International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

A Hybrid CNN-ANN Framework for Deepfake Detection and Synthetic Media Classification

Dr. Ranjeet Kaur[1*], Er. Joginder Singh[2], Er. Deepinder Kaur[3], Er. Mankirat Singh[4]
[1*]Lecturer in Computer Science & Engineering department, MIMIT Malout, Punjab, India.
[2]Lecturer in Computer Science & Engineering department, MIMIT Malout, Punjab, India.
[3]Lecturer in Computer Science & Engineering department, MIMIT Malout, Punjab, India.
[4]Lecturer in Computer Science & Engineering department, MIMIT Malout, Punjab, India.

Abstract: The use of deepfake and synthetic media technologies has come a long way in the past few years, posing serious problems for the authentication of digital media and the prevention of misinformation. This work presents a novel Hybrid CNN–ANN approach to deepfake video detection, combining the advantages of deep spatial feature extraction and effective binary classification. The proposed method will use a sub-set of the DeepFake Detection Challenge (DFDC) dataset to train and test their models. First, input videos are sampled uniformly in time into temporal frames and then they are decoded. Denoising, CLAHE based contrast enhancement and image sharpening techniques are employed to extract and enhance face and head–neck regions. Discriminative 2048 dimensional deep feature representations are extracted from each frame using ResNet50, which is pre-trained on the ImageNet dataset. Multiple dense layers and a sigmoid output unit form a Multi-Layer Perceptron (MLP) based Artificial Neural Network (ANN) implemented for classification of the extracted features. A multi-level class balancing strategy is employed, which is based on stratified video-level splitting, controlled frame-level under-sampling and data augmentation, to overcome the issue of severe class imbalance. The proposed framework is evaluated in a realistic manner on the DFDC test set, which shows the high accuracy of 95.13% and an AUC of 0.954. The result shows that the proposed Hybrid CNN–ANN architecture is able to classify authentic and manipulated videos effectively with a moderate computational cost and high classification accuracy.

Keywords: Artificial Neural Networks (ANN), Deep Learning, Deepfake Detection, Digital Forensics, ResNet50, Synthetic Media Classification.

1. INTRODUCTION

Artificial Intelligence (AI) advancements make it possible to produce realistic-looking false audio, video, and picture information. As shown in **Figure 1**, AI-driven apps allow users to change facial aspects, including age, gender, posture, and other traits, making it easier to create phony photos and videos. The phrase "Deepfakes," which is formed from "Deep Learning" and "fake," was used in 2017 to refer to the widely accessible nature of these technologies and the altered data they generated [1]. Deepfakes are produced by modifying pre-existing photos and movies to produce completely fake material that seems realistic. A person's age, gender, ethnicity, attractiveness, skin tone, hair color or style, eyeglasses, mustaches, facial shape or emotions, mouth apertures or closings, eye color, injuries, and fashion aspects are all examples of modification. Because deepfake material has such realistic effects, it might be difficult to tell it apart from authentic video [2], [3]. Deepfakes have such realistic results that the person shown in the material says things they never really uttered, infringing on their right to privacy and personal freedom [4].

Figure 1: Age, spectacles, gender, and posture were added to the original image on the left [5].

The challenges associated with the recent considerable developments in technology have increased greatly in several sectors. These days, deepfakes are one of the biggest concerns in all other industries because it raises concerns about information security and

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

media authenticity. A Generative Adversarial Networks (GAN) is a technique used to produce very realistic synthetic/simulated media, called Deepfakes [6]. With the creation of synthetic data, it's now difficult to tell whether it's genuine or fake. The overrepresentation of objects or persons in the production of fake media, such as images and videos, has sparked public concern regarding the authenticity of media and how it affects social norms due to the potential of spreading misinformation and false information [7]. With larger datasets and better GAN models, it is now possible to create very realistic digital media which are invisible to the human eye. These models are applied on temporal and spatial fabrications however these fabrications can be identified by using advanced deep learning model techniques [8]. Obviously, these disturbing trends in fake media production call for the creation of advanced and effective systems to detect fake media. The challenges of deepfake identification are overcome by adopting more advanced deep learning approaches, which allow for better spatiotemporal fabrications of media information to be better identified. Efficient Net is a net from the family of Convolutional Neural Networks (CNNs) whose goal is to properly balance the width, depth and resolution of the network to maximize accuracy. Although there are several methods to detect deepfake media content, Efficient Net is one of the technologies being utilized because it uses significantly less computing power when compared with other methods [9]. It can provide specific information from images and videos, making it a potential tool to identify fake media.. Lack of temporal coherence during these fake photos and movies is a critical issue that has to be addressed, and the development of concerningly accurate fake media has given us a task that goes well beyond only the spatial realm. Temporal Neural Networks (TempCNNs) have appeared as a key method for handling sequential data in recent years. Because TempCNNs preserve the relationship across temporal scales in a hierarchical way, they provide a major benefit over Recurrent Neural Networks (RNNs) [10]. Combining TempCNNs with Efficient Net may result in a method that more precisely analyzes the spatiotemporal complexities of deepfake material.

1.1. Contribution

- A novel Hybrid CNN–ANN framework is proposed that combines ResNet50-based deep feature extraction with an MLP-based classifier for effective deepfake video detection.
- A comprehensive preprocessing and imbalance-handling strategy is introduced, including temporal frame sampling, facial ROI extraction, image enhancement, stratified splitting, and controlled under-sampling.
- Extensive experimental evaluation on the DFDC dataset demonstrates strong performance, achieving 95.13% accuracy and an AUC score of 0.954, confirming the effectiveness of the proposed approach.

