electronics

MDPI

Article

Real-Time Deepfake Video Detection Using Eye Movement Analysis with a Hybrid Deep Learning Approach

Muhammad Javed [1], Zhaohui Zhang [1,*], Fida Hussain Dahri [2] and Asif Ali Laghari [3,*]

[1] Department of Computer Science and Technology, College of Computer Science, Donghua University, Shanghai 200022, China; 421017@mail.dhu.edu.cn
[2] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China; 223227084@seu.edu.cn
[3] Software College, Shenyang Normal University, Shenyang 110136, China
* Correspondence: zhzhang@dhu.edu.cn (Z.Z.); asif.laghari@smiu.edu.pk (A.A.L.)

Abstract: Deepfake technology uses artificial intelligence to create realistic but false audio, images, and videos. Deepfake technology poses a significant threat to the authenticity of visual content, particularly in live-stream scenarios where the immediacy of detection is crucial. Existing Deepfake detection approaches have limitations and challenges, prompting the need for more robust and accurate solutions. This research proposes an innovative approach: combining eye movement analysis with a hybrid deep learning model to address the need for real-time Deepfake detection. The proposed hybrid deep learning model integrates two deep neural network architectures, MesoNet4 and ResNet101, to leverage their respective architectures' strengths for effective Deepfake classification. MesoNet4 is a lightweight CNN model designed explicitly to detect subtle manipulations in facial images. At the same time, ResNet101 handles complex visual data and robust feature extraction. Combining the localized feature learning of MesoNet4 with the deeper, more comprehensive feature representations of ResNet101, our robust hybrid model achieves enhanced performance in distinguishing between manipulated and authentic videos, which cannot be conducted with the naked eye or traditional methods. The model is evaluated on diverse datasets, including FaceForensics++, CelebV1, and CelebV2, demonstrating compelling accuracy results, with the hybrid model attaining an accuracy of 0.9873 on FaceForensics++, 0.9689 on CelebV1, and 0.9790 on CelebV2, showcasing its robustness and potential for real-world deployment in content integrity verification and video forensics applications.

Keywords: Deepfake detection; real time; hybrid model; ResNet101; MesoNet4; video forensics

Citation: Javed, M.; Zhang, Z.; Dahri, F.H.; Laghari, A.A. Real-Time Deepfake Video Detection Using Eye Movement Analysis with a Hybrid Deep Learning Approach. *Electronics* **2024**, *13*, 2947. https://doi.org/10.3390/electronics13152947

Academic Editor: George A. Tsihrintzis

Received: 14 June 2024
Revised: 19 July 2024
Accepted: 22 July 2024
Published: 26 July 2024

1. Introduction

The growth of mobile devices and social networks in recent years has significantly enhanced the attraction and utilization of videos [1]. According to a survey from the premier digital marketing business, the internet sees over 1 billion hours of video being viewed daily [2]. Subsequently, an extensive display of videos came with intelligent methods and tools for manipulating video material, such as Deepfake and FaceSwap [3]. The term 'Deepfake' originated from combining 'Deep Learning' (DL) and 'Fake.' It refers to creating very realistic videos or images with the assistance of deep learning technology. Deepfakes refer to manipulated photos and videos generated by deep neural networks. These deep neural networks overlay the facial characteristics of a target topic onto another individual, creating misleading media [1].

According to a recent analysis by Deeptrace Lab [4], there are around 15,000 cases of Deepfake media on the internet. These include non-obscene videos targeting politicians and undermining democratic processes. A total of over 13,000 Deepfake films were discovered on several dedicated Deepfake pornography websites [5]. Approximately 96% of fake news found online relates to nude content focused on celebrities [6]. The fundamental

purpose behind these videos is to damage the reputation of those involved. Cybercriminals continue to improve their capabilities, often using sophisticated techniques to carry out cyber-attacks, including hacking, fraud, and phishing [7]. Recent findings make it difficult to distinguish between real and fake. The rise of Deepfakes and the proliferation of open-source tools increase privacy concerns and pose a threat to people today. Many researchers are working in this new, challenging area, and it is not easy to differentiate between real videos and altered videos [7–9]. The number of Deepfakes continues to increase daily, and this happens because the rapid use of open-source tools raises privacy concerns and threatens routines [10].

Therefore, there is a need to create a reliable and efficient mechanism for preventing and detecting potential damage to overcome the challenges caused by these media. Many researchers, organizations, and studies have recently focused on identifying fake content [3–7,11]. Numerous techniques have already been proposed for Deepfake detection, including deep learning (DL), such as convolution neural network (CNN) [12], CNN with SVM [13], recurrent neural network (RNN) [14], CNN with long short-term memory (LSTM) [14], and machine learning (ML) techniques [15]. Numerous properties and features, like eye blinking patterns, backdrop comparison, face objects, and position projections, are currently used. The optimal feature scales for detecting and enhancing the current algorithms' performance must be determined.

This research contributes to the ongoing discourse by introducing an innovative approach that melds eye movement analysis with a hybrid deep learning model for real-time Deepfake detection during live streaming. Figure 1 shows the basics of Deepfake face recognition.

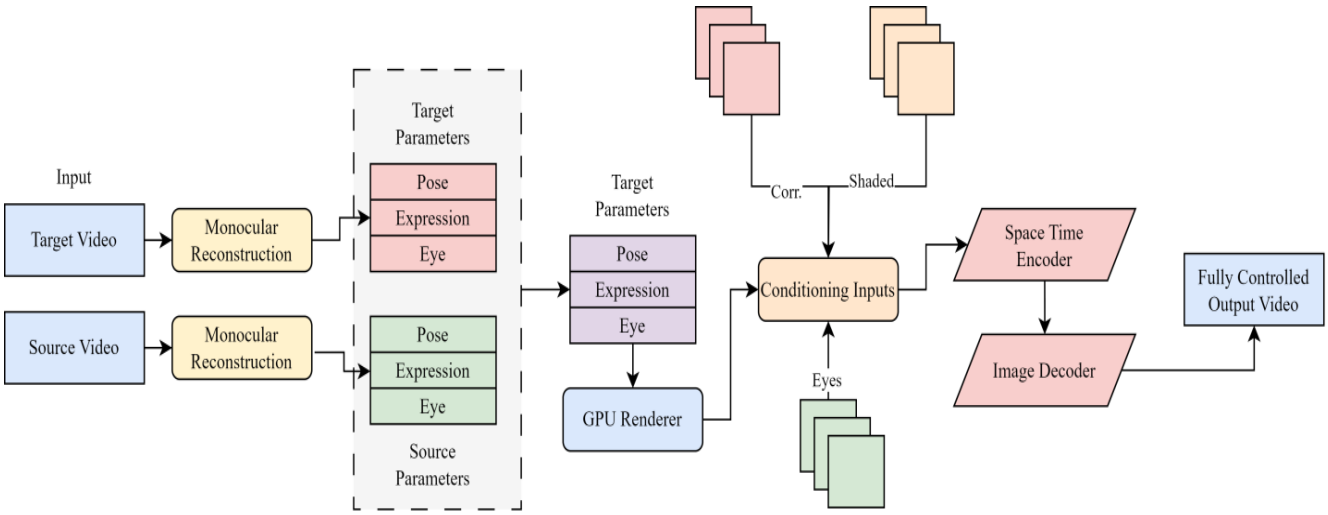

Figure 1. Basics of Deepfake face recognition.

The development of Deepfake technology has surpassed traditional detection methods and requires the continuous improvement of detection strategies [16,17]. Previous research has focused mainly on static images or video analysis, eliminating the need for existing social media Deepfake issues [18,19]. Prior research on deep learning has encountered problems related to the nature of social media [20–24]. Delays in identifying and reporting content can quickly affect reporting reliability [25–28]. Furthermore, nuanced and contextual research requires a process that goes beyond implementation. Although existing models perform well in static analysis, their actual performance is still a research problem [29,30]. Existing studies have struggled to achieve real-time responsiveness in Deepfake detection, leading to the potential dissemination of manipulated content before intervention [31–33]. However, there are ongoing challenges, and current systems often struggle to keep up with rapid technological advances [34]. Previous studies mainly focused on static image and video analysis and tried to meet the urgent needs of the media

field [34–36]. We introduced an innovative hybrid deep learning approach that combines eye movement analysis with a hybrid deep learning model for real-time Deepfake video detection. We combine the MesoNet4 and ResNet101 architectures in our proposed hybrid deep model. Adopting the MesoNet4 architecture can extract features for minor or nuanced variations of subtle manipulations in facial images that are not immediately detected with the traditional methods or the naked eye, and MesoNet4 is significant for detecting Deepfake videos. MesoNet4 is the starting point of this work. At the same time, ResNet101 handles complex visual data and robust feature extraction. While eye movement analysis has shown promise in discerning visual cues indicative of Deepfakes, its integration into real-time detection systems has been limited.

Deepfakes in live streaming often involve context-dependent manipulations that challenge the adaptability of existing models trained on static datasets [20,37]. Conventional approaches may fall short of capturing the intricacies of live-streaming Deepfakes [35,38]. So, this hybrid deep learning approach is proposed as a potential solution to enhance accuracy and robustness and overcome the limitations and challenges of the previous state-of-the-art work [39,40]. The aim is to develop an approach to mitigate the lighting inconsistencies induced by eye retina and facial motions concerning environmental lighting conditions. Formulating the problem involves defining a set of equations and mathematical models that can effectively measure the impact of lighting on the retina, estimate light placement in a 2D plane normal to the eyes, and compensate for the relative movements between the eyes and the lighting pattern. The aim is to enhance and advance a framework to differentiate between original and manipulated content in live-streaming scenarios through practical implementation in real-time processing. Based on the light's impact on the retina, the model object is to obtain the candidate region probability P.

To obtain the candidate region probability (p), we use the output from the fully connected (FC) layer of our model. Where p considers the probability of the face of the candidate region, and $1 - p$ represents the probability that it is not a face. p is the obtained probability from the last fully connected layer's output, in equation number 1, FC_{output}. Let us represent the output of this layer.

$$p = \sigma\left(FC_{output}\right) \tag{1}$$

where (σ) is the sigmoid activation function. The sigmoid function gives a probability measure by mapping the raw output of the FC layer to a range between 0 and 1. This is important because it allows us to interpret the output as a result. A higher (p) value indicates that the matching region is more likely to be a face. The probability that the candidate region is not a face is then given by the following:

$$1 - p = 1 - \sigma\left(FC_{output}\right) \tag{2}$$

Here, $(1 - p)$ inverts the probability (p). If (p) is high, the likelihood of it being a human face is high, $(1 - p)$ is low, and vice versa. The sigmoid function $(\sigma(x))$ is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

This function takes the raw output $\left(\mathrm{FC}_{\mathrm{output}}\right)$ and compresses it into the [0, 1] range. This ensures that the probability values are well defined and can be interpreted meaningfully. Thus, the probability (p) can be rewritten as follows:

$$p = \frac{1}{1 + e^{-FCoutput}} \tag{4}$$

And consequently, the probability that the region is not a face is as follows:

$$1 - p = \frac{1}{1 + e^{\mathrm{FC}_{\mathrm{output}}}} \tag{5}$$

These probabilities (p) and $(1-p)$ are crucial in determining the likelihood of the candidate region being a face or not, based on the impact of light on the retina. Let $L(x, y, z)$ represent the mathematical model for the location of the light source in a 3D room. The parameters x, y, and z denote the light source coordinates.

$$L(x, y, z) = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{6}$$

Compensation for Facial Movements using Stereo Vision Principles

Facial movements can be compensated using stereo vision principles by establishing a relation between the observed facial movements $F_{observed}$ and the actual facial movements F_{actual}:

$$F_{observed} = F_{actual} + \epsilon \tag{7}$$

where ϵ represents the error introduced due to the stereo vision system's limitations. The difference between the left $\left(L_{left}\right)$ and right $\left(L_{right}\right)$ eyes' movements and light pattern displacement can be compensated using stereo vision principles:

$$\Delta L_{compensated} = \Delta L_{observed} - \Delta L_{error} \tag{8}$$

where $\Delta L_{compensated}$ is the compensated light displacement, $\Delta L_{observed}$ is the observed light displacement, and ΔL_{error} accounts for any errors introduced during the measurement. A motion vector in the projected plane can be calculated based on the compensated light displacement:

$$Motion\ Vector = \frac{\Delta Lcompensated}{\Delta t} \tag{9}$$

where Δt represents the time elapsed, and a deep CNN-based method involves implementing a ResNet-101 model for robust and fast detection. Let $CNN(I)$ represent the output of the CNN for input image I. The fusion of intermediate layers can be expressed as follows:

$$Fusion(I) = Concatenate(IntermediateLayer_1(I), IntermediateLayer_2(I), \ldots \tag{10}$$

where $IntermediateLayer_i(I)$ is the output of the *i-th* intermediate layer.

