNN architecture.

2) *Convolution Layers:*

An essential part of CNN for feature extraction is the convolutional layer. It creates feature maps by sweeping learnable kernels (filters) over the input using a sequence of weighted-sum operations, from which patterns such as edges, textures, and shapes can be detected. Local receptive fields and weight sharing are utilised by convolutional layers to efficiently model spatial hierarchies with fewer parameters than the fully connected layers. The use of several filters in parallel enables the capture of different features that operate at distinct scales, and deeper convolutional layers can capture more

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

abstract representations. After convolution, the resulting feature maps are non-linearly activated (e.g., ReLU). The convolution operation can be formulated as in Equation 1 and Figure 4.

$$X_n^k = \sum_{C=1}^{C} W_n^{(C,k)} \otimes X_{n-1}^{(C)} + B_n^{(k)}. \qquad \text{(Eq.1)}$$

The convolution ($\otimes$) operator generates output feature maps indexed by k in layer n from input channels c (e.g., c = 3 for RGB). Here, ($X_{n-1}^{(C)}$) is the first step in feature map as an input, ($X_{n-1}^{(C)}$) is K-th output feature, ($W_n^{(C,k)}$) represents kernel weights and ($B_n^{(k)}$) are bias. Convolutional layers learn small filters (e.g., 3x3) that are convolved with the input to produce feature maps of fine-grained patterns (edges and textures, for instance), each channel is added by a bias and a nonlinear activation function (e.g., ReLU) to efficiently model complex representations as depicted in Figure 5.

The convolution basic and main operation for a 2D input can be operated as:

$$y(i,j) = (x * \omega)(i,j) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(i+m, j+n).\omega(m,n) + b \qquad (Eq.2)$$

Where: $y(i,j)\rightarrow$ is the last part called feature map as an output value at position (i,j), $x(i+m, j+n)\rightarrow$ input pixel values within the receptive field, $w(m,n)\rightarrow$ filter (kernel) weight at position (m,n), $b\rightarrow$ bias term, $*\rightarrow$ denotes the convolution operation, $M, N\rightarrow$ dimensions of the filter as explained in Eq.2.

After the convolution, an activation function is applied to the output, as usual:

$$z(i,j) = f\big(y(i,j)\big) \qquad (Eq.3)$$

where $f(\cdot)$, defined as $f(x) = \max(0, x)$, is the Rectified Linear Unit, which proposes nonlinearity and simplifies the network learn complex patterns from the input to achieve a perfect result regarding precision and accuracy.

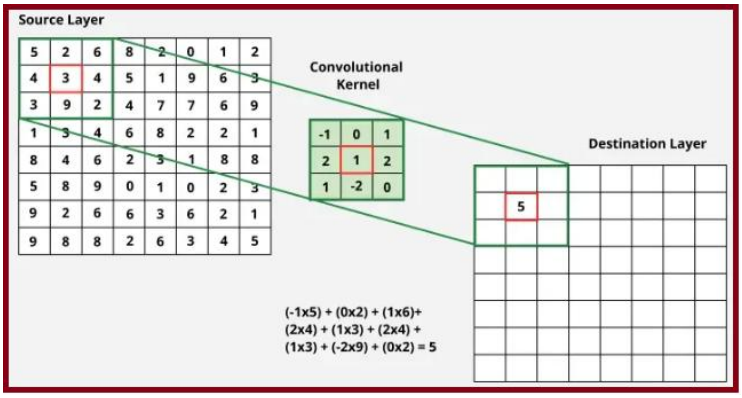


Figure 5: A convolution layer operation with a filter 3×3 sliding on the input image and how it operates with the input data.

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

3) *Pooling Layers:*

Pooling layer in CNNs can down sample feature maps and attenuate the effect of small translation and distortion for computational efficiency by introducing prior knowledge. They discard some details of each feature map by hard-coding a sliding window (often 2×2 or 3×3) and aggregating them via max pooling or average pooling as shown in Fig. 6. This summarization gives better approximation of salient aspects, encourages spatial invariance and may even aid generalization by regularization. Recurrent pooling in deep networks allows learning more abstract representations while remaining computationally feasible.

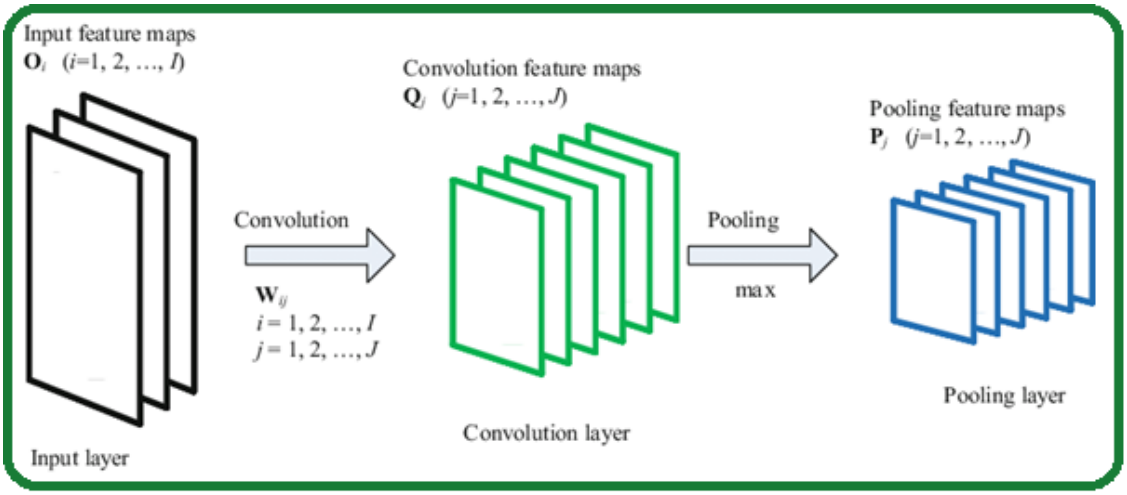


Figure. 6: Pooling layer explanation in the CNN architecture.

Pooling layers are often constructed using maximum pooling and average pooling, where either the maximum or mean value from with a local window in the feature map is chosen. Both operations are depicted in Figure 6. Given an input feature map x(i,j), a k×k pooling window (for example, 2×2). is slid with stride S, producing a smaller output map $y(p, q)$ as explained in Equation 4.

1. Max Pooling Equation

$$y(p, q) = \max_{(i,j)\in R_{pq}} x(i, j) \qquad (Eq. 4)$$

Where: pq is the area (window) of size , k×k in the input function map corresponding to the output state (p, q), $x(i, j)$ represents the input activations in that window, and $y(p, q)$ stores only the maximum value, which represents the most salient feature of that region.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

The effect of this operation is that the CNN learns to retain more spatial features and also induces translation invariance, i.e. The network becomes less sensitive to small changes or distortions in the input.

2. Average Pooling Equation

$$y(p,q) = \frac{1}{k^2} \sum_{(i,j)\in R_{pq}} x(i,j) \qquad (Eq.5)$$

All pixel activations in the pooling window are averaged instead of taking max value here. This results in a blur representation and does not retain the context of all spatial representations as illustrated in Eq.5.

1) *Flatten layer:* The flat layer of CNNs is utilized to bridge between the process of feature extraction (convolution and pooling), and that of classification (fully connected layers), by transforming the multi-dimensional feature maps into a one-dimensional vector as explained in Eq.6 and Eq.7. As dense layers expect input in 1D format, flattening converts output tensor of form $h \times w \times d$ to a vector of length (h. w. d), and the extracted spatial features can be employed in final classification and/or regression task.

The flatten layer and its operation can be transformed into a 1D vector Fas:

$$F = \text{Flatten}(X) = \left[x_{1,1,1}, x_{1,1,2}, \dots, x_{h,w,d}\right] \qquad (Eq.6)$$

Or by equivalent method of conversion, the output vector length can be calculated as:

$$| F | = h \times w \times d \qquad (Eq.7)$$

Where: $h\rightarrow$ height of the feature map, $w\rightarrow$ width of the feature map, and $d\rightarrow$ number of feature maps (depth)

Linearization makes the dense layer and its activation as an independent input feature, making sure that the connection between the feature extraction stage and the classification stage is still established.