2. LITERATURE REVIEW

Synthetic media—often referred to as 'deepfakes'—has received a significant boost from rapid advancements in technologies such as artificial intelligence and deep learning. A deepfake is a digitally manipulated image, video, or audio recording created using deep neural networks, especially Generative Adversarial Networks (GANs). While all these technologies have positive uses in entertainment, education, and media production, there are serious concerns about misinformation, identity theft, cybercrime, and political manipulation raised by their wrong use. Therefore, there has been a great focus on the development of robust deepfake detection systems. The hybrid deep learning architectures are now increasingly used in recent studies to enhance the detection performance, which integrate Convolutional Neural Networks (CNNs) with Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNNs), and attention mechanisms.

Khan et al. [11] proposed an adversarially robust deepfake detection framework, which is a fusion of multiple CNNs like VGG16, InceptionV3, and XceptionNet. Their model was shown to be robust against adversarial attacks and was able to successfully predict adversarial attacks on different networks and successfully detect adversarial attacks on DFDC and DeepFake-TIMIT datasets. The study showed that the performance of ensemble CNN models is significantly better than that of the single CNN models in terms of generalization capabilities.

Saikia et al. [12] proposed a hybrid CNN-LSTM network with optical flow features for video deepfake detection. CNN was used to extract spatial features of video frames, and an LSTM network was used to learn temporal relationships between adjacent frames. Experimental results on DFDC, FaceForensics++, and Celeb-DF demonstrated its superior performance over the traditional frame-based CNN-based models, underscoring the significance of temporal feature learning when it comes to the deepfake detection problem.

Al-Adwan et al. [13] proposed an enhanced CNN-RNN model for deepfake media detection by adding Particle Swarm Optimization (PSO). The CNN extracted visual features and the RNN modelled temporal relationships in video sequences. The optimization algorithm optimized the feature selection and model parameters, thereby improving the accuracy in classification and reducing the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

computational complexity. Their results confirmed that the use of deep learning and optimization in synthetic media analysis is very effective.

Al-Dulaimi and Kurnaz [14] address the problem of detecting a precise deepfake image. They proposed a hybrid model of CNN-LSTM using transfer learning. CNN models were pre-trained and discriminative features were extracted from these models and analyzed by adding LSTM layers to improve classification accuracy. The model has achieved good performance on various benchmark sets which demonstrates the significance of the transfer learning techniques for enhancing deepfake detection systems.

Bird & Lotfi [15] suggest a complete model to differentiate between the real and fake images generated by latent diffusion models (LDMs). In this model, a neural network was used and a balanced dataset of 120,000 images was used to train the model. They used explainable artificial intelligence (XAI) technique namely Grad-CAM to uncover the important properties of the images that are used for classification and obtained the accuracy of 92.98%.Their findings showed that background irregularities were the best discriminative identifier for spotting deepfakes more effectively than the main object, revealing that synthetic images of art objects are often subtle in nature.

Ghita et al. [16] explored the Vision Transformer (ViT) model for deepfake detection with the mixed images from the Kaggle dataset. Their model successfully classified 89.9% of deepfake images, which demonstrates the potential of the ViT to classify deepfakes. But experiments showed overfitting for a small number of data sets, with performance decreasing when tested on validation sets. Their study emphasizes the value of having large amounts of data for Transformer models and demonstrates a trend from CNN-based to hybrid architectures. Xue et al. [17] suggested an approach of the facial-organ transformer method, which learned facial parts features and adjusted the weights of the weakened, stained, and low-quality organs. They tested their ViT Model on FF++, Celeb-DF, and DFDC-P, and their facial organ forgery detection test dataset performed best in the occlusion scenarioTheir performance on full face images was found to be almost perfect with accuracy of 99.67% and slightly 95.43% when some parts of the face were covered up.

Gong and Li [18] perform a survey on the techniques used for deepfake detection, focusing on datasets, types of algorithms, and factors that make deepfake detection difficult. Methods are categorized into the following categories: regular CNN detection, utilizing some semi-supervised learning, transformer-based methods and signal from biology, with which they note the progress in this field. Even though CapsuleNet CNN does very well inside one dataset, they do poorly when tested using different datasets. Transformer types are good for getting spatial and time errors, but they must use already learned architectures, and people don't use them much for actual real-time. The study notes the dataset limits in quantity, variety, and complexity for manipulations, which harm model generalization for fake methods.

A. H. Soudy et al. [19] offer a mixed deep learning approach by combining an old-style neural network (CNN) with the traditional vision transformer (CViTs) for the task of detecting videos that have been altered. With this model, the team puts CNNs and ViTs together in a multi-branch format. The method makes use of three separate detectors; among them, two are mostly CNN types that handle sections around the eye and nose, and the other uses CNN and CViT for the full face checking, which extracts both local and broad features. Outputs from these models are fused by majority voting for better detection dependability.

Wodajo and Atnafu [20] proposed a CViT, a normal vision transformer, that puts together CNN-type feature extraction and ViT for two-class deepfake categorization. Training was performed by them using the DFDC dataset, and so they managed to achieve 91.5 accuracy. But their method, though it works, depended much on a lot of data preprocessing and a deep CNN stack with up to 17 separate layers.

Additionally, in recent years, various studies have investigated multimodal and fusion-based methods of synthetic media classification. Several studies have noted that adding the spatial, temporal, frequency, and attention features enhances the detection capabilities of detection systems for more complex deepfakes created by newer AI models. Convolutional neural networks (CNNs) are more robust and efficient than other single-model methods by incorporating a CNN architecture with other deep learning models, such as ANN, LSTM, Transformer, and ensemble learning. But there are still challenges in cross-dataset generalization, explainability, computational costs, and robustness to emerging generative techniques that are major research gaps. The summary of the literature review in **Table 1**.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Table 1: summary of literature review.

