Optimally Configured Generative Adversarial Networks to Distinguish Real and AI- Generated Human Faces

Kalaimani G

kalaimaniiphd@gmail.com

Karpagam College of Engineering

Kavitha G
Muthayammal Engineering College

Selvan Chinnaiyan
REVA University

Srikanth Mylapalli
Koneru Lakshmaiah Education Foundation

Research Article

Additional Declarations: No competing interests reported.

Optimally Configured Generative Adversarial Networks to Distinguish Real and AI-Generated Human Faces

Dr.G.Kalaimani*, Dr.G.Kavitha, Selvan Chinnaiyan , Srikanth Mylapalli

Professor, Computer Science and Engineering, Karpagam College of Engineering.
Professor, Computer Science and Engineering, Muthayammal Engineering College.
Professor, School of Computer Science and Engineering, REVA University, Bangalore, Karnataka, India.
Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation,
Andrapradesh, India.

*Corresponding author E-mail: kalaimaniiphd@gmail.com

Abstract: Artificial Intelligence (AI) has come a long way in the last several years, especially in terms of producing human-like faces with deep-fake technology. However, the challenge lies in accurately distinguishing between real and AI-generated human faces. As the applications of such technology continue to expand, the need for robust classification methods becomes crucial to ensure ethical and responsible use. Existing Generative Adversarial Networks (GANs) produce increasingly realistic synthetic faces, making it difficult for traditional methods to differentiate between real and generated faces. This poses potential risks in various domains, including security, identity verification, and misinformation. The primary objective of this research is to design an optimally configured GAN capable of distinguishing between real and generated faces and to develop a robust classifier that accurately classifies human faces as either real or generative. The results showcase the effectiveness of the optimally configured GAN model in achieving high accuracy, reaching 95%, in distinguishing between real and AI-generated faces across state-of-the-art techniques. The research contributes to the ethical deployment of AI technologies, safeguards security applications, strengthens identity verification systems, combats misinformation, and fosters public trust in the era of advanced AI.

Keywords: Artificial Intelligence, Generative Adversarial Networks, real and AI-generated human face, optimal configuration, Lyrebird Optimization Algorithm (LOA)

1. Introduction

Significant progress has been made in the field of AI, especially with regard to the creation of synthetic content, as demonstrated by GAN Dang et al (2018). This progress has facilitated the creation of increasingly realistic human faces, raising challenges in accurately distinguishing between authentic and AI-generated images Whittaker et al (2020); Monkam and Yan (2023). As these technologies become integral to various domains, the need for robust classification

methods becomes imperative to ensure ethical and responsible deployment Mirsky and Lee (2021). GAN have emerged as a cornerstone in the synthesis of realistic human faces Caramihale et al (2018); Alqahtani et al (2021). However, the growing sophistication of GANs presents a significant challenge in discerning between real and AI-generated faces Moshel et al (2022); Meyer (2022). This challenge is underscored by studies revealing the human perceptual limitations in discriminating Voss et al (2017) between the two, prompting the exploration of advanced methodologies for reliable classification Pataranutaporn et al (2021). The increasing realism of AI-generated Khoo et al (2022) faces poses challenges in reliable classification, particularly in security, identity verification, and misinformation detection Korshunov and Marcel (2018); Zhang and Ghorbani (2020). Traditional methods struggle to distinguish between real and synthetic faces Man and Chahl (2022), necessitating innovative approaches to address the limitations and uncertainties associated with this evolving technology Verdoliva (2020).

Creating and implementing an efficient system for the classification of genuine and AI-generated human faces is the main goal of this research Caporusso et al (2019). The proposed methodology involves a comprehensive preprocessing pipeline consisting of two pivotal stages. Firstly, Image Standardization is employed to ensure uniformity in image characteristics, mitigating biases and disparities. Subsequently, Feature Extraction via ResNet-50 Li and Lima (2021); Quach (2020) is utilized to capture intricate facial features crucial for discrimination. The classification process involves a GAN Shen et al (2021); Wang, et al (2021). with the discriminator playing a key role in distinguishing between real and AI-generated faces Tolosana et al (2020). The goal of the Lyrebird Optimization Algorithm (LOA) is to find the ideal weights in order to maximize the discriminator's performance. The findings of this study have important ramifications for improving our knowledge of the utilization of AI-generated information in society. The outcomes of this work support the ethical and responsible application of AI technology by overcoming the difficulties in identifying between generative and real-world human faces Guarnera et al (2022); Devi et al (2022). Moreover, the integration of the LOA adds a novel dimension, potentially enhancing the performance of discriminative models in the context of GANs. The findings are expected to guide future developments in the field, ensuring the reliability and trustworthiness of AI-based facial recognition systems Partadiredja et al (2020).

2. Literature Review

Yegemberdiyeva et al. Yegemberdiyeva and Amirgaliyev (2021) discuss the 2021 introduction of GAN, which creates artificial human-like faces for video processing, entertainment, and marketing campaigns. The study focuses on efficiency in identifying, differentiating, and memorizing actual and false faces. It covers human decision-making, facial recognition, and factors influencing face memorization.

Baraheem et al. (2023) developed a framework using convolutional neural networks to distinguish AI-generated images from actual ones. They used transfer learning and Class Activation Maps to identify discriminative cues. The method achieved 100% accuracy on a dataset, with EfficientNetB4 being the best detector. Adam was used as an optimizer Baraheem and Nguyen (2023).

The "forensic similarity method" is a 2021 detection approach that compares face and background similarity in video frames. The study uses FaceForensics and Celeb-DF datasets, showing improved generalization ability and 8-12% accuracy increase compared to Exception and Xception Pan et al (2021).

The study proposes a deepfake detection intelligence forensic technique using GANs in speech, audio, and image domains. The technique uses a guided filter to identify texture artefacts and probable forging elements, and the Resnet18 classification network to identify genuine and false images of faces. The method achieves the highest detection accuracy available Yang et al (2021).

Senapati et al. discuss common conditions like diabetes, high cholesterol, heart attacks, and cancer. They use food photos to create a system for recognizing and detecting food allergies. ResNet50, a transfer learning method, achieves 95% accuracy and is trained to deliver nutrients and food types Senapati et al (2023).

Dehghani et al. proposed the LOA, a bio-inspired metaheuristic algorithm that mimics the natural behavior of lyrebirds in the wild. The algorithm consists of exploration and exploitation phases, with the CEC 2017 test suite assessing its effectiveness. The simulations showed LOA's strong exploration, exploitation, and balancing capabilities, outperforming twelve popular metaheuristic algorithms in problem-solving. This innovative approach has the potential to improve optimization in practical situations Dehghani et al (2023).

3. Research Methodology

The increasing realism of GANs in creating synthetic human faces has made it challenging to accurately classify real and AI-generated faces. Studies show humans struggle to distinguish between real and synthetic faces and synthetic faces can be visually indistinguishable and elicit higher trustworthiness. Figure 1 illustrates the overview of the proposed approach in identifying real and AI-generated human faces.
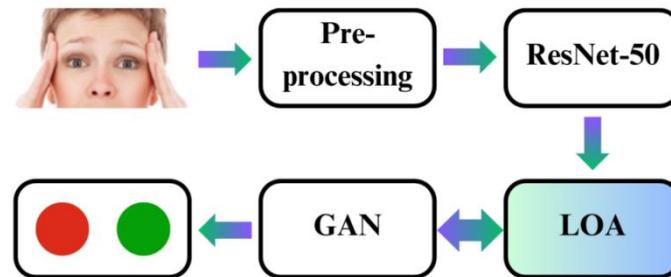
Fig. 1: Overview of the Research Methodology

3.1 Dataset Description

This research study used 500 real images from Kaggle and an additional set of AI-generated images from the Runway software application to create synthetic faces. The images underwent preprocessing to enhance discriminative features and prepare the data for classification. The preprocessing pipeline included Image Standardization and Feature Extraction to enhance the accuracy of the results. Figure 2 shows samples of real and AI-generated images.

**Real Human face Images**

AI-Generated Images

Fig. 2: Sample of real and the AI-Generated images

3.2 Image Standardization

Image standardization is a crucial pre-processing step that ensures consistent characteristics in all images in a dataset, including normalizing pixel values, adjusting image sizes, and applying other normalization techniques, thereby creating a uniform foundation for analysis and classification.

