IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 May 2024, accepted 17 June 2024, date of publication 19 June 2024, date of current version 26 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3416910

RESEARCH ARTICLE

Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces

SOHA SAFWAT[1], AYAT MAHMOUD[2], IBRAHIM ELDESOUKY FATTOH[3], AND FARID ALI[4]

[1]Software Engineering and Information Technology Department, Faculty of Engineering and Technology, The Egyptian Chinese University, Cairo 4541312, Egypt
[2]Department of Computer Science, Faculty of Computer Science, MSA University, Cairo 3750311, Egypt
[3]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni Suef 2722165, Egypt
[4]Department of Information Technology, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni Suef 2722165, Egypt

Corresponding author: Farid Ali (fared.ali@fcis.bsu.edu.eg)

ABSTRACT While human brains have the ability to distinguish face characteristics, the use of advanced technology and artificial intelligence blurs the difference between actual and modified images. The evolution of digital editing applications has led to the fabrication of very lifelike false faces, making it harder for humans to discriminate between real and made ones. Because of this, techniques like deep learning are being used increasingly to distinguish between real and artificial faces,producing more consistent and accurate results. In order to detect fraudulent faces, This paper introduces a pioneering hybrid deep learning model, which merges the capabilities of Generative Adversarial Networks (GANs) and the Residual Neural Network (RESNET) architecture, aimed at detecting fake faces. By integrating GANs' generative strength with RESNET's discriminative abilities, the proposed model offers a novel approach to discerning real from artificial faces. Through a comparative analysis, the performance of the hybrid model is evaluated against established pre-trained models such as VGG16 and RESNET 50. Results demonstrate the superior effectiveness of the hybrid model in accurately detecting fake faces, marking a notable advancement in facial image recognition and authentication. The findings on a benchmark dataset show that the proposed model obtains outstanding performance measures, including precision 0.79, recall 0.88, F1-score 0.83, accuracy 0.83, and ROC AUC Score 0.825. The study's conclusions highlight the hybrid model's strong performance in identifying fake faces, especially when it comes to accuracy, precision, and memory economy. By combining the generative capacity of GANs with the discriminative capabilities of RESNET, this solves the problems caused by more complex fake face generation approaches.With significant potential for use in identity verification, social media content moderation, cybersecurity, and other areas, the study seeks to advance the field of false face identification. In these situations, being able to accurately discriminate between real and altered faces is crucial. Notably, our suggested model adds Channel-Wise Attention Mechanisms to RESNET50 at the feature extraction phase, which increases its effectiveness and boosts its overall performance.

INDEX TERMS RESNET, generative adversarial networks, deep learning, real and fake faces, face detection, channel-wise attention.

I. INTRODUCTION

Images and movies with fake facial expressions produced through digital modification techniques have recently drawn increasing public criticism [1]. Deepfake is a term for artificial intelligence-produced, realistic-sounding, but fake, visuals, audio, and videos [2]. Deepfake is now more realistic and simpler to create because of recent improvements in deepfake generation. Deepfake has posed serious threat to society, and our right to privacy, necessitating the

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

development of deepfake detection techniques to counter these concerns [3], [4]. An individual known as Deepfakes [5] used publicly accessible artificial intelligence application to produce pornographic videos in December 2017 in which real faces were replaced with fake faces in photos and videos. Deepfakes is a user of the Reddit social media network [6]. The substitution of an individual's appearance, especially faces, using artificial intelligence algorithms is known as ''Deepfaking''. A particular type of synthetic media known as ''deepfake'' employs deep learning-based software to produce deceptive films, recordings, and/or photos. It entails swapping out one person's face in a photo or video with another person's likeness to produce a realistic imitation with the aim of deceiving viewers or altering content's genuine message [7]. The majority of deepfake detection techniques rely on features and machine learning techniques. Deepfake generation advances, a dearth of high-quality datasets, and a lack of benchmarks are some of the remaining difficulties in deepfake detection. Deepfake detection trends for the future may include robust, efficient, and systematic detection techniques as well as high-quality datasets [8]. GANs technology has made it possible to produce extremely lifelike face images that are visually challenging to differentiate real faces [9]. The generation process and discriminator, which are the two parts of a Generative Adversarial Network, collaborate to produce untrue photos which might be challenging to differentiate from real photos. As the discriminator is trained to distinguish between fake photos and real photos, the generator produces the fake pictures [10]. The generator tries to create more convincing photos with the aim of tricking the discriminator throughout training process, whereas the discriminator gets better at spotting untrue images. GANs are utilized for creating images of individuals, animals, and objects, but they may also be used to create fraudulent images for malicious purposes [11]. What is worse, humans struggle to recognize these convincing deep fake images, audios, and films. Therefore, it is crucial, imperative, and necessary to differentiate true media from deepfakes. Therefore, it is essential to create a reliable model that can precisely differentiate between real and fake photos. Due to the recent spike in the risk of fraudulent operations, numerous methods to identify phony face photos have been developed to solve this issue [12]. These techniques can be roughly divided into two groups: one group relies on manually created characteristics and depends on the statistical properties of the photos. The other group makes use of deep learning methods that utilize cutting-edge neural networks to find patterns and characteristics in the photos [13]. This paper is organized in six main sections. Section I states the research challenge, and emphasizes the importance of the subject and the goals of the investigation. An overview of the pertinent background information and associated studies is provided in Section II. The materials and methods used are presented in Section III. The suggested model is presented in Section IV, together with information on its architecture, design, and implementation. The implementation results and their discussion are presented in Section V. Conclusions and key contributions to the field and the directions for future work are outlined in Section VI.

II. BACKGROUND AND RELATED WORK

The deployment of realistic Deepfake images could be dangerous for people's privacy, democratic processes, and the nation's security [14]. The creation of trustworthy tools for spotting hazardous Deepfake material is essential. Machine learning methods and feature-based ones make up the two primary types of Deepfakes detection techniques [6]. To distinguish between deepfakes, machine learning methods, particularly deep learning, are frequently used. Feature-based algorithms exploit specific properties found in Deepfake media to identify them. As there is a critical need to stop the spread of damaging media, this study concentrates on machine learning methods to identify deepfakes. Machine learning methods are divided into two primary categories: standard techniques and deep techniques [6]. Traditional machine learning techniques involve strategies to analyze data along with producing predictions or classes depending on statistical models and algorithms [12]. It is used in SVM and RF-based Deepfake detection techniques. Based on statistical models, these methods seek to analyze the data and produce predictions or classes (groups). Traditional ML frequently necessitates hand-engineering features. However, due to their speed, ease of use, and robustness against noisy datasets, these techniques are still often used in numerous applications. Support Vector Machine (SVM) is a machine learning technique used for regression analysis and categorization. SVM can be used in Deepfake detection to discriminate between genuine and fake content. SVM may be trained using a dataset of actual and Deepfake photos and videos [7] for Deepfake identification, where it learns to differentiate between the two classes. Once taught, it can be used to determine the category of upcoming, undiscovered movies or photographs. To identify more than two classes of Deepfakes, several SVMs would need to be trained, which is one of the key drawbacks of this method. However, because SVM is a binary classifier which means it operates or differentiate between only two classes [15]. A machine learning approach called random forest (RF) can be used for classification, regression, and other applications. Random forest is used as a classifier in deep fake detection to differentiate between real and fraudulent content. Since it can handle an enormous number of characteristics and can determine which characteristic are considered more crucial for classification, random forest may serve as a beneficial method in deep fake detection. Furthermore, compared to other classifiers, it is less susceptible to overfitting, which makes it more resistant to noisy or defective data [16]