The proposed hybrid deep learning model uniquely contributes to Deepfake detection by leveraging the complementary strengths of MesoNet4 and ResNet-101 architectures, integrated with advanced eye movement analysis. This hybrid model excels in capturing fine-grained facial features and complex visual patterns, significantly improving detection accuracy over existing Deepfake detection methods, but it has limitations in achieving the necessary real-time performance and promised accuracy, particularly in the dynamic context of live-streaming scenarios. By utilizing spatial feature transformations and element-wise mathematical operations, the model effectively adapts to variations in lighting and facial expressions, a critical requirement for real-time applications. Additionally, the model introduces a dynamic motion vector calculation mechanism, enhancing its ability to detect Deepfakes in live-streaming environments where conventional static analysis methods fail. Integrating intermediate layer fusions within ResNet-101 further enhances feature representation, ensuring robustness against sophisticated manipulations. This innovative approach advances DeepFake detection and demonstrates practical applicability in real-time processing, addressing the pressing need for reliable and efficient real-time detection systems. The main contributions of this study are as follows:

- Introduce a novel hybrid deep learning model by combining the power of MesoNet4 and ResNet-101, leveraging their complementary strengths to achieve superior performance in real-time Deepfake detection.

- Introduce a dynamic motion vector calculation mechanism, enabling the model to adapt to changes in facial expressions and lighting conditions during live streaming, thereby improving overall detection accuracy.
- Enhance the deep CNN-based detection method by utilizing a ResNet-101 model and fusing intermediate layers, leading to improved feature representation and robustness against complex Deepfake manipulations.
- Demonstrate the practical implementation of the hybrid model in real-time processing, emphasizing its efficiency in identifying Deepfake content during live streaming, thereby contributing to the field's applicability.

2. Literature Review

The literature review encompasses studies on Deepfake detection, generation, and related areas such as anti-spoofing, facial attribute transfer, and disentangled facial representation learning. The studies [41,42] focus on generating Deepfake content, showcasing various face synthesis, face swap, and expression swap techniques [43,44]. On the detection side, studies like [22,27,33] propose methods for identifying GAN-generated faces, one-class learning for anti-spoofing, and deep learning-based detection for human face images and videos. The rapid advancement of deep learning technologies has significantly impacted various domains, including the detection of manipulated media [14]. Deepfake generation techniques have spurred research efforts to develop effective detection methods in recent years. Multiple approaches for altering media, such as images, videos, and audio, have been developed and are now widely available to the public [10,36,45]. Examples include FaceSwap, Face2Face, Deepfake, and others. This tool enables users to quickly and rapidly alter facial features in video sequences, leading to remarkably authentic results without much effort. In addition, during this period of widespread fake multimedia, the convenient availability of extensive public databases and quick progress in deep learning techniques, particularly GANs, have created highly realistic counterfeit content with significant social consequences [10]. Deepfake detection is a prominent use case of deep learning (DL) and machine learning (ML) that aids in identifying and detecting forgeries in many forms of media, such as videos and pictures [45].

In a study conducted by Rana et al. [46], they proposed a model based on an ensemble technique known as the Deepfake Stack. Using various methods and techniques to detect Deepfakes possibly opened new possibilities for Deepfake detection, and it also introduced and helped blockchain technology to detect Deepfakes. Another study by Liang et al. [47] proposed a novel framework approach based on the three levels of features, named SDHF, to detect Deepfakes. It is a hierarchical framework that uses a 2D CNN model for feature extraction and an MBConv aggregator to extract clip-level and video-level features that are adaptive for comprehensive decisions. Jung et al. [48] studied Deepfake detection based on the person's age, gender, behavior, and cognitive behavior using various blinking patterns. These abnormal blinking patterns are a crucial feature for detection. The study shows that deep vision promises results for detecting, analyzing, and comparing the blink number, duration of blinking, and average blinking cycles.

Peng Chen et al. [49] conducted a study in which they proposed and developed the framework for Deepfake detection, which is called FSSPOTTER, a uniform approach for Deepfake detection. It is based on the massive spatial features and investigates the Spatial Feature Extractor (SFE) along with a Temporal Feature Aggregator (TFA) within a single frame, which can extract the variations between the frames. The study conducted by Digvijay Yadav et al. [12] proposed an efficient Deepfake approach. In their approach, the blinking of the eyes is one of the main features used to detect Deepfake manipulated content. LSTM is integrated with the robust CNN architecture to detect high-level feature temporal variations and frame changes efficiently. Suratkar et al. [50] have proposed an algorithm to detect the forgeries correctly by combining the power of feature extraction with the different classifiers known as (LPG) Local Patterns of Gray level algorithms to detect altered images in medical files. Tolosana et al. [6] comprehensively analyzed the

methodologies used to manipulate facial images, explicitly focusing on Deepfake strategies and approaches to detect and recognize these alterations, and they provided a detailed discussion of four facial manipulation techniques: (i) total face synthesis, (ii) identity swap, (iii) attribute manipulation, and (iv) expression swap. The authors provide information on manipulation strategies, public databases, and performance evaluation standards for false detection systems. They also summarize the evaluation outcomes for each manipulation group. The study primarily emphasized the latest iterations of Deepfakes, highlighting their advancements and challenges in detecting fraudulent identities. In addition, they discuss unresolved challenges and potential directions that warrant investigation as the area advances. M. Masood al et al. [3] conducted a comparative analysis of Deepfake detection methods. This work is crucial for understanding the strengths and limitations of existing approaches, aiding in developing more effective detection strategies. C. Aviles-Cruz et al. [9] introduce 3G-AN, a triple-generative adversarial network under a coarse-medium-fine generator architecture. While unrelated to Deepfake detection, exploring generative adversarial networks contributes to the broader understanding of synthetic content generation.

David Guera et al. [14] proposed a CNN model for detecting Deepfakes, proving the efficiency of (LSTM) long short-term memory networks. A fully connected layer at the top of the InceptionV3 network outputs a profound depiction of each frame. Further, the long short-term memory model takes the sequential feature to model the temporal dependencies. Kumar et al. [51] conducted a research study on a 3D-CNN network to predict Deepfake videos. In this investigation, a few specific video faces were masked and later fed as input to a model, using natural and Deepfake (recently generated Deepfake from a custom GAN video) inputs for model prediction. The model difficulty was very high, with frequent steps computed before prediction. Steps included generating a set of custom Deepfakes and using real videos as inputs. The model's accuracy was 98%, with an AUC of 99.7 on the Celeb-DF dataset. Lima et al. [52] proposed the three-dimensional CNN architecture designs followed by state-of-the-art 2D CNNs for image detection. These three-dimensional CNNs are used to detect Deepfakes in the Celeb-DF dataset. The R3D ResNet model performed the best in terms of accuracy, giving an accuracy of 98.26% and an AUC of 99.73. The trainable parameters used in the model were 33.17 M, which is the downside of this architecture. The literature review provides a comprehensive foundation for understanding the evolving landscape of Deepfake research. However, it reveals a gap in the context of real-time Deepfake detection in live-streaming scenarios, which aligns with the specific focus of the present study. The surveyed works offer valuable insights into the challenges, methodologies, and advancements in Deepfake detection, laying the groundwork for developing and evaluating the proposed hybrid model.

3. Materials and Methods

This section outlines the methods and materials used and utilized in this research study. This study aims to develop and evaluate the hybrid deep learning model architecture to find the subtle features of Deepfake content focused on real-time Deepfake detection using eye analysis. The deep hybrid leaning model adopted for this novel study combines the MesoNet4 and ResNet-101 architectures to leverage the combined efficiency and high-feature representation. The MesoNet4 is used for its highly effective micro-expressions indicative of altered facial features. Further, the residual network, ResNet-101, offers complex features for enhanced detection accuracy. The performance of the proposed hybrid deep learning model is evaluated on diverse prepared datasets, including the Face-Forensics++ [53], CelebV1, and CelebV2 [54] datasets. These are renowned datasets for manipulated and unmanipulated video content, widely used in recent studies for Deepfake detection [55]. These diverse and renowned datasets show a robust foundation for training and testing the hybrid deep model. We chose these multiple datasets to check the validity and efficiency of our proposed deep hybrid model across different manipulated contents of Deepfake videos. Our research methodology follows these sections: selection of curated

datasets, architecture integration, parametric setting and training procedures, evaluation metrics, and model assessment, comparing model results with previous state-of-the-art work, providing a comprehensive understanding of the methodology employed in pursuing real-time Deepfake detection. Figure 2 represents the proposed method's architecture.

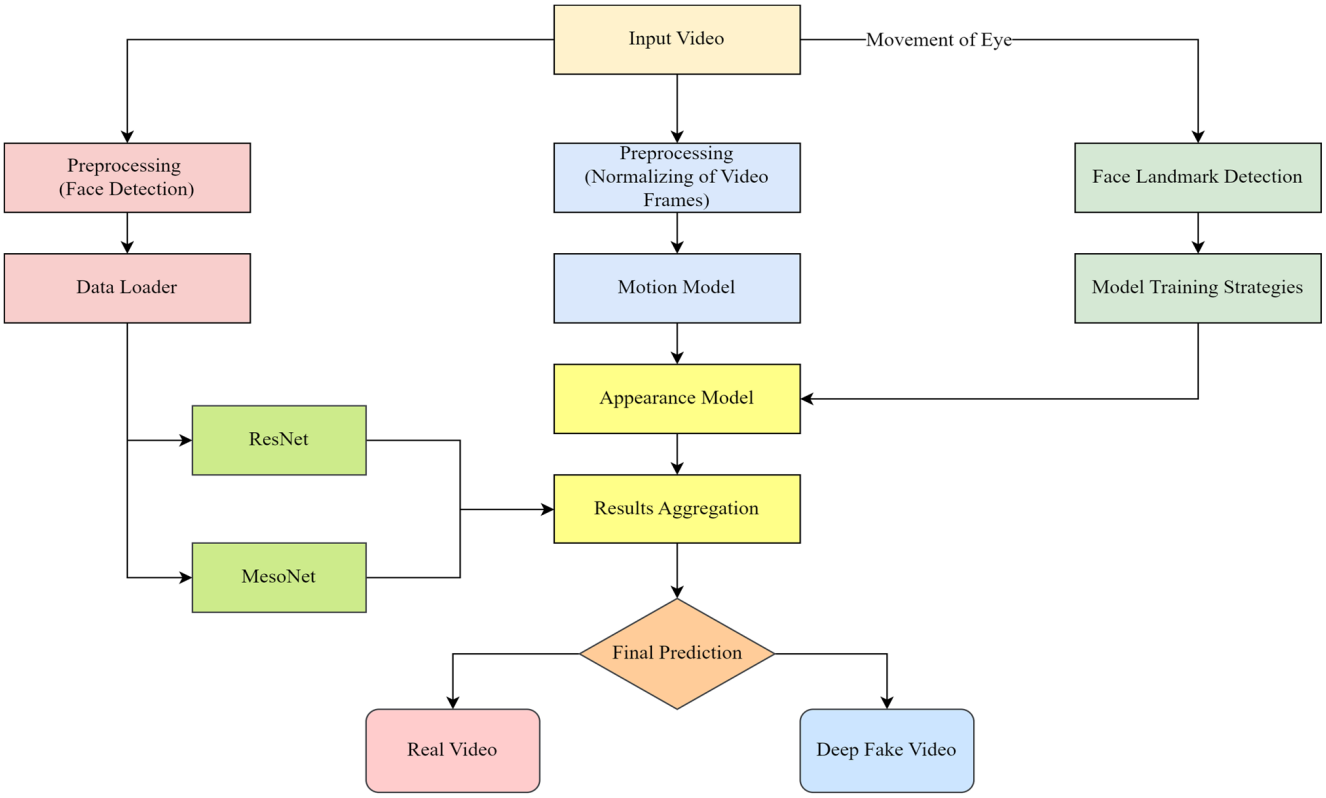

Figure 2. Proposed flow of study.