3.3 Feature Extraction using ResNet-50

The research used ResNet-50, a reliable CNN architecture, to extract high-level features from images. This method is ideal for distinguishing between real and AI-generated faces. ResNet-50's robust architecture and autonomous learning of hierarchical features make it superior to other CNN techniques and traditional methods, making it a preferred choice for various applications. Figure 3 provides an overview of the feature extraction process using ResNet-50.
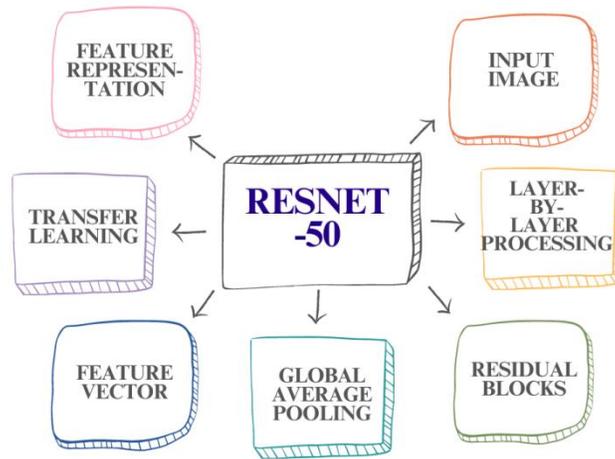
Fig. 3: ResNet-50 Feature Extraction Overview

3.3.1 Input Image

The process begins with the input of an image into the ResNet-50 architecture. Typically, ResNet-50 is pretrained on large image datasets, such as ImageNet, to learn a diverse set of features. This pretrained model can then be used as a feature extractor for various downstream tasks, including classifying real and AI-generated faces.

3.3.2 Layer-by-Layer Processing
The input image undergoes a series of convolutional and pooling layers in a hierarchical fashion. Each layer extracts increasingly abstract and complex features from the image. The depth of ResNet-50 allows it to capture intricate patterns and representations in the data.

3.3.3 Residual Blocks

Multiple convolutional layers are present in each of the residual blocks that make up ResNet-50. These blocks have skip connections that bypass one or more layers, allowing the network to directly learn the residual features. The skip connections mitigate the vanishing gradient problem, facilitating the training of deep networks.

3.3.4 Global Average Pooling

It is common practice for utilizing a global average pooling layer at the end of the network. This layer reduces the spatial dimensions of the features to a single value per feature map. It helps in aggregating the learned features across the entire spatial extent of the input.

3.3.5 Feature Vector

A feature vector representing the high-level features taken out of the input image is the result of the global average pooling layer. This vector encapsulates the most discriminative information learned by ResNet-50 during its hierarchical processing.

3.3.6 Transfer Learning

The pretrained ResNet-50 model is a feature extractor that may be applied to feature extraction for a particular purpose, like identifying between genuine and AI-generated faces. The weights learned during the initial training on a large dataset (like ImageNet) are retained, and only the final fully connected layers are replaced or fine-tuned for the target task.

3.3.7 Feature Representation

The feature vector extracted by ResNet-50 serves as a rich representation of the input image, capturing intricate details, textures, and patterns that are crucial for subsequent classification tasks.

ResNet-50 operates by leveraging its deep architecture with residual connections to hierarchically extract increasingly complex features from input images. The resulting feature vector serves as a high-level representation that can be employed in downstream tasks, such as classifying real and AI-generated faces, through the process of transfer learning. The images underwent Image Standardization and Feature Extraction through ResNet-50, culminating in a dataset enriched with standardized and feature-enhanced images.

3.4 Generative Adversarial Networks (GAN)

GAN play a pivotal role in distinguishing between real and AI-generated human faces, offering a powerful framework for both the creation and identification of synthetic content. Whereas the discriminator seeks to accurately discriminate between actual and created images, the generator works to create artificial faces that are identical to genuine ones. The capacity of the generator to produce realistic faces and the discriminator to differentiate between real and synthetic samples are both improved by this adversarial interaction. The effectiveness of GANs in generating realistic faces presents a challenge in discrimination. As GANs advance, they produce synthetic faces that closely resemble real faces, making it increasingly problematic for traditional methods to differentiate between the two. This challenge underscores the need for innovative approaches to accurately classify faces as either real or AI-generated. Researchers often focus on optimizing GANs specifically for the task of discrimination between real and generated faces. This research integrates the GAN architecture optimally to enhance the discriminator's discriminatory capabilities. The sequential procedure for the GAN in identifying actual and AI-generated human faces is depicted in Figure 4.
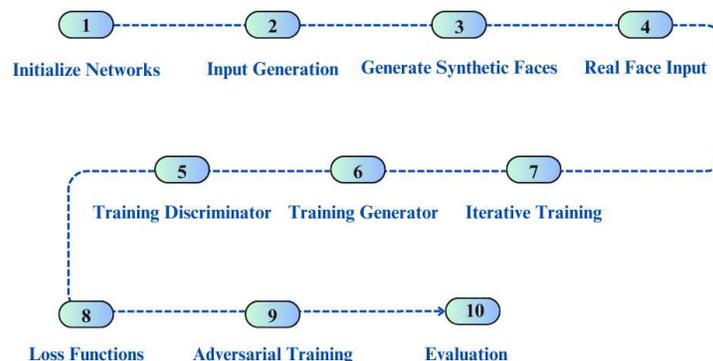
Fig. 4: step-by-step process for the GAN in classifying real and AI-generated human faces

3.4.1 Initializing Networks

In the initialization phase of a GAN, the process involves configuring the parameters and architecture of both the Generator (G) and Discriminator (D). For the Generator, which is tasked with producing synthetic human faces, research considers optimization techniques to identify the optimal number of weights. The choice of activation functions, ReLU for hidden layers and sigmoid for the output layer, is essential, as is the definition of the loss function, often employing Mean Squared Error (MSE). Simultaneously, the Discriminator, designed to discern between real and generated faces, undergoes a similar initialization procedure. Its neural network layers' weights and biases are initialized, activation functions are chosen (typically ReLU for hidden layers and sigmoid for the output layer), and a binary cross-entropy loss function is specified. These careful configurations lay the foundation for stable and effective learning in the subsequent training phases of the GAN. The choice of optimizers and hyperparameters further contributes to the overall success of the network initialization.

3.4.2 Input Generation

Generate random noise z from a probability distribution

3.4.3 Generate Synthetic Faces

Utilize the G to transform the input noise z into synthetic faces $\hat{x} = G(z)$.

3.4.4 Real Face Input

Providing real human face images x as input to both the G and D is a fundamental aspect of GAN training, as it establishes a benchmark for the G to replicate and for the D to differentiate. This dynamic interaction forms the basis for the GAN to learn and generate synthetic faces that closely resemble real-world examples.

3.4.5 Training Discriminator

The Discriminator is essentially a binary classifier with the goal of correctly classifying input images into two categories real or generated human face. It assigns probabilities to input images being real (from the dataset) or generated (produced by the Generator). The loss function used for training the D is commonly the binary cross-entropy loss. This loss quantifies how well the D is able to classify the human face correctly. The formula for binary cross-entropy loss is often expressed as Equation (1):

$$L_D = -[E_{real}[\log D(x)] + E_{generated}[\log(1 - D(G(z)))]]$$

(1)

Where, E_{real} and $E_{generated}$ denote the expectation over real and generated data samples respectively. Finding optimal weights for the discriminator is an essential component of successful GAN training. It helps maintain the generative-adversarial equilibrium, improves the quality of generated data, stabilizes the training process, and enables transferability and generalizability. Additionally, it can provide valuable insights into the underlying data distribution. In the optimization phase of the GAN training process, the research incorporates the LOA to identify optimal weights for the Discriminator (θ_D). The Figure 5 exhibits the overview of GAN for Classifying Real and AI-Generated Images.
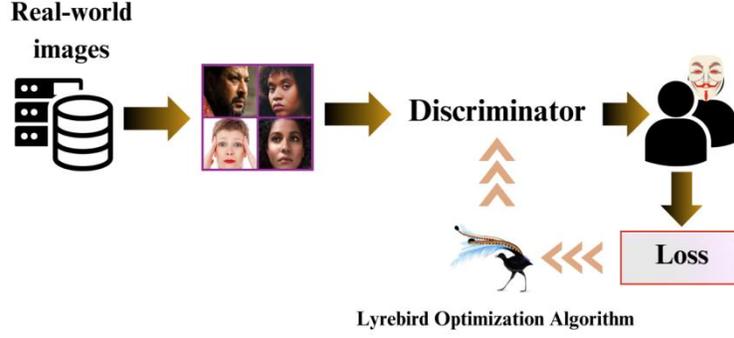
Fig. 5: Overview of GAN for Classifying Real and AI-Generated Images

3.4.6 Training Generator

In the Training Generator phase, the objective is to train the G to minimize the binary cross-entropy loss by maximizing the expected value of the logarithm of the Discriminator's probability of correctly classifying generated samples. Mathematically, this is expressed as Equation (2):

$$Maximize\ E_z\left[\log\left(D(G(z))\right)\right] \tag{2}$$

Here, z represents random noise sampled from a probability distribution. The goal is for the Generator to produce synthetic samples $G(z)$ that are convincing enough to deceive the Discriminator. The Generator's parameters are updated using gradient ascent, wherein the gradients of the expected value with respect to the Generator's parameters are computed and used to adjust the weights. This process encourages the Generator to generate samples that are increasingly difficult for the Discriminator to distinguish from real data. The iterative nature of this training process contributes to the overall improvement in the quality and realism of the synthetic data produced by the Generator.