DeepFaceLab (2019) [17] is software application used to manipulate facial images. A Russian smartphone application named FaceApp, for instance, has the capability to generate deceptive photographs that appear older than the subjects actually are. A piece of software called Deepfakes can be used to swap out a human face with that of any other person or

animal. With the aid of machine learning and human image synthesis, DeepFaceLab is a Windows program that lets users replace faces in videos [18]. The article investigates how undiscovered medical deepfakes might affect patient safety as well as the assets of hospitals. To create techniques for identifying such attacks, the researchers carried out a case study. Support Vector Machine, Random Forest, and Decision Tree were among the eight machine learning algorithms that were put to the test [19]. Deep learning techniques, as opposed to traditional machine learning models, can discover Deepfake properties and have grown to be a popular way for identifying Deepfakes. These techniques include GAN, CNN, and RNN as examples. Furthermore, compared to other techniques, deep learning-based detection algorithms typically produce higher levels of accuracy [13]. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks are only a few of the deep learning methods that are presented in the article, cited in [7] for various applications. By identifying genuine from false photos, these techniques can be utilized to identify Deepfakes. Below is an overview of how various techniques can be used to identify Deepfake content. A deep neural network model called the CNN comprises some hidden layers, an input layer, and output layer. The hidden layers take inputs from top layer and convolution the input values. The matrix multiplication or dot product is used in this convolution procedure. Then, further transformations like pooling layers are used together with a nonlinearity activation function like the Rectified Linear Unit (RELU). By computing the outputs using functions like maximum pooling or average pooling, pooling layers seek to reduce the complexity of the input data [20]. Multiple layers make up ANNs, involving one input layer, some hidden layers, and one output layer.Input data sets are utilized as inputs in Artificial Neural Networks, which the network endeavors to classify. Signal spread occurs via connections, known as edges, between the interconnected points or synthetic neurons in ANNs, which has an architecture like that of the human brain. After processing the signals, each neuron sends the signals received to the neurons connected to it. An edge and neuron-related weight is used to modify the intensity of the signal at a link [16]. Therefore, it is crucial to understand not only the deep learning methods stated before, but also the traditional neural network (NN) and how it relates to traditional machine learning. The traditional NN is a popular variety of neural network that is used in tasks involving supervised learning like classification and regression. Traditional neural network (NN) is made up of some hidden layers, one input layer, and one output layer. The hidden layers contain nodes which calculate weighted inputs and provide an output. Artificial Neural Networks (ANNs) are based on the core principle that the human brain functions in a similar manner.

The Deep InceptionNet Learning Algorithm Introduced by [21], is used to detect deepfake images. The study achieves a noteworthy accuracy of 93% when compared to other convolutional networks, demonstrating the algorithm's effectiveness in differentiating between true and altered content.

In [22], the author reviews the literature on several deep learning strategies for identifying created fake faces. The author highlights the importance of reliable detection methods given the quick advancement of AI-driven multimedia alteration. In order to create a more precise and succinct deepfake detection system, methods including CNN, Xception Network, Recurrent Neural Networks (RNN), and Long Short Term Memory (LSTM) are investigated.

The goal of the DeepFakeDG project by [23] was to create a web application that uses machine learning and deep learning techniques to identify falsified information. The study tackles the issues raised by deepfake algorithms by utilizing methods like face swapping and behavioral analysis, highlighting the possible uses of deepfake detection in legal and law enforcement settings.

Examining Vision Transformers (ViTs) for multiclass deepfake picture detection is a unique approach to the rapidly changing field of facial modification technology, as suggested by [24]. The study is the first to take into account the StyleGAN2 and Stable Diffusion problems. ViTs outperform conventional CNN-based models in terms of detection accuracy, precision, and recall.

The authors of [25] concentrate on the use of artificial intelligence (AI), machine learning, and neural networks in conjunction with deep learning approaches to classify actual and fake human faces. The study's impressive accuracy, attained by using deep learning algorithms like ResNet50, highlights the promise of these methods in differentiating between real and fake facial photos. In summary, the literature review highlights the ongoing progress in deepfake detection techniques, tackling the various issues brought about by developing multimedia manipulation technologies.

III. MATERIALS AND METHODS

A. RESIDUAL NEURAL NETWORK

Residual Neural Network, or ResNet, is a deep learning architecture that was proposed by [26]. It is widely used in computer vision tasks and has achieved state-of-the-art performance on various image recognition challenges [27], [28]. The main idea behind ResNet is the introduction of residual connections [29], which allow for the efficient training of very deep neural networks. ResNet architecture typically consists of several convolutional layers followed by residual blocks. A residual block is composed of multiple convolutional layers with shortcut connections bypassing these layers [30]. This structure enables the network to learn residual functions representing the difference between the input and the desired output, making the learning process more efficient [30]. The ResNet architecture consists of multiple layers [31], including convolutional layers, residual blocks, and an output layer as Figure 1 shows. The input represents initial input image or feature map. The input

passes through a convolutional layer, which applies a set of learnable filters to extract features from the input. A residual block consists of two or more convolutional layers with shortcut connections. The input to the block is passed through the convolutional layers, and the output is added to the original input through the shortcut connection. This bypass allows the network to learn the residual function—the difference between the input and the desired output. The residual function makes it easier to train very deep networks.

1) CHANNEL-WISE ATTENTION MECHANISMS

An important development in deep learning architectures is Channel-Wise Attention Mechanisms, especially in convolutional neural networks (CNNs), where the ability to recognize complex patterns is critical. During the feature extraction process, these methods selectively highlight pertinent feature channels while suppressing noise and unnecessary data. Channel-Wise Attention Mechanisms provide numerous benefits to performance when they are incorporated into the feature extraction stage of RESNET50, a well-known CNN architecture that is distinguished by its deep layers and skip connections. First of all, they allow for selective feature focus, which makes sure the network highlights important characteristics that are essential for differentiating between real and modified images, such those found in false face identification tasks. Additionally, these technologies support adaptive feature representation, which enables the model to dynamically modify feature representations in response to input data, hence boosting discriminative skills and capturing subtle variations.Moreover, by reducing the impact of unimportant changes, their integration strengthens generalization, promoting robustness against adversarial perturbations and enhancing performance on unknown data. Surprisingly, these performance gains are attained with merely a slight rise in computing complexity, making Channel-Wise Attention Mechanisms suitable for practical implementation in real-world scenarios without substantial overhead. As a result, its incorporation into RESNET50 greatly increases its performance in jobs requiring accurate feature extraction, such as false face identification, among others [32].

After several residual blocks, the network typically applies global average pooling, which computes the average value of each feature map. This reduces the spatial dimensions of the feature maps and aggregates the learned information across the entire image. Finally, the global average pooled features are passed through a fully connected layer or a softmax layer to produce the desired output, such as class probabilities.

B. GENERATIVE ADVERSARIAL NETWORKS (GANs)

GAN is a type of deep learning architecture that is used for generating new data samples, such as images introduced in [17]. A typical GAN consists of two components: generator and discriminator, where both networks compete with each other. The generator is the heart of the GAN, where it attempts to generate fake data that looks real by learning the features from the real data. The discriminator evaluates the generated data with the real data and classifies whether the generated data looks real or not and provides feedback to the generator to improve its data generation. The goal of the generator is to generate data that can trick the discriminator. The architecture of the basic model of GAN is shown in Figure 2.