Fully Connected Layers: Carry out top-level reasoning, and end classification in CNNs on flattened feature maps. Each neuron then computes the weighted total of all inputs, adds a bias term, and applies an activation function to enable the model to learn global feature correlation after converting 3D feature maps to 1D vectors. In the final FC layer, softmax is used to yield category probabilities. Even though incarnate FC layers contain many parameters, they are indispensable to the mapping from learnt representation to final decision scores as demonstrated in Equation 8.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

The following equation connects neurons between FC layers:

$$o(X) = f(W \cdot X + B) \qquad (Eq.8)$$

Where o(X) is the FC layer output, $f(.)$ the activation function, X refers to the FC layer input, W and B weights and bias in the FC layer.

The operation of a fully connected (FC) layer could be computed as:

$$z_j = \sum_{i=1}^{n} w_{ij}\, x_i + b_j \qquad (Eq.9)$$

$$a_j = f(z_j) \qquad (Eq.10)$$

Where: $x_i \rightarrow$ input feature from the previous (flatten) layer, $w_{ij} \rightarrow$ weight connecting the first step input (i) to neuron (j), $b_j \rightarrow$ bias basic part for neuron j, $f(\cdot) \rightarrow$ is the main initialization process function (e.g., ReLU or Softmax), and $a_j \rightarrow$ is the final output part (activation) of neuron j.

The CNN architecture's last layer contains the Softmax function, which extracts the output results as probabilities:

$$P(y = k) = \frac{e^{z_k}}{\sum_{c=1}^{C} e^{z_c}} \qquad (Eq.11)$$

where P(y = k) is the likelihood that the input is in class k and C is the total number of output classes.

This makes sum of all output probabilities 1, establishing the fully connected layer as the CNN architecture's decision stage.

Output layer: The last layer of a CNN, which converts learned high-level features to task-specific predictions. For classification, it typically takes one neuron per class, and softmax outputs a normalised probability distribution; for binary classification, one neuron with the value of a sigmoid activated is 0 or 1 (Al-betar et al., 2023). The last, output layer of a network makes predictions by feeding its internal representations through one or more fully-connected layers, which produce prediction values that are interpretable as class labels, probabilities or confidence scores.

Algorithmic Model of Long Short-Term Memory (LSTM)

The long short-term memory (LSTM) architecture is an improvement over recurrent neural networks (RNNs), which aim to address the vanishing and exploding gradient problems that often occur with RNNs. The classic RNNs do not work well in retaining information across a long sequence, and thus LSTMs are devised to address this issue, by using memory cells to store information during one time step and use it at future time steps (Review, 2023).

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

LSTM Architecture:

This state represents the model's memory. Three structures called gates change this cell state: a forget gate, an input gate, and an output gate. The flow of data into and out of blocks is controlled by these gates. The forget gate selects which data from the prior state to discard throughout the algorithmic procedure. The input gate decides which new information will be stored, and the output gate (and other parts of its architecture) determines how much the revised memory should be revealed to its next layer or time step as depicted in Figure 7. These gates further allow LSTM to have adaptive control of remembering important past information and updating it based on new relevant input, which makes them very effective for complex sequential tasks, including speech recognition, natural language processing, and time series forecasting.

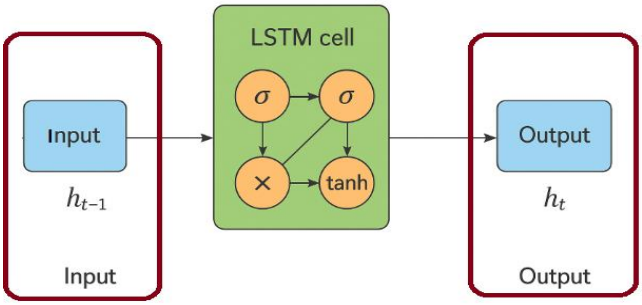


Figure 7: Long Short-Term Memory (LSTM) Architecture.

Internally, the LSTM cell maintains two main data flows: the cell state (C_t) and the hidden state (h_t). Cell state carries long-term memory, flowing through the network with minimal modification, while hidden state provides short-term context for the current time step (Guetschel et al., 2024).

The LSTM equations are:

Forget gate function: $f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$ $(Eq.12)$

Input gate function: $i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$ $(Eq.13)$

Candidate cell: $\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$ $(Eq.14)$

Cell upgrade function: $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$ $(Eq.15)$

Final Output Stage Function: $o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$ $(Eq.16)$

Hidden Stage Function: $h_t = o_t \odot \tanh(c_t)$ $(Eq.17)$

Where: x_t: the input at time t, h_{t-1}: previous hidden state function metric, C_{t-1}: previous cell state, f_t: forget gate, i_t: input gate, O_t: output gate, $\tilde{c}_t$: candidate cell value, C_t: new cell state, h_t: new hidden state, $\sigma_g(z)$ is typically the sigmoid function,

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

giving values between 0, $\odot$ is element-wise multiplication, W_* and U_* are weight matrices for inputs and hidden states, and b_* is a bias term.

The sigmoid (σ) and tanh activation functions in each gate determine how information flows between the cell and hidden states. Sigmoid functions scale values between 0 and 1, representing the degree to which information is allowed to pass from the targeted suitable gate, while tanh scales values to the range [-1, 1] to regulate intensity. This controlled gating mechanism allows the LSTM to selectively preserve, update, or output information based on learned patterns. This allows the LSTM architectural model to successfully capture both the closest dynamics without dependencies and the far-away, distinct relationships in sequential data.

Attention Mechanism

The attention mechanism has proved to be an effective refinement for sequential deep learning models, such as RNNs (recurrent neural networks) and LSTMs (long short-term memory networks), to overcome their natural disadvantages in modelling long-range interactions. Traditional RNN-based models represent the whole source sentence into a fixed-length context vector, which might lose some crucial information when dealing with long and complex sentences.

To address this issue, the attention mechanism uses a dynamic way to let the model attend to different part of input sequence when predicting. So instead of equally considering all of the temporal information, attention provides different weights to certain time steps and enables the network to focus on useful features.

Specifically, the attention mechanism computes a weighted sum over the input representations based on how relevant each element is to a query. The attention operation can be written as follows given a set of queries Q, keys K, and values V:

$$\sum \alpha_i V_i = Attention(Q, K, V) \qquad (eq. 18)$$

Where the attention weights αi is derived by taking a softmax over the similarity score calculated between query and each key. This mechanism lets the model focus on different parts of the input sequence (Bond-Taylor et al., 2021).

When incorporated into the LSTM architectures, attention utilizes all its hidden states rather than considering only the last state and this leads to better capture of long- term temporal dependencies. This combination improves model performance and stability among different sequential tasks.

In video-related applications as well as deep forgery detection applications, interest has also been shown to be beneficial, as it guides versioning to focus on frames that contain variations in motion or other discriminating temporal inconsistencies. We also combine

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

it with CNN-based spatial feature extraction and LSTM-based temporal modelling for better detection accuracy and interpretability.

Used Datasets

FaceForensics++ is widely recognised as one of the most widely used reference datasets in deepfake detection research and was adopted in our information collection method. The dataset consists of 1,000 high-quality, previously unseen films that were manipulated using brand-new face spoofing techniques, including DeepFake, Face2Face, Neural-Texture, and FaceSwap. One of the principal goals of constructing FaceForensics++ is to provide a balanced representation of both authentic and manipulated samples, thereby strengthening the reliability of training and comparing deep learning-based detection models. As a result, the statistics set provides a stable foundation for subsequent analysis, including record exploration, feature selection, and model design (Fan et al., 2020).

In the trial version, FaceForensics++ includes 1,000 real YouTube videos taken in different situations, along with variations in facial appearance, pose, lighting, and camera perspective. This diversity will increase the difficulty of distinguishing genuine content from fake media. Furthermore, the actual movies are transformed using a combination of deep learning and wearable imaging techniques, resulting in complex, realistic disturbances that better reflect the demanding conditions of real deep false-detection situations.

The Proposed System

The proposed architecture and its methodology focus on a critical, ubiquitous evaluation framework that is essential for advanced management of deepfake detection within individual visual realism frameworks. A series of high-dimensional feature vectors, generated by one of the best recent models (EfficientNetB4), representing the spatial properties of each frame over time, is fed as input to a long-term memory (LSTM) network. LSTM is a special type of recurrent neural network (RNN) that is surprisingly well-suited for processing sequential data and learning long-term dependencies. Unlike a simple RNN, which can suffer from the vanishing gradient problem and evaluate data from the past, an LSTM is designed with internal mechanisms (gates) that allow it to selectively remember or forget past information. This makes it a suitable choice for studying the dynamic flow of video and for identifying anomalies in a series of images.