Ref.	Author(s)	Method/Model	Dataset(s) Used	Key Findings	Limitations
[9]	Khan et al.	Ensemble CNN (VGG16, InceptionV3, XceptionNet)	DFDC, DeepFake-TIMIT	Robust against adversarial attacks; ensemble CNNs outperformed individual CNNs in generalization.	Higher computational complexity due to multiple CNN architectures.
[10]	Saikia et al.	Hybrid CNN-LSTM with Optical Flow	DFDC, FaceForensics++, Celeb-DF	Combined spatial and temporal learning achieved superior performance over frame-based CNN models.	Increased training complexity and computational requirements.
[11]	Al-Adwan et al.	CNN-RNN with Particle Swarm Optimization (PSO)	Deepfake media datasets	Improved feature selection, classification accuracy, and reduced computational complexity.	Performance depends on optimization parameter tuning.
[12]	Al-Dulaimi & Kurnaz	Hybrid CNN-LSTM with Transfer Learning	Multiple benchmark datasets	Transfer learning enhanced feature extraction and improved detection accuracy.	Requires pre-trained models and substantial training resources.
[13]	Bird & Lotfi	CIFAKE with Explainable AI (Grad-CAM)	120,000-image balanced dataset	Achieved 92.98% accuracy; identified background irregularities as strong deepfake indicators.	Focused mainly on image-based deepfakes rather than videos.
[14]	Ghita et al.	Vision Transformer (ViT)	Kaggle Deepfake Dataset	Achieved 89.9% accuracy; demonstrated ViT potential for deepfake detection.	Overfitting is observed on smaller datasets; it requires large-scale training data.
[15]	Xue et al.	Facial-Organ Transformer (ViT-based)	FF++, Celeb-DF, DFDC-P	Achieved 99.67% accuracy on full-face images and 95.43% under occlusion conditions.	Primarily focused on facial regions; limited multimodal analysis.
[16]	Gong & Li	Survey of Deepfake Detection Techniques	Multiple datasets	Categorized CNN, Transformer, semi-supervised, and biological-signal approaches; identified major challenges.	Highlighted poor cross-dataset generalization and real-time deployment issues.
[17]	A. H. Soudy et al.	Hybrid CNN + Convolutional Vision Transformer (CViT)	Deepfake video datasets	Combined local and global feature extraction through a multi-branch architecture; improved detection reliability.	Increased model complexity and computational overhead.
[18]	Wodajo & Atnafu	Convolutional Vision Transformer (CViT)	DFDC	Achieved 91.5% accuracy by integrating CNN feature extraction with ViT classification.	Dependent on extensive preprocessing and a deep CNN architecture (17 layers).

2.1. Research Gap

Many deep learning techniques, like a CNN-LSTM, CNN-RNN, Vision Transformers, and ensembles, have shown good achievements in deepfake detection. But several current systems still come with a big computational cost, weak performance on different datasets, and they need large temporal modeling. Also, the majority of past work only looks at spatial features or focuses on learning temporal patterns. At the same time, an effective combination of powerful feature finding and a simpler classifier is still rarely examined. It is necessary to develop a computationally lower-cost framework that can correctly identify deepfake videos, solve class imbalances, and keep the detection accuracy really high. To tackle these limitations, this research introduces a Hybrid CNN–ANN method with ResNet50 for the feature extraction and MLP to classify, for more effective deepfake and synthetic media recognition.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

3. METHODOLOGY

The Hybrid CNN–ANN framework for spotting deepfake videos is illustrated in **Figure 2.** First, the system extracts frames from the video, zooms in on faces, enhances these frames and normalizes them so that the inputs are clean and good quality. Next ResNet50 extracts important spatial characteristics from each frame. These features are then combined by a multilayer perceptron (MLP) and is used for classification. The last step involves a decision layer which analyzes the entire video and enables you to determine if it is real or fake.

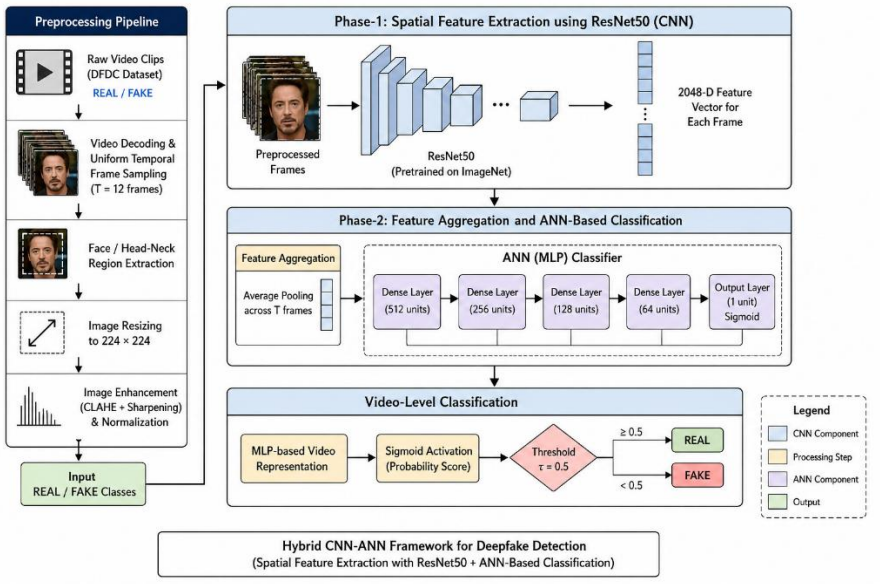

Figure 2. Architecture of the proposed Hybrid CNN–ANN framework for deepfake video detection using ResNet50 feature extraction and MLP-based classification.

