IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2024, accepted 18 April 2024, date of publication 25 April 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393842

RESEARCH ARTICLE

MMGANGuard: A Robust Approach for Detecting Fake Images Generated by GANs Using Multi-Model Techniques

SYED ALI RAZA[1], USMAN HABIB[1], (Senior Member, IEEE), MUHAMMAD USMAN[2], ADEEL ASHRAF CHEEMA[2], AND MUHAMMAD SAJID KHAN[3]

[1]FAST School of Computing, National University of Computer and Emerging Sciences, Islamabad 44000, Pakistan
[2]Department of Computer Science, National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus, Chiniot, Islamabad 35400, Pakistan
[3]Wales Institute of Digital Information, University of South Wales, CF37 1DL Pontypridd, U.K.

Corresponding author: Muhammad Sajid Khan (muhammad.khan@southwales.ac.uk)

ABSTRACT Recent advances in Generative Adversarial Networks (GANs) have produced synthetic images with high visual fidelity, making them nearly indistinguishable from human-created images. These synthetic images referred to as deepfakes, have become a major source of misinformation due to social media. Technology is advancing rapidly, so reliable methods for distinguishing real from fake images are needed. The current detection mechanisms require image forensics tools such as error level analysis (ELA), and clone detection to detect manipulated images. These approaches are limited because they require forensics expertise to use, are manual in application nature, and are unscalable, creating a need for a framework for a scalable tool that experts and non-experts can use to combat the spread of manipulated images and preserve digital visual information authenticity. We approach this problem with a multi-model ensemble framework using the transfer learning method to effectively detect fake images. The proposed approach named Multi-Model GAN Guard (MMGANGuard)integrates four models into an ensemble framework to identify GAN-generated image characteristics to improve deepfake detection. The Gram-Net architecture, ResNet50V2, and DenseNet201 models are used with co-occurrence matrices using transfer learning for MMGANGuard. Through comprehensive experiments, the proposed model demonstrates promising results in detecting the deepfake with high accuracy on the StyleGAN dataset. For automated detection of deepfake-generated images, the proposed model exceeded 97% accuracy, 98.5% TPR, 98.4% TPR, and 95.6% TPR in these evaluations, eliminating the need for manual assessment which is promising for future research in this domain.

INDEX TERMS
Deep fake, data analytics, deep learning, GANs, StyleGAN, detection, multi-model.

I. INTRODUCTION

The rise in the usage of smartphones and the widespread availability of low-cost digital devices like mobile phones, laptops, and tablets has abruptly increased multimedia consumption and the availability of digital images [1]. Advancement in artificial intelligence (AI) [2] has provided immense benefits to humankind but it also has its drawbacks. AI models like (GANs) [3] have now been intelligent enough to create deepfake [4], depending on the analysis of large amounts of data to learn how to generate new examples that are excruciatingly accurate in comparison with the original thing as shown in Figure 1. The field of machine learning has witnessed notable advancements, particularly in the development of sophisticated algorithms capable of efficiently manipulating multimedia content to disseminate disinformation on social networking platforms. The term deepfake is coined from the technological framework known as deep learning, which encompasses a form of artificial intelligence [5]. The utilization of deep fakes

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

FIGURE 1. The image on the left is real while the image on the right is fake.

in contemporary society has given rise to their application for unethical and illicit intentions, such as the fabrication of counterfeit profiles, the production of falsified images, and the dissemination of unfounded rumors through social media platforms. The prevalence of targeted campaigns has significantly increased in contemporary times, thereby raising concerns regarding potential adverse consequences. Misinformation [6] can be categorized as information that is factually incorrect or misleading, disseminated without the intention of deceiving. On the other hand, disinformation [6] refers to a deliberate tactic employed to manipulate individuals by fabricating and spreading false information to achieve predetermined political or financial objectives [7]. As per expert analysis, there is a projection that the deliberate dissemination of distorted news through various channels will increasingly serve as the primary means to influence public opinion or conceal information. At present, the dissemination of false information is facilitated by the extensive utilization of social media platforms.

Deepfake has a long history of being used to make famous people controversial among their supporters. Deepfake images can be utilized to harm people's reputations, such as character assassination of well-known personalities to defame them, trying to blackmail individuals for monetary gain, or causing religious or political unrest by targeting famous personalities with fake images [8]. Deepfake technology/applications, such as FakeApp [9], FaceSwap [10], DALL-E [11], and Midjourney [12], are now widely available, and anyone with no prior knowledge of computer science may generate a fake video or image in seconds. Furthermore, YouTube provides easy access to open-source projects on GitHub. Many experts believe that as technology advances, deepfakes will become considerably more sophisticated, causing more substantial dangers to the public, such as election interference, political conflict, and increased criminal activities.

In recent times, the emergence of deepfake technology has brought both awe and apprehension. While deepfakes have showcased their potential for creative expression, it is crucial to acknowledge the unsettling misuses associated with these manipulated images. This article aims to shed light on the various dangers posed by deepfake images and the growing concerns surrounding their malicious exploitation. They can be used in a variety of ways for malicious intent like Exploitation and Personal Harassment, Fraudulent Activities and Scams, and Political Manipulation [13], [14].

There have been many attempts to resolve this issue with different approaches [5] including manual and automated solutions. The manual detection of deepfake content requires an extensive amount of knowledge to operate such tools reducing the broader scope to ordinary people. Some automated techniques for fake image detection have also been proposed lately which use several methods such as detecting an image using the pupil of the eye [15], which comes with several steps of masking and filtering the eye area which detects the irregularity in the pupil but this approach is prone to noise and depends on the specific angle of face to detect the eyes. Other techniques like Robust Hashing [16] are also being proposed but the main problem is they lose their performance when tested on multiple datasets of GANs. To resolve this problem effectively in an automated fashion, we propose a multi-model ensemble framework named MMGANGaurd as shown in Figure 2

The Multi-Model GAN Guard aims to develop a framework that integrates advanced techniques such as Global Texture Enhancement [17], [18] Co-occurrence matrices on RGB channels [19], [20], [21], DenseNet201 [22], [23], and ResNet50V2 [24], [25]. The primary objective is to detect StyleGAN-generated [26], [27], images and provide a prediction indicating whether the image is fake or real. The implementation of this framework can bring several benefits, including:

- Debunking fake images shared on social media platforms, helping to reduce the spread of misinformation and misleading content.
- Preventing targeted campaigns aimed at defaming or harming individuals by identifying and exposing manipulated images.
- Detecting and addressing controversial fake images, particularly those that involve the exploitation and potential blackmailing of women, thus promoting safety and protecting individuals from harm.