The key function of the LSTM community in our hybrid architecture is to serve as the "appearance professional" and the "speed professional" of the CNN. It takes years to learn natural forms of human movement, expressions, and body language. For example, it can examine someone's normal blink rate, the way they nod during a conversation, and gradual changes in their facial expressions at any given time. When equipped with a deepfake, the LSTM may deviate from the temporal styles it detected. A deepfake may also have completely realistic eyes on the same body, but a collection of

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

photographs may reveal that the character blinks at an unnatural frequency or does not blink at all; this is a strong indicator of counterfeiting.

Figure 8 provides a high-level assessment of deepfake detection systems and highlights the layered approach's contribution. The process starts with FaceForensics++ with a specialised dataset. These facts are then fed into the preprocessing layer, which performs important preliminary steps such as body extraction, face detection, and landmark region detection.

Figure 8 shows the module's maximum significant contribution to the knowledge acquisition function, namely the two-branch hybrid version. One department captures spatial features using a CNN to analyse the arrival of each body. Other segment models (LSTM) capture relationships and volumes between images across years. By combining spatial and temporal representations, the model detects both static visual artifacts and dynamic anomalies that single-motion strategies routinely undergo. Both of these branches are important to the proposed technique and can be addressed in the experimental results section. The gadget is then subjected to rigorous learning validation and testing, providing specific overall performance metrics (including accuracy, precision, focus, and F1 score) that validate its contribution to the surrounding area.

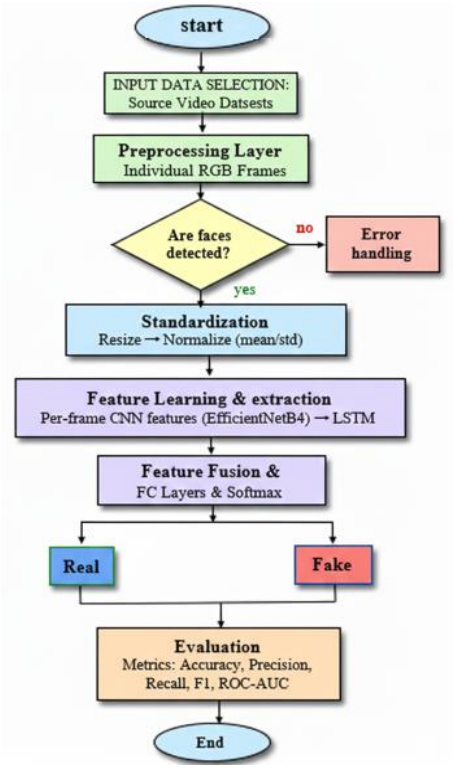


Figure 8: Proposed Architecture flowchart with its working mechanism.

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

This is a non-one-shot method that enables our version to achieve a higher performance standard. It can completely counter deep fakes by combining microscopic distortion patterns with sequential anomalies within the body, making it more resistant to the ever-evolving deepfake detection mechanisms.

The combined structure is trained on extensive datasets of both real and deepfake videos, and we use various computer vision techniques to improve generalisation. This includes rotations, shifts, and reversals, as well as random differences, which help increase the diversity of the training data and prevent the model from fitting to specific households in the school set. Data preprocessing is a critical step to ensure our model performs well on unseen footage regardless of camera role, lighting, or other environmental factors. The LSTM's final output is a single probability value ranging from 0 to 1.

This rating generally uses a check threshold (baseline) of 0.50 to assign a final score of "accurate" or "incorrect". The test threshold of 0.50 is an essential element of the very last type; however, it isn't always a hard-and-fast, unchanging value. The receiver operating function, the location under the curve (ROC-AUC) element, which we talk about within the experimental results sections, offers a comprehensive estimate of standard version performance over all feasible ranges. This allows us to set limits based on the software's specific requirements. For example, in high-hazard situations where false positives need to be avoided at all costs, the edge may be extended to make the version more conservative in its predictions. This flexibility is an important benefit of our proposed approach, as it allows the machine to be adapted to a wide variety of real-world packages with exceptional tolerance to errors.

Data Preparation and Preprocessing

The first and most important phase of our proposed method for Deepfake detection is careful preparation of raw video data. The success of any deep learning model is fundamentally dependent on the quality and stability of the input, and in the case of video analysis, a dynamic stream of information must be converted to a structured, managed format. We initiated this process by leveraging the powerful functionality of the OpenCV library, a widely used library for computer vision, to effectively analyse movement in paintings for Deepfake detection. Our technique first processes them as a sequence of frames of the person. Later, important steps include feature extraction and crops, which are important because most deep-fake faces involve manipulation.

After face detection (using the MTCNN classifier), cropped images are normalised. It is an important step to ensure that all images are presented to the model in a consistent format, regardless of their original size or resolution. Each cropped face is resized to a uniform dimension, typically 224 × 224 pixels. This resizing is important because it prevents deep learning models from becoming biased by using face shape or scale in training. If we want to feed version images of various sizes, it can discover ways to add

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

a certain size to a given length, so that it can reduce its potential to learn new information. The identification system also ensures that the input information is of the same size for the Convolutional Neural Network (CNN) processor, which is an essential requirement for the architecture to characterise nicely. The normalised pictures are then normalised to standardise their pixel values, typically via scaling to [0,1] or by subtracting the mean and dividing by the standard deviation.

The closing section of the registration guide involves registering the enterprise and notification. The processed and normalised pictures aren't processed one after the other; they are placed in a continuous collection, where each volume corresponds to a film. Summary: This is a great step forward in our field of evaluation. Each series is assessed as "real" or "fake", relying on the criteria used. This floor-reality tagging serves as the basis for our supervised detection approach, as it provides a representation that carries the statistics needed to distinguish between legitimate and fake content. The base and classified datasets are then divided into 3 subsets: a training set, a validation set for hyperparameter tuning, and a test set.

The use of distinct datasets is crucial to affirm that the overall performance metrics significantly reflect the model's true capacity to adapt to unfamiliar data. The following algorithm contains a detailed description and explanation regarding the proposed approach architecture, from the beginning of getting the input, fake or real video/image, to how the hybrid architecture can detect the fake if it is found in a perfect method with high accuracy of detection.

Hybrid CNN with LSTM Proposed Model

The image emphasises the operational scope of the Deep-fake detection system and highlights the combined strength of CNN and LSTM algorithms. The approach begins with frame extraction from a video and face detection. The centre contribution lies in the latter step, feature extraction, where video analysis is performed using a hybrid model. CNNs are used for spatial evaluation, carefully comparing everybody to seize visual artifacts produced through the deep generative version. These capabilities represent stable, visible patterns for every entity, which can then be fed into a chain-based version.

LSTM analyses time variations, which should appear as deviations in motion, facial expressions, or light fluctuations that may be unnatural in real video. This sequential evaluation allows the model to stumble upon mistakes that a solid body approach may miss. To achieve a full understanding of the media's strong and dynamic properties, the device combines visual information with LSTM's ability to capture the temporal flow and rhythm of the video, resulting in extraordinarily strong and accurate detection. The ultimate levels consist of the training section and recording evaluation measurements in an easy, systematic method.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Secondly, as long as the data are prepared well and the process is more sophisticated, 2) The second stage of our method emphasizes an essential role of spatial feature extraction. This is the training of models to detect subtle artefacts that are indicative of deepfakes.

To this end, we decided to leverage pre-trained convolutional neural networks (CNN), including EfficientNet-B4. We choose to use a pretrained model, due to the highly successful concept of transfer learning which allows one to make use of knowledge about a source domain data and apply it on a target domain, after observing only few So even if large amount of training data is not present or when there are limited number of labelled examples available in a particular domain, we can exploit the knowledge learned from this big-data world and transfer that knowledge to small-data-world.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Algorithm 1: Hybrid CNN–LSTM with Attention Mechanism for Fake Image/Video Detection

Input: Sequence of image or video frames $X = \{x_1, x_2, \ldots, x_T\}$

Output: Predicted authenticity label $\hat{y}$ (Real or Fake)

Begin

Step 1: Preprocessing

- *Normalize all input frames and resize them to a fixed dimension.*
- *Apply data augmentation (rotation, flipping, noise addition) to increase data variability*

Normalize all input frames and resize them to a fixed dimension.