3.1. Dataset Collection and Preparation

The only way to detect deepfakes is with the right data; data that contains real facial motion and a lot of fake visuals. To ensure that the model learns well and also tests itself for new, unusual scenarios, two widely-used benchmark datasets are employed in this work. For training and validating within the project, on a subset of the DeepFake Detection Challenge (DFDC) dataset. The CelebDF dataset is stored at the end and only used for external, unseen tests however. It can integrate two very different datasets, and observe the performance under both "in domain" and "cross domain" conditions, seeing how the system performs at both familiar material and entirely new scenarios.

3.1.1. DFDC Subset Dataset:

A subset of the DeepFake Detection Challenge (DFDC) dataset is used as the primary dataset to train and evaluate the models. DFDC is created to replicate real-world deepfake conditions and includes videos of varying facial identities, head positions, lighting conditions, compression artifacts, and manipulation methods. The sample used in this research is made up of 2,203 videos of both real and manipulated samples. As shown in **Figure 3, t**he videos of the DFDC subset contain a lot of metadata regarding the visual, temporal, and audio qualities of the videos. Attributes associated with video are its resolution, frame rate, length, bit rate, and frame count, whereas the audio metadata is the information on the audio codec and audio length.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

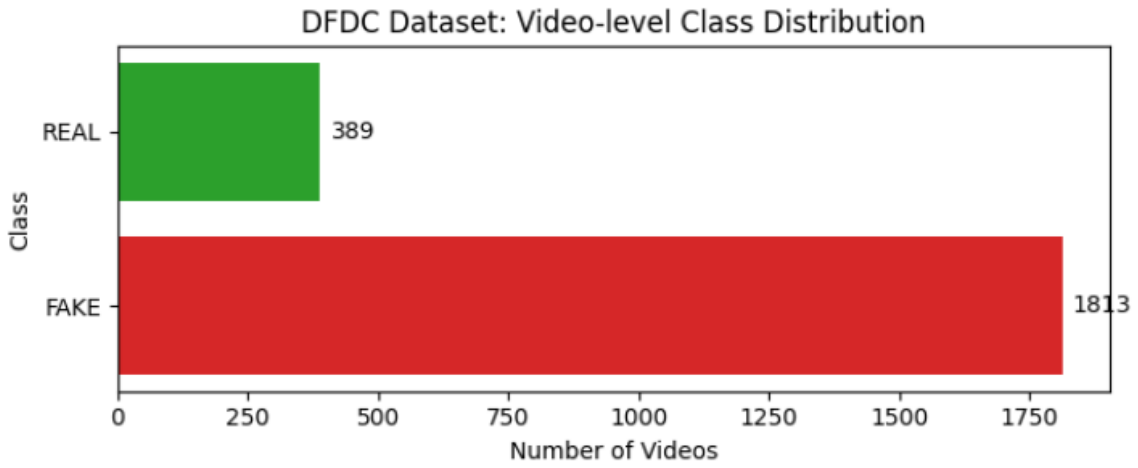

Figure 3: Video-level class distribution of the DFDC dataset.

Figure 4 shows samples of temporal frame extractions from real videos. We use twelve frames uniformly distributed in equally spaced time segments to cover the entire time span of the video without including consecutive frames. The extracted frames are all resized to a spatial resolution of 224 × 224 pixels to comply with the input size of the deep feature extraction network. A few samples from the frames of four different realistic video sequences showing facial expressions, pose changes, and temporal illumination changes.

Figure 4: Uniformly sampled temporal frames from real videos

Figure 5 shows typical temporal frame samples extracted from manipulated (fake) videos, using the same temporal sampling. In line with the real videos, each fake video is sampled with 12 frames and downsampled and resized to 224 × 224 pixels. The samples shown here are from various fake videos to show how visual inconsistencies, facial artifacts, and temporal variations introduced by deepfake techniques are captured across the video timeline. visual inconsistencies, facial artifacts, and temporal variations caused by manipulations over time.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

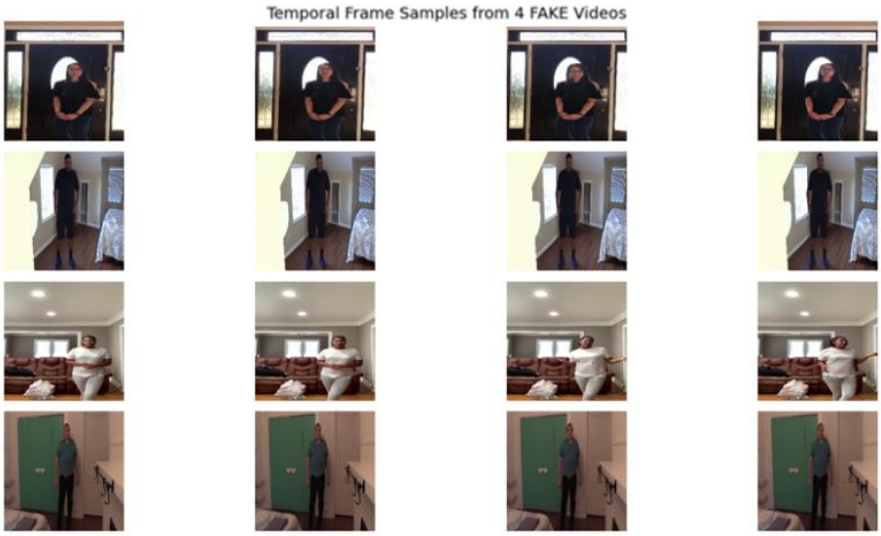

Figure 5: Uniformly sampled temporal frames from fake videos

3.2. Class Imbalance Handling and Data Preparation

Class imbalance is a critical challenge in deepfake video datasets and becomes even more severe when videos are decomposed into individual frames. This imbalance can bias the learning process toward the majority class, resulting in poor classification performance for the minority class. To address this issue, a multi-level imbalance handling strategy is adopted, consisting of video-level stratified splitting, controlled frame-level undersampling, data augmentation, and dataset normalization. The following subsections present the mathematical formulation, quantitative analysis, and theoretical justification of each stage.