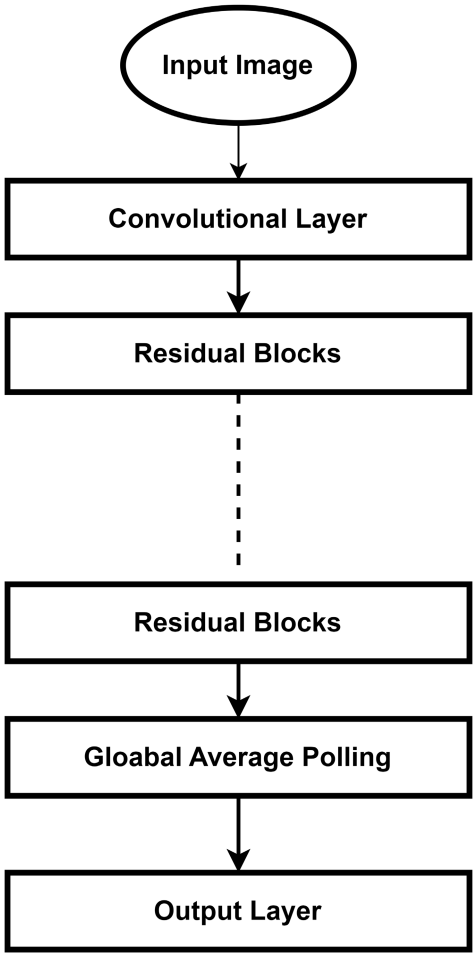

FIGURE 1. ResNet architecture.

IV. PROPOSED MODEL AND DATASET

A thorough explanation of the proposed model and its method for identifying real and fake faces is provided in this section. A critical problem in face recognition is addressed by the proposed model in this study. The model can be useful in areas like security and criminal investigation because it can distinguish between actual and fraudulent photos effectively. The important elements of the proposed model, including the use of machine learning, deep learning methods, and spatial domain features, will be briefly discussed. The methodologies used in the study will also be

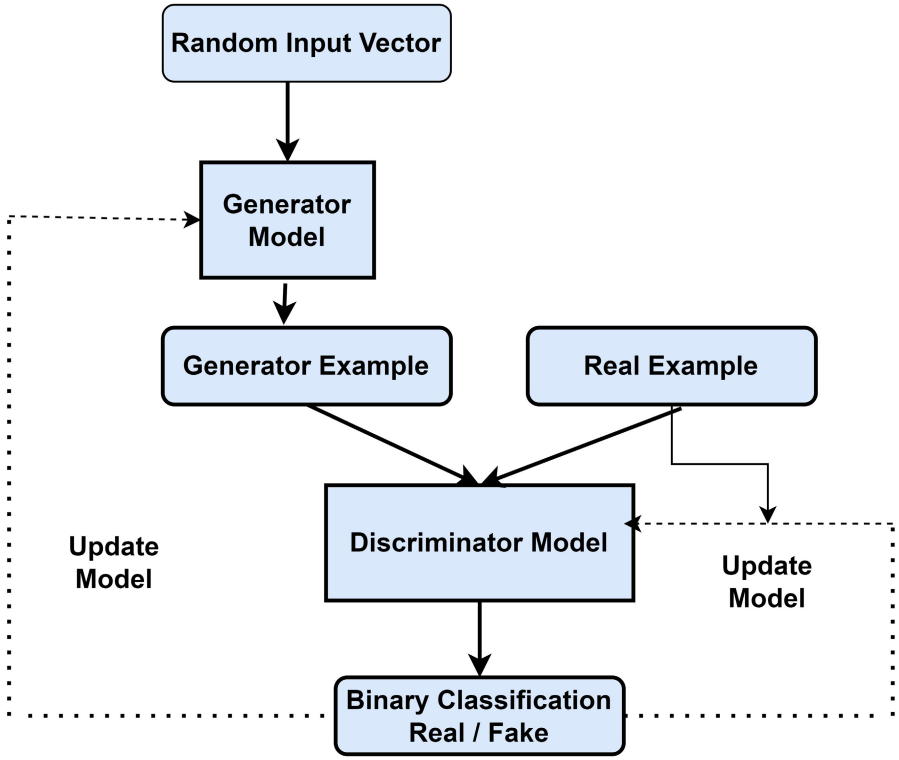

FIGURE 2. Architecture of GAN.

examined, with a focus on their advantages and potential drawbacks.

A. DATASET

The dataset used in this study was introduced in [1], The Real and Fake Face Detection dataset is a widely used benchmark dataset contains 2,041 face images, 1,081 images labeled as real images such as Figure 3(a) and 960 images are labeled as fake images such as Figure 3(b). The benchmark is used for assessing the effectiveness of various face detection models in distinguishing between real and fake images. Different digital image manipulation methodologies are also used to generate the fake images in this dataset like face swap, face2face, and Deepfakes.

FIGURE 3. Example of real and fake face used in training phase.

B. PROPOSED MODEL STEPS AND ARCHITECTURE

The proposed architecture consists of six phases. The overall architecture of the proposed model is shown in Figure 4. The first phase is data preprocessing. First, data cleaning is applied to check the dataset for any corrupt or mislabeled images. Any problematic images are removed to ensure data integrity and prevent the model from learning from incorrect or noisy samples. Then, data augmentation techniques are applied to increase the dataset's size and diversity. Common augmentations include rotation, flipping, scaling, and random crops. This step helps the model become more robust and more able to generalize better on unseen data. Next, all images are resized to a consistent size that can be fed into the deep learning model. Deep learning models, such as ResNet, typically require images of fixed dimensions. Common choices are 224 $\times$ 224 pixels. At last, the pixel values of the images are normalized to bring them to a common scale. The most common approach is to scale the pixel values to the range [0, 1]. This step helps the model converge faster during training and prevents issues related to different pixel value scales. In this research endeavor, modifications were introduced to the standard ResNet model architecture to optimize its suitability for a specific analytical task. Initially, the ResNet50 model, a pre-trained Convolutional Neural Network (CNN), was selected as the base framework. To tailor the model to feature

extraction objectives, a departure was made from the standard practice of solely discarding the final classification layers. A ResNet50 model serves as the foundational Convolutional Neural Network (CNN) architecture throughout the feature extraction stage. On the other hand, attention mechanisms are incorporated into the ResNet backbone, in contrast to conventional methods that discard the final classification layers. In particular, to capture attention-weighted feature representations, attention modules are introduced after particular convolutional layers. By dynamically adjusting the significance of various spatial regions within the feature maps, these attention modules allow the model to concentrate on pertinent facial characteristics and manipulation artifacts.

Instead, substantive enhancements were introduced to augment both training efficiency and predictive accuracy. This involved incorporating additional convolutional layers into the ResNet architecture, fine-tuning them to discern intricate features within the image dataset. Additionally, optimization of kernel sizes was performed to better capture nuanced patterns in the data. Innovative regularization techniques, including dropout and batch normalization, were deployed to mitigate overfitting risks and enhance the model's generalization ability. The efficacy of these modifications was rigorously evaluated through systematic experimentation, providing empirical evidence of their substantial impact on training efficacy and predictive performance. Subsequently, the modified ResNet model, enriched with bespoke alterations, was employed for feature extraction, resulting in robust feature representations. These extracted features, imbued with tailored modifications, served as input for downstream analytical tasks, including classification, clustering, and feature similarity analysis. Through orchestrated enhancements, this study distinguishes itself from the conventional ResNet framework, underscoring its superior efficacy and adaptability for the targeted analytical domain.