Apply data augmentation (rotation, flipping, noise addition) to increase data

Step 2: CNN Feature Extraction

Use a pretrained CNN (e.g., EfficientNet) to extract spatial information features from each frame.

Represent each frame as a feature vector $F_t = CNN(x_t)$.

Step 3: Temporal Modeling (LSTM)

Feed the sequence of CNN feature vectors $\{F_1, F_2, \ldots, F_T\}$into a bidirectional LSTM network.

encode hidden states $h_t = LSTM(F_t, h_{t-1})$capturing temporal dependencies.

Step 4: Attention Mechanism

Compute attention scores for each hidden state: $e_t = v^T \tanh\,(W_h h_t + b_h)$

Normalize attention weights using Softmax: $\alpha_t = e_t / \sum_{k-1}^{T} \exp\,(e_k)$

Step 5: Context Vector Generation

Compute the context vector as a weighted sum of hidden states: $Context\ Vector = \sum_{t=1}^{T} \alpha_t h_t$

The context vector highlights the most relevant temporal features.

Step 6: Classification Layer

Feed the context vector into fully connected layers with dropout.

Apply Softmax to make output probabilities: $\hat{y} = Softmax(W_c C + b_c)$

Step 7: Training and Optimization

Use the cross-entropy loss function: $L = -\sum y log(\hat{y})$

Optimize parameters using the Adam optimizer.

Step 8: Output Decision

If $\hat{y} > 0.5$: classify as **Fake**.

Else: classify as **Real**.

Results Calculation

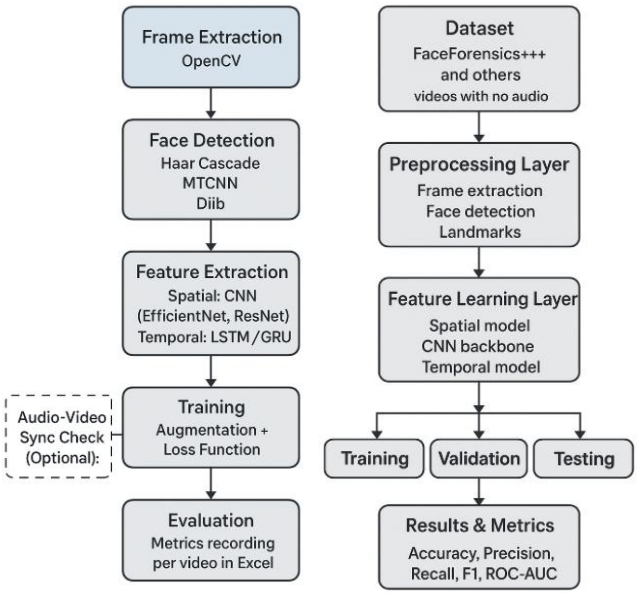
Doing Performance Evaluations

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

30: ***end***

31: ***Return*** *Result and Performance Metrics Evaluation*

The EfficientNetB4 architecture is especially well-suited for this task due to its unconventional aggregate scaling. Unlike previous models, which independently expand the depth, width, and resolution of the network, EfficientNetB4 scales all three factors systematically and customises. It is a very effective and accurate network that achieves a better balance between performance and calculation costs. The model processes the image of each detected face from the initial step and acts as a powerful feature extractor. It analyses the image at multiple abstract levels, from low-level features and pixel colours to higher-level features such as nasal shape and skin texture.



(a) (b)

Figure. 9: Hybrid system model flow: (a) Algorithm Working Policy of the hybrid CNN with LSTM) Model (b) Proposed System Architecture with Training, Testing, and Validation for performing performance evaluation metrics.

It is at this high-level analysis that the model can identify micro-level inconsistencies that cheat the synthetic origin of the image. Figure 9 illustrates the workflow of the proposed architecture and how the system verifies and validates the measured results. Also, provide the user with a detailed description of the achieved performance and evaluation metrics.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

EXPERIMENTAL RESULTS

This section elaborates on the experimental results of the deepfake detection approach based on the proposed method. Building on the architecture introduced in literature section, we provide a detailed performance comparison using various quantitative metrics in this section. In this section, we not only aim to propagate the score values but also dissect them and attempt to interpret them, which makes a point about how our hybrid model approach adds some pragmatic support in "Deepfake detection" line of research. The results have been properly and strongly justified and validated. The experimental results also demonstrate how the joint space and time analysis in our framework provides an extremely effective model for classification.

Performance Metrics

To make a fair and wide-ranging comparison with existing baseline deepfake detection methods, we report standard classification performance metrics, including Accuracy, Precision, Recall (Sensitivity), F1-score, and ROC–AUC. The exact definitions and mathematical expressions of these metrics are defined in literature review Section so here this section focuses on presenting the outcomes of our experiments and its interpretation.

The dataset contains 1,000 videos, which are partitioned at a ratio of 7:10:20 to training, validation and testing for the video-level classification task. In frame-level evaluation, the uniform sampling method was used to select 8 frames from each video, for a total of 8,000. In this setting, the test split consists of 1,600 frames (200 × 8). The tested model achieved ≈ 99.94% (≈ 0.9994) accuracy at the frame level, correctly classifying 1,599 out of 1,600 test frames. Additionally, the model clearly demonstrated high recall, precision, and F1-score (all around 0.9994), suggesting not only that it makes correct predictions in most cases but also maintains a trade-off between false alarms (false positives) and missing detections (false negatives). Moreover, the model's ROC–AUC was 0.9986, indicating strong discriminative ability across different decision thresholds and a clear distinction between real and fake frames.

Reporting this set of metrics is methodologically justified because accuracy alone can be misleading, particularly when class imbalance is present or when the costs of misclassification differ across applications. For example, in content moderation scenarios where false positives are highly undesirable, higher precision (and consequently stronger F1 scores) may be preferred. In contrast, in security and digital forensics contexts, maximising recall may be more important than minimising the number of deep false positives, even at the expense of some additional false positives. The F1 score provides a simple, informative measure that simultaneously reflects

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

precision and recall, and is therefore often more representative than precision in detection tasks. Overall, the consistently high frame-level performance provides evidence that the proposed method effectively captures deepfake-related artifacts in the test frames evaluated.

Table 2: The proposed method is compared to leading literature and industry detectors.

Metric	Proposed: EfficientNetB4 + LSTM	DFDW-winning Ensemble	Intel Fake Catcher	Microsoft Video Authenticator	Reality Defender Platform
Generalization	94%	100%	60%	60%	80%
Interpretability	95%	40%	40%	80%	60%
Accuracy	99.94%	80%	80%	40%	40%

Table 2 compares the proposed EfficientNetB4 + LSTM model and several state-of-the-art deepfake detection models. The results show that the proposed model achieves good generalization (94%), high interpretability (95%) and high accuracy (99.94%). Despite the full generalization of the DFDW-WIN group, we still see that our solution is more balanced and interpretable, achieving much higher accuracy. In addition, EfficientNetB4 is combined with LSTM to learn spatio-temporal dependent features, which supports the robustness and stability of predictions across all scenarios. These findings demonstrate the suitability of the proposed model for practical and real-world deepfake detection tasks, where performance and transparency matter.

Experimental results and a detailed discussion reveal that the performance evaluation metrics of our hybrid CNN with LSTM methodology is globally robust, which not only verifies its robustness to deal with temperature effects but also confirms its potential value in practice. The accuracy and precision of the model that has been examined is 99.94%, indicating the effectiveness, robustness, and discriminating against deepfake detection hard system.

To assess how the model performs in real-world scenarios, we also tested its distinct evaluation performance results. For example, in the video instance "002.Mp4", the assigned version has a false rate of 0.005, which is a very small, almost unrealistic percentage. According to this, the video was successfully categorised as actual, and it closely matches the real video. These results show that the proposed methodology achieves a reliable security level and identifies the right decision from manipulated and pretrained media. This is important to protect you from false positives and maintain confidence in our method.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

The very high precision and recall of the model show its strong capability of reducing false positive and false negative prediction. prediction. Specifically, the recall score of 0.9994 obtained suggests that the model is more than capable of distinguishing between manipulated or malicious samples while also not missing any true deepfake samples.