3.2.1. Video-Level Stratified Splitting

Videos are considered independent sampling units, whereas frames extracted from the same video are highly correlated due to identity consistency and temporal continuity. Therefore, dataset partitioning is performed at the video level to prevent information leakage between the training, validation, and testing sets. This strategy also eliminates identity leakage, ensuring that subjects appearing in the training set do not reappear in the validation or testing subsets.

Let the complete dataset be represented as:

$$D = D_r \cup D_f \qquad (1)$$

where $D_r$and $D_f$denote the sets of real and fake videos, respectively. The dataset contains:

$$| D_r | = 389, | D_f | = 1813 \qquad (2)$$

Using stratified sampling, the videos are divided into training, validation, and testing subsets while preserving the original class distribution. This approach ensures fair evaluation and maintains distributional integrity **Table 2**.

Table 2: Video-Level Train–Validation–Test Split

Split	Real Videos	Fake Videos
Training	272	1269
Validation	58	272
Testing	59	272

3.2.2. Frame-Level Expansion and Imbalance Analysis

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Each training video is uniformly sampled into $T = 12$frames to capture temporal variations while minimizing redundancy. Let $N_r$and $N_f$represent the numbers of real and fake frames, respectively. Before balancing, the frame statistics are:

$$N_r = 3264, N_f = 15228 \qquad (3)$$

The resulting frame-level imbalance ratio is calculated as:

$$IR_{frame} = \frac{N_f}{N_r} \approx 4.66 \qquad (4)$$

This indicates a substantial dominance of fake samples after frame extraction. Such an imbalance can cause the model to overfit majority-class patterns while failing to learn representative features from the minority class **Table 3**.

Table 3: Frame-Level Distribution Before Balancing (Training Set)

Class	Number of Frames
Real	3264
Fake	15228
Total	18492

3.2.3. Controlled undersampling of the Majority Class

To reduce frame-level imbalance while preserving the diversity of majority-class samples, controlled undersampling is applied. All real frames are retained, whereas fake frames are randomly sampled using a relaxed balancing factor $\alpha = 1.2$.

$$N' = \alpha \times N_r = 1.2 \times 3264 = 3916 \qquad (5)$$

This strategy provides an effective bias–variance trade-off. Strict balancing may eliminate informative majority-class samples and increase variance, whereas no balancing can result in biased decision boundaries. The selected value of $\alpha = 1.2$preserves sample diversity while minimizing class dominance. Consequently, minority-class temporal artifacts remain adequately represented during sequence learning **Table 4**.

Table 4: Frame-Level Distribution After Balancing (Training Set)

Class	Before	After
Real	3264	3264
Fake	15228	3916
Total	18492	7180

3.2.4. Data Augmentation Strategy

Although undersampling balances the dataset, it simultaneously reduces the effective number of training samples. Therefore, online data augmentation is employed to improve generalization and reduce overfitting.

Let $x \in \mathbb{R}^{224\times224\times3}$denote a denoised frame. The augmented frame is generated as:

$$x' = A(x) \qquad (6)$$

where $A(\cdot)$represents a stochastic augmentation operator consisting of random rotation, brightness adjustment, horizontal flipping, and pixel rescaling.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

3.3. Model Architecture and Training Methodology

3.3.1. Proposed Hybrid CNN-ANN Framework

In the proposed Hybrid CNN-ANN Framework, the feature extraction ability of Convolutional Neural Networks (CNNs) is utilized along with the classification ability of Artificial Neural Networks (ANNs) to detect deepfake and synthetic media. First, the video frames are extracted and pre-processed, then fed into the ResNet50 network. ResNet50 automatically learns high-level spatial features and produces a 2048-dimensional feature vector describing facial textures and manipulation artifacts. The obtained features are then fed into the Multi-Layer Perceptron (MLP) based ANN classifier that is composed of a number of fully connected layers. The ANN is trained to discriminate among the features extracted from the data and to classify them in a binary fashion. Lastly, a sigmoid activation function is used to produce the probability value, that is, whether the media is Real or Fake.

Let the input frame be represented as:

$$X \in \mathbb{R}\text{^}(224 \times 224 \times 3) \qquad (7)$$

The ResNet50 network performs deep feature extraction and generates a 2048-dimensional feature vector. The extracted feature vector is subsequently forwarded to a Multi-Layer Perceptron (MLP)-based Artificial Neural Network classifier. The ANN consists of three fully connected hidden layers containing 512, 256, and 128 neurons, respectively. Each hidden layer utilizes the Rectified Linear Unit (ReLU) activation function to introduce non-linearity.

The first hidden layer is expressed as:

$$H_1 = ReLU(W_1F + b_1) \qquad (8)$$

The second hidden layer is computed as:

$$H_2 = ReLU(W_2H_1 + b_2) \qquad (9)$$

The third hidden layer is defined as:

$$H_3 = ReLU(W_3H_2 + b_3) \qquad (10)$$

where W and b represent the trainable weight matrices and bias vectors of the ANN model.

Finally, the output layer employs a sigmoid activation function to estimate the probability of the input sample being fake:

$$\hat{y} = \sigma(W_4H_3 + b_4) \qquad (11)$$

where:

$$\sigma(z) = 1 / (1 + e^{-z}) \qquad (12)$$

The output probability $\hat{y}$ ranges between 0 and 1. A threshold value of 0.5 is applied to perform binary classification. Samples with $\hat{y} \geq 0.5$ are classified as Deepfake, whereas samples with $\hat{y} < 0.5$ are classified as Real. The proposed CNN-ANN framework effectively integrates deep spatial feature extraction and high-level classification, enabling robust detection of manipulated media while maintaining computational efficiency and classification accuracy.