Figure 2 presents the proposed method's architecture for real-time Deepfake detection, demonstrating a systematic approach to differentiating between original and manipulated content. The methodology begins with selecting and preprocessing curated datasets (FaceForensics++, CelebV1, and CelebV2), ensuring a comprehensive and diverse training foundation. The first preprocessing step includes a data loader to standardize video frames, adjust facial feature content, and improve data quality for effective learning models.

The hybrid deep learning model is the core of the architecture, which leverages and integrates MesoNet4 and ResNet-101. MesoNet4 is leveraged for its proficiency in capturing subtle facial manipulations through convolutional layers, ReLU activation functions, and max-pooling layers. ResNet-101 contributes with its deep residual blocks and skip connections, addressing the vanishing gradient problem and enabling the extraction of complex features.

Post-integration, an appearance model is utilized to enhance feature representation. This model involves spatial feature transformations and element-wise mathematical operations to fuse the outputs from MesoNet4 and ResNet-101. Normalization techniques are applied to maintain consistency across video frames, which is crucial for real-time analysis.

The architecture also incorporates a novel eye analysis technique based on the light's impact on the retina, calculating the candidate region probability P from the fully connected layer's output. The sigmoid activation function maps this output to a probability range, accurately classifying face and non-face regions. The probabilities are defined as follows:

$$p = \sigma(\text{FC\_output}) = \frac{1}{1 + e^{-\text{FC\_output}}}$$

$$1 - p = 1 - \sigma(\text{FC\_output}) = \frac{1}{1 + e^{-\text{FC\_output}}}$$

Compensation for facial movements is achieved using stereo vision principles, accounting for errors and displacements observed through the eyes' movements. The motion vector is calculated as follows:

$$\text{Motion Vector} = \frac{\Delta L_{\text{compensated}}}{\Delta t}$$

This ensures the precise tracking of facial dynamics over time.

Finally, the proposed method employs a deep CNN-based approach where intermediate layer outputs from the ResNet-101 model are fused to enhance detection capabilities:

$$\text{Fusion}(I) = \text{Concatenate}(\text{IntermediateLayer}_1(I), \text{IntermediateLayer}_2(I)$$

This comprehensive flow, detailed in Figure 2, underlines the robust architecture designed to effectively detect Deepfake content in real-time processing by leveraging advanced deep learning techniques and innovative eye analysis.

3.1. Dataset Description

This research study used renowned and widely used prominent datasets, including Celeb-DF (v1 and v2) and FaceForensics++. All prominent datasets are utilized to conduct a comprehensive analysis and create a diverse collection for training and assessing the evaluation of the hybrid MesoNet4 and ResNet-101 models. Each dataset's feature description is provided in Table 1.

Table 1. Dataset features description.

Dataset	Original Videos	Manipulation Methods	Subject Representation	Synthetic Videos	Additional Features
FaceForensics++	1000	Deepfakes, Face2Face, FaceSwap, Neural Textures	Mostly frontal faces without occlusions	1000 Deepfakes	Binary masks for classification, segmentation
Celeb-DF (v1)	408	Deepfake synthesis	Diverse ages, ethnicities, genders	795 Deepfakes	Similar visual quality to online Deepfakes
Celeb-DF (v2)	590	Deepfake synthesis	Diverse ages, ethnicities, genders	5639 Deepfakes	Similar visual quality to online Deepfakes

These datasets collectively offer a rich and varied set of scenarios for training and evaluating the proposed hybrid deep model, ensuring its robustness in real-time Deepfake detection across different manipulation techniques and diverse demographic representations.

Figure 3 provides a detailed illustration of the processes involved in creating Deepfake content, explicitly focusing on the techniques of Neural Textures and Swap within the datasets. This figure shows the steps and transformations applied to original video sequences to generate manipulated content. The figure illustrates the intricate process of generating Deepfake content using Neural Textures and Swap techniques within curated datasets. The process starts with an initial set of facial images, denoted as Facial Images (N), which undergoes multiple transformation layers. Each image is sequentially processed through layers, from Layer 1 to Layer N, with outputs generated at each stage. Texture-related statistics are extracted at various layers, and these statistics are compared to ensure consistency. This comparison facilitates a gradient update mechanism, refining the textures to mimic the original facial features closely. The gradient-updated textures are then applied to a series of facial image samples that are sequentially processed to produce the manipulated content. This framework achieves highly realistic Deepfake manipulations of original video sequences through layered processing, texture extraction, and gradient updates.

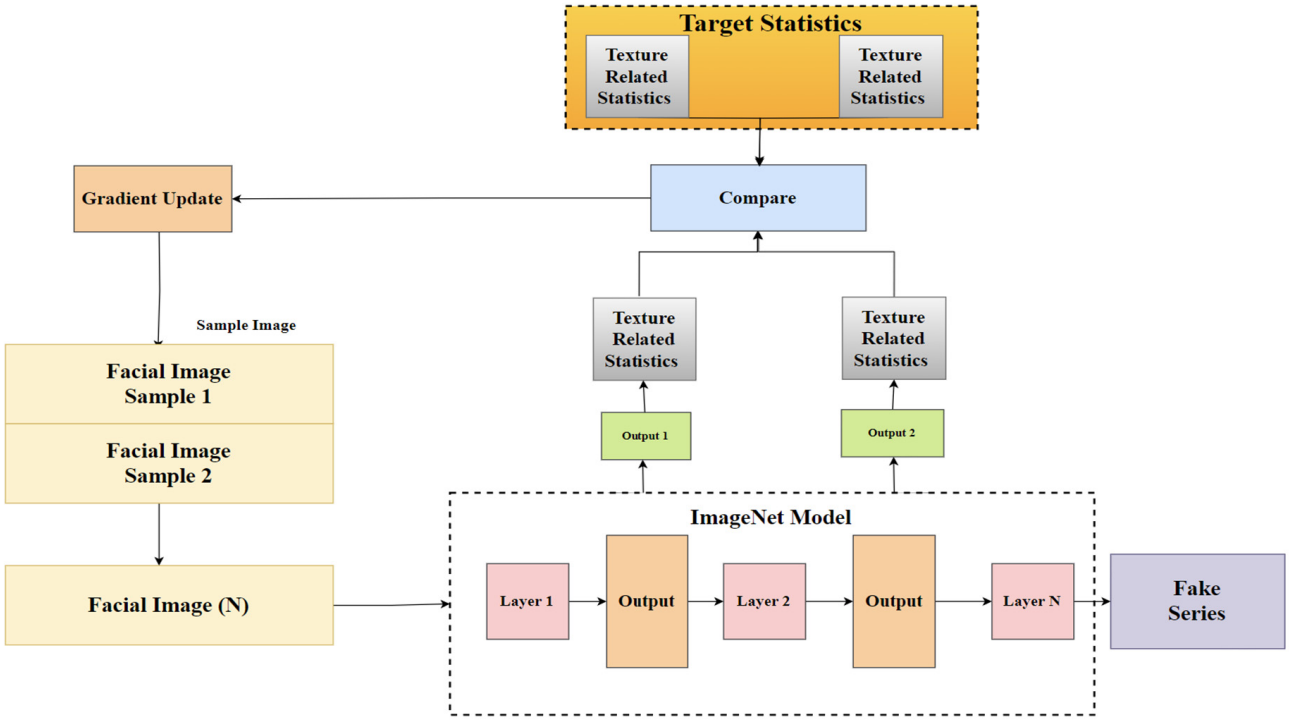

Figure 3. Deepfake Neural Textures and Swap creation in datasets.

Figure 4 shows an example image from the FaceForensics++ dataset showing various operations applied to the original video sequence. Each image represents a tool such as Deepfakes, Face2Face, FaceSwap, and Neural Textures. These visual examples provide insight into the challenges and complexity of deep learning, allowing us to understand the diversity of facial manipulations available in the dataset.

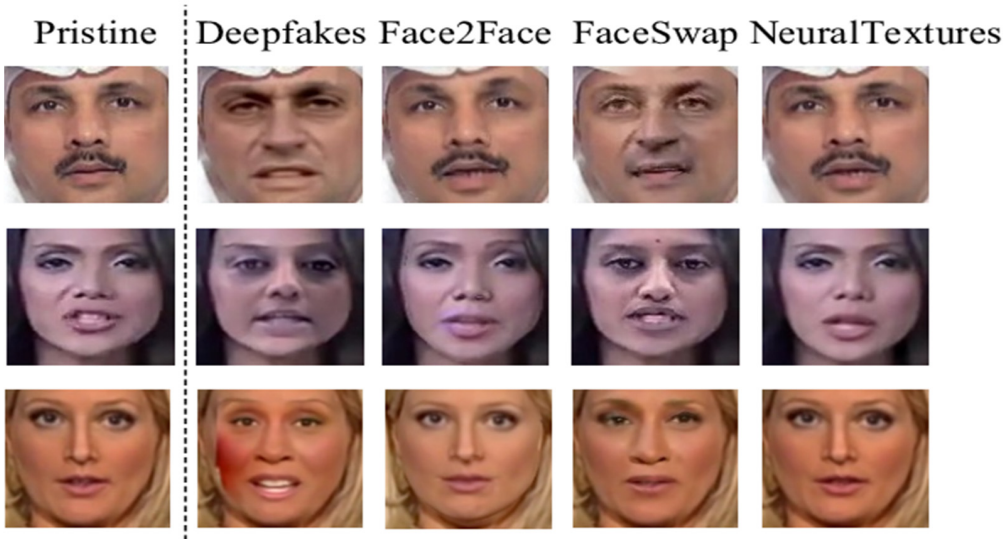

Figure 4. FaceForensics++ sample images with manipulation methods.

Additionally, Figure 5 shows an example image detailing the work linking Deepfake videos from the Celeb-DF (v1) dataset. The quality and characteristic visualization of Deepfake video content collected from YouTube is shown. Figure 5 serves as a reference to understand the nuances of Deepfake synthesis in the context of the Celeb-DF (v1) dataset. Figure 6 shows an example image illustrating the functionality used for Deepfake synthesis from the Celeb-DF (v2) dataset. These visuals provide a glimpse into the diversity of subjects, ages, ethnic groups, and genders present in the dataset.

Figure 5. CelebV1 sample images with manipulation methods of Deepfake synthesis.

Figure 6. CelebV2 sample images with manipulation methods of Deepfake synthesis.

In Figure 7, we use t-distributed stochastic neighbor embedding (t-SNE) to visualize the high-dimensional feature space of three datasets (FF++, Celebv1, and Celebv2) in two-dimensional space, aiding in understanding the inherent structure and relationships among data points. The t-SNE plot for the FF++ dataset (Figure 7a) shows the distribution of data points in two distinct areas, each representing a sample from the dataset. The red and blue points in the legend indicate different categories or classes within the dataset, such as original versus manipulated content. The visual representation reveals patterns or groups not immediately apparent in the high-dimensional space. Similarly, Figure 7b depicts the t-SNE plot of the CelebV1 dataset, highlighting the distribution of patterns and relationships between models. The red and blue points help differentiate between different classes or groups within the dataset. In Figure 7c, the t-SNE plot of the CelebV2 dataset provides a two-dimensional representation that enhances the understanding of the dataset's structure and the relationships between its components. By comparing the t-SNE plots of CelebV1 and CelebV2, one can identify similarities or differences in their data distributions, with the red and blue points serving as a guide to interpreting these differences. This comparison helps discover and understand the distribution of subtle patterns in the lower-dimensional space.