3.5 Lyrebird Optimization Algorithm (LOA)

The LOA draws inspiration from the natural behavior of lyrebirds, specifically manifesting a key behavioral characteristic observed when these birds sense potential danger [30]. In response to a perceived threat, the lyrebird exhibits a distinctive strategy: it momentarily pauses, conducts a meticulous scan of its surroundings, and subsequently makes a decision to either escape from the immediate environment or seek refuge in a suitable hiding place. This innate survival strategy of lyrebirds during moments of danger serves as a foundation for the mathematical modeling incorporated into the design of the LOA. In the context of the research, this behavioral emulation is leveraged to enhance the optimization process for identifying optimal weights in the Discriminator of a GAN. The LOA's utilization of lyrebird-inspired strategies, such as careful exploration and decisive exploitation, reflects a unique and bio-inspired approach to algorithmic optimization, offering a promising avenue for achieving efficient and effective parameter tuning in complex computational tasks. Figure 6 shows the flow Chart of LOA.

3.5.1 Initial Solution Generation
Generating diverse and effective initial Discriminator weight sets is a crucial step for identifying optimal weights in a GAN. The LOA's initial solution generation is inspired by the behavioral traits of lyrebirds during moments of perceived danger. By incorporating randomness within specified bounds, the algorithm establishes an initial set of solutions that emulate the careful exploration observed in lyrebirds. Let X_i denote a solution vector in the ith iteration, corresponding to a set of weights for the Discriminator θ_D. Each component of X_i is initialized randomly within specified bounds in Equation (3) and (4)

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{N \times m} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots & \cdot^{\cdot^{\cdot}} & \vdots \\ x_{i,1} & \cdots & x_{i,d} & \cdots & x_{i,m} \\ \vdots & \cdot^{\cdot^{\cdot}} & \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,d} & \cdots & x_{N,m} \end{bmatrix}_{N \times m} \tag{3}$$

$$x_{i,d} = lb_d + r \cdot (ub_d - lb_d) \tag{4}$$

Where each X_i indicates the ith member or potential solution inside the population, and X is the matrix of the LOA population. The variable $x_{i,d}$ corresponds to the dth dimension or decision variable of the ith member in the search space. N signifies the total number of lyrebirds in the population, while m denotes the count of decision variables. The variable r takes on a random value within the range [0, 1]. Additionally, lbd and ubd represent the lower and upper bounds, respectively, of the dth decision variable.

Since every member of the LOA represents a potential solution to the issue at hand, it is possible to evaluate the objective function linked to each member of the LOA. Consequently, for every LOA member, an objective function value is obtainable. This collection of evaluated values, equivalent to the population size, can be succinctly represented as a vector in accordance with Equation (5).

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} F(X_1) \\ \vdots \\ F(X_i) \\ \vdots \\ F(X_N) \end{bmatrix}_{N \times 1} \tag{5}$$

Here, the evaluated objective function based on the ith LOA member is denoted by Fi, and the evaluated objective function vector is represented by F. The primary objective is to inject diversity into the initial solutions, encouraging exploration across the parameter space. This diversity sets the foundation for the subsequent optimization process. The initially generated solutions serve as the starting point for the optimization process. The LOA will subsequently iteratively refine and optimize these solutions based on the Discriminator's performance in distinguishing between real and generated samples.

3.5.2 Fitness Computation

In the context of GAN training, the fitness function measures the discriminator's capacity to discern between generated and real data. The fitness metric employed in the Discriminator formula is the binary cross-entropy loss, and the fitness function may be written as follows Equation (6):

$$F_D = -[E_{real}[\log D(x)] + E_{generated}[\log(1 - D(G(z)))]]$$

(6)

Here, the fitness function is the negative of the binary cross-entropy loss, which is computed by evaluating the Discriminator's performance on both real (x) and generated G(z) samples. The expectations E_{real} and $E_{generated}$ represent the averages over real and generated samples, respectively. The terms log D(x) and log (1-D (G (z))) is the log probabilities assigned by the Discriminator to real and generated samples, respectively. The objective of the LOA is to iteratively adjust the weights of the Discriminator (θ_D) based on the computed fitness values, aiming to minimize this binary cross-entropy loss. The optimization process involves searching for weight configurations that enhance the Discriminator's ability to accurately classify real and generated samples, contributing to the overall effectiveness of the GAN in generating realistic synthetic data.

3.5.3 Solution Updating

Utilising the mathematical modelling of the lyrebird strategy when it detects danger, the positions of the population members are updated in each iteration. The two stages of this population update process—running away and hiding—mirror the choices made by lyrebirds in

similar circumstances. In formulating LOA, Equation (7) is employed to simulate the lyrebird's decision-making process when choosing between escape and hiding strategies during times of danger. Consequently, during each iteration, the update of each LOA member's position is exclusively influenced by either the first or second phase, emulating the lyrebird's distinct decision in the given situation.

$$Update\,process\,for\,X_i : \begin{cases} based\,on\,phase\,1, & r_p \leq 0.5 \\ based\,on\,phase\,2, & else \end{cases} \quad (7)$$

In this case, r_p is a random number within the range [0, 1].

3.5.3.1 Escaping Strategy (Exploration Phase)

Utilising a model of the lyrebird's escape from the danger position to the safe zones, the population member's position is updated in the search space throughout this phase of LOA. The lyrebird's capacity to explore new locations in the problem-solving space and make large positional changes after moving to a safe place is indicative of LOA's global search exploration capability. The positions of other population members with higher objective function values are regarded as safe areas for each member in the LOA design. Equation (8) is thus able to be utilised to find the set of safe regions for each LOA member.

$$SA_i = \{X_k, \quad F_k < F_i\,and\,k \in \{1,2,\ldots,N\}\}, \quad where\,i = 1,2,\ldots,N,$$
(8)

In this case, X_k is the kth row of the X matrix, which has a better objective function value (i.e., F_k) than the ith LOA member (i.e., $F_k < F_i$). SA_i is the set of safe areas for the ith lyrebird.

It is anticipated in the LOA design that the lyrebird sometimes makes its way to one of these secure locations. Equation (9), which is based on the lyrebird displacement modelling in this phase, is used to determine a new position for each LOA member. Then, in accordance with Equation (10), this new position takes the place of the relevant member's prior position if the value of the objective function is enhanced.

$$x_{i,j}^{P1} = x_{i,j} + r_{i,j} \cdot \left(SSA_{i,j} - I_{i,j} \cdot x_{i,j}\right), \quad (9)$$

$$X_i = \begin{cases} x_i^{P1}, & F_i^{P1} \leq F_i, \\ X_i, & elsse \end{cases} \quad (10)$$

Here, SSA_i is the selected safe area for ith lyrebird, $SSA_{i,j}$ is its jth dimension, X_i^{P1} is the new position calculated for the ith lyrebird based on escaping strategy of the proposed LOA, $X_{i,j}^{P1}$ is its jth dimension, F_i^{P1} is its objective function value, $r_{i,j}$ are random numbers from the interval [0, 1], and $I_{i,j}$ are numbers that are randomly selected as 1 or 2.

3.5.3.2 Hiding Strategy (Exploitation Phase)

During this stage of LOA, the population member's position is modified in the search space according to the lyrebird's modeled approach of hiding in its immediate safe area. The lyrebird's position changes somewhat when it moves in little steps to find a good hiding place and accurately scans its surroundings, demonstrating the LOA's potential for use in local search.