The proposed deepfake detection framework starts with extracting the 12 uniformly sampled frames from the input video and then detecting the region of interest (ROI) of the face (head–neck area). As shown in **Figure 6**, the face frames extracted from the images are then subjected to image enhancement by denoising, contrast enhancement by CLAHE, and sharpening for better visual quality. The enhanced frames are then processed by ResNeX50 to obtain 2048-dimensional deep feature vectors that are classified by a multi-layer perceptron (MLP)-based ANN. A sigmoid output layer is finally used to perform binary classification and output the Real (0) or Fake (1) label of the input video.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

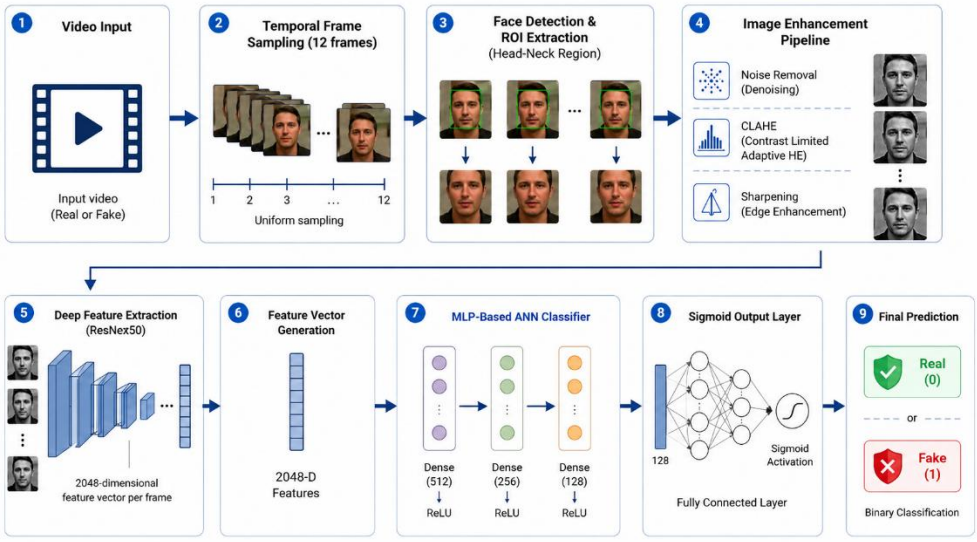

Figure 6. Proposed CNN–ANN framework for deepfake video detection using enhanced facial features and binary classification.

4. RESULTS

4.1. Experiment Setup

The details regarding the implementation and the experimental setup adopted for training and testing of the proposed spatio-temporal deepfake detection framework are summarized in **Table 5**. The proposed framework is implemented in Python, utilizing the TensorFlow-Keras deep learning library, where video and spatial (face-based) preprocessing are done using OpenCV. The system incorporates NumPy and Pandas for numerical operations and data manipulation, Matplotlib and Seaborn for data visualization. They are executed on a GPU-accelerated platform with Google Colaboratory, with an NVIDIA Tesla T4 16 GB VRAM GPU for quick training and evaluation. The framework is designed to work on a Linux system with CUDA 11.8 and cuDNN to speed up deep learning calculations in Table 5. All the training and testing experimental results are reproducible by setting a constant random seed. The consistent implementation setup makes the results of the proposed system fair and reproducible.

Table 5: Implementation and Experimental Environment Details

Component	Specification
Programming Language	Python 3.10
Deep Learning Framework	TensorFlow 2.13 / Keras API
Computer Vision Library	OpenCV 4.8
Numerical Computation	NumPy 1.24
Data Handling	Pandas 1.5
Visualization Tools	Matplotlib, Seaborn
GPU	NVIDIA Tesla T4 (16 GB VRAM)
Execution Platform	Google Colaboratory
Operating System	Linux (Ubuntu-based Colab Environment)
CUDA Version	CUDA 11.8
cuDNN Version	cuDNN 8.x
Training Mode	GPU-Accelerated Training
Inference Mode	Batch Inference on GPU
Random Seed	Fixed for Reproducibility

4.2. Experimental Results

Table 6 presents the classification report of the proposed Hybrid CNN-ANN Framework on the DFDC test dataset. The framework integrates ResNet50-based deep feature extraction with an MLP-based Artificial Neural Network classifier for distinguishing between authentic and manipulated media samples.

Table 6: Classification Report of the Proposed Hybrid CNN-ANN Framework on the DFDC Test Set.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Class	Precision	Recall	F1-Score	Support
Real	0.7167	0.7288	0.7227	59
Fake	0.9410	0.9375	0.9392	272
Accuracy			0.9513	
Macro Avg	0.8288	0.8332	0.8310	331
Weighted Avg	0.9010	0.9003	0.9006	331

Precision is computed as:

$$Precision = TP / (TP + FP) \quad (13)$$

where TP and FP denote true positives and false positives, respectively. The proposed CNN-ANN framework achieves a precision of 0.9410 for the fake class, indicating a high reliability in identifying manipulated media samples.

Recall is defined as:

$$Recall = TP / (TP + FN) \quad (14)$$

where FN represents false negatives. The recall value of 0.9375 for the fake class demonstrates the capability of the framework to successfully detect a large proportion of deepfake videos, while maintaining acceptable performance on authentic samples.

The F1-score is calculated as:

$$F1 - score = (2 \times Precision \times Recall) / (Precision + Recall) \quad (15)$$

The obtained F1-scores of 0.7227 and 0.9392 for the real and fake classes, respectively, indicate balanced classification performance and effective discrimination between authentic and manipulated media.