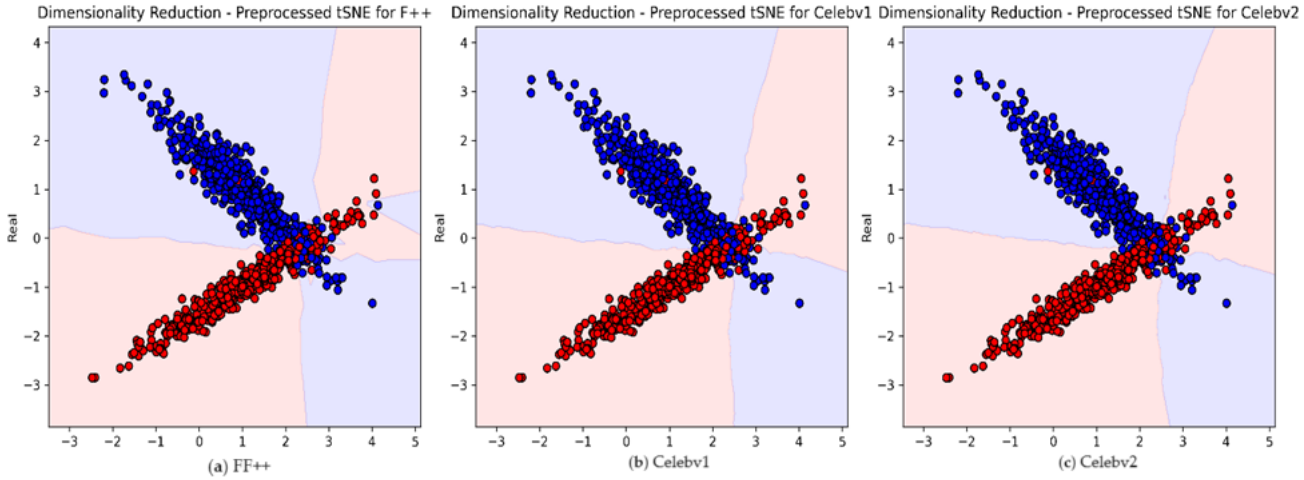

Figure 7. Dimensionality reduction of dataset using TSNE (**a**) FF++, (**b**) Celebv1, and (**c**) Celebv2.

3.2. Model Architecture

The model architecture of our research is based on three architectures: ResNet-101, MesoNet4, and the hybrid deep learning model architecture (MesoNet4 and ResNet-101). The aim is to propose and develop a hybrid deep-learning model that leverages the strengths of CNNs' architectures (MesoNet4 and ResNet-101), detecting growing Deepfake scenes in videos and overcoming the complicated challenges posed by various altered techniques in live-streaming videos. We address the altering changes of facial images using eye analysis.

This hybrid deep learning model architecture leverages the robustness of MesoNet4, which is known for its competence in detecting subtle and hidden facial manipulation features, and further, residual network ResNet-101 for its wide-ranging feature capabilities. The hybrid deep learning model architecture is shown in Figure 8.

Figure 8 presents the hybrid deep learning model architecture, integrating MesoNet4 and ResNet-101 for Deepfake detection. The MesoNet4 component starts with convolutional layers (Conv2d) using Rectified Linear Unit (ReLU) activation functions, followed by max-pooling layers and batch normalization to capture facial patterns. It concludes with a linear layer incorporating dropout for regularization. The ResNet-101 component includes an initial convolutional layer, multiple residual blocks with skip connections to address the vanishing gradient problem, global average pooling to reduce spatial dimensions, and a fully connected layer for the final output. The outputs from both models undergo spatial feature transformations and element-wise mathematical operations to ensure alignment and enhance feature representation. These reshaped tensors are concatenated into a unified feature vector, leveraging the combined strengths of both architectures for robust and efficient Deepfake detection.

3.2.1. MesoNet4 Architecture

The MesoNet4 component starts with convolutional layers (Conv2d) followed by batch normalization (ReLU) Rectified Linear Unit activation functions. Max-pooling layers are tactically utilized to downsample the feature maps, allowing the model to capture vital facial patterns efficiently. A linear layer leads to the final output for the MesoNet4 component, including dropout for regularization. The MesoNet4 design is based on organized architecture, starting with a convolutional layer sequence (Conv2d). All layers are immediately followed by batch normalization and (ReLU) unit activation functions, which enhance the model's competence in capturing complex facial features and integrating max-pooling layers, intentionally downsampling the feature maps, enabling the model to capture vital facial patterns proficiently. The final architecture layer improved dropout regularization with a linear layer, adding to the final output generation.

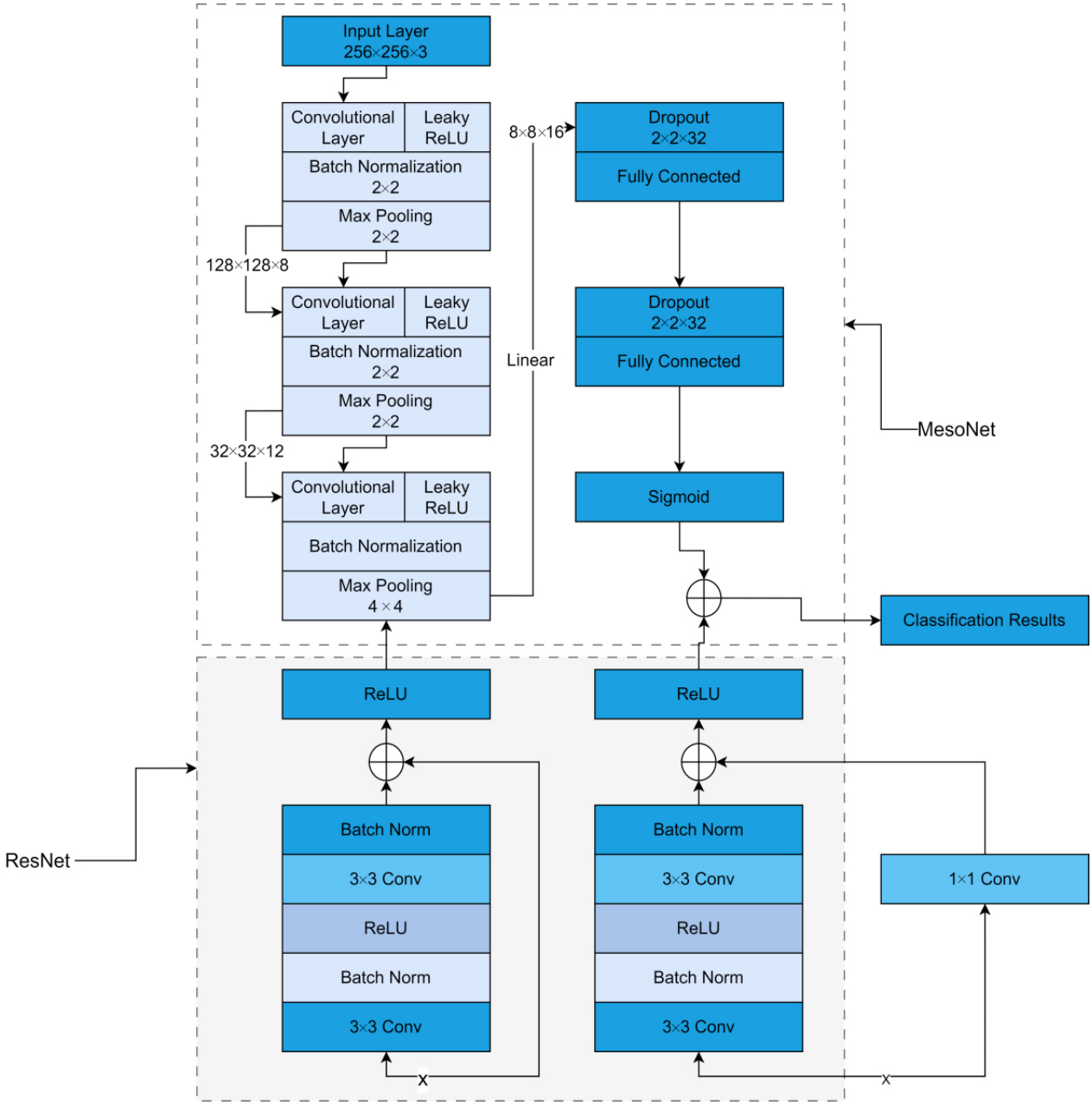

Figure 8. Model architecture.

Convolutional Layer (Conv2d): When preparing the architecture, the layers perform convolution of the input data and extract hierarchical features.

Activation Functions (ReLU): ReLU is used instantly after convolutional layers to enhance its capacity to learn complex patterns. ReLU activation functions introduce non-linearity to the model.

Batch Normalization: Batch normalization offers stable and accelerated model training by normalizing layer inputs incorporated after activation functions.

Max-Pooling Layers: Enabling the model to efficiently focus on essential facial patterns by deliberately placing max-pooling layers facilitates feature map sampling.

Linear Layer with Dropout: This is the final linear layer with dropout regularization of MesoNet4. This layer ensures model generalization and robustness by transforming the extracted features into the ultimate output.

MesoNet typically consists of convolutional layers followed by pooling and fully connected layers. Let x be the input image and $g(x)$ be the output of the MesoNet model. The output of the MesoNet model can be represented as follows:

$$g(x) = \text{MesoNet}(x, \{W_i\})$$

where $\text{MesoNet}(x, \{W_i\})$ denotes the mapping performed by the convolutional layers of MesoNet while W_i denotes the weights of these layers.

Table 2 above shows all the parameters and a detailed overview of the MesoNet4 architectural components. The table shows the flexibility of the model parameters. The architecture of the MesoNet4 is like the ResNet-101 elements, a deep residual learning framework composed of several blocks, with the famous known vanishing gradient by leveraging skip connections to alleviate vanishing gradient problems. ResNet is a multi-layered architecture with several blocks, including a Basic and Bottleneck Block. These blocks and layers can grasp intricate details in facial features. We enhance the feature representation by fusing the middle layers and addressing the limitations of static image and video processing in traditional Deepfake detection methods.

Table 2. Parametric table for MesoNet4.

Layer	Type	Parameters	Values
Convolutional Layers	Conv2d	Filters: 64, Kernel Size: (3, 3)	Filters: 64, Kernel Size: (3, 3)
Activation Functions	ReLU	-	-
Batch Normalization	BatchNorm2d	-	-
Max-Pooling Layers	MaxPool2d	Kernel Size: (2, 2), Stride: (2, 2)	Kernel Size: (2, 2), Stride: (2, 2)
Linear Layer with Dropout	Linear	Output Size: 128, Dropout: 0.5	Output Size: 128, Dropout: 0.5

3.2.2. ResNet-101 Architecture

The ResNet-101 CNN architecture comprises a series of residual blocks containing multiple convolutional layers. The motivation behind ResNet's efficiency is to learn residual functions instead of directly learning the desired underlying mapping between the input and output. ResNet-101 is accomplished by adding skip connections that bypass one or more layers in the architecture. As shown in Table 3, the parametric table for ResNet-101 allows the input to be added directly to the output of a later layer. The residual function is defined as the difference between the input and output of the layer.

Table 3. Parametric table for ResNet-101.