The third phase involves generating fake faces using GAN. This is done through training a GAN to generate realistic fake face images. The generator network takes random noise as input and generates fake face images. The discriminator network tries to distinguish between real and fake faces. The GAN is then trained using a combination of adversarial and reconstruction losses to ensure realistic fake face generation. The fourth phase includes the proposed hybrid model. First, it takes the feature representations obtained from the CNN as input. Then additional layers (e.g., fully connected layers) are added to the CNN's feature representation. The output of the additional layers is then connected to the GAN's discriminator network. Next, the combined model is trained by freezing the CNN layers and updating the GAN's discriminator and additional layers. If necessary, the entire hybrid model is fine-tuned. The fifth phase is Training, in which a labeled dataset of real and fake face images is used for training. The hybrid model is trained using a suitable loss function (e.g., binary cross-entropy) to classify real and fake faces. Accordingly, the hybrid model's weights are updated using backpropagation and gradient descent. The sixth phase is Evaluation, in which the performance of the hybrid model is assessed on a separate validation or test dataset. Metrics such as accuracy, precision, recall, and F1-score to are calculated to evaluate the model's effectiveness in fake face detection.

V. RESULTS AND DISCUSSIONS

In this section, the obtained results of many experiments and the proposed model are introduced, but initially the section briefly describes the different measures used to evaluate the performance of these models.

Sensitivity (Recall): sensitivity measures the proportion of true positives that are correctly identified as such. In other words, it is the probability that a test will correctly identify a positive case.

$$Sensitivity = TP/(TP + FN). \tag{1}$$

where TP is True Positive, FN is False Negative.

Precision: Precision measures the fraction of positive predictions that are actually positive

$$Precision = TP/TP + FP. \tag{2}$$

where FP is False Positive.

Accuracy: accuracy measures the fraction of predictions that are correct, regardless of whether they are positive or negative.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN). \tag{3}$$

F1 Measure: is a weighted average of precision and recall. It is calculated by taking the harmonic mean of precision and recall.

$$F1Measure = 2 * (precision * Recall)/(precision + Recall). \tag{4}$$

In order to guarantee the stability and applicability of the suggested hybrid deep learning model, a careful data division strategy is adopted, dividing the dataset into 70% for training and 30% for testing. This partitioning technique allowed for thorough evaluation of our model's performance, letting it learn from most of the data while undergoing a thorough analysis on a different, unseen pieces. During the training phase, we also used a k-fold cross-validation (CV) technique is also employed, which entailed splitting the training data into several folds and training and verifying the model iteratively. The model's capacity to generalize across various training data subsets was further guaranteed by this method. The 30% set aside for testing functioned as an independent dataset, unaltered during model development, to simulate real-world circumstances and improve the model's applicability. The Objective of the integration of data splitting and k-fold CV is reinforce the dependability of the results and highlight the model's efficiency in a variety of situations.

In this research, many experiments were applied on the dataset to compare their results with the proposed model. Firstly, we applied the VGG16 which is a deep convolutional

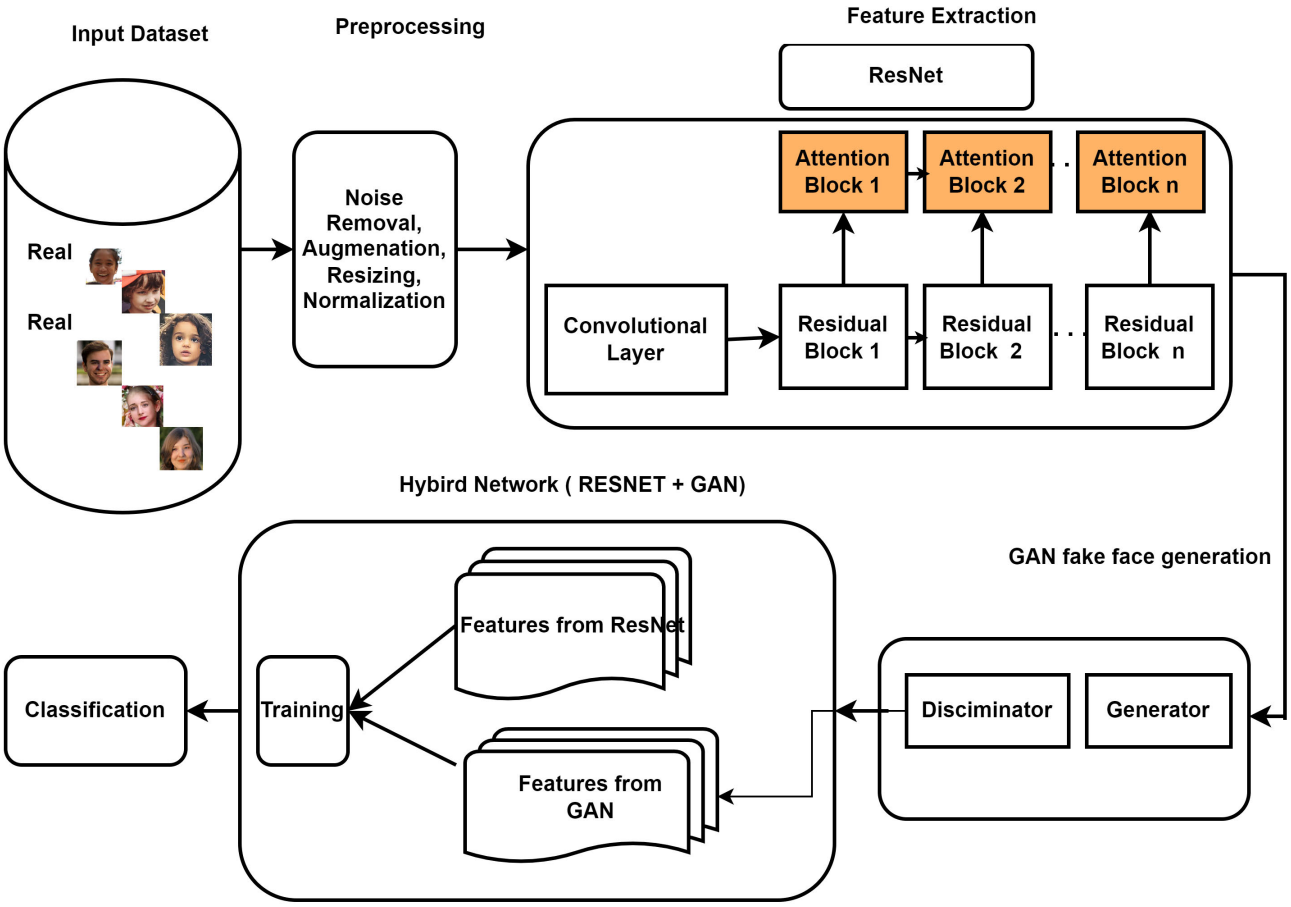

FIGURE 4. The overall architecture of the proposed model.

TABLE 1. VGG16 results.

Precision	Recall	F1-score	Accuracy	ROC AUC Score
0.6232	0.6224	0.6226	0.6260	0.6421

TABLE 2. ResNet-50 results.

Precision	Recall	F1-score	Accuracy	ROC AUC Score
0.7264	0.7265	0.7260	0.7263	0.6991

neural network architecture known for its simplicity and effectiveness, is applied. It consists of 16 weight layers, including convolutional and fully connected layers. It follows a repeated pattern of using small 3×3 convolutional filters followed by max-pooling layers. VGG16's main contribution is in demonstrating the benefits of using deep networks for image classification. The results obtained from the VGG16 network is reported in Table 1 and Figure 5.