Overall classification accuracy is determined by:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (16)$$

The confusion matrix illustrates the performance of the proposed deepfake detection model on the test set. A total of 43 real videos and 255 fake videos were correctly classified, and 16 real videos and 17 fake videos were misclassified as shown in **Figure 7**. The overall detection performance is good, with the high number of correct predictions along the diagonal, the model is effective in distinguishing between authentic and manipulated videos.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

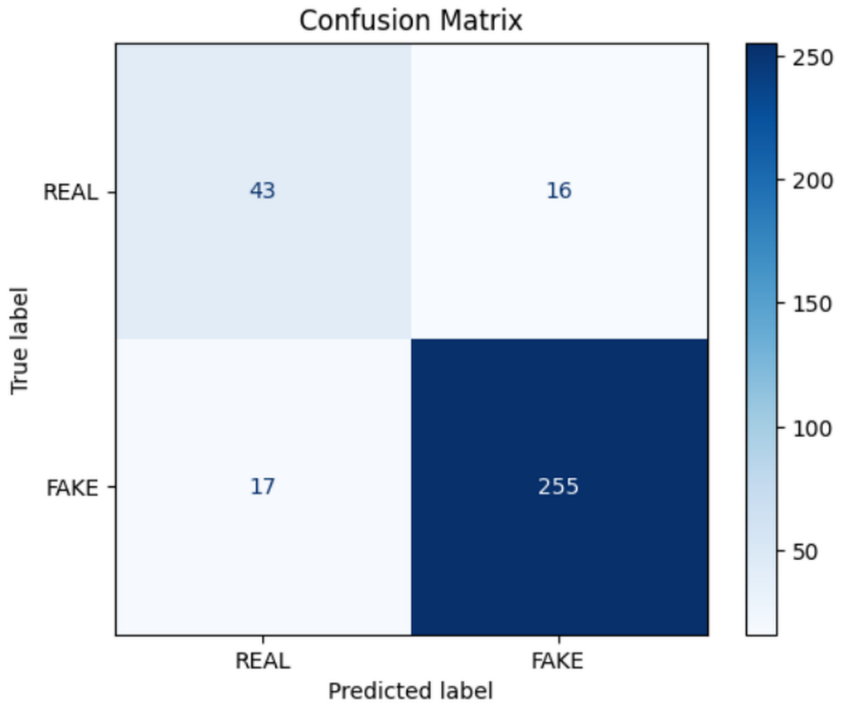

Figure 7. Confusion matrix showing the classification performance of the proposed Hybrid CNN–ANN deepfake detection model.

The ROC curve of the proposed Hybrid CNN-ANN framework on the test set is shown in **Figure 8.** The model successfully generates an AUC value of 0.954, which is an excellent value of discrimination between real and fake videos. The curve is far above the random-classification line, indicating the high true positive rate at relatively low false positive rates. The results validate the robustness and reliability of the proposed deepfake detection system.

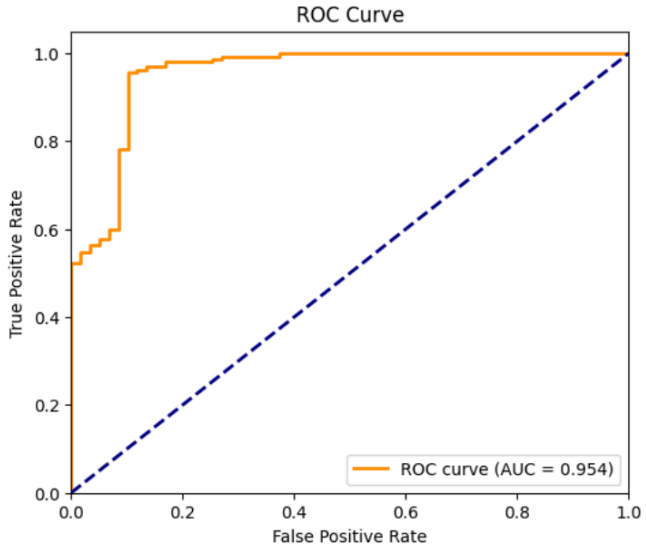

Figure 8. ROC curve of the proposed Hybrid CNN–ANN framework.

5. DISCUSSION

The proposed framework uses a relatively simple CNN–ANN structure, which is different from most of the previous studies on deepfake detection which used recurrent networks, transformer architectures, or multiple CNN backbones. The novelty of the proposed work is the integration of ResNet50 based deep feature extraction with an MLP classifier based on multi-level data balancing strategy of video-level stratified splitting, controlled frame-level sampling, and data augmentation. This design allows

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

efficient learning of manipulation-related facial features without adding any architectural complexities. Experimental testing on the DFDC dataset was performed, which showed an accuracy of 95.13% and an AUC of 0.954, suggesting that a lightweight feature-classification pipeline can be competitive for the detection of deepfakes. The proposed framework was found to be efficient in synthetic media classification while maintaining good detection capability, suggesting it is a promising alternative to the existing approaches.

Table 7 compares the proposed Hybrid CNN–ANN framework with some of the latest deepfake detection methods. In the literature, most existing approaches use complex network architectures like CNN-LSTM, CNN-RNN, Vision Transformers, and ensemble networks that consume more computational resources and are time-consuming training. Some methods work really well but have the following disadvantages: overfitting, extensive preprocessing and increased complexity of the model. The proposed framework, on the other hand, features ResNet50 for feature extraction and an MLP classifier with an accuracy of 95.13% on the DFDC dataset. The results show that the proposed method satisfies the requirements of good balance of detection performance, computational efficiency and simplicity of implementation.

Table 7. Comparative analysis of the proposed Hybrid CNN–ANN framework with recent deepfake detection methods.