In LOA design, a new position is determined for each LOA member utilizing Equation (11) based on the modeling of the lyrebird's migration towards the nearby appropriate region for concealment. If, in accordance with Equation (12), this new position enhances the value of the objective function, it replaces the prior position of the associated member.

$$x_{i,j}^{P2} = x_{i,j} + \left(1 - 2r_{i,j} \cdot x_{i,j}\right) \cdot \frac{ub_j - lb_j}{t} \quad (11)$$

$$X_i = \begin{cases} x_i^{P2}, & F_i^{P2} \leq F_i, \\ X_i, & elsse \end{cases} \quad (12)$$

Here, X_i^{P2} is the new position calculated for the ith lyrebird based on the hiding strategy of the proposed LOA, $x_{i,j}^{P2}$ is its jth dimension, F_i^{P2} is its objective function value, $r_{i,j}$ are random numbers from the interval [0, 1], and t is the iteration counter.
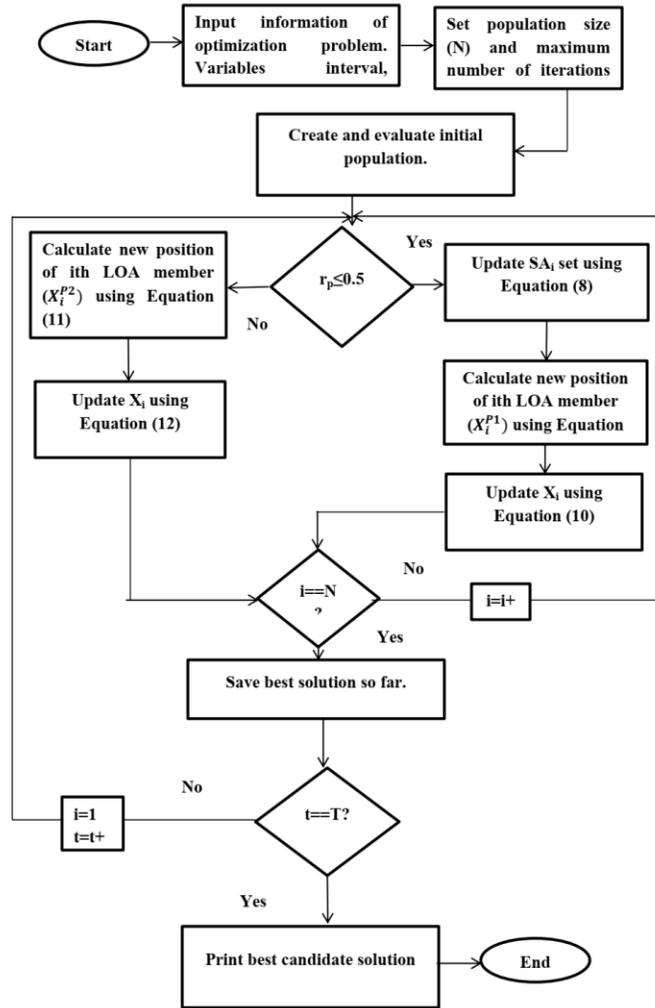
Fig. 6: Flow Chart of LOA

By integrating the LOA, the research aims to enhance the efficiency of weight optimization for the Discriminator, leveraging the algorithm's capabilities in balancing exploration and exploitation during the search process. This innovative approach holds the potential to contribute to improved adversarial learning dynamics within the GAN framework, leading to enhanced discriminative capabilities of the Discriminator and, consequently, improved overall GAN performance.

3.5.3.3 *Iterative Training*

Iterative Training is a pivotal concept in the framework of a GAN, where the process revolves around the continuous refinement of both the G and D. The iterative cycle begins with the training of the Generator to craft synthetic samples closely resembling real data, aiming to minimize the binary cross-entropy loss. Subsequently, the Discriminator undergoes training to adeptly distinguish between real and generated samples, working to minimize the classification error. This process alternates between training the G and D, fostering an adversarial relationship. The iterative dynamics aim to achieve equilibrium, where the Generator produces synthetic samples indistinguishable from real data, and the Discriminator struggles to reliably differentiate between the two. The iterative training process involves a delicate balancing act, ensuring improvements in one component prompt adjustments in the other. This dynamic equilibrium is crucial for the GAN to generate high-quality, realistic synthetic data. Training iterations persist

until a convergence criterion is met, signaling the Generator's proficiency in generating synthetic data and the Discriminator's challenge in discerning between real and generated samples.

3.5.3.4 *Loss Functions*

The Loss Functions are critical components that guide the training process by quantifying the performance of both the D and G. The GAN framework involves two primary loss functions: D Loss and G Loss.

3.5.3.5 Discriminator Loss

The D Loss assesses how well the D is able to distinguish between real and generated samples. It is typically formulated as binary cross-entropy loss. Mathematically, the Discriminator Loss (L_D) is expressed as Equation (13)

$$L_D = -\frac{1}{m}\sum_{i=1}^{m}(\log D(x^{(i)}) + \log(1 + D(G(z^{(i)}))))$$

(13)

Where, m is the batch size, $D(x^{(i)})$ is the probability assigned by the Discriminator to the real sample $x^{(i)}$ and $D(G(z^{(i)}))$ is the probability assigned to the generated sample $G(z^{(i)})$.

3.5.3.6 Generator Loss

The Generator Loss evaluates how well the Generator can deceive the Discriminator by producing synthetic samples that resemble real data. The Generator Loss L_G is formulated as Equation (14):

$$L_G = -\frac{1}{m}\sum_{i=1}^{m}\log(D(G(z^{(i)})))$$
(14)

Here, D (G (z $^{(i)}$)) represents the probability assigned by the Discriminator to the generated sample $G(z^{(i)})$.

These loss functions are integral to the training process in a GAN. During training iterations, the Discriminator and Generator aim to minimize and maximize their respective loss functions, leading to an adversarial dynamic that drives the GAN towards generating increasingly realistic synthetic data. The careful balance between D and G Loss is crucial for the GAN to reach equilibrium, where the Generator produces high-quality synthetic samples, and the Discriminator struggles to differentiate between real and generated data.

3.5.3.7 *Adversarial Training*

Adversarial Training is a cornerstone in GAN, involving an ongoing dynamic interaction between the G and D. The Generator continually refines its ability to create realistic synthetic samples, challenging the Discriminator to distinguish between real and generated data. Simultaneously, the Discriminator evolves to accurately classify samples, prompting an iterative refinement process. This delicate balancing act between the two components continues until a state of equilibrium is reached, where the Generator generates convincing samples, and the Discriminator struggles to discern between real and synthetic data. Adversarial Training is pivotal for achieving the GAN's goal of producing high-quality, realistic synthetic data.

3.5.3.8 Evaluation

The Evaluation phase, in the context of a GAN, is a pivotal step aimed at assessing the GAN's performance on a distinct dataset, different from the one used during training. This process is essential for determining the GAN's ability to effectively classify between real and AI-generated human faces in real-world scenarios. Utilizing various performance metrics such as Sensitivity, Specificity, Accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), and False Discovery Rate (FDR), the GAN undergoes scrutiny to gauge its generalization capability and discriminate between diverse and unseen data. The overall research objective is to validate the GAN's effectiveness, reliability, and ethical considerations for accurate classification in real-world scenarios.

4. Results and Discussion

This preliminary research considers four traditional CNN techniques for identifying real and AI-generated human faces: ResNet-50, VGG-16, GoogleNet, and AlexNet. The reason for considering CNN techniques is that these methods are capable of extracting features from given images and further utilizing them for classification. The feature extraction capabilities embedded within the CNN architecture, combined with the ability to learn discriminative features directly from data, make CNNs a powerful and advantageous choice for classifying real and AI-generated human faces compared to using separate algorithms for feature extraction.

In subsequent research, we consider the extracted features from CNN techniques and further utilize them as input for classification techniques such as GAN, General Regression Neural Network (GRNN), Artificial Neural Network (ANN), and Radial Basis Neural Network (RBNN). The results demonstrate performance improvement, prompting further research to optimally configure the top-performing technique.

Eventually, the research integrates optimization techniques to configure the hyperparameters of GAN for performance improvement. The optimization techniques involved in this process include LOA, Grey Wolf Optimization (GWO), Cat Swarm Optimization (CSO), and Particle Swarm Optimization (PSO).

The research considers eight performance evaluation metrics, namely Accuracy, False Discovery Rate (FDR), False Negative Rate (FNR), False Positive Rate (FPR), Negative Predictive Value (NPV), Positive Predictive Value (PPV), Sensitivity, and Specificity, to assess the effectiveness of the employed techniques in classifying human faces and AI-generated faces. Additionally, Receiver Operating Characteristic (ROC) is considered to exhibit the performance of employed techniques. A confusion matrix is also employed in this research to showcase the performance of the proposed method.
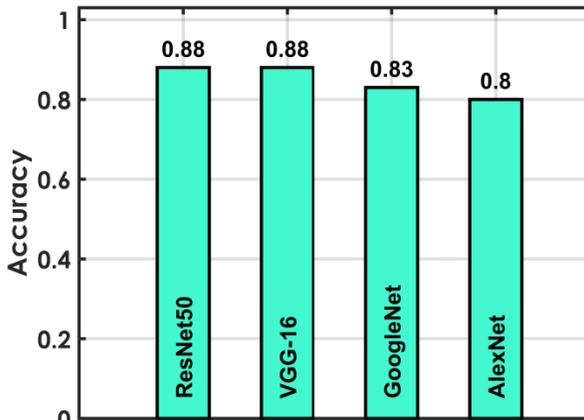
4.1 Distinguish Real and AI-Generated Human Faces Through CNN Techniques