Table 2 and Figure 6 show the results yielded from the second experiment, in this experiment, ResNet-50 is used to classify the real faces and fake faces

The results of the third experiment that yielded from the proposed model, which is the hybrid between the ResNET-50 and the GAN algorithm is presented in Table 3 and Figure 7. The hybrid model attempts to find the optimum generated

TABLE 3. Results of the proposed model.

Precision	Recall	F1-score	Accuracy	ROC AUC Score
0.7916	0.8824	0.8345	0.8298	0.825

images from GAN starting from 100 image and the best value when using 400 images. The training and testing ratio used 70 % and 30 %.

The model underwent exhaustive training over 100 epochs, each spanning approximately 10 hours, on a workstation equipped with a single NVIDIA GPU, 16 GB of RAM, and a 6-core Intel i7 processor. Parameter selection, including an input size of (224, 224) and a batch size of 64, was guided by rigorous experimentation. Custom layers seamlessly integrated into the model augmented its discriminative capabilities, leveraging the robustness of the ResNet-50 architecture pretrained on ImageNet. Additionally, data augmentation techniques, such as rotation, width and height changes, and horizontal flips, were employed to enhance the model's ability to identify complex elements in facial photographs. The utilization of the Adam optimizer with binary crossentropy loss contributed to improved accuracy. Despite the ResNet model's known demand for a substantial number of parameters, resulting in a bulky size, the proposed

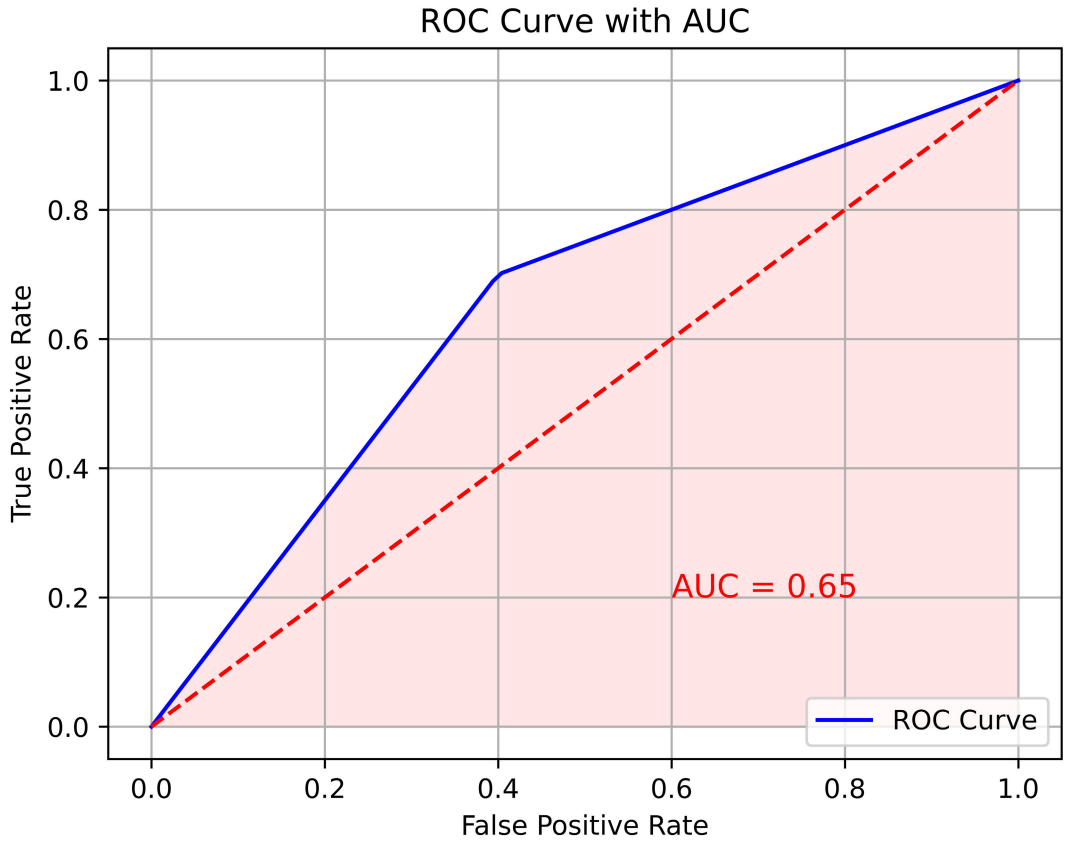

FIGURE 5. False positive rate vs true positive rate for VGG 16 network.

TABLE 4. Architecture of VGG16 and ResNET-50.

Architecture	VGG16	ResNET-50
Batch size	64	64
Number of Epochs	100	100
Learning Rate	1e-4	1e-3
Optimizer	SGD optimizer	Adam optimizer
Total parameters	134 M	25.6 M

model structure and parameters represent the culmination of iterative refinement and experimentation, reflecting the best configuration achieved through exhaustive optimization efforts. Acknowledging the potential for further enhancements in overall accuracy and research outcomes through modifications to the ResNet model, future work will explore the application of optimization techniques or heuristic methods to systematically identify optimal parameters.

From the previous results, it can be noticed that the results obtained from ResNET- 50 are better than VGG16 due to residual connections. As a way to enhance the results of the ResNET-50, the proposed model was applied by hybridizing ResNET-50 with GAN algorithms. As shown in table 3, the accuracy results of the proposed model reached above 83% which is better than the results obtained from the ResNET-50 network by nearly 10%. Figure 8 shows an overall comparison between the proposed hybrid model and VGG16 and ResNET-50.

A comparison between the proposed model in this research and a model that uses ResNET 18 implemented in [1] is presented in Table 5.

When compared to previous studies, the third model—a hybrid that combines RESNET50 with a GAN algorithm—shows better results, especially when compared to more conventional models like VGG16 and stand-alone deep architectures like RESNET50. The hybrid model's noteworthy success can be ascribed to a number of important features that also improve its fake face detecting ability.

Firstly, the hybrid model makes use of both the generative and discriminative components' advantages. The discriminative core is RESNET50, which is renowned for its deep and efficient feature extraction capabilities. As a result, the model can distinguish between minute characteristics and patterns linked to both authentic and synthetic facial features. In addition to introducing a generative component, the addition of a GAN allows the model to identify fake faces that already exist as well as potential variants or new instances of synthetic faces that might appear in the future. Second, the GAN component improves the model's generalization over a wide variety of fictitious face variants by introducing a novel type of data augmentation during training. The hybrid model gains exposure to a wider dataset by producing realistic synthetic faces. This can potentially mitigate the risk of overfitting and enhance its resilience in real-world situations