By leveraging the combined power of these techniques, the approach aims to contribute to a more trustworthy and secure digital environment, fostering transparency and safeguarding individuals from the negative consequences of fake images. The contribution of the proposed model is as follows:

1) We propose a multi-model approach using ensemble modeling to address the crucial issue of deepfake detection which is a novel and unique idea.
2) The proposed model leverages the transfer learning method to boost the performance of the detection. We use four pre-trained models trained on a StyleGAN [26] dataset and fine-tune them to come up with an ensemble model.
3) We provide a complete prototype of the ensemble model with a soft voting approach to boost the detection accuracy using different datasets.

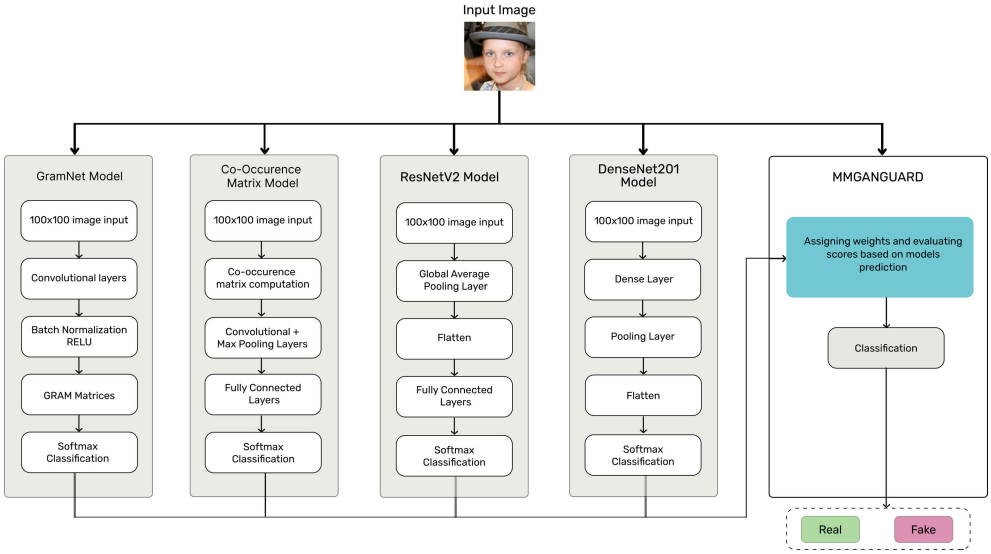


FIGURE 2. Flowchart MMGANGUARD.

4) We demonstrate the model's strength and scalability to deepfake detection using the StyleGAN dataset [26]. The proposed framework represents a robust performance to address the issue with good generalizability and adaptability to the problem.

The rest of the paper organization includes a detailed summary of related work followed by a discussion on the MMGANGuard in the proposed methodology section. This section provides a detailed overview of the architecture and training process of the proposed model. Following this, a comparative evaluation of the model is presented in the implementation and experiments section. Finally, the paper concludes by outlining open gaps for future research.

II. RELATED WORK

Identifying fake faces from real ones has become a challenging problem as more development is happening in the GANs enhancement but there has been a lesser focus on identifying real and fake images generated by GANs. The authors in [17] have discussed state-of-the-art techniques for detecting fake faces generated from multiple GAN architectures like FaceGAN, StarGAN, DRAGAN, and PGGAN, etc. The paper focuses on a texture-based technique for detecting fake and real images by calculating global texture statistics as a robust measure for fake and real faces. Different experiments have been performed on different techniques. A model called GramNet architecture has been introduced which acts as a backbone for CNN to detect the texture differences between a fake and a real face and outperforms all the previous techniques on different GANs generated datasets [17].

Another study proposed in [28] introduced a method called fake images discriminator (FID) for detecting that GAN-generated fake pictures make use of strong spectral correlation, which can be defined as the correlation among the three-color components in finite neighborhood pixels of the images. There are two approaches mentioned for detecting fake images passive forensics and active forensics. The suggested approach comes under passive forensics which involves automated detection of images based on certain features, this begins by converting the color picture into its three component colors of RGB. Then on the RGB components, Discrete wavelet transform (DWT) is applied. The suggested FID approach demonstrates remarkable performance on faces generated by StyleGAN2. Additionally, the FID approach is quite resilient against the four most typical perturbation assaults which involve compression, adding blur, noise, and resizing. This work hasn't performed more experiments with additional datasets which can be termed as one of the drawbacks [28].

Similarly, authors in [19] proposed a framework by using a mix of co-occurrence matrices and deep learning to recognize GAN-produced fake pictures. The paper used a deep convolutional neural network (CNN) architecture to extract co-occurrence matrices on three color channels in the pixel domain and trained a model. The proposed method is promising and achieves more than 99 percent classification accuracy in both datasets, with more than 56,000 pictures based on unpaired image-to-image translations using cycleGAN [29]) and face attributes/expressions using StarGAN [30].

Chih-Chung Hsu et, al. proposed to recognize computer-generated pictures quickly and accurately. Simply learning a binary classifier is difficult due to the challenges in identifying common discriminative characteristics for assessing the fake pictures created by various GANs. To solve this problem, the authors use contrastive loss to find the characteristic properties of synthetic pictures created by various GANs and then combine a classifier to recognize such computer-generated images. The suggested technique

effectively recognized 94.7% of fake pictures created by multiple state-of-the-art GANs [31].

Another study in [16] proposed a novel approach for detecting fake images using robust hashing. While various hashing methods have been developed for image retrieval, the authors specifically chose this approach for its remarkable resilience against image compression and resizing. Furthermore, it exhibits high sensitivity to the manipulations typically employed in generating fake images. Particularly, when an original hash code is employed for compression, the suggested approach outperforms standard methods in detecting fake or tampered photos. However, it is worth noting that the approach has not been tested on multiple datasets generated by GANs, which may potentially impact its overall performance [16]. The authors in [32] an approach for detecting a GAN-generated image through convolutional neural networks. They proposed an architecture for detecting neuron behavior to identify fake faces. Different experiments were performed, and he concluded that by monitoring neuron behavior we can detect a fake face. Mean neuron coverage (MNC) is proposed for capturing the layered neuron activation behavior. This approach has proven good against the four common perturbation attacks such as compression, resizing, light, and noise but one of the drawbacks of this approach is it may not perform well on the random image from the latest GANs like StarGAN-v2 [32].