The results of the face classification task, aimed at distinguishing between real and AI-generated human faces, reveal varying performances among the four different techniques: ResNet50, VGG 16, Google Net, and AlexNet. ResNet50 and VGG 16 demonstrated strong and balanced performance, achieving an overall accuracy of 88%. ResNet50 exhibited equal sensitivity and specificity at 88%, while VGG 16 showcased higher specificity (90%) but with a slightly lower sensitivity of 86%. Both models maintained low false positive and FNR, indicating a robust ability to discern real and AI-generated faces. Google Net displayed a balanced but slightly lower accuracy of 83%, with sensitivity and specificity at 84% and 82%, respectively. The model, however, showed a relatively higher false positive rate compared to ResNet50 and VGG 16. In contrast, AlexNet presented the lowest overall accuracy at 80%, with equal sensitivity and specificity of 80%. The model demonstrated higher false positive and FNR, suggesting potential challenges in accurately distinguishing between the real and AI-generated human faces. Figure 7 exhibits the performance comparison of distinguish Real and AI-generated human faces through CNN techniques. Table 1 shows the performance evaluation matrices for four different CNN models.
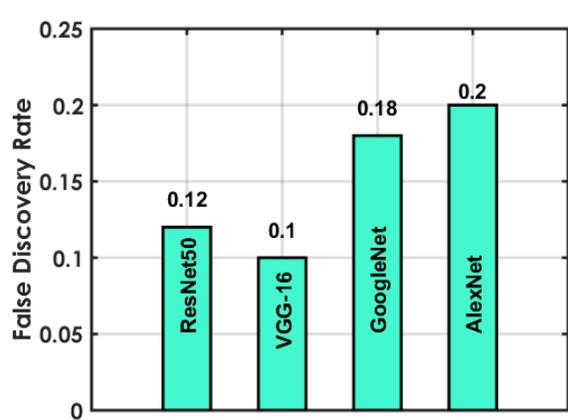
Table 1: Metrics to evaluate the performance of CNN model

Techniques	TP	TN	FP	FN	Total Images
ResNet50	44	44	6	6	100
VGG 16	43	45	5	7	100
Google net	42	41	9	8	100

Alex net	40	40	10	10	**100**

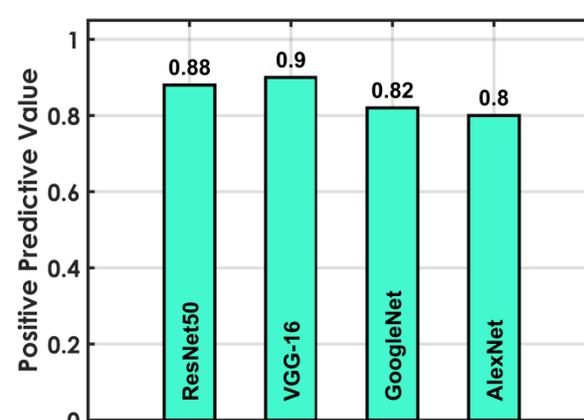

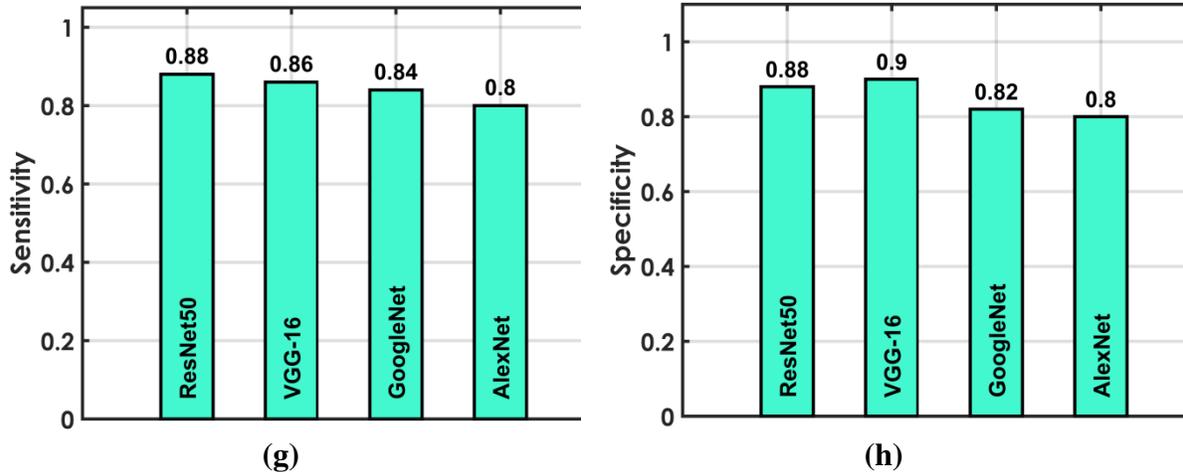**(a)**

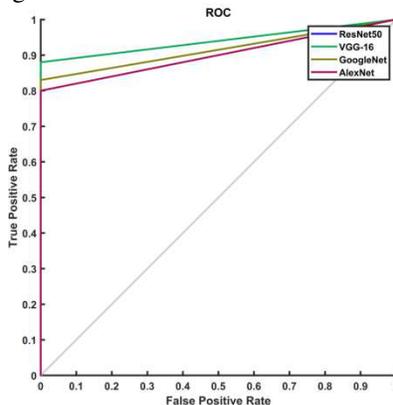**(b)**

(c)

(d)

(e)

(f)

(g) **(h)**
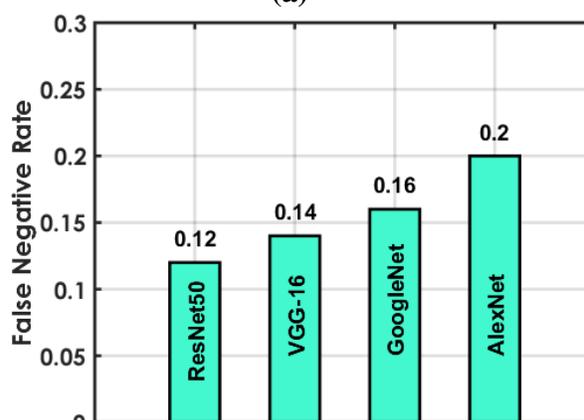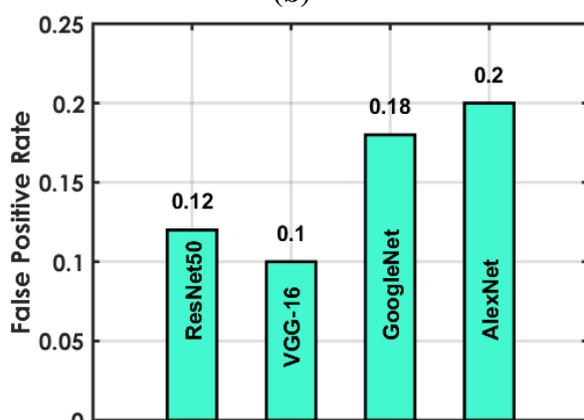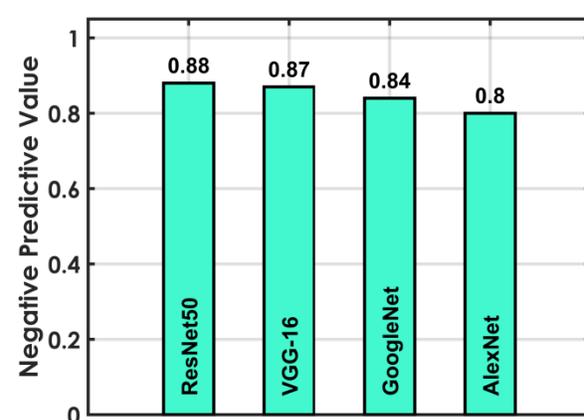
Fig. 7: performance comparison of Distinguish Real and AI-Generated Human Faces through CNN Techniques

4.2 Receiver Operating Characteristic Analysis for CNN techniques

The Receiver Operating Characteristic (ROC) analysis for the given Figure 8 reveals the discrimination performance of different techniques namely ResNet50, VGG 16, Google Net, and AlexNet in distinguishing real and AI-generated human faces. ResNet50 and VGG 16 exhibit strong ROC curves, showcasing high sensitivity at 88% and 86%, respectively, while maintaining relatively low False Positive Rates (FPR) at 12% and 10%, respectively. Google Net follows closely with 84% sensitivity and 18% FPR, indicating a balanced trade-off. AlexNet, though with a lower sensitivity of 80%, demonstrates a 20% FPR. In general, ROC curves for ResNet50 and VGG 16 are expected to be closer to the upper-left corner, signifying superior discriminatory ability, while Google Net and AlexNet's curves would illustrate their respective trade-offs between sensitivity and FPR at dissimilar decision thresholds. The area under the ROC curve (AUC-ROC) for each model quantifies its overall performance, with a higher AUC-ROC indicating more effective discrimination.