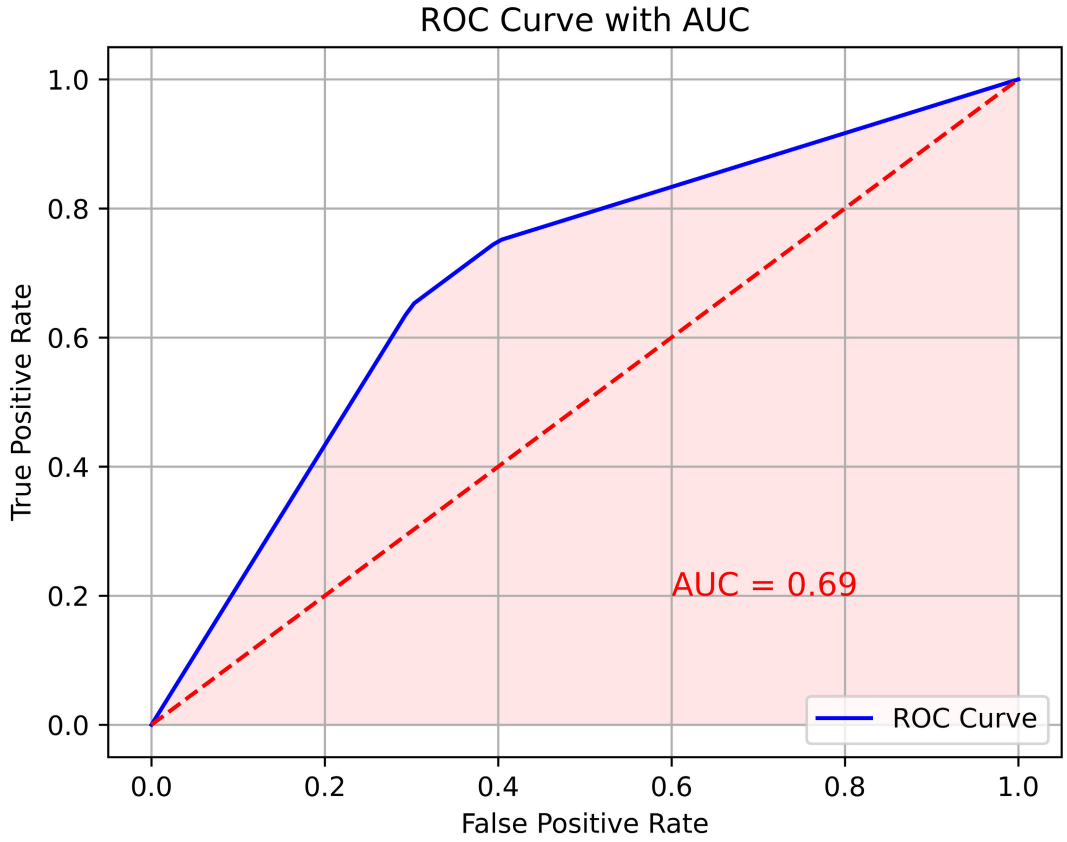

FIGURE 6. False positive rate vs true positive rate for ResNet-50 network.

where the emergence of false faces can be very unpredictable and dynamic. In addition, the effectiveness of the hybrid model emphasizes how crucial it is to take the complete context of false face detection into account. The hybrid model's incorporation of GAN-generated images helps it to better understand the subtleties of facial structure, expression, and realism—factors crucial in distinguishing sophisticated fake faces that may elude the detection capabilities of simpler models—even though deep learning architectures like VGG16 and RESNET50 are skilled at capturing intricate features. Moreover, the success of the hybrid model implies that constraints seen in traditional models may be addressed by carefully combining discriminative and generative approaches. This discovery highlights the value of hybrid architectures in pushing the limits of accuracy and dependability in fake face identification and creates new research opportunities. In summary, the third model outperforms the others because it combines the generative skills of a GAN with the discriminative power of RESNET50 in a synergistic manner. This special combination not only improves feature discrimination but also presents a fresh way to deal with the problems caused by constantly changing fake face creation methods. The hybrid model's effectiveness offers important insights for next image processing and artificial intelligence research and applications as the field of fake face identification advances.

TABLE 5. Comparison between the proposed model and model that uses ResNET 18.

Evaluations	Proposed hybrid model	Model in [1] ResNet 18
Precision	0.7916	0.79
Sensitivity	0.8824	0.73
Accuracy	0.8298	0.77

In many cases, using a single architecture like RESNET50 or VGG16 can produce poor results compared to a hybrid model that combines a GAN and a RESNET50 (Residual Network). The qualities of both components working together give rise to this advantage. Due to its deep design, RESNET50 excels in extracting detailed features from images, whereas GANs have the capacity to create synthetic data instances that mimic the training dataset. This hybridization combines the data generating power of GANs with the feature extraction power of RESNET50 to provide more diversified and informative features for classification problems. The GAN-RESNET50 hybrid model also contributes to the augmentation and improvement of the data. GANs can produce extra synthetic data, resolving problems with insufficient training data and improving the model's capacity to generalize to new data. The retrieved characteristics are also refined by the GAN's ability to differentiate between real and produced data, potentially improving their suitability for classification tasks. The hybrid

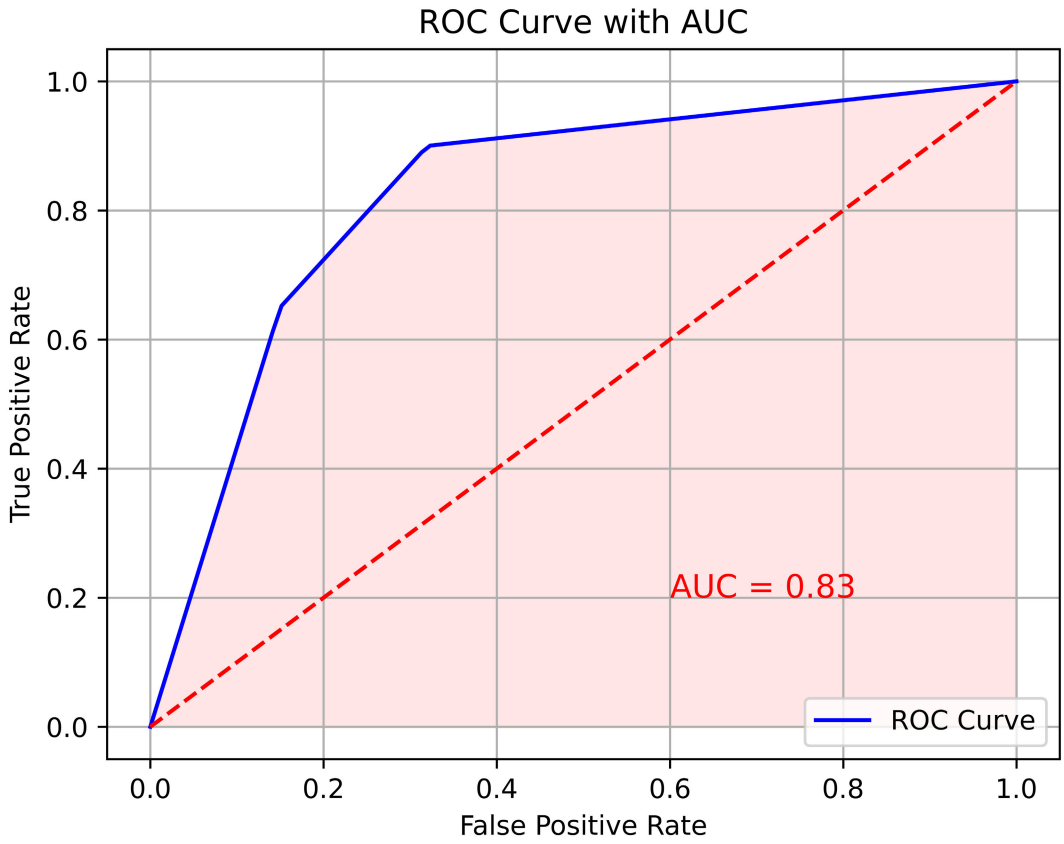

FIGURE 7. False positive rate vs true positive rate for the proposed model.