Guo et al. [15] proposed a method to detect faces generated by GAN models based on irregular pupil shapes. The authors highlight that GAN-generated faces often lack physiological constraints, leading to distinctive features in the pupils. They introduce a technique to automatically extract pupils from both eyes and calculate the boundary intersection-over-union (BIoU) scores. These scores are used to assess and identify if the pupil shapes resemble ellipses, thus revealing GAN-generated faces. The proposed method achieved an impressive AUC of 0.94, indicating the effectiveness of using irregular pupil shapes as anomalies for identifying GAN-generated faces [15], [33].

III. MMGANGUARD

The MMGANGaurd solution is a novel framework that leverages an ensemble approach integrating four different models that each process the image independently. We use the pre-trained models of DenseNet, GramNet, Co-occurrence matrices, and ResNet to come up with an ensemble model that uses soft voting to finalize the binary classification for a particular data observation. We focus on the strengths of each model for its robustness and adaptability to combine and fine-tune it for better detection accuracy. Figure 3 presents an overview of the system architecture in which four different models are designed to detect fake images. The Gram-Net Model utilizes Gram-Net Architecture, which captures the global texture statistics of an image. By analyzing these statistics, the model generates a prediction score. The Co-Occurrence Model, on the other hand, focuses on examining the relationships between pixels within the image to contribute to the overall assessment of its authenticity. The ResNet50 Model is a deep residual network that performs a more in-depth analysis of the image, allowing for a more comprehensive evaluation. Lastly, the DenseNet201 Model, known for its dense connections, adds further depth to the image analysis, enhancing the model's ability to detect fake images. Each of these models brings its approach to the task, contributing to the overall combined model score for predicting real and fake images The weighted average is computed by the following equation:

$$\begin{aligned} MMGANGuard &= ResNet * 2 + DenseNet * 6 \\ &= +(CoOccurrence + GramNet) * 0.1 \end{aligned} \tag{1}$$

As per Figure 3, a set of weights is defined: [2, 6, 0.1]. These weights determine the relative importance of each model's prediction in the outcome. The ResNet model weights 2, the DenseNet model has a weight of 6, and the combined predictions from the Co-Occurrence and GramNet models are given a weight of 0.1. To obtain the MMGAN-Guard prediction, the individual predictions from each model are multiplied by their respective weights. The weighted predictions from the Co-Occurrence and GramNet models are summed together. Then, all the weighted predictions are averaged by dividing them by 4 (the total number of models) to normalize the results. Finally, the result is generated to determine the class with the highest probability from the predictions. The model also calculates the confidence level as a percentage by dividing the value of the predicted class by the sum of all classes' values and rounding it to two decimal places.

Comparing MMGANGuard to existing deepfake detection methods in terms of computational efficiency and scalability is essential for assessing its practical applicability. MMGAN-Guard's efficiency can be evaluated based on factors such as inference speed, memory usage, and model size. Compared to traditional deepfake detection methods that may rely on complex handcrafted features or computationally intensive algorithms, MMGANGuard, with its fusion of deep learning architectures and transfer learning, offers advantages in terms of computational efficiency. By leveraging pre-trained models like ResNet50V2 and DenseNet201, MMGANGuard can achieve high detection accuracy with reduced computational resources and inference time. Additionally, the use of transfer learning allows MMGANGuard to adapt and generalize well to new datasets or scenarios, enhancing its scalability. Furthermore, MMGANGuard's modular design facilitates easy integration with existing deepfake detection pipelines, making it accessible and adaptable for deployment in real-world applications.

MMGANGuard, the combination of Gram-Net, ResNet50V2, DenseNet201, and co-occurrence matrices, represents a comprehensive approach to deepfake detection within the StyleGAN dataset. Gram-Net's specialized architecture excels in capturing style and texture variations,

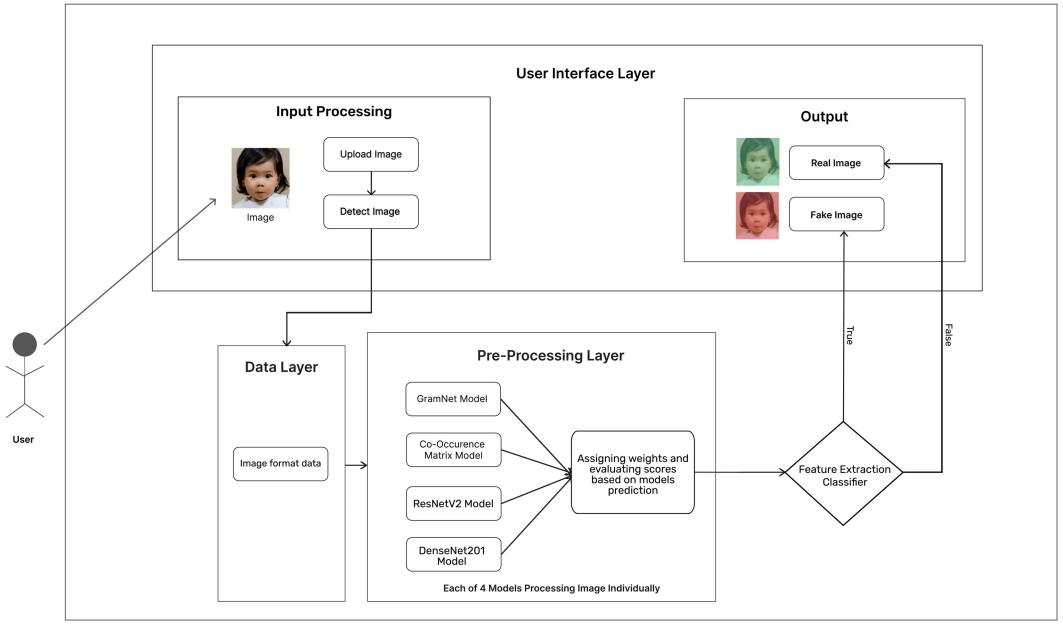


FIGURE 3. System architecture of MMGANGuard.

complemented by ResNet50V2's depth and DenseNet201's dense connectivity, which together enable the extraction of intricate hierarchical features essential for discerning manipulated content. Leveraging transfer learning, these models benefit from pre-trained weights on diverse datasets, enhancing their generalization capabilities. Furthermore, the integration of co-occurrence matrices provides additional insights into spatial relationships within images, enriching the model's understanding of textural patterns and anomalies. Through this synergistic fusion of diverse architectures and feature representations, MMGANGuard achieves robust and reliable detection of deepfake images, bolstering the integrity and authenticity of digital media content.