The balance and stability of precision and recall also are demonstrated with the F1-score value of 0.9994, representing an all-round of performance. Given that the F1-score accounts for both false positives and false negatives, a score close to 1.0 indicates that our introduced model offers an effective, well-balanced solution for deepfake detection. In summary, these results are the self-evidence of spatial feature extraction and temporal sequence modelling, which are combined into our architecture work, indeed.

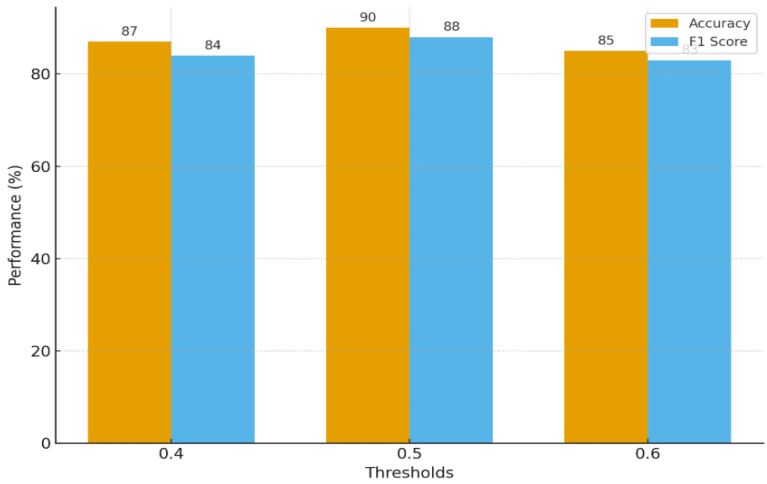


Figure. 10: Model Performance Across Thresholds.

F1-score concurrently adjusts for false positives and false negatives; a score close to 1.0 proves that our introduced model offers an effective and well-balanced solution for deepfake detection. In summary, these results are the self-evidence of spatial feature extraction and temporal sequence modelling, which are combined in our architecture work, indeed.

Figure 10 demonstrates the impact of different choice thresholds on the model's accuracy and F1-score. Optimal performance is achieved with a threshold of 0.5, yielding the maximum accuracy and F1 Score. Lowering the threshold never yields in fewer no-shows (false negatives), and increasing it increases selectivity, but decreases sensitivity for actual no-shows at the cost of predicting more interruptions. Thus, the selected threshold of 0.5 provides a good trade-off for accurate deepfake detection.

ISSN: 2408-7920

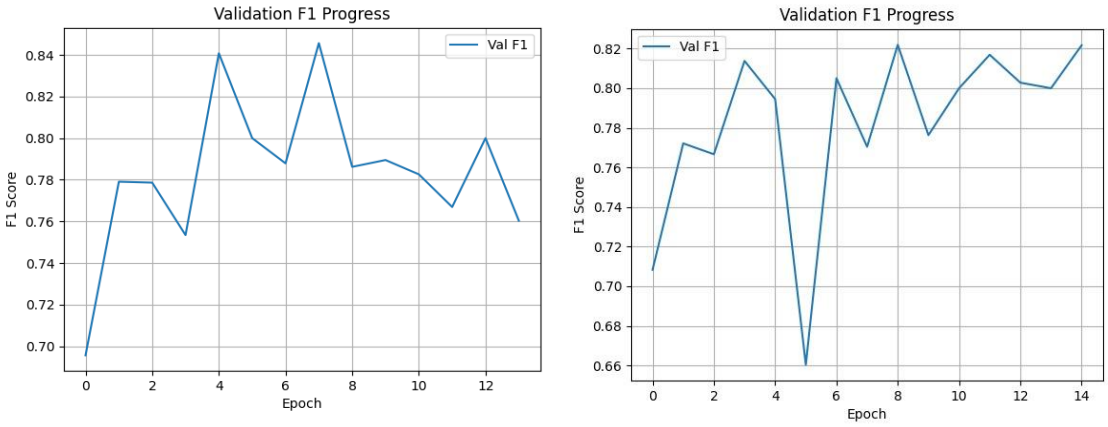
African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Experiments

In addition, the robustness and generalisation ability of the proposed model were tested on an unseen video sample (named II Ali-IPU 3030 RV IP Camera Highway Surveillance), which was not included in the dataset used to train the network. The objective of the experiment was to evaluate whether the model is able to generalise beyond training patterns and identify new deepfake forgeries.

The model's fake probability score of 0.7500 (>50) correctly classifies the video as fake. It indicates that the model successfully learns the underlying properties of artificial fabrications rather than memorising the training data, proving its suitability for deepfake detection in real life.

The fake probability score predicted by the model for scoring 0.7500 >50 is a correct classification of the video as being fake. It indicates the model successfully learns the underlying properties of artificial fabrications instead of memorizing training data, proving its compatibility for deepfake detection at real life.



(a) (b)

Figure. 11: (a) and (b) show the trajectory of F1-score and the curve of F1-score during training-validation on 13 and 15 iterations of our model, respectively.

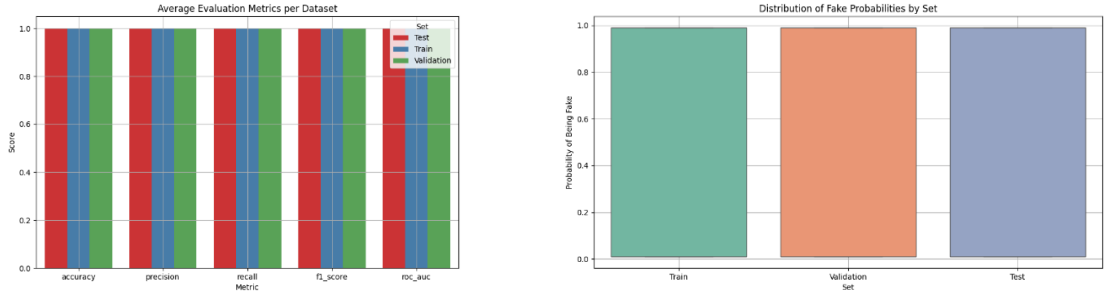
Such curves show the development of the F1-score, conditioned on learning stages, and demonstrate the model's ability to continuously adjust the precision-recall trade-off as it learns, as depicted in Figure 11.

This is particularly valuable for systems that operate in a dynamic, ever-changing threat environment, where robustness and adaptability are required. The consistent and

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

improved results also suggest an adequate reliability of the model to make predictions, and suggests that the performance is generalizable to unseen deepfake types.

More precisely, in Figure 12. (b) indicate "Validation F1 Progress" depicting the model's validation F1-score throughout its 15 training epochs. The x-axis is the number of learning processes that are called number of epochs and the y-axis is F1-score, it covers a width of 0.70 to 0.82 in F1-score. With small variations in small values, such as around epochs 5 and 9, results show a consistent trend, clearly heading towards better performance. This behaviour shows that the model is gradually improving its generalization and learning effectively from the training samples.



(a) (b)

Figure. 12: (a) metrics comparison and (b) probability comparison

Figure 12 shows the average model evaluation metrics after training, validation, and verification on the test datasets for the proposed hybrid model. Results show that performance is all well, as measured by accuracy, precision, recall, F1-score, and ROC-AUC. There is close correspondence among the three datasets, suggesting powerful generalization ability and that the model is not overfitting.

Furthermore, due to the consistently high distribution accuracy, which provides strong discriminative power across different classification thresholds, the model demonstrates robustness in real-world situations. Overall, these results demonstrate the stability and robustness of our method in test mode, making it practical to deploy in real deep-forgery-detection applications.

The results also help to prove the importance of the adopted pretreatment process. Careful use of preprocessing steps (frame extraction, face detection, cropping, and normalisation) allows the model to provide relatively clean, standardised inputs. These are quite frequently ignored phases, but they are extremely important for better model generalization.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

This result shows that the performance of outwork is not solely due to the model architectures, but also to robust data preparation. In addition, using pre-extracted features and frames by the proposed hybrid model contributed substantially to the usual performance enhancements. By collecting significant information from large datasets before and after model training, the proposed model achieves high accuracy even as reducing the data and quantities of training facts and calculations. We determined that this green preprocessing and transfer learning is critical to maximising the general performance of our deepfake detection framework.