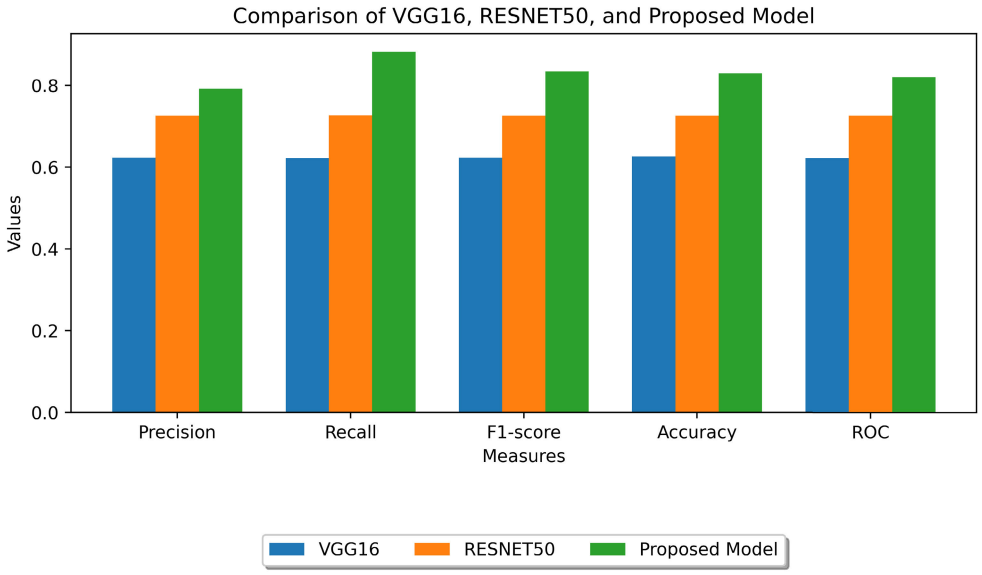

FIGURE 8. Comparison between proposed model vs VGG16 and ResNET-50.

model captures underlying data distributions by utilizing GANs for unsupervised pretraining, leading to more efficient feature extraction during the next fine-tuning stage. This method is very useful in situations where there are noisy or unbalanced datasets. To improve class separation during classification, the GAN can produce synthetic samples for minority classes or clean noisy training data. The hybrid technique is also effective for domain adaptation tasks where there are distribution mismatches between the source and target domains. The model can improve its performance in the target domain by domain adapting features from the source domain to it. A powerful ensemble effect results from the interaction

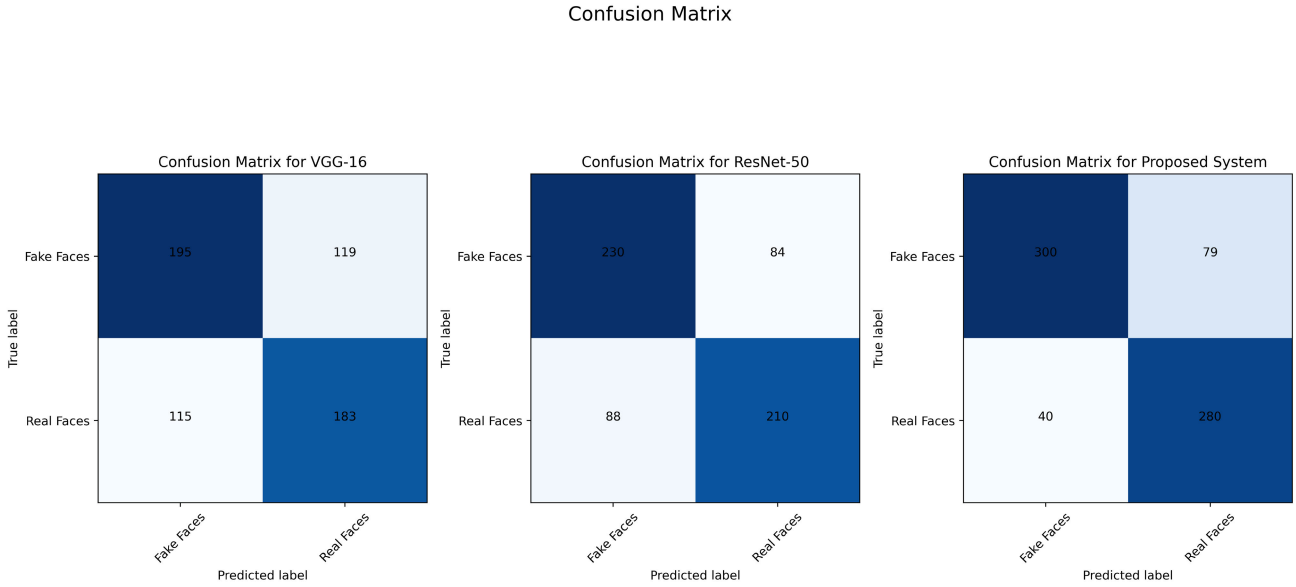

FIGURE 9. Confusion matrix.

of GANs with RESNET50. When discriminative features learnt by RESNET50 are paired with the diversity added by GAN-generated data, model performance is frequently improved. It is important to stress that the effectiveness of this hybrid technique depends on a number of variables, including dataset qualities, task difficulty, the success of GAN training, and architectural decisions. To determine whether a hybrid GAN-RESNET50 model actually performs better than standalone models like RESNET50 or VGG16 for a particular job, careful testing and analysis are required. There are a number of reasons for the suggested hybrid deep learning model's performance, and ways to make it even better are constantly sought. The model's initial success can be attributed to its clever use of GANs and the RESNET architecture, which combines the advantages of both technologies. The synergistic effect between the generative capability of GANs and the discriminative skills of RESNET improves the model's ability to distinguish between authentic and fake faces.

There are strong benefits to improving detection performance when attention processes are included into a ResNet-based phony face detection model. Selective focus is made possible by attention mechanisms, which let the model focus on important parts of the input image while ignoring unimportant parts. This improves prediction accuracy by allowing the model to prioritize the examination of particular visual features or manipulation artifacts indicative of forging in the context of fake face detection. Moreover, by highlighting pertinent areas of the input image, attention techniques improve feature representation and help conventional CNN architectures pick up on minute details or patterns connected to phony faces. This feature augmentation increases the model's resilience to changes in lighting, facial emotions, and image quality in addition to improving detection accuracy. Furthermore, by producing attention maps that clarify which elements of the input image impact predictions and facilitate comprehension of the model's decision-making process, attention mechanisms enhance interpretability. Furthermore, attention mechanisms serve as a regularization strategy, preventing overfitting by motivating the model to suppress noise or extraneous data and concentrate on relevant features. This improves generalization capacity and guarantees more dependable detection performance across a variety of datasets and real-world situations.

Confusion matrices were used to assess the effectiveness of three classification techniques: VGG-16, ResNet-50, and the Proposed System, as shown in Figure 9. With 195 true positives and 183 true negatives, VGG-16 showed a balanced performance; nevertheless, its false positive and false negative rates were rather high at 119 and 115, respectively. With 230 true positives, ResNet-50 showed better sensitivity; however, this came at the expense of a larger false negative rate of 88. On the other hand, the Proposed System demonstrated significant improvements in accuracy and sensitivity, obtaining 300 true positives and 280 true negatives. Furthermore, the Proposed System showed a better trade-off between true positive and false positive rates than both VGG-16 and ResNet-50, despite having a little higher false positive rate of 90. These results indicate that the Proposed System has a strong advantage over existing approaches such as VGG-16 and ResNet-50 in classification tasks, demonstrating its potential for enhanced performance.