Qualitative examples of correctly identified deepfakes, false positives, and false negatives provide valuable insights into the performance of deepfake detection models. Correctly identified deepfakes showcase the model's ability to discern subtle anomalies and manipulation artifacts, such as unnatural facial expressions and misaligned features. Conversely, false positives highlight instances where authentic images are incorrectly flagged as deepfakes, often due to high visual fidelity or resemblance to synthetic content. False negatives, on the other hand, represent deepfakes that evade detection, often employing advanced manipulation techniques or subtle alterations that elude the model's detection capabilities. By analyzing these examples, researchers can identify patterns, challenges, and potential areas for improvement in deepfake detection algorithms, ultimately enhancing the accuracy and reliability of these systems in real-world scenarios.

MMGANGuard is flexible in terms of adapting to new patterns from GANs and we can set the weights of different models to increase performance on certain GAN types and it is scalable. Currently, we have to define the weights manually for each model based on how well it performs on the data set, and moving forward in the future we will work on assigning weights based on the AI model so that it will automatically decide the weights. Another limitation is that GANs are rapidly evolving and every GAN architecture has a separate set of features so it won't perform well on unknown GAN data.

A. ARCHITECTURE OVERVIEW

Figure 3 describes the system architecture of MMGANGuard designed to detect deepfake images. The process allows users to upload images for analysis, processes these images using a combination of four different models, and provides an evaluation of whether the image is fake or real. The User Interface layer is the front end of the system where users can interact with the system. It consists of an input box for image upload and a 'Detect' button to initiate the deepfake detection process. Once the user uploads an image and clicks the 'Detect' button, the image is sent to the Data Layer for further processing. Once the classification is complete, the result is sent back to the user interface layer for display to the user. This deepfake detection framework provides a user-friendly interface for users to upload and analyze images using advanced deep-learning models. By utilizing the strengths of the Gram-Net Model, Co-Occurrence Model, ResNet50 Model, and DenseNet201 Model, it delivers a comprehensive analysis for deepfake detection.

B. MODELS EXPLANATION

MMGANGuard is developed with the assistance of the following four models: Co-occurrence Matrix, Global

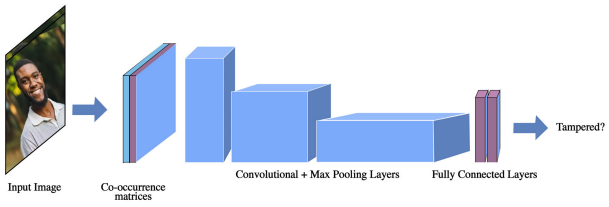


FIGURE 4. Co-occurrences Matrices.

Texture Statistics: Gram-Net Model, ResNet50V2 Model, and DenseNet201 Model.

MMGANGuard's decision-making process relies on identifying specific characteristics indicative of GAN-generated images, which contribute to its interpretability. For instance, the model may focus on subtle inconsistencies in facial features, such as unrealistic proportions or blending artifacts, which are common in deepfake images but less prevalent in authentic ones. Additionally, MMGANGuard may analyze texture patterns and style variations that deviate from natural image distributions, leveraging insights from Gram-Net and co-occurrence matrices to identify anomalous regions within the image. By elucidating these underlying characteristics, MMGANGuard provides users with insights into why certain images are flagged as potential deepfakes, enhancing the transparency and interpretability of its results.

1) CO-OCCURRENCE MATRICES

The function takes an array of images, X, and computes co-occurrence matrices for each image. These matrices measure the similarity between pixels in the image. The function creates a 3D array with the same dimensions as X, representing each image with its RGB channels and pixel dimensions. It iterates over each image, channel, and row, creating a 2D histogram of adjacent pixels. This histogram is then normalized and added to the co-occurrence matrix for the current image and channel. The resulting 3D array of co-occurrence matrices is passed to a convolutional neural network (CNN) for further processing. Refer to Figure 4 for a visual representation of the model.

The CNN consists of convolutional, pooling, fully connected, and output layers. It is trained using input data and labels, and evaluated using test data and labels. The model optimization employs binary cross-entropy as the loss function and the Adam optimizer with a learning rate of 0.001. The model achieves an accuracy of 97% on the dataset.

2) GLOBAL TEXTURE STATISTICS: GRAM-NET MODEL

Gram-Net, outlined in Figure 5, introduces Gram Blocks into the ResNet architecture. These blocks are strategically placed before each down-sampling layer and at the input image. They effectively integrate global image texture information across different semantic levels. Each Gram Block encompasses key elements such as a dimension-aligning convolutional layer, a Gram matrix calculation layer for extracting comprehensive texture features, a pair of conv-bn-relu layers for refining representations, and a global pooling layer for optimal alignment with the ResNet backbone. The Gram matrix calculation is as follows:

$$G_{ij}^{\ l} = \sum_k F_{ik}^{\ l} F_{jk}^{\ l} \tag{2}$$

In Equation 2, F^l denotes the l-th feature map that has been transformed into a vectorized spatial dimension. Meanwhile, $F_{ik}^{\ l}$ signifies the kth element in the ith feature map of layer l. The subsequent demonstration illustrates the effectiveness of the Gram matrix as a robust descriptor for capturing global or extensive-range texture characteristics. The Gram-Net model is a powerful deep-learning architecture designed for image classification tasks. It is composed of various layers, including convolutional layers, batch normalization layers, pooling layers, and dense layers. As shown in Figure 6:

The model takes 100×100 images with 3 color channels as input. It begins with a convolutional layer, Conv $7 \times 7 \times 1$, followed by batch normalization and ReLU activation. A max pooling layer reduces spatial dimensions. Residual blocks with two convolutional layers capture hierarchical features. Gram matrices are used to capture style information, obtained through convolutional layers and processed further. Global average pooling aggregates spatial information. Additional convolutional layers process the gram matrices and style information. Dense layers handle classification by concatenating features and producing the final output.

When tested on the 140k Real and Fake Face dataset [34], the model achieves an accuracy of 93.95%. With approximately 13 million trainable parameters, Gram-Net is a powerful model for image classification. Its architecture allows it to capture both spatial and style information, facilitating the learning of rich representations from images [17].