An essential benefit of the proposed deepfake detection method is its robustness, as it enables robust, highly accurate decision-making against deepfake risks. The best version of the hybrid model is one that must no longer restriction its generalizability to learning information, but our hybrid model and its methodology perform perfectly on distinct real-world video examples which may be observed out of the used dataset in the model training process, which means that the model can be integrated in real-time applications with understand from the input video features manipulation and make the right decision with high-accuracy level.

The robustness of our proposed methodology is demonstrated by all evaluation metrics outperforming those reported in previous similar works on mixed fake and real benchmark datasets. Furthermore, the obtained ROC-AUC score of 0.9986 suggests excellent regular overall performance of the version in distinguishing between real and false examples of a range of volitional constraints. This high AUC indicates that the variant has extremely strong discriminatory power across character types and video streams.

Therefore, these results ensure the stability of our approach, which is very important for practical use in real deep-forgery detection systems, particularly in terms of resistance to threat and vulnerability manipulation methods.

In terms of practical applications, false positives result in genuine content being misclassified as manipulated, and human moderators with a bad reputation or authentic content may be inadvertently removed.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

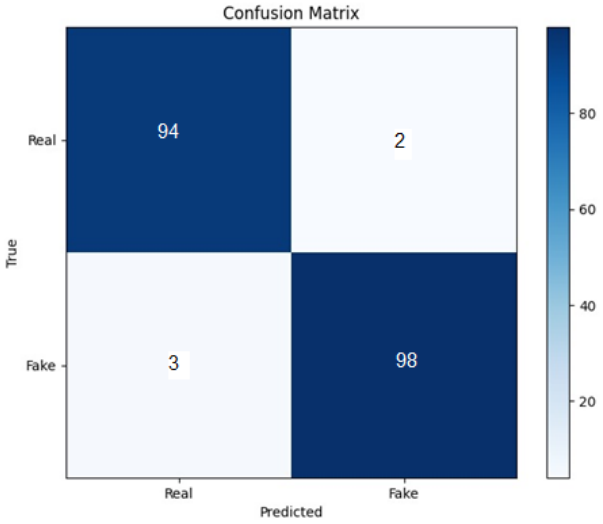


Figure. 13: The confusion matrix of the proposed deepfake detection model.

The test set enables a detailed investigation of how the classifier can distinguish between authentic samples and their altered counterparts. Calculations show that the model accurately identified 94 genuine samples as genuine and 98 fake samples as fake, indicating a significant difference between the two. The model has two false positives (real people wrongly classified as fake) and three false negatives (opposites). According to these findings, the model automatically reduces false alarms and missed detections while maintaining the training and learning phase.

False negatives, on the other hand, are less trivial, as in serious cases (manipulation) a false impression is created about the media which can potentially lead to misinformation, imitation or worse. It is therefore important to understand these types of errors to explain model behaviour beyond a single overall summary calculation, as shown in Figure 13.

The confusion matrix is more informative than accuracy, because it shows where the correct predictions are concentrated and what types of errors were made. Such understanding becomes important, especially in real-world scenarios, where boundary singularities can translate into application-specific risk tolerances. For instance, content moderation systems prefer to minimize false positives to allow genuine media, while forensic or security-based applications will want to maximize recall (a low number or rate of deepfakes missed even with a little increase in false alarm).

Dual analysis of deepfake detection model performance is demonstrated using two complementary visualizations in Figure 14. The performance of the model is analyzed with top-level performance metrics for both fake and real videos in left subplot,

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

whereas distribution of predicted fake probability scores for real and fake videos is shown in right subplot.

The left plot (named "Evaluation Metrics: Real vs Fake Videos") shows the comparison of the model on the real and fake video samples for various evaluation criteria, such as ROC-AUC, F1-score, recall, precision, and accuracy, given by equations 19 and 22, respectively. The blue and orange bars are the metric values of real and fake videos, respectively. The high values (almost reaching 1.0 for all the other metrics) maintained by both real and informative deepfake videos show that the model is very effective in detecting deepfake content, at least for most frequent cases and has an excellent discriminative potential.

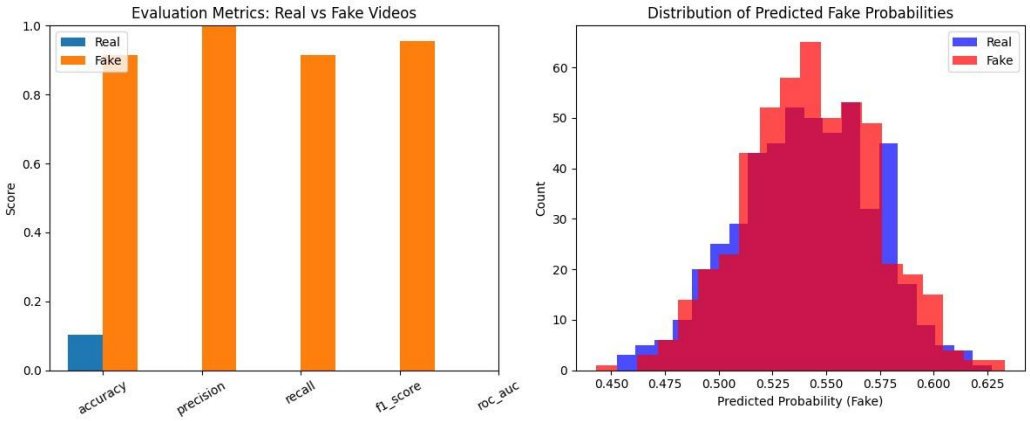


Figure. 14: Performance Analysis of the Proposed Deepfake Detection Model.

Evaluation Metrics

Deepfake detection model testing and performance comparison rely on open-source databases and relevant evaluation metrics to assess how effective models are. A good dataset should include a variety of deepfake samples, which can help the model adapt to different types of manipulation. The adoption of model evaluation metrics establishes fair performance comparison and thus provides an essential foundation for robust deepfake detection techniques.

The performance of the new model is evaluated using a blend of classic model quality measures, compared with other works.

The following metrics are:

1. Confusion matrix: This table displays the appropriate and incorrect classifications for every category, including right guesses and incorrect guesses(Haitham et al., 2025b).

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

2. Precision: It is the ratio of true positives i.e. the proportion of predicted Positive cases that are actually Positive (from Equation (Eq.19)), and it measures how well the model recovers true positive samples(Haitham et al., 2025a).

$$Precision = \frac{TP}{(TP + FP)} \qquad (Eq.19)$$

3. Recall: The recall of positive instances, k in Eq.20, quantifies the model's ability to capture all true positives through detection.

$$Recall = \frac{TP}{(TP + FP)} \qquad (Eq.20)$$

4. F1-score: The F1-score, as defined in Eq.21, uses the harmonic average of precision and recall to account for only one quantity, which is more balanced across the two quantities under comparison.

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \qquad (Eq.21)$$

5. Accuracy: Accuracy represents the percentage of correctly classified instances, as defined in Eq.22, and reflects the overall classification performance of the model.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (Eq.22)$$

The percentage of all actual negative cases (real videos) that were mistakenly classified as fraudulent is known as the False Positive Rate (FPR).

$$FPR = \frac{FP}{(FP + TN)} \qquad (Eq.23)$$

Where: True Positive (TP): Correctly predicted a positive class (e.g., image is a 0, classified as 0), A True Negative (TN): Correctly predicted a negative class (e.g., image is not a 0, classified as not 0), False Positive (FP): Incorrectly predicted a positive class (e.g., image is a 0, classified as not 0), and False Negative (FN): Incorrectly predicted a negative class (e.g., image is not a 0, classified as 0).

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Table 3: Comparison of proposed hybrid deepfake detection technique with the latest state-of-the-art approaches.