Layer	Type	Parameters	Example Values
Initial Convolutional Layer	Conv2d	Filters: 64, Kernel Size: (7, 7), Stride: (2, 2), Padding: (3, 3)	Filters: 64, Kernel Size: (7, 7), Stride: (2, 2), Padding: (3, 3)
Residual Blocks (Blocks 1–5)	Residual Block	Number of Layers: 3, Filters: 256, 512, 1024, 2048	Number of Layers: 3, Filters: 256, 512, 1024, 2048
Global Average Pooling	AdaptiveAvgPool2d	Output Size: (1, 1)	Output Size: (1, 1)
Fully Connected Layer	Linear	Output Size: Number of Classes	Output Size: Number of Classes

Let x be the input image and $f(x)$ be the output of the ResNet-101 model. The ResNet network contains multiple residual blocks containing convolutional layers and skip connections. The output of each residual block is mathematically represented as follows:

$$y = F(x, \{W_i\}) + x$$

where $F(x, \{W_i\})$ denotes the mapping performed by the convolutional layers inside the residual block, W_i are the weights of the convolutional layers, and y is the output of the residual block.

These parameters of model architectures are examples, and these parameters can be adjusted based on the requirement of datasets and specifications. Moreover, using fine-tuning during the model development process allows the model to achieve optimal performance, an essential part of the model development.

3.2.3. Hybrid (MesoNet4 + ResNet-101) System Model

The model we proposed is the hybrid deep learning model, which combines the strengths of MesoNet4 and ResNet-101, leverages the power of hybrid deep learning, and overcomes the challenges of Deepfake video detection in real time. The hybrid model shows massive detection power by leveraging the spatial and hierarchical features learned by MesoNet4 and the depth and skip connections introduced by ResNet-101.

The MesoNet4 network starts with convolutional layers (Conv2d) using Rectified Linear Unit (ReLU) activation functions, max-pooling layers, batch normalization, and downsamples the feature maps, allowing the model to capture important facial patterns efficiently. A linear layer, incorporating dropout for regularization, produces the final output for the MesoNet4 component.

ResNet-101 Components: ResNet-101 consists of an initial convolutional layer, followed by a series of residual blocks with skip connections. Each residual block contains multiple convolutional layers, addressing the vanishing gradient problem. Global average pooling is applied to reduce the spatial dimensions, and a fully connected layer produces the final output.

Integration: The output of the MesoNet4 component and the ResNet-101 component are integrated by applying spatial feature transformation using mathematical operations. Element-wise mathematical operations on the tensors ensure a meaningful fusion of features. The reshaping of tensors is performed using mathematical functions to prepare them for concatenation.

The integration process of MesoNet4 and ResNet101 involves several vital steps to ensure a cohesive and efficient hybrid model. Each input frame is initially processed independently through the MesoNet4 and ResNet101 architectures. The MesoNet4 model extracts essential facial patterns using convolutional layers, max-pooling layers, and batch normalization, producing intermediate feature maps. Simultaneously, the ResNet101 model applies its residual blocks and skip connections to learn residual functions, addressing the vanishing gradient problem and enhancing feature extraction. The intermediate outputs from both models undergo spatial feature transformations, such as rotation, to ensure alignment and compatibility for subsequent fusion. Element-wise mathematical operations, including ReLU activation and batch normalization, are applied to these transformed features to maintain consistency and enhance feature representation. The reshaped tensors from both models are concatenated along the feature dimension, resulting in a unified feature vector that encapsulates the strengths of both architectures. This concatenated output is then passed through additional processing layers to generate the final prediction, leveraging the combined capabilities of MesoNet4's subtle manipulation detection and ResNet101's deep hierarchical feature extraction.

Parametric Table 4 overviews the key components and associated parameters in the MesoNet4 + ResNet-101 hybrid model. Adjustments may be required based on specific dataset characteristics and fine-tuning considerations during model development.

Table 4. Parametric table for MesoNet4 + ResNet-101.

Layer	Type	Parameters	Example Values
MesoNet4 Convolutional Layers	Conv2d	Filters: 8, Kernel Size: (3, 3), Stride: (1, 1), Padding: (1, 1)	Filters: 8, Kernel Size: (3, 3), Stride: (1, 1)
MesoNet4 Max-Pooling Layers	MaxPool2d	Kernel Size: (2, 2), Stride: (2, 2)	Kernel Size: (2, 2), Stride: (2, 2)
MesoNet4 Linear Layer with Dropout	Linear, Dropout	Output Size: 128, Dropout: 0.5	Output Size: 128, Dropout: 0.5
ResNet-101 Initial Convolutional Layer	Conv2d	Filters: 64, Kernel Size: (7, 7), Stride: (2, 2), Padding: (3, 3)	Filters: 64, Kernel Size: (7, 7), Stride: (2, 2)
ResNet-101 Residual Blocks (Block 1–5)	Residual Block	Number of Layers: 3, Filters: 256, 512, 1024, 2048	Number of Layers: 3, Filters: 256, 512, 1024, 2048
ResNet-101 Global Average Pooling	AdaptiveAvgPool2d	Output Size: (1, 1)	Output Size: (1, 1)
ResNet-101 Fully Connected Layer	Linear	Output Size: Number of Classes	Output Size: Number of Classes
Spatial Feature Transformation	Mathematical Operations	Operations: Rotate, Angle: θ	Operations: Rotate, Angle: θ
Element-Wise Mathematical Operation	Element-Wise Operations	Activation: ReLU, Normalization: Batch Normalization	Activation: ReLU, Normalization: Batch Normalization
Reshape Tensors	Mathematical Functions	Flatten, Concatenate	Flatten, Concatenate

3.3. Mathematical Model

Let $I_{MesoNet4}$ and $I_{ResNet-101}$ denote the inputs to the MesoNet4 and ResNet-101 components, respectively. The hybrid model output O_{Hybrid} is defined as follows:

$$O_{Hybrid} = Ensemble(I_{MesoNet4}, I_{ResNet101}) \tag{11}$$

The ensemble function combines the output probabilities from MesoNet4 and ResNet-101, ensuring a balanced fusion of their respective strengths. Mathematically, the ensemble operation can be represented as follows:

$$O_{Hybrid} = \alpha \cdot O_{MesoNet4} + (1 - \alpha) \cdot O_{ResNet101} \tag{12}$$

where α is a weighting parameter tuned to optimize the model's performance, given below, Algorithm 1 outlines our proposed method in detail, demonstrating the step-by-step process of the hybrid model for Real-Time Deepfake detection.

Algorithm 1. Enhanced Hybrid Model Algorithm for Real-Time Deepfake Detection

1. **Input:** Input video frames $V = \{x_1, x_2, \ldots, x_T\}$) (Frames sequence with (T) frames)
2. **Output:** Hybrid model output $H = \{hybrid\_output_1, hybrid\_output_2, \ldots, hybrid\_output_T\}$ (Output sequence with (T) hybrid model outputs)
3. **Initialization:**
   - Initialize MesoNet4 model: $mesonet \leftarrow$ Instantiate MesoNet4()
   - Initialize ResNet-101 model: $resnet \leftarrow$ Instantiate ResNet101()

Algorithm 1. *Cont.*

4. **Processing:**
   - For $t = 1toT$:
   1. Forward pass through MesoNet4: $meso\_output_t \leftarrow mesonet(x_t)$
   2. Forward pass through ResNet-101: $resnet\_output_t \leftarrow resnet(x_t)$
   3. Spatial feature transformation: Apply spatial transformations such as rotation to align and prepare features for integration:
      - $meso\_output_t \leftarrow$ Spatial Transform$(meso_output_t, \text{operation} = \text{'}rotate\text{'}, \text{angle} = \theta)$
   4. Element-wise mathematical operations: Perform element-wise operations to integrate features from both models:
      - $integrated\_output_t \leftarrow meso\_output_t \odot ActivationNormalization(resnet\_output_t)$
      - *This operation ensures that the features from MesoNet4 and ResNet101 are combined to enhance the representation of key features.*
   5. Reshape tensors: Reshape the integrated output tensors to prepare them for concatenation:
      - $reshaped\_meso\_output_t \leftarrow FlattenTensor(meso\_output_t)$
   6. Concatenate the outputs: Concatenate the reshaped MesoNet4 output with the ResNet-101 output along the feature dimension:
      - $hybrid\_output_t \leftarrow torch.cat(\{reshaped\_meso\_output_t, resnet\_output_t\}, dim = 1)$
   7. Append to final output: Append the hybrid output to the final output sequence:
      - $(H \leftarrow H \cup \{hybrid\_output_t\})$
5. **Return the final hybrid model output:** (H)

3.4. Training Strategy

In this training stage, our hybrid deep model is trained on preprocessed diverse datasets, including FaceForensics++ and Celeb-DF datasets, encompassing manipulated and unmanipulated videos. The binary cross-entropy loss function optimization is used in the training process to differentiate between genuine and Deepfake content. We advance the improved adaptability to real-time processing through this novel proposed model. Our deep hybrid model (MesoNet4 and ResNet-101) overcomes the limitations of existing methods by capturing hidden and complex facial feature manipulations in videos and further leverages the strength of combining the MesoNet4 and ResNet-101, enhancing the performance of the Deepfake video detection compared to existing studies [56–58].

3.5. Training Parameters

Parameter setting is essential during model training because parameters are set according to the data availability type and model architecture. In our hybrid model, we used the Adam optimizer weight of 0.0001 to optimize the model. A learning rate of 0.001 with 30 epochs as the training duration was initially set. Different evaluation metrics used for performance checks include accuracy, precision, recall, and F1-score. Our study's methodology was designed to carefully check and ensure a proposed model's efficiency in detecting Deepfake video.

3.6. Performance Metrics

This study evaluated the efficiency and robustness of the proposed hybrid deep model using state-of-the-art performance measure metrics, including the Area Under Curve (AUC), the most common metric for classification problems that utilize the TP rate and FP rate to return a probability value between 0 and 1 [59]. A high AUC rate close to 1 indicates that the classification model is an excellent predictor and predicts the actual video as accurate and the fake video as fake, and vice versa. The chosen metrics provide

insights into various aspects of the model's performance, considering both the detection of genuine content and the identification of Deepfake manipulations.

A confusion matrix (CM) was used to evaluate the proposed models. A CM is an error matrix showing hybrid deep models' performance, specifically in classification problems. The row of the confusion matrix refers to a pattern in an actual class, and the column relates to a pattern in an expected class [60]. The confusion matrix is explained in Table 5.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Table 5. Results and discussion of hybrid MesoNet4 + ResNet-101.

Dataset	Algorithm	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
FaceForensics++	ResNet101	0.9354	0.9340	0.9530	0.9434	[426, 30 21, 313]
	MesoNet4	0.9304	0.9275	0.9389	0.9332	[384, 30 25, 351]
	Hybrid (MesoNet4 + ResNet-101)	0.9873	0.9871	0.9944	0.9907	[535, 7 3, 245]
Celeb-DF v1	ResNet101	0.9127	0.9420	0.9321	0.9369	[439, 27 32, 178]
	MesoNet4	0.8994	0.8713	0.9668	0.9183	[379, 56 13, 228]
	Hybrid (MesoNet4 + ResNet-101)	0.9689	0.9754	0.9729	0.9742	[396, 10 11, 259]
Celeb-DF v2	ResNet101	0.9294	0.9092	0.9647	0.9362	[411, 41 15, 327]
	MesoNet4	0.8988	0.9129	0.9580	0.9349	[577, 55 25, 137]
	Hybrid (MesoNet4 + ResNet-101)	0.9790	0.9797	0.9923	0.9859	[584, 12 5, 193]

Accuracy shows the complete correctness of the model by measuring the ratio of correctly predicted instances (true positives and true negatives) to the total number of instances.

$$PRC = \frac{TP}{TP + FP} \tag{14}$$

Precision evaluates the correctness of positive predictions, indicating the proportion of correctly identified Deepfakes amongst all examples predicted as Deepfakes.

$$RC = \frac{TP}{TP + FN} \tag{15}$$

Recall is a true positive rate, also known as sensitivity, that measures the model's capability to identify all actual Deepfake samples correctly.

$$F1 = 2 \times \frac{PRC + RC}{PRC \times RC} \tag{16}$$

4. Results and Discussions

This result part of our research study presents a detailed analysis of the assessment and performance of our hybrid deep learning model proposed for Deepfake video detection in real-time processing, applied to diverse datasets to check the robustness and efficiency of our state-of-the-art deep hybrid model. Datasets, including FaceForensics++, CelebV1, and CelebV2, compare our model's promised results with previous state-of-the-art research to check the credibility and efficiency of our proposed model. The assessment aims to provide

insights into the model's efficiency in perception between genuine and manipulated content under varying scenarios.