3) RESNET50V2 MODEL

The ResNet50V2 model is used for feature extraction. It is loaded without the top layer to allow for the addition of custom classification layers suitable for our binary classification task. The output of the ResNet50V2 model is then passed through a Global Average Pooling layer, flattened, and passed through two Dense layers with ReLU activation. The final layer is a Dense layer with two nodes (corresponding to our classes: real and fake), which uses a softmax activation function to yield the classification output this model has 23.5 million parameters. The model is compiled using the Adam optimizer and categorical cross-entropy as the loss function. We monitor several metrics during training, including categorical accuracy, precision, recall, and the area under the Receiver Operating Characteristic and Precision-Recall curves.

The provided model explained in Figure 6 is a deep learning architecture designed for binary image classification. It consists of an input layer, a ResNet50V2 layer, a global average pooling layer, a flattened layer, dense layers, and a classification layer. The first layer, ResNet50V2, takes features from the images that are fed to it. The features are

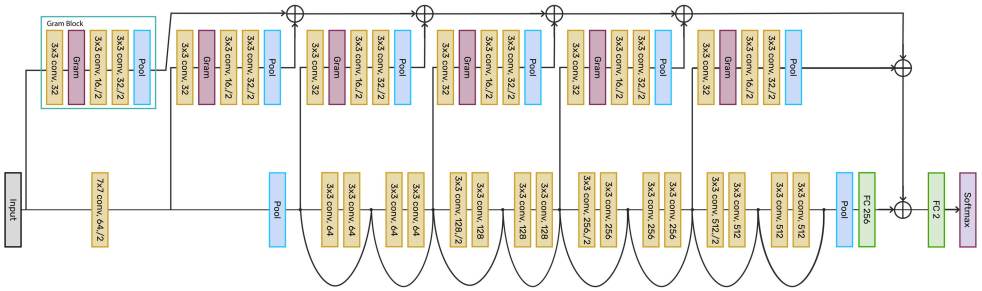


FIGURE 5. Gram-net Architecture.

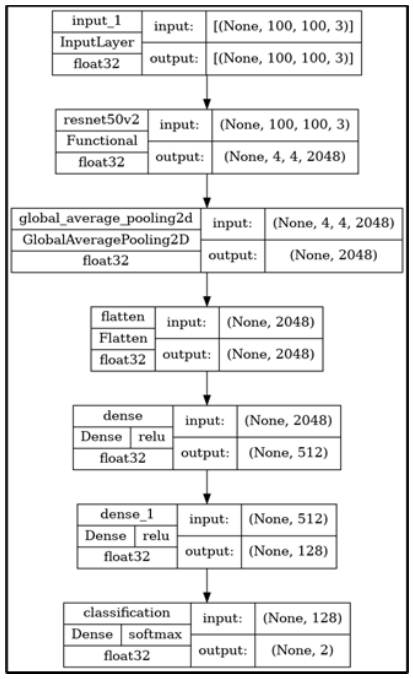


FIGURE 6. ResNet50V2 model with deepfake classification.

then put through pooling and dense layers to be classified. The model has a total of 24,679,810 parameters, with 24,634,370 being trainable. Training on the 140k Real and Fake Faces [34] dataset achieves an accuracy of 98.26%. In summary, this model employs ResNet50V2 for feature extraction and dense layers for classification, and it achieves high accuracy in differentiating between real and fake images.

4) DENSENET201 MODEL

The DenseNet201 model is based on the original DenseNet architecture proposed by Huang et al. [22] DenseNet introduces the concept of dense connections, where each layer is connected to every other layer in a feed-forward fashion. This connectivity pattern improves gradient flow, encourages feature reuse, and reduces the number of parameters. DenseNet201 extends this idea by introducing a deeper network with 201 layers, including dense blocks and transition layers. Overview of the model is shown in Figure 8 below.

The architecture shown in Figure 7 comprises three dense blocks. Between each pair of adjacent blocks, there exist transition layers. These transition layers play a crucial role in altering the feature-map sizes using convolution and pooling operations. By incorporating these transition layers, the model can effectively manage the flow of information and adapt the feature-map sizes to facilitate efficient learning and information propagation throughout the network. This approach allows DenseNet to leverage the benefits of dense connections and adaptively adjust the feature dimensions, leading to improved performance in various deep-learning tasks.

The provided model architecture is designed for real/fake image classification, specifically for detecting deepfake images. It includes an input layer, DenseNet201 layer, a global average pooling layer, flatten layer, dense layers, and a classification layer. The model has a total of 19,371,458 parameters, with 19,142,402 being trainable. It is compiled with the Adam optimizer using a learning rate of 0.001 and employs categorical cross-entropy as the loss function. The model achieves an accuracy of 95.4% on the 140k Real and Fake Faces dataset. In summary, this architecture combines DenseNet201 for feature extraction and dense layers for classification, resulting in a model that detects deepfake images with high accuracy.

IV. IMPLEMENTATION AND EXPERIMENTS

This section includes the evaluations, results, and the dataset used for training and testing. It presents a comprehensive analysis of the obtained results, providing insights into the performance and effectiveness of the implemented system. The experiments were performed on Google Colab T4 with 15GB GPU RAM and 12.7GB System RAM.

A. DATASET

140k Real and Fake Faces dataset [34] is used for training and testing, the dataset comprises a collection of 70k real images sourced from the Nvidia Flickr dataset and an additional 70k fake images randomly sampled from the 1 Million FAKE images dataset generated using StyleGAN [26]. To create this dataset, both the real and fake image datasets were merged, as shown in Figure 10. All images were resized

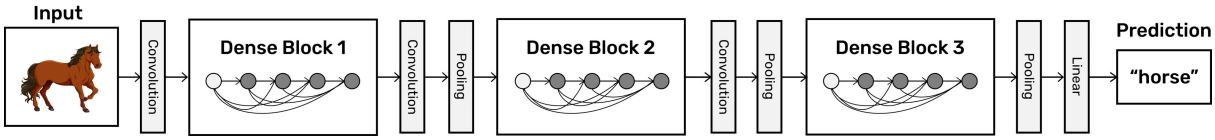


FIGURE 7. DenseNet model with deep fake classification.

TABLE 1. Overview of the performance on the '140k real and fake faces' dataset.