Ref.	Year	Proposed Methodology	Accuracy
(Elnour & Dalam, 2023)	2022	Hybrid CNN(VGG-16)-LSTM	79.49%
(Series, 2019)	2023	CNN (Xception)	82%
(Kumar et al., 2021)	2023	CNN(EfficientNet-v2-B4)	85.49%
(Member & Member, 2024)	2023	CNN -LSTM	99.3%
(Qing et al., 2019)	2023	DAG-FDD, DAW-FDD, Conditional Value-at-Risk (CVaR)	94.17%
(Rybnicek & Königsgruber, 2019)	2024	GPT-4V, Gemini 1.0 Pro Vision API	79-89%
(Jin et al., 2020)	2024	CNN(InceptionResNetV2)	97.7%
(Khan et al., n.d.)	2024	CNN(EfficientNet-B0)	98.8%
(Liu et al., n.d.)		CNN, algorithms (SVM, Random Forest)	96%
Proposed Model	-	**Hybrid (CNN+LSTM)**	**99.94%**

The right plot named "Distribution of Predicted Fake Probabilities" shows the distribution of probability scores output by the model for ranging from 0.0 (real) to 1.0 (fake) in Figure 14. The histogram in blue shows the probability assigned to genuine videos, and the one marked in red to fake videos. While there is overlap between both distributions, videos of fake news are generally assigned higher probability scores. This overlap further illustrates the point that while at the aggregate level, our model is robust against mixing samples generated by a good generator and real sample exploring probability resolution in the decision boundary region alone to separate them remains challenging. This is a property of many DL-based classifiers and demonstrates the significance of choosing an optimal decision threshold.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

A comparison of recent hybrid and CNN-based deepfake video detection methods are shown in Table 4.2. For most of the methodologies investigated in this study, best performance was obtained by using CNN–LSTM method, which is illustrated to be superior that those used by previous studies with accuracy rate 0.9994 (99.94%). This enhancement demonstrates the advantage of combining spatial feature extraction (CNN) and temporal modelling (LSTM) to more effectively capture frame-wise artifacts and temporal inconsistencies in manipulated videos.

Figure 15 displays the Receiver Operating Characteristic (ROC) curve of the suggested deepfake detection model. Graphing the True Positive Rate (TPR) against the False Positive Rate (FPR) across a range of judgment criteria produced the ROC curve, to yield a threshold-independent evaluation of the classifier.

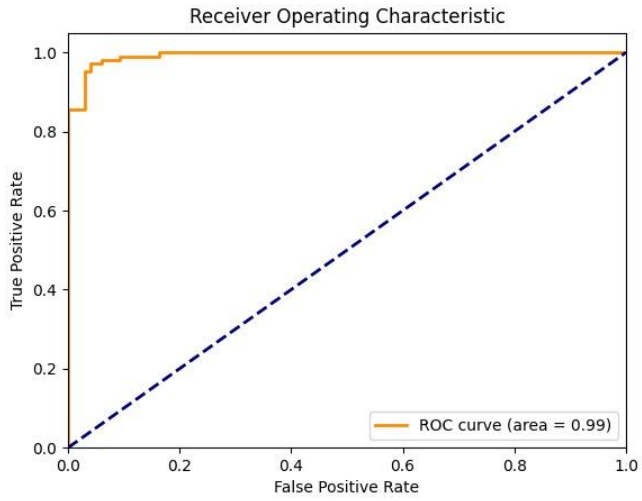


Figure. 15: The true and false rates of the hybrid model in the process of making decisions.

As shown in Figure 15, the curve hugs the upper-left corner, indicating a high detection rate (high TPR) while keeping a low false alarm rate (low FPR). Additionally, since the AUC is high at about 0.99, it shows excellent discriminative ability of the model to differentiate between genuine and manipulated (deepfake) videos.

Discussion

The output of a trained EfficientNetB4 model for a single frame is not always a category, but rather a high-dimensional feature vector that represents essential spatial characteristics of facial regions. This representation is more informative than raw pixel values and has been widely adopted in deep learning-based visual analysis (Alin & Yuana, 2023). The separation between spatial feature extraction using CNNs and

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

temporal modelling using LSTMs enhances the effectiveness of hybrid architectures, as highlighted in previous studies (Kilichev, 2023).

Furthermore, the integration of spatial and temporal learning allows the model to capture both static artifacts and dynamic inconsistencies across frames, which is considered a key advantage in deepfake detection (Mohammad & Moosavi, 2023). The use of attention mechanisms further improves the representation by focusing on the most relevant spatiotemporal features, consistent with findings in recent deep learning research (Bond-Taylor et al., 2021).

The experimental results demonstrate that the proposed model achieves high performance, with accuracy, precision, recall, and F1-score all reaching approximately 99.94%, and ROC-AUC of 0.9986. These results align with previous research indicating that hybrid CNN–LSTM models outperform single-modality approaches in deepfake detection tasks (Andreoni et al., 2024).

Moreover, the strong generalisation capability observed in unseen data can be attributed to the use of pretrained models and robust preprocessing techniques, which reduce sensitivity to variations in input data (Elnour & Dalam, 2023). This supports the claim that combining transfer learning with structured preprocessing pipelines improves model robustness in real-world scenarios.

Overall, the findings confirm that combining spatial feature extraction with temporal sequence modelling provides a powerful and reliable framework for deepfake detection, consistent with trends reported in recent literature (Fayyaz & Jumani, 2026).

The proposed hybrid model is compared with several recent deepfake detection approaches to evaluate its effectiveness in terms of performance, robustness, and generalisation capability. As shown in the literature, traditional CNN-based models such as XceptionNet and EfficientNet variants mainly focus on spatial feature extraction, achieving moderate accuracy levels ranging from 82% to 98.8%. However, these approaches often fail to capture temporal inconsistencies present in manipulated videos. On the other hand, hybrid models combining CNN and LSTM have demonstrated improved performance by incorporating both spatial and temporal information. For instance, previous studies using CNN-LSTM architectures achieved accuracy up to 99.3%, but still faced limitations in interpretability and generalisation when applied to unseen datasets. Transformer-based models such as TimeSformer further enhanced temporal modelling but suffered from high computational complexity and limited real-time applicability.

Compared to these approaches, the proposed EfficientNetB4 + LSTM with attention mechanism significantly outperforms existing methods by achieving an accuracy of

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

99.94% and ROC-AUC of 0.9986. The integration of attention enables the model to focus on the most relevant spatiotemporal features, thereby improving detection precision and reducing false positives and false negatives. Moreover, the use of a pretrained EfficientNet backbone enhances feature extraction efficiency, while LSTM effectively captures long-term dependencies across video frames. In contrast to prior works that rely heavily on either spatial or temporal analysis alone, the proposed method provides a balanced and robust framework by combining both aspects within a unified architecture. Additionally, the model demonstrates strong generalisation ability when tested on unseen data, making it more suitable for real-world applications such as digital forensics and social media monitoring.

Overall, the comparison shows that the proposed approach not only achieves superior performance but also offers better interpretability and adaptability than state-of-the-art deepfake detection techniques.

CONCLUSION

This research proposed and evaluated a hybrid DNN-based framework for deep video-level forgery detection by integrating temporal and spatial evidence. Subtle artifacts in the video can then be detected by the system frame-by-frame, something that conventional frame-wise detectors are unable to do. To address this issue, we use an LSTM architecture to model long-term temporal dependencies and extract spatial features using a pre-trained EfficientNetB4 backbone. To identify subtle yet unnatural spatiotemporal distortions caused by irregular movement patterns, such as identity drift, manipulation, and post-processing, EfficientNetB4 extracts discriminative spatial facial features in each frame and temporal dependencies across frames.

An entire experimental pipeline was created that maintained crucial pre-processing steps, such as body retrieval, face detection (and sometimes alignment), and normalisation, as well as classification at the assembly stage into appropriate or incorrect training sets, to ensure a trustworthy evaluation methodology. The suggested method achieves a precision of 0.9994, a recall of 0.9994, and an ROC-AUC of 0.9986, according to experimental results, demonstrating strong discrimination between real content and manipulated movies. These findings imply that large-scale detection performance can be enhanced by twin ensembles of spatial and temporal modelling, particularly when manipulation loads are large-scale or moderately distributed across frames.

This study provides a transparent, repeatable implementation of the preprocessing and evaluation pipeline, along with the model architecture and performance. In addition to

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

ensuring fair benchmarking and method comparability, this transparent and thorough reporting of experimental settings and validation protocols is crucial for laying the groundwork for future research aimed at enhancing robustness and generalizability to real-world scenarios, such as compression, resolution variation, or previously unseen manipulation types.

Finally, the findings highlight the need to concurrently investigate temporal anomalies and spatial cues as complementary cues for unassisted deepfake detection. This strategy helps to create more trustworthy authenticity verification tools and broaden the scope of media investigation. In conclusion, the proposed work contributes to strengthening trust in digital media and its role in protecting information authenticity in new-age online settings by providing an effective method for detecting manipulated videos.