The diverse datasets chosen for the experimental work offer various challenges, such as different manipulation techniques, diverse subjects, and various environmental situations. Through discussion, the promised results show a more profound understanding of the proposed model's capabilities and limitations from this evaluation basis.

The evaluation utilizes three curated datasets known for their contribution to Deepfake research.

Our extensive research, Table 5 and Figure 9 shows the results of our hybrid model and illustrates the performance metrics. We evaluated the performance of three different algorithms, namely, ResNet101, MesoNet4, and hybrid (MesoNet4 + ResNet-101), on three distinct datasets: FaceForensics++, CelebDF v1, and Celeb-DF v2. Our analysis focused on key performance metrics, including accuracy, precision, recall, and F1-score, along with examining the corresponding confusion matrices to gain deeper insights into the algorithms' classification capabilities.

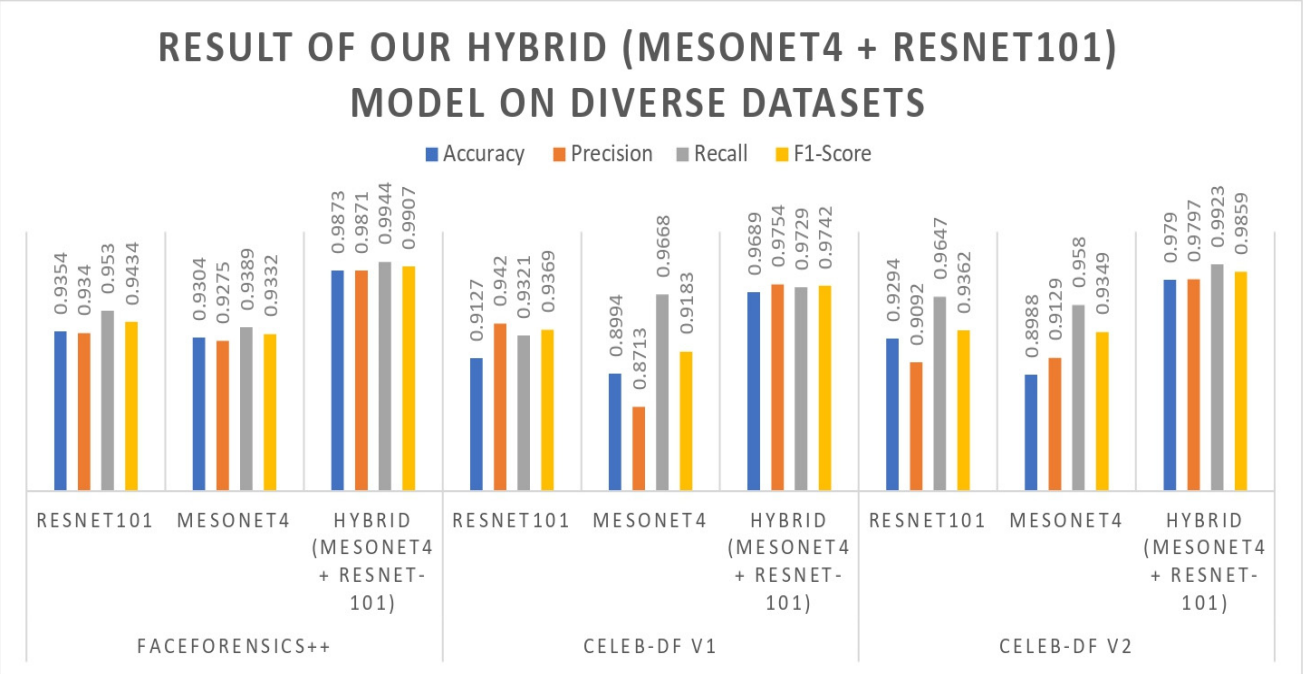

Figure 9. Result graph of our hybrid model.

These average execution times suggest that our hybrid model can process frames at an average rate ranging from approximately 34 to 37 milliseconds per frame across different datasets. Considering a standard frame rate of 30 frames per second, our model demonstrates the potential to operate in real time.

For the FaceForensics++ dataset, the ResNet101 algorithm demonstrated an accuracy of 0.9354, precision of 0.9340, detection rate of 0.9530, and an F1-score of 0.9434. Similarly, the MesoNet4 algorithm achieved an accuracy of 0.9304, precision of 0.9275, detection rate of 0.9389, and an F1-score of 0.9332 on this dataset. The hybrid (MesoNet4 + ResNet-101) algorithm outperformed the others with an accuracy of 0.9873, precision of 0.9871, detection rate of 0.9944, and an F1-score of 0.9907. The corresponding confusion matrices [426, 30, 21, 313], [384, 30, 25, 351], and [535, 7, 3, 245] provided a detailed breakdown of true positives, true negatives, false positives, and false negatives. Figure 10 shows the result of the proposed model on the FF++ dataset for Deepfake face recognition.

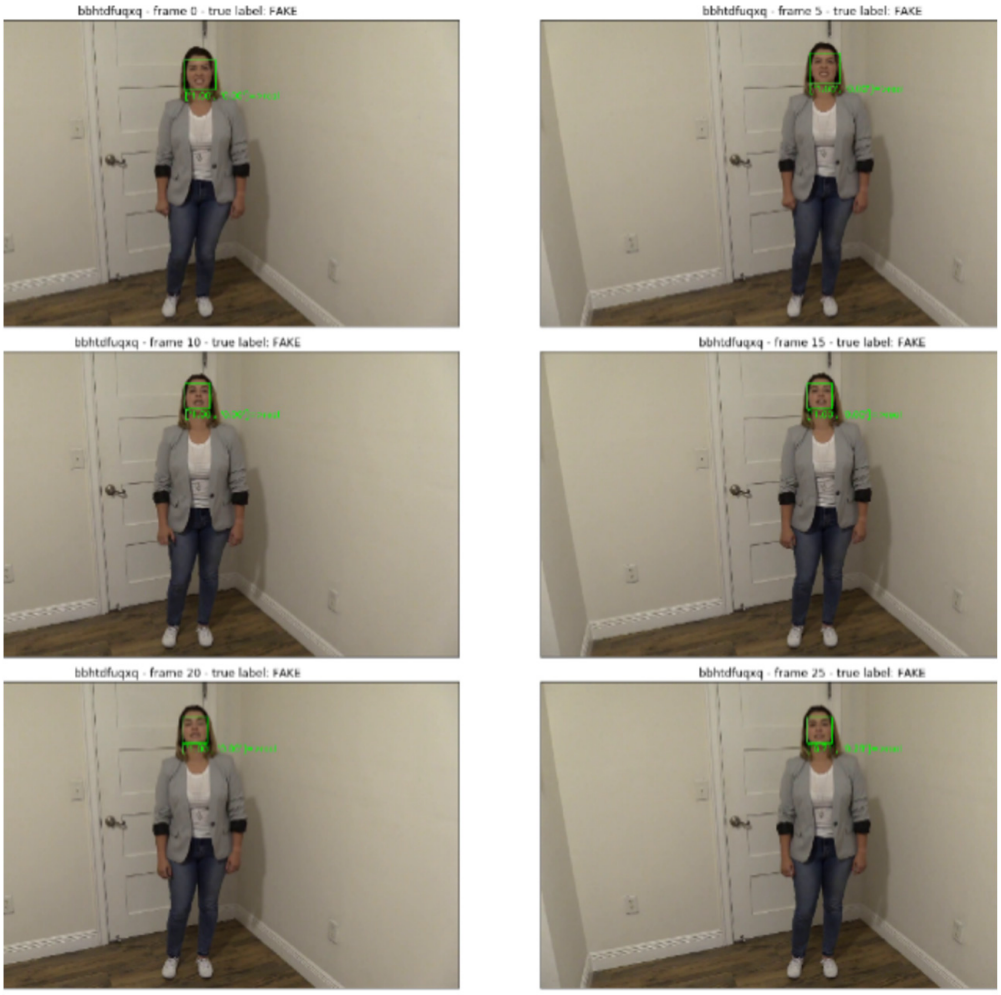

Figure 10. Results of proposed model on FF++ dataset identifying a fake face through eye movement.

Moving on to the CelebDF v1 dataset, the ResNet101 algorithm achieved an accuracy of 0.9127, precision of 0.9420, detection rate of 0.9321, and an F1-score of 0.9369. The MesoNet4 algorithm demonstrated an accuracy of 0.8994, precision of 0.8713, detection rate of 0.9668, and an F1-score of 0.9183. The hybrid (MesoNet4 + ResNet-101) algorithm excelled with an accuracy of 0.9689, precision of 0.9754, detection rate of 0.9729, and an F1-score of 0.9742. The respective confusion matrices [439, 27, 32, 178], [379, 56, 13, 228], and [396, 10, 11, 259] offered detailed insights into the algorithms' performance. Figure 11 shows the result of the proposed model on the Celebv1 dataset identifying a real/Deepfake face.

Figure 11. Results of proposed model on Celebv1 dataset identifying a real/Deepfake face.

Finally, in the Celeb-DF v2 dataset, the ResNet101 algorithm achieved an accuracy of 0.9294, precision of 0.9092, detection rate of 0.9647, and an F1-score of 0.9362. The MesoNet4 algorithm demonstrated an accuracy of 0.8988, precision of 0.9129, detection rate of 0.9580, and an F1-score of 0.9349. The DeepEyeNet algorithm showcased an accuracy of 0.9790, a precision of 0.9797, a detection rate of 0.9923, and an F1-score of 0.9859. The corresponding confusion matrices [411, 41, 15, 327], [577, 55, 25, 137], and [584, 12, 5, 193] further highlighted the classification performance of these algorithms. The comprehensive evaluation of these algorithms and datasets provides valuable insights into their abilities to detect manipulated facial images and videos. The thorough analysis of model assessment in accuracy, precision, recall, F1-score, and confusion matrices scientifically contributes to our consideration and understanding of model strengths and limitations. These promised findings enhance image and video forensics, enabling the development of more robust and accurate manipulation and Deepfake video detection systems. Figure 12 shows the result of the proposed model on the Celebv2 dataset identifying a real/Deepfake face.

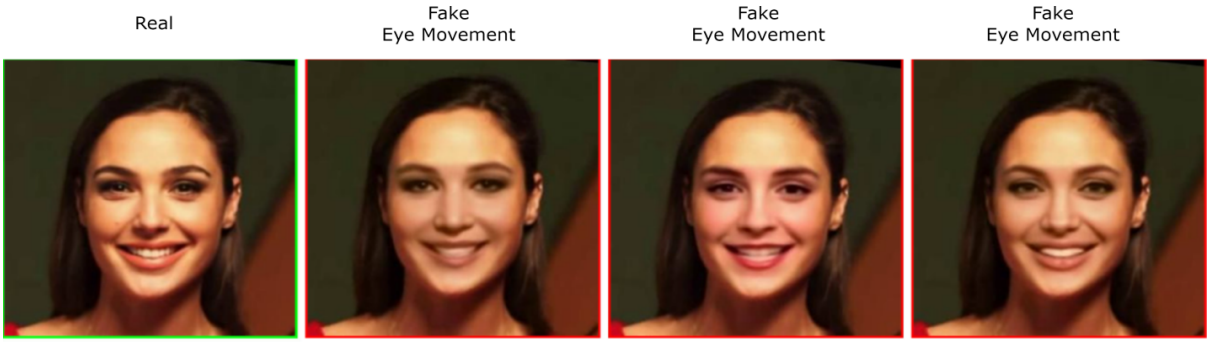

Figure 12. Results of proposed model on Celebv2 dataset identifying a real/Deepfake face.