Model	Precision (Real)	Recall (Real)	F1-Score (Real)	Precision (Fake)	Recall (Fake)	F1-Score (Fake)	Accuracy
GRAM Net	0.93	0.97	0.95	0.97	0.92	0.95	0.946
Co-Occurrence	0.91	0.94	0.93	0.94	0.91	0.93	0.926
ResNet50V2	0.98	0.95	0.96	0.95	0.98	0.96	0.96
DenseNet201	0.95	0.98	0.96	0.98	0.95	0.96	0.96
MMGANGuard	**0.96**	**0.98**	**0.97**	**0.98**	**0.95**	**0.97**	**0.97**

FIGURE 8. 140K real and fake images dataset.

to a resolution of 256 pixels for consistency. The data was then divided into train, validation, and test sets to facilitate model training and evaluation. The training and validation datasets used in the experiments consist of 100,000 and 20,000 samples, respectively. Additionally, there are 20,000 samples reserved for testing. To enhance the model's generalization and mitigate overfitting, augmentation techniques are applied during training. Three augmentation operations are employed: random translation, random zoom, and random rotation. These operations introduce variations to the images, helping the model learn from diverse perspectives and reducing the risk of memorizing specific features of the training set. Additionally, the dataset includes several CSV files that provide convenient auxiliary information. By combining these two datasets and preprocessing the images, this dataset offers a diverse range of real and fake faces for various applications in computer vision and machine learning research.

B. EVALUATION METRICS

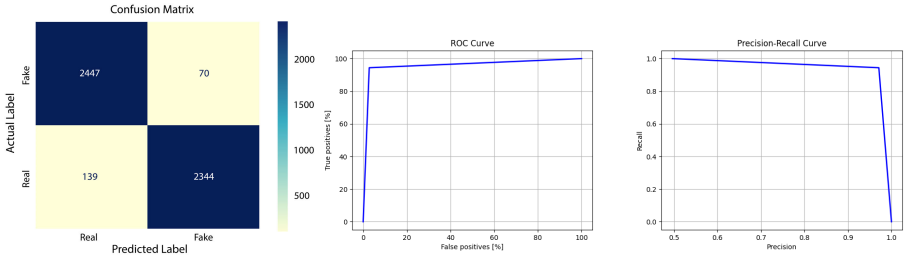
To evaluate the performance of the deepfake image detection application, various metrics are employed, including accuracy, precision, recall, F1-score, and the confusion matrix. By running the models on the 140k Real and Fake Faces dataset, the following results were obtained for each model.

Table 1 presents the performance metrics of four different models (GRAM Net, Co-Occurrence, ResNet50V2, DenseNet201) along with a MMGANGuard on the 140k real and fake faces dataset. The dataset is used for classifying images as either real or fake faces. The models were evaluated based on several performance measures, including precision, recall, F1-score, and accuracy. Precision measures the ability of a model to correctly identify true positives (real or fake faces) out of all the samples it classified as positive. Recall measures the ability of a model to identify all true positives, out of all the actual positive samples in the dataset. F1-score is the harmonic mean of precision and recall, providing an overall measure of a model's performance. Figure 9 represents the validations curves of the model including the confusion matrix, ROC and precision-recall curve. Similarly, Figure 10, and 11 represents the curves for ResNet and GramNet respectively.

In Table 1, the results for each model are presented separately for real and fake faces, along with an overall accuracy score. The values in the Precision (Real), Recall (Real), and F1-Score (Real) columns represent the model's performance in correctly identifying real faces. Similarly, the values in the Precision (Fake), Recall (Fake), and F1-Score (Fake) columns represent the model's performance in correctly identifying fake faces. The Accuracy column provides the overall accuracy of each model in classifying both real and fake faces.

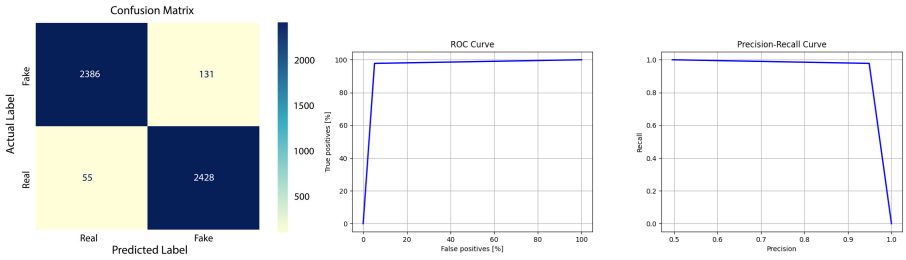
In Table 3 the computation time of the existing four models is compared on 10 epochs each and shows that these models comparatively trains better than the existing state of the art models.

Based on Table 1, the MMGANGuard outperforms individual models in most of the metrics. It achieves high precision, recall, and F1-score for both real and fake faces, indicating its effectiveness in correctly classifying both categories. Moreover, the overall accuracy of the MMGANGuard is also higher compared to the individual models. Therefore, the table suggests that the MMGANGuard performs better



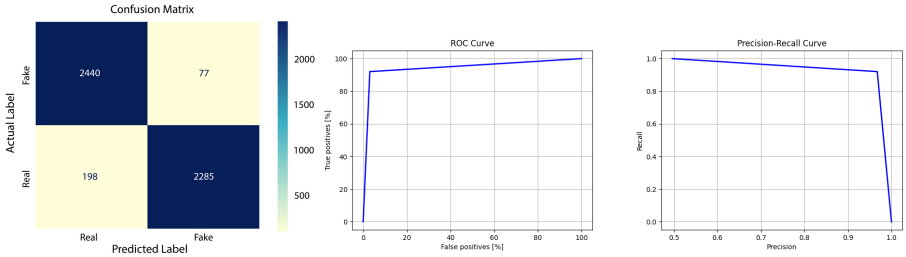
(a) Confusion Matrix (b) ROC Curve (c) Precion Recall

FIGURE 9. **Dense Net Curves.**



(a) Confusion Matrix (b) ROC Curve (c) Precion Recall

FIGURE 10. **ResNet Curves.**



(a) Confusion Matrix (b) ROC Curve (c) Precion Recall

FIGURE 11. **Gram Net Curves.**

overall in classifying real and fake faces compared to the other models evaluated in this dataset.

C. ACCURACY CURVE

Accuracy is a common evaluation metric used to measure the performance of a classification model. It represents the proportion of correct predictions made by the model out of the total number of predictions. On the other hand, validation accuracy refers to the accuracy of a model's predictions on a validation dataset, which is a separate dataset used to assess the performance of the model during training. It serves as an estimate of how well the model generalizes to the unseen. The accuracy curve for the Gram-Net Figures 12a, ResNet50V2 Figures 12b, DenseNet201 Figures 12c, Co-Occurrence Figures 12d trained models are represented.