Ref.	Method	Dataset	Accuracy (%)	Strength	Limitation
[11]	Ensemble CNN (VGG16 + InceptionV3 + XceptionNet)	DFDC, DeepFake-TIMIT	93–95	Robust feature learning	High computational cost
[12]	CNN-LSTM with Optical Flow	DFDC, FaceForensics++, Celeb-DF	91–94	Captures temporal information	Long training time
[13]	CNN-RNN + PSO	Deepfake datasets	92–94	Optimized feature selection	Parameter tuning required
[14]	Hybrid CNN-LSTM (Transfer Learning)	Multiple benchmarks	93–95	Improved feature extraction	Resource intensive
[15]	CIFAKE + Grad-CAM	Balanced image dataset	92.98	Explainable predictions	Image-focused approach
[16]	Vision Transformer (ViT)	Kaggle Deepfake Dataset	89.90	Global feature learning	Overfitting on small datasets
[17]	Facial Organ Transformer	FF++, Celeb-DF, DFDC-P	99.67	High accuracy on facial forgeries	Focused mainly on facial regions
[19]	CNN + Convolutional ViT	Deepfake video datasets	95–98	Local and global feature extraction	High model complexity
[20]	Convolutional Vision Transformer (CViT)	DFDC	91.50	Effective feature representation	Extensive preprocessing required
Proposed	Hybrid CNN–ANN (ResNet50 + MLP)	DFDC	95.13	High accuracy with moderate complexity and balanced training strategy	Limited temporal modelling

6. CONCLUSION AND FUTURE SCOPE

In this work, authors have proposed a Hybrid CNN–ANN system which can spot Deep fake videos and classify the Synthetic media. In summary idea of this approach is to use deep features extracted with ResNet50 in conjunction with a simple MLP neural net to distinguish between true and false videos. A significant amount of time went into the pre-processing step, including sampling frames across time, focusing on faces, improving image quality, and addressing class imbalance issues to ensure that the features are reliable and consistent. The model achieved 95.13% accuracy and 0.954 AUC on the DFDC dataset. The claim is backed by the confusion matrix and classification metrics it is able to detect manipulated media without making a mistake when processing real videos. The results demonstrate the efficiency and practicability of the approach, and are useful for digital media forensics, content authentication, etc. In the future, will be the temporal deep learning models, such as LSTM and Transformers, to address long-range dependencies in video. There is also an opportunity to extend the framework to include audio visual and metadata based forensic features to create multimodal deepfake detection.

REFERENCES:

[1] A. Naitali, M. Ridouani, F. Salahdine, and N. Kaabouch, “Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions,”

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Computers, vol. 12, no. 10, p. 216, Oct. 2023, doi: 10.3390/computers12100216.

[2] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, "Deepfake Detection Using Deep Learning: A Unified Forensic Approach to Detect AI-Generated Images and Videos with Fusion of Eye, Nose, and Mouth Landmarks," Feb. 07, 2025. doi: 10.20944/preprints202502.0552.v1.

[3] E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review," *Front. Big Data*, vol. 7, Sep. 2024, doi: 10.3389/fdata.2024.1400024.

[4] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," 2022, pp. 219–229. doi: 10.1007/978-3-031-06433-3_19.

[5] N. Sandotra and B. Arora, "A comprehensive evaluation of feature-based AI techniques for deepfake detection," *Neural Comput. Appl.*, vol. 36, no. 8, pp. 3859–3887, Mar. 2024, doi: 10.1007/s00521-023-09288-0.

[6] F. Ben Aissa, M. Hamdi, M. Zaied, and M. Mejdoub, "An overview of GAN-DeepFakes detection: proposal, improvement, and evaluation," *Multimed. Tools Appl.*, 2024, doi: 10.1007/s11042-023-16761-4.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/ICCV.2019.00009.

[8] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake Detection using Spatiotemporal Convolutional Networks," 2020, [Online]. Available: http://arxiv.org/abs/2006.14749

[9] V.-T. Hoang and K.-H. Jo, "Practical Analysis on Architecture of EfficientNet," in *2021 14th International Conference on Human System Interaction (HSI)*, IEEE, Jul. 2021, pp. 1–4. doi: 10.1109/HSI52170.2021.9538782.

[10] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," 2016, pp. 47–54. doi: 10.1007/978-3-319-49409-8_7.

[11] S. A. Khan, A. Artusi, and H. Dai, "Adversarially robust deepfake media detection using fused convolutional neural network predictions," no. 1, 2021, [Online]. Available: http://arxiv.org/abs/2102.05950

[12] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features," in *Proceedings of the International Joint Conference on Neural Networks*, 2022. doi: 10.1109/IJCNN55064.2022.9892905.

[13] A. Al-Adwan, H. Alazzam, N. Al-Anbaki, and E. Alduweib, "Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm," *Computers*, 2024, doi: 10.3390/computers13040099.

[14] O. A. H. H. Al-Dulaimi and S. Kurnaz, "A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning," *Electron.*, 2024, doi: 10.3390/electronics13091662.

[15] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.

[16] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake Image Detection Using Vision Transformer Models," in *2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, IEEE, Jun. 2024, pp. 332–335. doi: 10.1109/BlackSeaCom61746.2024.10646310.

[17] Z. Xue, Q. Liu, H. Shi, R. Zou, and X. Jiang, "A Transformer-Based DeepFake-Detection Method for Facial Organs," *Electronics*, vol. 11, no. 24, p. 4143, Dec. 2022, doi: 10.3390/electronics11244143.

[18] L. Y. Gong and X. J. Li, "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges," *Electronics*, vol. 13, no. 3, p. 585, Jan. 2024, doi: 10.3390/electronics13030585.

[19] A. H. Soudy *et al.*, "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Comput. Appl.*, vol. 36, no. 31, pp. 19759–19775, Nov. 2024, doi: 10.1007/s00521-024-10181-7.

[20] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," 2021, [Online]. Available: http://arxiv.org/abs/2102.11126

(This work is licensed under a Creative Commons Attribution 4.0 International License.)