The dedication to ongoing development is in line with how fake face production technology are developing. The goal is to improve the model's performance and dependability in real-world applications including identity confirmation, social media content moderation, and cybersecurity by fine-tuning its decision-making processes, addressing

potential biases, and increasing the diversity of training data.

VI. CONCLUSION

In this research, the authors provide a novel hybrid deep learning model to tackle the rising problem of recognizing fake faces in an era of deepfake technology and increasingly sophisticated picture alteration techniques. In order to develop a reliable and precise method for distinguishing real from fake facial photos, the study made use of the features of the RESNET architecture after applying Channel-Wise Attention Mechanisms and GANs. On a benchmark dataset, the suggested model performed superbly, obtaining high precision, recall, F1-score, accuracy, and ROC AUC score. These findings highlight the model's efficiency and dependability in the critical task of detecting fake faces. The contribution is significant because it has the potential to be used in many other fields, such as cybersecurity, identity verification, and social media content control. In these domains, the ability to discriminate between real and altered faces is crucial, and our hybrid model provides a potent tool for tackling this problem. Future research in the field of fake face detection should focus on a few crucial areas to further improve the capabilities of hybrid deep learning models. For example, new deep learning architectures should be investigated, and optimization techniques should be investigated to increase the model's precision, recall, and overall accuracy. Increased detection performance can be facilitated by state-of-the-art structures and well calibrated parameters. Finally; future work could certainly explore cross database evaluations to further validate the generalizability of the proposed model across different datasets and scenarios.

REFERENCES

[1] N. A. S. Eldien, R. E. Ali, and F. A. Moussa, "Real and fake face detection: A comprehensive evaluation of machine learning and deep learning techniques for improved performance," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jul. 2023, pp. 315–320.

[2] Y. Zhu, C. Zhang, J. Gao, X. Sun, Z. Rui, and X. Zhou, "High-compressed deepfake video detection with contrastive spatiotemporal distillation," *Neurocomputing*, vol. 565, Jan. 2024, Art. no. 126872.

[3] A. Gandhi and S. Jain, "Adversarial perturbations fool deepfake detectors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[4] M. M. El-Gayar, M. Abouhawwash, S. S. Askar, and S. Sweidan, "A novel approach for detecting deep fake videos using graph neural network," *J. Big Data*, vol. 11, no. 1, p. 22, Feb. 2024.

[5] O. B. Newton and M. Stanfill, "My NSFW video has partial occlusion: Deepfakes and the technological production of non-consensual pornography," *Porn Stud.*, vol. 7, no. 4, pp. 398–414, Oct. 2020.

[6] W.-D. Zhou, L. Dong, K. Zhang, Q. Wang, L. Shao, Q. Yang, Y.-M. Liu, L.-J. Fang, X.-H. Shi, C. Zhang, R.-H. Zhang, H.-Y. Li, H.-T. Wu, and W.-B. Wei, "Deep learning for automatic detection of recurrent retinal detachment after surgery using ultra-widefield fundus images: A single-center study," *Adv. Intell. Syst.*, vol. 4, no. 9, Sep. 2022, Art. no. 2200067.

[7] A. M. Almars, "Deepfakes detection techniques using deep learning: A survey," *J. Comput. Commun.*, vol. 9, no. 5, pp. 20–35, 2021.

[8] X. Chang, J. Wu, T. Yang, and G. Feng, "DeepFake face image detection based on improved VGG convolutional neural network," in *Proc. 39th Chin. Control Conf. (CCC)*, Jul. 2020, pp. 7252–7256.

[9] Y. Fu, T. Sun, X. Jiang, K. Xu, and P. He, "Robust GAN-face detection based on dual-channel CNN network," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–5.

[10] N.-T. Do, I.-S. Na, and S.-H. Kim, "Forensics face detection from GANs using convolutional neural network," in *Proc. ISITC*, 2018, pp. 376–379.

[11] F. F. Kharbat, T. Elamsy, A. Mahmoud, and R. Abdullah, "Image feature detectors for deepfake video detection," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–4.

[12] J. Parmar, S. Chouhan, V. Raychoudhury, and S. Rathore, "Open-world machine learning: Applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 10, pp. 1–37, Oct. 2023.

[13] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2016, pp. 5–10.

[14] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, Jan. 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of artificial neural networks and deep learning," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham, Switzerland: Springer, 2022, pp. 379–425.

[17] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[18] C. Clarke, J. Xu, Y. Zhu, K. Dharamshi, H. McGill, S. Black, and C. Lutteroth, "FakeForward: Using deepfake technology for feedforward learning," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2023, pp. 1–17.

[19] S. Solaiyappan and Y. Wen, "Machine learning based medical image deepfake detection: A comparative study," *Mach. Learn. Appl.*, vol. 8, Jun. 2022, Art. no. 100298.

[20] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics*, vol. 12, no. 8, p. 1789, Apr. 2023.

[21] P. Theerthagiri and G. B. Nagaladinne, "Deepfake face detection using deep InceptionNet learning algorithm," in *Proc. IEEE Int. Students' Conf. Electr., Electron. Comput. Sci. (SCEECS)*, Feb. 2023, pp. 1–6.

[22] R. Chauhan, "Deep learning-based methods for detecting generated fake faces," *Authorea Preprints*, 2023.

[23] D. AbdElminaam, N. Sherif, Z. Ayman, M. Mohamed, and M. Hazem, "DeepFakeDG: A deep learning approach for deep fake detection and generation," *J. Comput. Commun.*, vol. 2, no. 2, pp. 31–37, Jul. 2023.

[24] M. A. Arshed, S. Mumtaz, M. Ibrahim, C. Dewi, M. Tanveer, and S. Ahmed, "Multiclass AI-generated deepfake face detection using patch-wise deep learning model," *Computers*, vol. 13, no. 1, p. 31, Jan. 2024.

[25] F. M. Salman and S. S. Abu-Naser, "Classification of real and fake human faces using deep learning," Tech. Rep., 2022.

[26] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," 2019, *arXiv:1911.05351*.

[27] Z. Zhang, Z. Lei, M. Omura, H. Hasegawa, and S. Gao, "Dendritic learning-incorporated vision transformer for image recognition," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 2, pp. 539–541, Feb. 2024.

[28] H. M. T. Khushi, T. Masood, A. Jaffar, S. Akram, and S. M. Bhatti, "Performance analysis of state-of-the-art CNN architectures for brain tumour detection," *Int. J. Imag. Syst. Technol.*, vol. 34, no. 1, Jan. 2024, Art. no. e22949.

[29] E. Hassan, M. S. Hossain, A. Saber, S. Elmougy, A. Ghoneim, and G. Muhammad, "A quantum convolutional network and ResNet (50)-based classification architecture for the MNIST medical dataset," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105560.

[30] H. Wang and L. Ma, "Image generation and recognition technology based on attention residual GAN," *IEEE Access*, vol. 11, pp. 61855–61865, 2023.

[31] S. Duan, W. Pan, Y. Leng, and X. Zhang, "Two ResNet mini architectures for aircraft wake vortex identification," *IEEE Access*, vol. 11, pp. 20515–20523, 2023.

[32] C. Chen and B. Li, "An interpretable channelwise attention mechanism based on asymmetric and skewed Gaussian distribution," *Pattern Recognit.*, vol. 139, Jul. 2023, Art. no. 109467.