Fig. 8: Receiver Operating Characteristic for CNN techniques
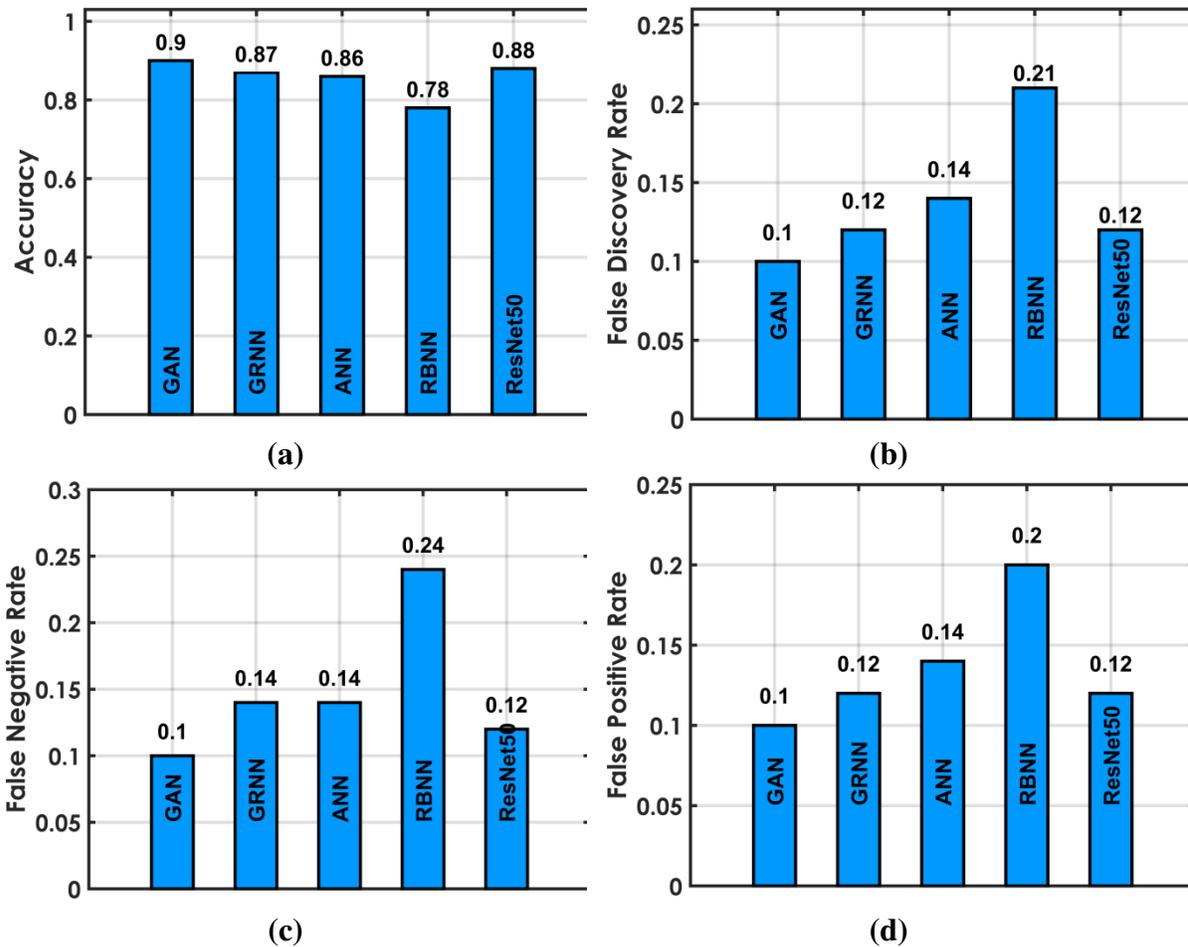
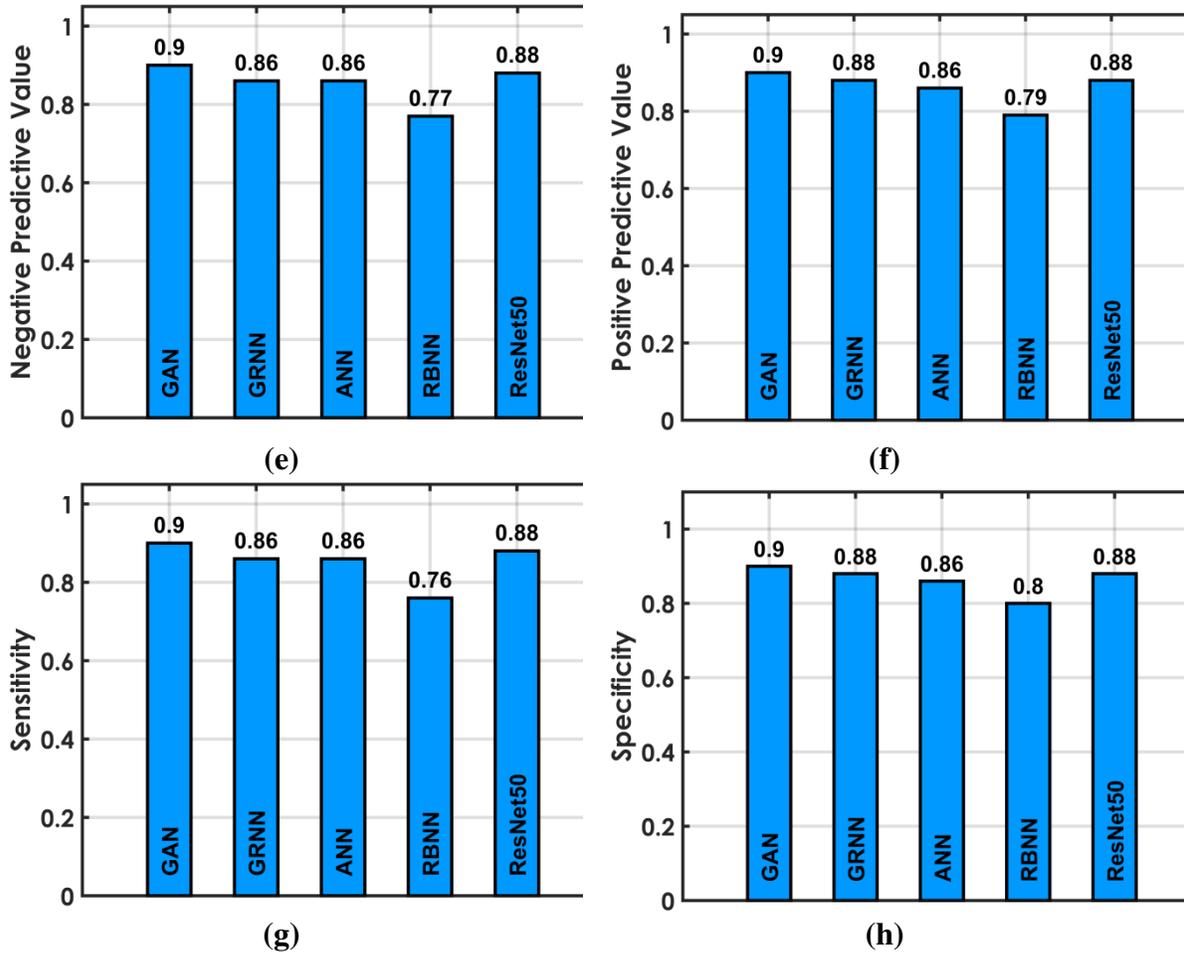## 4.3 Distinguish Real and AI-Generated Human Faces Through traditional Approach

The input for these classification techniques comprises the features extracted from ResNet-50, used to discern between real and AI-generated images. In the comprehensive comparison of the four techniques GAN, GRNN, ANN, and RBNN evaluating their performance in distinguishing between real and AI-generated human faces shown in Figure 9, GAN stands out with robust results, achieving an accuracy of 89%. This was accomplished through a balanced

combination of high sensitivity (88%) and specificity (90%). Following closely, GRNN demonstrates strong and balanced performance, boasting an accuracy of 87%. Meanwhile, ANN maintains consistency in accuracy at 86%, showing comparable but slightly lower sensitivity and specificity. On the other hand, RBNN exhibits a lower overall accuracy of 78%, marked by diminished sensitivity and specificity. Notably, GAN and GRNN excel in minimizing false positives and false negatives, emphasizing their effectiveness in accurate classification. Table 2 displays the performance assessment matrices for four distinct classical classification models. The optimal technique to utilize will rely on the particular needs of the application and the intended trade-off between sensitivity and specificity.