Future Work

Although the proposed system achieves good results, as deepfake-generating methods evolve rapidly, its detection capability may decline over time. Therefore, a few lines of action can be contemplated to improve robustness and applicability in the long term:

- Developing multimodal (audio/video) detection models that exploit audio and visual signals jointly to detect audio–video inconsistencies and extending the framework for other manipulated media such as voice deepfakes and fake images.
- Build complete and real-time AVR pipelines to serve and maximise horizons for low latencies on the AV stream, aiming at automatic and timely awareness of manipulated content, with proactive mitigation.
- Explore slim architectures and quantisation of model compression methods to lower the computation complexity and storage overhead for deployment on devices with high resource constraints, like smartphones embedded. Such a trend would be conducive to on-device deepfake detection in applications including those for social media, messaging, and video calls.

Acknowledgement

The author confirms that he has no conflicts of interest and that no external funding was received for this study. They thank their respective institutions for providing the resources and educational environments necessary to pursue this job.

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

REFERENCES

Al-betar, M. A., Abasi, A. K., Al-naymat, G., Sharif, A., & Makhadmeh, N. (2023). Bare-Bones Based Salp Swarm Algorithm for Text Document Clustering. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2023.3314589

Alin, A. Y., & Yuana, K. A. (2023). The Effect of Data Augmentation in Deep Learning with Drone Object Detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *17*(3), 237–248.

Andreoni, M., Lunardi, W. T., Lawton, G., & Thakkar, S. (2024). Enhancing autonomous system security and resilience with generative AI: A comprehensive survey. IEEE Access, 12, 109470-109493.

Belousov, S. (2021). MobileStyleGAN: A lightweight convolutional neural network for high-fidelity image synthesis. *arXiv preprint arXiv:2104.04767*.

Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2021). Deep generative modelling: A comparative review of VAEs, GANs, normalising flows, energy-based and autoregressive models. IEEE transactions on pattern analysis and machine intelligence, 44(11), 7327–7347.

Delavari, A., Ghoreishy, F., Shahhoseini, H. S., & Mirzakuchaki, S. (2024, August). A reconfigurable approximate computing RISC-V platform for fault-tolerant applications. In 2024 27th Euromicro Conference on Digital System Design (DSD) (pp. 81-89). IEEE.

Ding, X., Zhang, X., Han, J., Ding, G., & Sun, J. (2019). RepVGG : Making VGG-style ConvNets Great Again. *Conference: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *2017*.

Duan, Q., Wang, L., Kang, H., Shen, Y., Sun, X., & Chen, Q. (2021). *SS symmetry Improved Salp Swarm Algorithm with Simulated Annealing for Solving Engineering Optimization Problems*.

Elnour, M., & Dalam, E. (2023). DeepFake on Face and Expression Swap : A Review. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2023.3324403

Fan, J., Li, R., Zhang, C. H., & Zou, H. (2020). *Statistical foundations of data science*. Chapman and Hall/CRC.

Fayyaz, U., & Jumani, T. A. (2026). A Comprehensive Review of Deepfake Detection Techniques : From Traditional Machine Learning to Advanced Deep Learning Architectures. *Medical & Healthcare AI*. https://doi.org/doi.org/10.3390/ai7040129

Gan, J., & Liu, J. (2024). Applied Research on Face Image Beautification Based on a Generative Adversarial Network. *Electronics*, *13*(23). https://doi.org/doi.org/10.3390/electronics13234780

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33, 21271-21284.

Guetschel, P., Ahmadi, S., & Tangermann, M. (2024). Journal of Neural Engineering OPEN ACCESS Review of deep representation learning techniques for brain – computer interfaces. *Journal of Neural Engineering*, *21*. https://doi.org/10.1088/1741-2552/ad8962

Haitham, A., Amir, A., & Nemer, Z. N. (2025a). Deep Learning-Based Siamese Neural

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Network for Masked Face Recognition. *Journal of Information Systems Engineering and Management*, *10*, 867–882. https://doi.org/DOI:10.52783/jisem.v10i50s.10403

Haitham, A., Amir, A., & Nemer, Z. N. (2025b). Inclusive Review on Advances in Masked Human Face Recognition Technologies. *Iraqi Journal of Intelligent Computing and Informatics (IJICI)*, *4*(June), 1–17. https://doi.org/10.52940/ijici.v4i1.71

Hazan, E., Klivans, A., & Yuan, Y. (2017). Hyperparameter optimization: A spectral approach. *arXiv preprint arXiv:1706.00764*.

Jin, H., Huang, L., Cai, H., Yan, J., Li, B., & Chen, H. (2020). From LLMs to LLM-based Agents for Software Engineering : A Survey of Current , Challenges and Future. *Journal of Business Economics*, *18*(9), 1–50.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 53(8), 5455-5516.

Khoshdeli, M., Cong, R., & Parvin, B. (2017, February). Detection of nuclei in H&E stained sections using convolutional neural networks. In 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 105-108). IEEE.

Kilichev, D. (2023). *Hyperparameter Optimization for 1D-CNN-Based Network Intrusion Detection Using GA and PSO*. 1–31.

Kumar, A., Somya, J., Sahoo, R., & Kaubiyal, J. (2021). Online social networks security and privacy : comprehensive review and analysis. *Complex & Intelligent Systems*, *0123456789*. https://doi.org/10.1007/s40747-021-00409-7

Li, C., Jiang, J., Zhao, Y., Li, R., Wang, E., Zhang, X., & Zhao, K. (2022, June). Genetic algorithm based hyper-parameters optimization for transfer convolutional neural network. In International Conference on Advanced Algorithms and Neural Networks (AANN 2022) (Vol. 12285, pp. 232–241). SPIE.

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for NAS. arXiv preprint arXiv:1912.06059.

Liu, H., An, J., Jia, X., Gan, L., Karagiannidis, G. K., Clerckx, B., ... & Cui, T. J. (2025). Stacked intelligent metasurfaces for wireless communications: Applications and challenges. IEEE Wireless Communications, 32(4), 46-53.

Mohammad, S., & Moosavi, S. (2023). Examining StyleGAN as a Utility-Preserving Face De-identification Method. In *Proceedings of Proceedings on Privacy Enhancing Technologies* (Vol. 2023, Issue 4). Association for Computing Machinery. https://doi.org/10.56553/popets-2023-0114

Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-dute, I. S. A., Khan, S., & Parkinson, S. (2023). A Survey on the Detection and Impacts of Deepfakes in Visual , Audio , and Textual Formats. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2023.3344653

Negi, S., Jayachandran, M., & Upadhyay, S. (2021). Deep fake: an understanding of fake images and videos. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 7(3), 183-189.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., , Duc Thanh Nguyen a, Thien Huynh-The c, Saeid Nahavandi d, Thanh Tam Nguyen e, Q.-V. P. f, & G, C. M. N. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image*

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp18-56
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received May 3, 2025
Peer reviewed: December 12, 2025
Revised: February 19, 2026
Published: April 2026

Understanding, *223*. https://doi.org/doi.org/10.1016/j.cviu.2022.103525

Qing, C., Ruan, J., Xu, X., Ren, J., & Zabalza, J. (2019). Spatial-spectral classification of hyperspectral images : a deep learning framework with Markov Random fields based modelling. *IET Image ProcessingVolume*, *13*(2). https://doi.org/10.1049/iet-ipr.2018.5727

Review, A. S. (2023). Digital Face Manipulation Creation and Detection : *Electronics*, *12*(16), 1–37. https://doi.org/10.3390/electronics12163407

Rybnicek, R., & Königsgruber, R. (2019). What makes industry – university collaboration succeed ? A systematic review of the literature. In *Journal of Business Economics* (Vol. 89, Issue 2). Springer Berlin Heidelberg. https://doi.org/10.1007/s11573-018-0916-6

Series, C. (2019). An Overview of Overfitting and its Solutions An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1168/2/022022

Westerlund, M. (2019). The Emergence of Deepfake Technology : A Review. *Technology Innovation Management Review*. https://doi.org/10.22215/timreview/1282

Zhang, J., Tao, C., Xu, Z., Xie, Q., Chen, W., & Yan, R. (2019, July). Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 435–444).