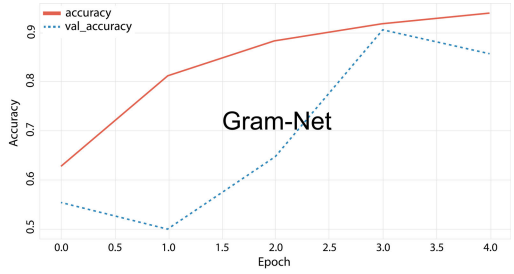
Overall, the DenseNet201 model demonstrates strong performance with high precision, recall, and F1-score values for both the real and fake classes. The model achieves an accuracy of 96%, indicating its effectiveness in correctly classifying instances from the dataset.

TABLE 2. **Detailed performance metrics.**

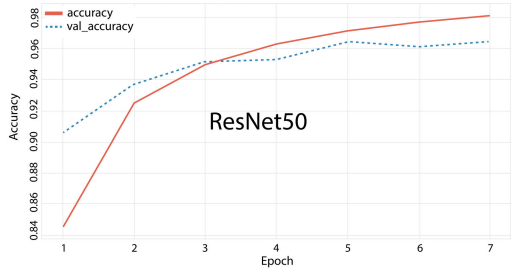
Model	TPR	FPR	TNR	FNR
ResNet50V2	97.85%	5.02%	94.98%	2.15%
DenseNet201	94.96%	2.38%	97.62%	5.04%
Gram-Net	92.74%	3.16%	96.84%	7.26%
Co-Occurrence	91.09%	5.72%	94.28%	8.91%
MMGANGuard	**98.4%**	**4.40%**	**95.6%**	**1.50%**

D. CONFUSION MATRIX

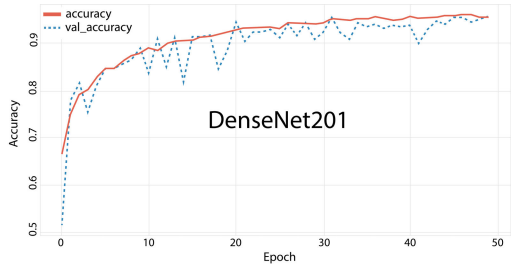
A confusion matrix in Table 2 helps evaluate classification model performance. Table 2 compares the True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR).



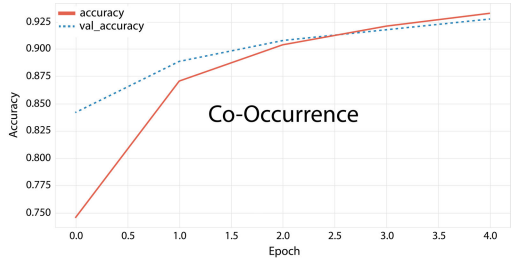
(a) Gram-net



(b) ResNet50V2



(c) DenseNet201



(d) Co-occurrence

FIGURE 12. Accuracy curves of models.

TABLE 3. Comparison of computation time of models.

Model	Epochs	Duration
Co-Occurrence	10	20min 9sec
DenseNET201	10	10min 3sec
RESNET50V2	10	1hr 9min
Gram-Net	10	8.5hrs

Across all the models, there is a consistent pattern of higher accuracy in classifying fake images (higher TN rates) compared to real images. The models generally exhibit a good ability to detect fake images with high True Negative Rates (TNR) ranging from approximately 94% to 97%. However, the performance in identifying real images (True Positive Rates or TPR) varies, with values ranging from approximately 91% to 98%. The False Positive Rates (FPR) are relatively low, indicating a low proportion of fake images being incorrectly classified as real. The False negative Rates (FNR) are also generally low, suggesting a reasonable ability to correctly classify real images as real.

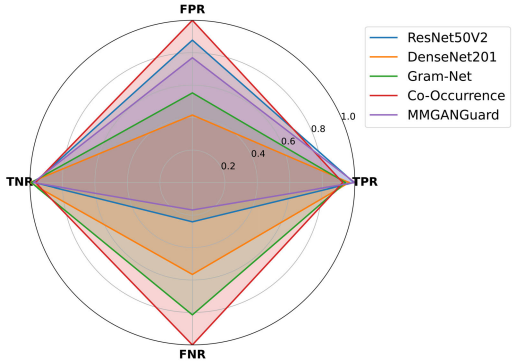


FIGURE 13. Radar chart of the results.

The MMGANGuard achieves a higher TPR of 98.4% compared to all the individual models. This indicates a superior ability to correctly identify true positives. However, the FPR of 4.40% suggests a slightly higher rate of false positives compared to DenseNet201 and Gram-Net. Nevertheless, the MMGANGuard model demonstrates a relatively better balance between true positive identification and avoiding false positives. The following Figure 13 also shows the comparison of the models. In summary, the MMGANGuard outperforms the individual models in terms of true positive identification (TPR). However, they exhibit a slightly higher false positive rate (FPR) compared to some of the individual models. Overall, the MMGANGuard model shows promising performance in detecting fake images.

V. CONCLUSION AND FUTURE WORK

In conclusion, the deepfake image detection MMGANGuard utilizing a combination of four different models, namely GRAM Net [17], Co-Occurrence [19], ResNet50V2 [24], and DenseNet201 [22], has demonstrated promising results in combating the proliferation of manipulated media on the internet. The high accuracy achieved by these models highlights their effectiveness in detecting deepfake images generated by StyleGANs.

To further enhance the deepfake image detection application, several avenues for future work can be explored. Firstly, expanding the dataset to include a wider range of deepfake types, such as videos, audio recordings, and images, would enable the models to learn and detect a broader spectrum of manipulated media, currently it only works for StyleGAN generated dataset. This would contribute to improving the application's performance in detecting deepfakes across various modalities. Additionally, considering the evolving

landscape of deepfake generation, exploring different types of GANs that produce deepfakes with distinct sets of features and characteristics could help ensure the application's effectiveness against emerging manipulation techniques. Adapting the models to handle GANs with different architectural variations and training methodologies would strengthen the application's ability to detect increasingly sophisticated deepfake content. Moreover, integrating the deepfake image detection application with popular social media platforms would provide real-time deepfake detection and flagging capabilities. This proactive approach could help mitigate the rapid spread of deepfake content by promptly identifying and alerting users to the presence of manipulated media. By partnering with social media platforms, the application can contribute to a safer online environment and empower users to make informed decisions about the authenticity of shared content.