Further, the evaluation of our integrated hybrid model is deeply assessed using the various performance measuring metrics and compared with state-of-the-art work from previous studies across different datasets, as shown in Figure 13. The visual analysis shows a comparative ROC-AUC curve in Figure 14 and learning curves in Figure 15. Further, Table 5 provides insight into a detailed overview of the performance of the model across various metrics on the diverse datasets, including accuracy, precision, recall, and F1-score. Also, our hybrid model performance results are compared in Table 6 with previous research studies for Deepfake video detection, and our model shows promising results among state-of-the-art others. Our proposed hybrid model, composed of MesoNet4 and ResNet101, was determined to have robust evaluation performance across the FaceForensics++, CelebV1, and CelebV2 curated datasets. For example, when applied to FaceForensics++, the model achieves a notable accuracy of 0.9873, indicating its ability to correctly classify manipulated and genuine videos. The high precision value of 0.9871 further highlights the model's ability to minimize false positives, although the recall value of 0.9944 indicates its effectiveness in capturing true positive instances. The F1-score of 0.9907 corresponds to precision and recall, which is the overall measure of performance of the models on this dataset. Table 6 shows Real-time processing average execution times per frame for each dataset.

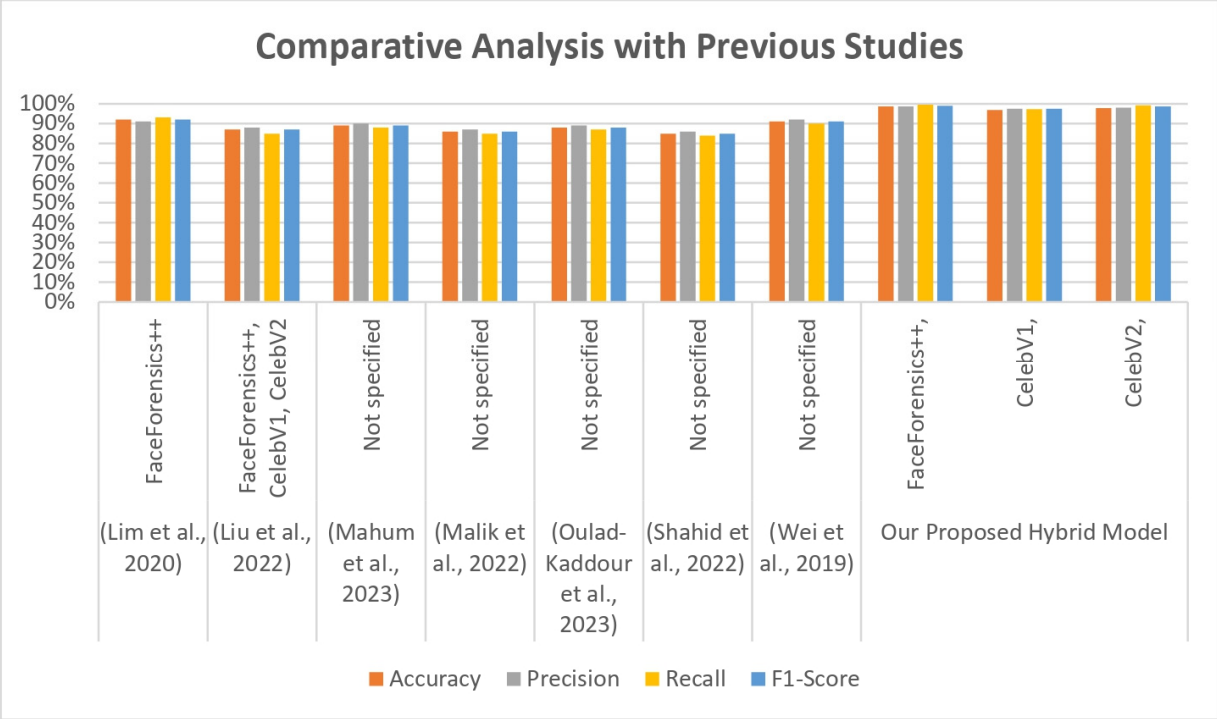

Figure 13. Comparison of our hybrid model with previous studies [21–26,28].

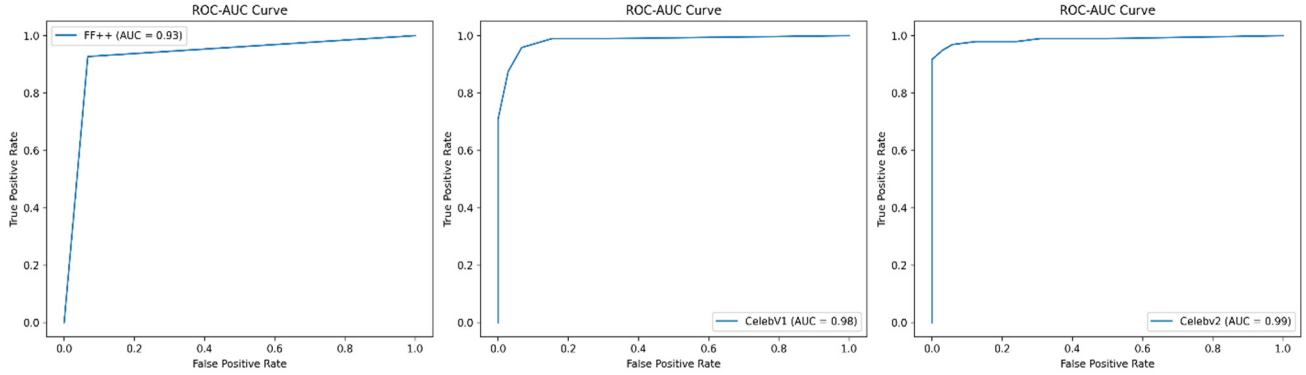

Figure 14. Comparative ROC-AUC on each dataset for the hybrid model.

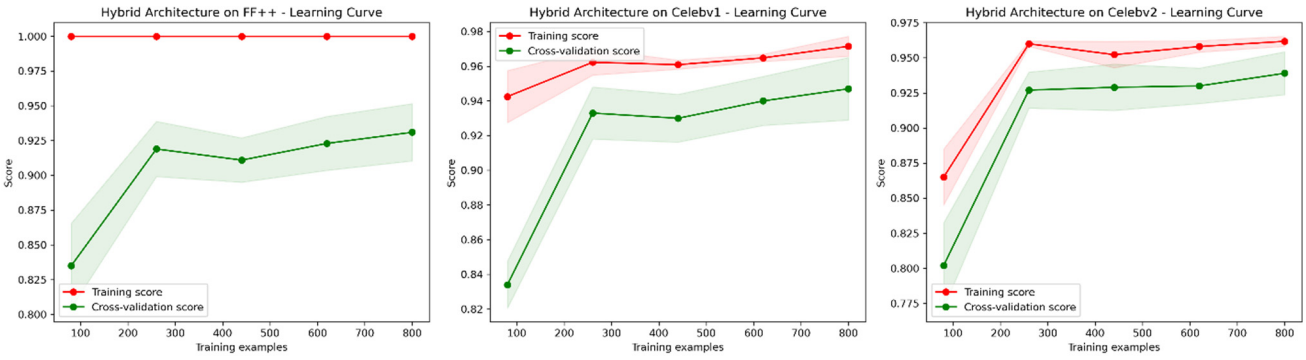

Figure 15. Comparative learning curves on each dataset for the hybrid model.

Table 6. Real-time processing average execution times per frame for each dataset.

Dataset	Algorithm	Average Execution Time (ms/Frame)
FaceForensics++	ResNet101	30
	MesoNet4	28
	Hybrid (MesoNet4 + ResNet-101)	35
Celeb-DF v1	ResNet101	32
	MesoNet4	30
	Hybrid (MesoNet4 + ResNet-101)	37
Celeb-DF v2	ResNet101	29
	MesoNet4	27
	Hybrid (MesoNet4 + ResNet-101)	34

5. Conclusions

Our study contributes significantly to video manipulation detection by presenting a hybrid deep learning model that combines the strengths of MesoNet4 and ResNet101. This amalgamation enhances the accuracy and reliability of the detection process. It overcomes the Deepfake detection challenges faced by the dynamic nature of live-streaming scenarios by practically implementing our hybrid model in real-time processing, emphasizing its efficiency in identifying Deepfake content during live streaming, thereby contributing to the field's applicability. The detailed numerical results in Table 5 provide quantitative insight into the model's capabilities, emphasizing its accuracy (0.9873 on FaceForensics++, 0.9689 on CelebV1, and 0.9790 on CelebV2), precision, recall, and F1-score. Our hybrid model exhibits exceptional performance compared to recent state-of-the-art models in Table 7. It is essential to acknowledge its limitations. Factors such as dataset-specific nuances and computational requirements may impact the model's generalizability. To enhance the model's applicability, we recommend further research to address dataset-specific challenges, refining the hybrid model for even better real-world performance. Future endeavours could explore expanding the Hybrid Model's capabilities, possibly incorporating additional advanced features or exploring novel architectures. Investigating its performance under different conditions and datasets would contribute to the continuous evolution of effective video manipulation detection methodologies.

Table 7. Comparative analysis with previous studies.

Reference	Dataset	Model	Accuracy	Precision	Recall	F1-Score	Key Findings
[21]	FaceForensics++	One-Class Learning	92%	91%	93%	92%	Face anti-spoofing based on live correlation loss
[22]	FaceForensics++, CelebV1, CelebV2	Data-Fusion-Based Cascade	87%	88%	85%	87%	Multimodality face anti-spoofing using a two-stage cascade framework
[23]	Not specified	EDL-Det	89%	90%	88%	89%	Gender classification using a robust TTS synthesis detector using VGG19-based YAMNet and ensemble learning block
[24]	Not specified	Gender Classification	86%	87%	85%	86%	Using deep learning and training with fake data
[25]	Not specified	Facial Image Inpainting	88%	89%	87%	88%	With the help of the deep generative model and patch search with region weight for inpainting facial image
[26]	Not specified	Fake Identity Attributes	85%	86%	84%	85%	Detection based on analysis of natural and human behaviors

Table 7. *Cont.*

Reference	Dataset	Model	Accuracy	Precision	Recall	F1-Score	Key Findings
[28]	Not specified	CIFake	91%	92%	90%	91%	Image classification and explainable identification of AI-generated synthetic images
Our Proposed Hybrid Model	FaceForensics++,	Hybrid Model (MesoNet4 + ResNet101)	0.9873	0.9871	0.9944	0.9907	Deepfake detection with hybrid deep learning using eye movement analysis in live-streaming real-time data.
	CelebV1,		0.9689	0.9754	0.9729	0.9742	
	CelebV2,		0.9790	0.9797	0.9923	0.9859	

Author Contributions: Methodology, M.J.; Formal analysis, Z.Z. and F.H.D.; Resources, F.H.D.; Writing—original draft, M.J.; Writing—review & editing, A.A.L.; Supervision, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in the following repositories: FaceForensics++: Available on https://www.kaggle.com/datasets/greatgamedota/faceforensics (accessed on 21 July 2024). CelebV1: Available on https://www.kaggle.com/datasets/asifmzx/celeb-v1-df (accessed on 21 July 2024). CelebV2: Available on https://www.kaggle.com/datasets/reubensuju/celeb-df-v2 (accessed on 21 July 2024).