Table 2: Metrics to evaluate the performance of traditional model

| Techniques	TP	TN	FP	FN	Total Images
GAN	44	45	5	6	100
GRNN	43	44	6	7	100
ANN	43	43	7	7	100
RBNN	38	40	10	12	100

Fig. 9: Performance comparison of Distinguish Real and AI-Generated Human Faces through traditional approach

4.4 Receiver Operating Characteristic Analysis for traditional techniques

The ROC analysis for the presented Figure 10, encompassing various classification techniques including GAN, GRNN, ANN, and RBNN, provides insights into their discrimination capabilities for distinguishing between real and AI-generated images. Each technique's ROC curve showcases the trade-off between sensitivity (the ability to correctly identify real images) and the false positive rate (FPR, indicating the misclassification of AI-generated images as real) at different decision thresholds. GAN exhibits a strong ROC curve with high sensitivity (88%) and specificity (90%), indicating an effective discrimination between the two classes. GRNN closely follows, demonstrating a well-balanced performance with 87% accuracy. ANN maintains a consistent accuracy of 86%, while RBNN exhibits a lower overall accuracy of 78%. The ROC curves for these techniques visualize their ability to minimize both false positives and false negatives, providing valuable insights into their performance across different classification thresholds.
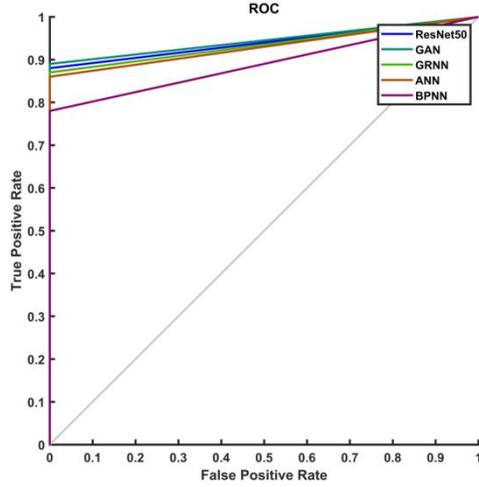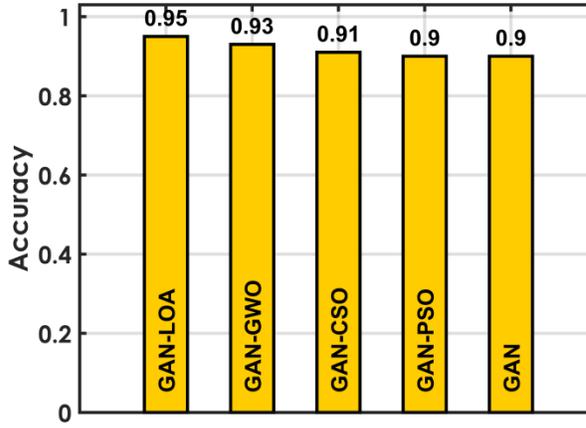
Fig. 10: Receiver Operating Characteristic from traditional approach

4.5 Distinguish real and AI-Generated Human Faces Through optimally configured GAN

In a comprehensive evaluation of various GAN models employing different optimization algorithms LOA, CSO, GWO, and PSO, the distinct performance characteristics emerge in Figure 11. GAN-LOA excels with a remarkable sensitivity of 96%, specificity of 94%, and an overall accuracy of 95%, demonstrating a robust ability to distinguish between real and AI-generated images. GAN-CSO follows closely with a balanced sensitivity and specificity, yielding an accuracy of 93%. GAN-GWO and GAN-PSO both maintain competitive performances, achieving accuracies of 91% and 90%, respectively. The overall GAN model, combining results from these variants, exhibits strong discrimination capabilities, attaining an accuracy of 89% with a well-balanced trade-off between sensitivity and specificity. Notably, all GAN models minimize false positives and false negatives effectively, emphasizing their collective efficacy in image classification tasks. The optimal choice among these models depends on specific application requirements and the desired balance between sensitivity and specificity. Table 3 displays the performance evaluation metrics for optimally configured GAN models through optimization techniques.

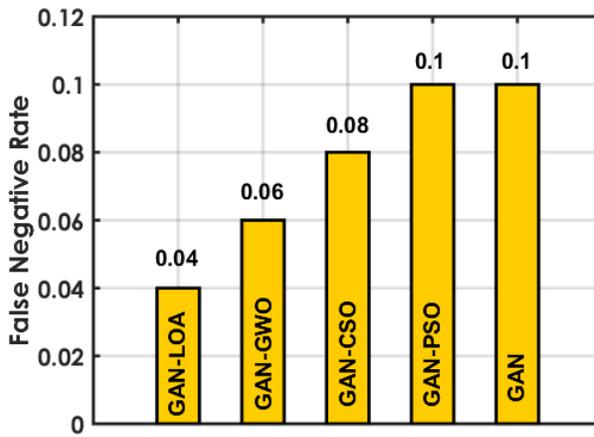**Table 3:** Metrics to evaluate the performance of optimally configured GAN model

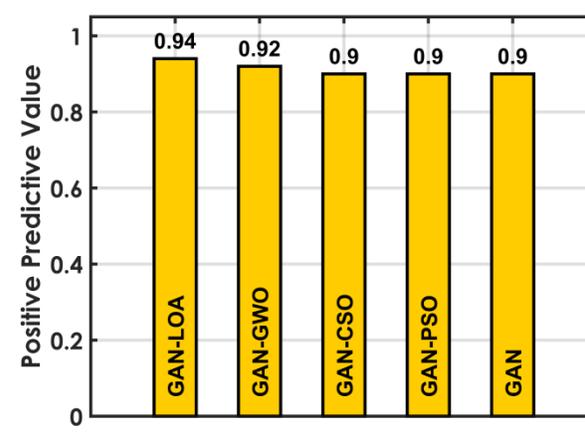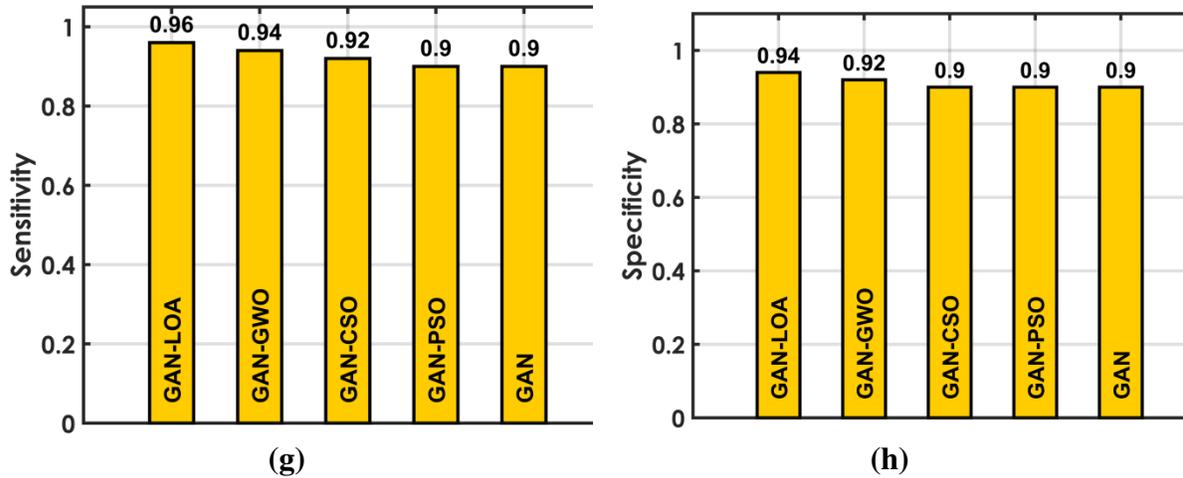| Techniques	TP	TN	FP	FN	Total Images
GAN-LOA	48	47	3	2	**100**
GAN-CSO	47	46	4	3	**100**
GAN-GWO	46	45	5	4	**100**
GAN-PSO	45	45	5	5	**100**
GAN	44	45	5	6	**100**

(a)

**(b)**

(c)

(d)

(e)

(f)