Lastly, the deepfake image detection application holds potential beyond its current scope. It could be adapted and extended for use in other domains where image manipulation is prevalent, such as:

- Journalism and forensics. The app could help journalists check the authenticity of visual content or help forensic investigators look into and find cases of image tampering by including features and requirements that are specific to those fields.
- This can be embedded into our social media platforms and can have a flag if the uploaded image is fake or real, this would hugely impact the community as it will help in controling deepfakes.
- Corporate Security: Companies can implement the tool to safeguard against malicious actors attempting to spread fabricated videos to damage reputations or manipulate stock prices.

Overall, the deepfake image detection application's current success and potential for future advancements make it a valuable tool in the ongoing fight against the misuse and spread of manipulated media.

REFERENCES

[1] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023.

[2] J.-P. Haton, "A brief introduction to artificial intelligence," *IFAC Proc. Volumes*, vol. 39, no. 4, pp. 8–16, 2006.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances Neural Information Processing Systems*, 2014, pp. 1–9.

[4] D. Johnson, "What are deepfakes? How fake AI-powered audio and video warps our perception of reality," Jun. 2023. [Online]. Available: https://www.businessinsider.com/guides/tech/what-is-deepfake

[5] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," *Proc. Comput. Sci.*, vol. 219, pp. 211–219, Sep. 2023.

[6] A. M. Guess and B. A. Lyons, "Misinformation, disinformation, and online propaganda," in *Social Media Democracy: The State Field, Prospects for Reform*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[7] D. Freelon and C. Wells, *Disinformation As Political Communication*. Milton Park, U.K.: Routledge, 2020.

[8] A. de Ruiter, "The distinct wrong of deepfakes," *Philosophy Technol.*, vol. 34, no. 4, pp. 1311–1332, Dec. 2021.

[9] J. Pu, N. Mangaokar, L. Kelly, P. Bhattacharya, K. Sundaram, M. Javed, B. Wang, and B. Viswanath, "Deepfake videos in the wild: Analysis and detection," in *Proc. Web Conf.*, Apr. 2021, pp. 981–992.

[10] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4832–4842.

[11] G. Daras and A. G. Dimakis, "Discovering the hidden vocabulary of DALLE-2," 2022, *arXiv:2206.00169*.

[12] A. Borji, "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and DALL-E 2," 2022, *arXiv:2210.00586*.

[13] H. Farid, "Creating, using, misusing, and detecting deep fakes," *J. Online Trust Saf.*, vol. 1, no. 4, pp. 1–12, Sep. 2022.

[14] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114155.

[15] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, "Eyes tell all: Irregular pupil shapes reveal GAN-generated faces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2904–2908.

[16] M. Tanaka, S. Shiota, and H. Kiya, "A detection method of operated fake-images using robust hashing," *J. Imag.*, vol. 7, no. 8, p. 134, Aug. 2021.

[17] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8057–8066.

[18] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, "MTD-net: Learning to detect deepfakes images by multi-scale texture difference," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4234–4245, Aug. 2021.

[19] L. Nataraj, T. Manhar Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. S. Manjunath, "Detecting GAN generated fake images using co-occurrence matrices," 2019, *arXiv:1903.06836*.

[20] G. Li, B. Li, S. Tan, and G. Qiu, "Learning deep co-occurrence features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1610–1623, Apr. 2023.

[21] H. Chi and M. Peng, "Toward robust deep learning systems against deepfake for digital forensics," in *Cybersecurity and High-Performance Computing Environments*. Boca Raton, FL, USA: CRC Press, 2022, pp. 309–331.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[23] 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.

[24] 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.

[25] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools Appl.*, vol. 81, no. 5, pp. 6259–6276, Feb. 2022.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[27] A. H. Bermano, R. Gal, Y. Alaluf, R. Mokady, Y. Nitzan, O. Tov, O. Patashnik, and D. Cohen-Or, "State-of-the-art in the architecture, methods and applications of stylegan," in *Computer Graphics Forum*, vol. 41, 2022, pp. 591–611.

[28] G. Tang, L. Sun, X. Mao, S. Guo, H. Zhang, and X. Wang, "Detection of GAN-synthesized image based on discrete wavelet transform," *Secur. Commun. Netw.*, vol. 2021, pp. 1–10, Jun. 2021.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020, *arXiv:1703.10593*.

[30] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[31] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in *Proc. Int. Symp. Comput., Consum. Control (IS3C)*, Dec. 2018, pp. 388–391.

[32] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," 2019, *arXiv:1909.06122*.

[33] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15334–15342.
[34] Xhlulu. (2020). *140k Real and Fake Faces*. Accessed: Mar. 28, 2024. [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

SYED ALI RAZA received the bachelor's degree in software engineering from Bahria University, Islamabad, Pakistan, in 2018, and the master's degree in data science from the Department of Data Science, National University of Computer and Emerging Sciences, Islamabad Campus. Currently, he is the Founder of the Thriving Data Analytics Agency. His focus lies in steering various projects encompassing data analytics and data engineering. His research interests include generative adversarial networks (GANs) and the critical area of identifying counterfeit images produced by such systems. This specialized focus complements his broader interests in data science and artificial intelligence.

USMAN HABIB (Senior Member, IEEE) received the master's degree from the Norwegian University of Science and Technology (NTNU), Norway, and the Ph.D. degree from the ICT Department, Technical University of Vienna, Austria. He is currently an Associate Professor and the Head of the Software Department, FAST School of Computing, National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan. Before joining FAST-NUCES, he was with the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI), Swabi, and COMSATS University Islamabad, Abbottabad Campus. With over 18 years of teaching and research experience, since 2006, he has successfully completed various industrial projects along with serving in academia. He also actively engages in research and has authored numerous conferences and journal publications. His current research interests include machine learning, data analytics, pattern recognition, security, and medical image processing.

MUHAMMAD USMAN is currently pursuing the Ph.D. degree with the FAST National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus, Pakistan. He is a Lecturer with the Department of Computer Science, FAST National University of Computer and Emerging Sciences. His research interests include predictive analytics and machine learning applications for cross-domain applications.

ADEEL ASHRAF CHEEMA is currently pursuing the Ph.D. degree with the FAST National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus. He is an Assistant Professor with the Department of Computer Science, FAST National University of Computer and Emerging Sciences, Pakistan. His research interests include recommender systems and information retrieval techniques.

MUHAMMAD SAJID KHAN is currently a Collaborator with the Wales Institute of Digital Information, U.K., and a Researcher with the University of South Wales, U.K. His research interests include computer vision (3D image processing, biometrics, optics, robotics, and human–computer interaction) to help solve real-world problems. His work's focus is on applications associated with tracking human movement and identifying individuals in a closed environment.