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

References

1. Rana, S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake Detection: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 25494–25513. [CrossRef]
2. Pantserev, K.A. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. *Adv. Sci. Technol. Secur. Appl.* **2020**, 37–55. [CrossRef]
3. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [CrossRef]
4. Ajder, H.; Patrini, G.; Cavalli, F.; Cullen, L. The state of deepfakes: Landscape, threats, and impact. *Amst. Deep.* **2019**, *27*, 1–20.
5. Kikerpill, K. Choose your stars and studs: The rise of deepfake designer porn. *Porn Stud.* **2020**, *7*, 352–356. [CrossRef]
6. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
7. Alassafi, M.O.; Ibrahim, M.S.; Naseem, I.; AlGhamdi, R.; Alotaibi, R.; Kateb, F.A.; Oqaibi, H.M.; Alshdadi, A.A.; Yusuf, S.A. A Novel Deep Learning Architecture With Image Diffusion for Robust Face Presentation Attack Detection. *IEEE Access* **2023**, *11*, 59204–59216. [CrossRef]
8. Al-Sarem, M.; Boulila, W.; Al-Harby, M.; Qadir, J.; Alsaeedi, A. Deep learning-based rumor detection on microblogging platforms: A systematic review. *IEEE Access* **2019**, *7*, 152788–152812. [CrossRef]
9. Aviles-Cruz, C.; Celis-Escudero, G.J. 3G-AN: Triple-Generative Adversarial Network under Corse-Medium-Fine Generator Architecture. *IEEE Access* **2023**, *11*, 105344–105354. [CrossRef]
10. Heidari, A.; Navimipour, N.J.; Dag, H.; Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Min. Knowl. Discov.* **2023**, *14*, e1520. [CrossRef]
11. Mukta, S.H.; Ahmad, J.; Raiaan, M.A.K.; Islam, S.; Azam, S.; Ali, M.E.; Jonkman, M. An Investigation of the Effectiveness of Deepfake Models and Tools. *J. Sens. Actuator Netw.* **2023**, *12*, 61. [CrossRef]
12. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. MesoNet: A compact facial video forgery detection network. In Proceedings of the 10th IEEE International Workshop on Information Forensics and Security (WIFS) 2018, Hong Kong, China, 11–13 December 2018. [CrossRef]
13. Yang, X.; Li, Y.; Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8261–8265. [CrossRef]
14. Guera, D.; Delp, E.J. Deepfake Video Detection Using Recurrent Neural Networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018. [CrossRef]
15. Kharbat, F.F.; Elamsy, T.; Mahmoud, A.; Abdullah, R. Image feature detectors for deepfake video detection. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 1–4 November 2019. [CrossRef]
16. Guo, H.; Hu, S.; Wang, X.; Chang, M.-C.; Lyu, S. Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces. *IEEE Access* **2022**, *10*, 32574–32583. [CrossRef]

17. Hoque, M.A.; Ferdous, S.; Khan, M.; Tarkoma, S. Real, Forged or Deep Fake? Enabling the Ground Truth on the Internet. *IEEE Access* **2021**, *9*, 160471–160484. [CrossRef]
18. Hu, C.; Feng, Z.; Wu, X.; Kittler, J. Dual Encoder-Decoder Based Generative Adversarial Networks for Disentangled Facial Representation Learning. *IEEE Access* **2020**, *8*, 130159–130171. [CrossRef]
19. Huang, D.; Tao, X.; Lu, J.; Do, M.N. Geometry-Aware GAN for Face Attribute Transfer. *IEEE Access* **2019**, *7*, 145953–145969. [CrossRef]
20. Laishram, L.; Lee, J.T.; Jung, S.K. Face De-Identification Using Face Caricature. *IEEE Access* **2024**, *12*, 19344–19354. [CrossRef]
21. Lim, S.; Gwak, Y.; Kim, W.; Roh, J.-H.; Cho, S. One-class learning method based on live correlation loss for face anti-spoofing. *IEEE Access* **2020**, *8*, 201635–201648. [CrossRef]
22. Liu, W.; Wei, X.; Lei, T.; Wang, X.; Meng, H.; Nandi, A.K. Data-Fusion-Based Two-Stage Cascade Framework for Multimodality Face Anti-Spoofing. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 672–683. [CrossRef]
23. Mahum, R.; Irtaza, A.; Javed, A. EDL-Det: A Robust TTS Synthesis Detector Using VGG19-Based YAMNet and Ensemble Learning Block. *IEEE Access* **2023**, *11*, 134701–134716. [CrossRef]
24. Malik, A.; Kuribayashi, M.; Abdullahi, S.M.; Khan, A.N. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* **2022**, *10*, 18757–18775. [CrossRef]
25. Oulad-Kaddour, M.; Haddadou, H.; Vilda, C.C.; Palacios-Alonso, D.; Benatchba, K.; Cabello, E. Deep Learning-Based Gender Classification by Training With Fake Data. *IEEE Access* **2023**, *11*, 120766–120779. [CrossRef]
26. Shahid, W.; Li, Y.; Staples, D.; Amin, G.; Hakak, S.; Ghorbani, A. Are You a Cyborg, Bot or Human?—A Survey on Detecting Fake News Spreaders. *IEEE Access* **2022**, *10*, 27069–27083. [CrossRef]
27. Waseem, S.; Abu Bakar, S.A.R.S.; Ahmed, B.A.; Omar, Z.; Eisa, T.A.E.; Dalam, M.E.E. DeepFake on Face and Expression Swap: A Review. *IEEE Access* **2023**, *11*, 117865–117906. [CrossRef]
28. Wei, J.; Lu, G.; Liu, H.; Yan, J. Facial Image Inpainting With Deep Generative Model and Patch Search Using Region Weight. *IEEE Access* **2019**, *7*, 67456–67468. [CrossRef]
29. Zhang, Y.; Hu, R.; Li, D.; Wang, X. Fake identity attributes detection based on analysis of natural and human behaviors. *IEEE Access* **2020**, *8*, 78901–78911. [CrossRef]
30. Bird, J.J.; Lotfi, A. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access* **2024**, *12*, 15642–15650. [CrossRef]
31. Fang, M.; Yang, W.; Kuijper, A.; Struc, V.; Damer, N. Fairness in face presentation attack detection. *Pattern Recognit.* **2024**, *147*, 110002. [CrossRef]
32. Habbal, A.; Ali, M.K.; Abuzaraida, M.A. Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Syst. Appl.* **2024**, *240*, 122442. [CrossRef]
33. Joshi, I.; Grimmer, M.; Rathgeb, C.; Busch, C.; Bremond, F.; Dantcheva, A. Synthetic Data in Human Analysis: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4957–4976. [CrossRef]
34. Kim, E.; Cho, S. Exposing Fake Faces Through Deep Neural Networks Combining Content and Trace Feature Extractors. *IEEE Access* **2021**, *9*, 123493–123503. [CrossRef]
35. Melnik, A.; Miasayedzenkau, M.; Makaravets, D.; Pirshtuk, D.; Akbulut, E.; Holzmann, D.; Renusch, T.; Reichert, G.; Ritter, H. Face Generation and Editing With StyleGAN: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3557–3576. [CrossRef]
36. Abbas, F.; Taeihagh, A. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Syst. Appl.* **2024**, *252*, 124260. [CrossRef]
37. Leyva, R.; Sanchez, V.; Epiphaniou, G.; Maple, C.; Leyva, R.; Sanchez, V.; Epiphaniou, G.; Maple, C. Data-agnostic Face Image Synthesis Detection using Bayesian CNNs. *Pattern Recognit. Lett.* **2024**, *183*, 64–70. [CrossRef]
38. Mania, K. Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study. *Trauma Violence Abus.* **2022**, *25*, 117–129. [CrossRef]
39. Oladoyinbo, T.O.; Olabanji, S.O.; Olaniyi, O.O.; Adebiyi, O.O.; Okunleye, O.J.; Alao, A.I. Exploring the Challenges of Artificial Intelligence in Data Integrity and its Influence on Social Dynamics. *Asian J. Adv. Res. Rep.* **2024**, *18*, 1–23. [CrossRef]
40. Thakur, R. Introduction to artificial intelligence and its importance in modern business management. *Leveraging AI Emot. Intell. Contemp. Bus. Organ.* **2023**, 133–165. [CrossRef]
41. Uddin, Z.; Shahriar, A.; Mahamood, N.; Alnajjar, F.; Pramanik, I.; Ahad, A.R. Deep learning with image-based autism spectrum disorder analysis: A systematic review. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107185. [CrossRef]
42. Wang, T.; Wu, D. Computer-Aided Traditional Art Design Based on Artificial Intelligence and Human-Computer Interaction. *Comput. Aided Des. Appl.* **2024**, *21*, 59–73. [CrossRef]
43. Yu, J.; Dickinger, A.; So, K.K.F.; Egger, R. Artificial intelligence-generated virtual influencer: Examining the effects of emotional display on user engagement. *J. Retail. Consum. Serv.* **2024**, *76*, 103560. [CrossRef]
44. Yu, Z.; Cai, R.; Li, Z.; Yang, W.; Shi, J.; Kot, A.C. Benchmarking Joint Face Spoofing and Forgery Detection With Visual and Physiological Cues. *IEEE Trans. Dependable Secur. Comput.* **2024**, 1–15. [CrossRef]
45. Rehaan, M.; Kaur, N.; Kingra, S. Face manipulated deepfake generation and recognition approaches: A survey. *Smart Sci.* **2024**, *12*, 53–73. [CrossRef]

46. Rana, S.; Sung, A.H. DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection. In Proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), New York, NY, USA, 1–3 August 2020; pp. 70–75. [CrossRef]
47. Liang, T.; Chen, P.; Zhou, G.; Gao, H.; Liu, J.; Li, Z.; Dai, J. SDHF: Spotting DeepFakes with Hierarchical Features. In Proceedings of the 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, 9–11 November 2020; pp. 675–680. [CrossRef]
48. Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
49. Chen, P.; Liu, J.; Liang, T.; Zhou, G.; Gao, H.; Dai, J.; Han, J. FSSPOTTER: Spotting face-swapped video by spatial and temporal clues. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6. [CrossRef]
50. Suratkar, S.; Bhiungade, S.; Pitale, J.; Soni, K.; Badgujar, T.; Kazi, F. Deep-fake video detection approaches using convolutional–recurrent neural networks. *J. Control Decis.* **2023**, *10*, 198–214. [CrossRef]
51. Kumar, P.; Vatsa, M.; Singh, R. Detecting Face2Face facial reenactment in videos. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 2578–2586. [CrossRef]
52. De Lima, O.; Franklin, S.; Basu, S.; Karwoski, B.; George, A. Deepfake Detection Using Spatiotemporal Convolutional Networks. *arXiv* **2020**, arXiv:2006.14749.
53. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11. [CrossRef]
54. Hong, S.-Y.; Park, D.; Yi, G. Disentanglement of Latent Factors of Real and Fake Appearance for Deepfake Face Manipulation Detection. *J. Stud. Res.* **2023**, *12*. [CrossRef]
55. Khan, S.A.; Dang-Nguyen, D.-T. Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms. *IEEE Access* **2023**, *12*, 1880–1908. [CrossRef]
56. Ramadhani, K.N.; Munir, R.; Utama, N.P. Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention. *IEEE Access* **2024**, *12*, 8932–8939. [CrossRef]
57. Guarnera, L.; Giudice, O.; Battiato, S. Mastering Deepfake Detection: A Cutting-Edge Approach to Distinguish GAN and Diffusion-Model Images. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–23. [CrossRef]
58. Xu, P.; Ma, Z.; Mei, X.; Shen, J. Detecting facial manipulated images via one-class domain generalization. *Multimed. Syst.* **2024**, *30*, 33. [CrossRef]
59. Jayashre, K.; Amsaprabhaa, M. Safeguarding media integrity: A hybrid optimized deep feature fusion based deepfake detection in videos. *Comput. Secur.* **2024**, *142*, 103860. [CrossRef]
60. Aloke, E.J.; Abah, J. Enhancing the Fight against Social Media Misinformation: An Ensemble Deep Learning Framework for Detecting Deepfakes. *Int. J. Appl. Inf. Syst.* **2023**, *12*, 1–14. [CrossRef]