(g) **(h)**
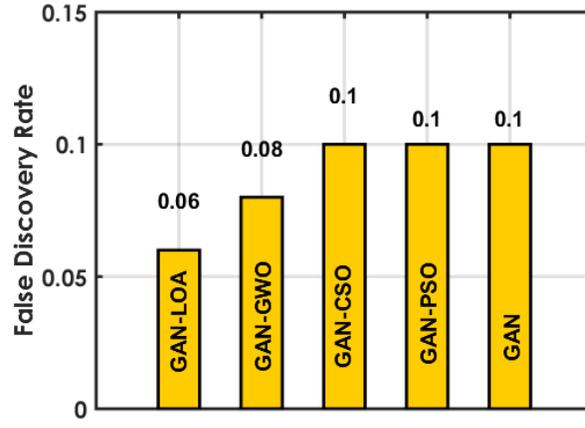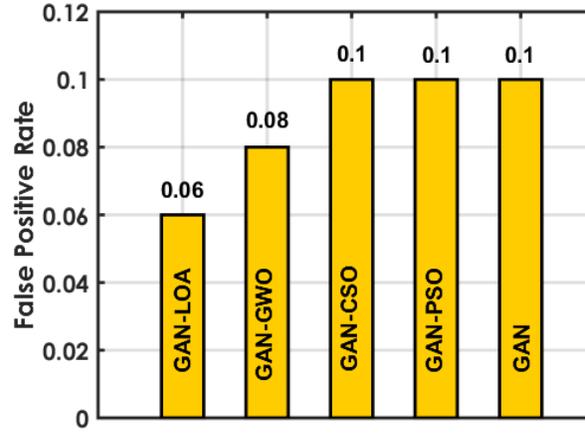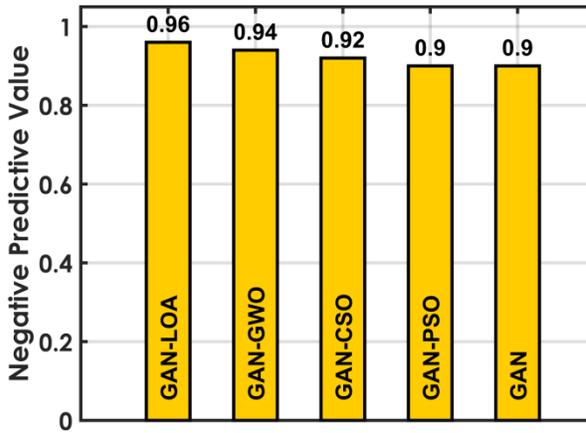
Fig. 11: performance comparison of Distinguish Real and AI-Generated Human Faces through optimally configured GAN

4.6 Receiver Operating Characteristic (ROC) Analysis for optimally configured GAN techniques

The ROC analysis for the diverse GAN models employing distinct optimization algorithms LOA, CSO, GWO, PSO, and the overarching GAN reveals nuanced performance characteristics in distinguishing between real and AI-generated images shown in Figure 12. GAN-LOA stands out with an exceptional ROC curve, boasting a high sensitivity of 96% and specificity of 94%, indicative of its superior ability to accurately classify both classes. GAN-CSO follows suit with a robust ROC curve, demonstrating a balanced sensitivity of 94% and specificity of 92%. GAN-GWO and GAN-PSO exhibit competitive ROC curves, balancing sensitivity and specificity at 92%, 90% and 90%, 90%, respectively. The overall GAN model showcases a strong ROC curve, achieving a balance between sensitivity (88%) and specificity (90%). These curves collectively illustrate the trade-off between true positive and false positive rates, providing valuable insights into the models' discrimination capabilities across different decision thresholds. The choice of the most suitable GAN model hinges on specific application requirements and the preferred balance between sensitivity and specificity.
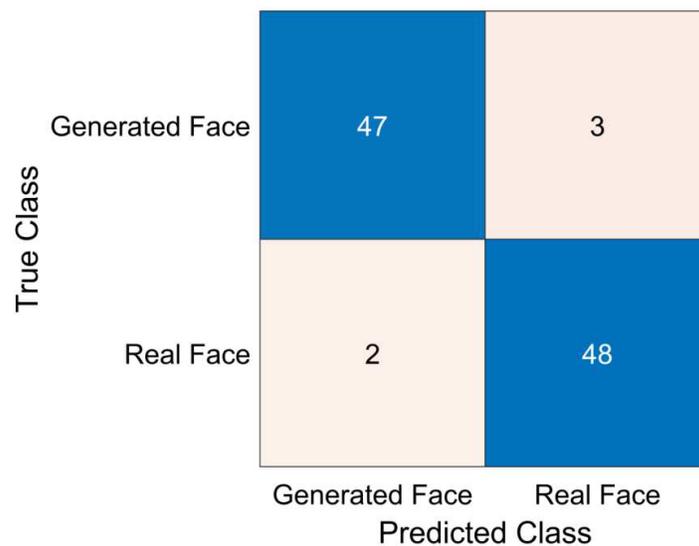
4.7 Receiver Operating Characteristic Analysis for Proposed approach
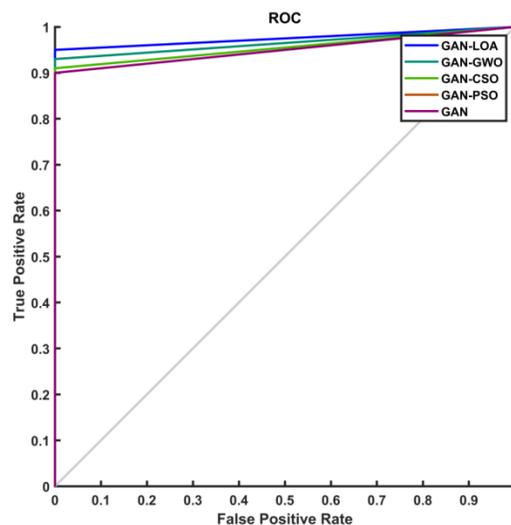
4.8 Confusion Matrix

The confusion matrix provided encapsulates the performance of a binary classification model in distinguishing between AI-generated and real faces, as illustrated in Figure 13. In this matrix, the diagonal elements reveal instances of correct predictions, with 47 AI-generated faces and 48 real faces accurately classified. The off-diagonal elements provide insights into misclassifications, including 3 cases where AI-generated faces were erroneously predicted as real and 2 instances where real faces were mistakenly classified as AI-generated. This information is instrumental in computing various evaluation metrics. Precision, calculated as TP/(TP+FP), signifies the proportion of correctly predicted real faces out of all instances predicted as real, while recall, computed as TP/(TP+FN), measures the fraction of real faces correctly identified out of all actual real faces. The overall accuracy of the model, expressed as (TP+TN)/(TP+TN+FP+FN), reflects the proportion of correctly classified instances across both classes. This comprehensive analysis aids in a nuanced understanding of the model's strengths and areas for improvement in discerning between AI-generated and real faces.

Fig. 13: Confusion matrix for the proposed LOA configured GAN

5. Conclusion

The strides in AI technology, particularly the development of realistic human faces through deep-fake techniques, present considerable challenges in accurately discerning between real and AI-generated faces. This underscores the necessity for robust classification methods, ensuring the ethical and responsible deployment of such technology. Traditional approaches face difficulties in distinguishing between authentic and AI-generated faces, magnifying potential risks in security, identity verification, and misinformation. This study addressed these challenges by designing an optimally configured GAN, surpassing current limitations and providing a robust classifier for accurately distinguishing real human face from Ai-generated human faces images. The results, with an impressive 95% accuracy in distinguishing between real and AI-generated faces, affirm the effectiveness of the proposed GAN model. Future research directions include enhancing generalizability across diverse datasets and face types, exploring internal decision-making mechanisms for interpretability, and developing real-time face classification for live video streams. These advancements will expand the model's impact, advancing the ethical deployment of AI, reinforcing security and identity verification measures, countering

misinformation, and ultimately solidifying public trust in the responsible use of this potent technology.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Alqahtani H, Kavakli-Thorne M, Kumar G (2021) Applications of generative adversarial networks (gans): An updated review. *Arch Comput Methods Eng* 28: 525-552 https://doi.org/10.1007/s11831-019-09388-y

Baraheem SS, Nguyen TV. (2023) AI vs. AI: Can AI Detect AI-Generated Images?. *J. Imaging* 9(10): 199 https://doi.org/10.3390/jimaging9100199

Caramihale T, Popescu D, Ichim L (2018) Emotion classification using a tensorflow generative adversarial network implementation. *Symmetry* 10(9): 414 https://doi.org/10.3390/sym10090414

Caporusso N, Zhang K, Carlson G, Jachetta D, Patchin D, Romeiser S, Vaughn N, Walters A (2019) User discrimination of content produced by generative adversarial networks. *In Human Interaction and Emerging Technologies: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019),* 22-24, Nice, France, 725-730 Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-25629-6_113

Dang LM, Hassan SI, Im S, Lee J, Lee S, Moon H (2018) Deep learning based computer generated face identification using convolutional neural network. *Appl. Sci* 8(12): 2610 https://doi.org/10.3390/app8122610

Dehghani M, Bektemyssova G, Montazeri Z, Shaikemelev G, Malik OP, Dhiman G (2023) Lyrebird Optimization Algorithm: A New Bio-Inspired Metaheuristic Algorithm for Solving Optimization Problems. *Biomimetics* 8(6): 507. https://doi.org/10.3390/biomimetics8060507

Devi AG, Thota A, Nithya G, Majji S, Gopatoti A, Dhavamani L (2022) Advancement of Digital Image Steganography using Deep Convolutional Neural Networks. *In 2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)* 250-254 IEEE, DOI: 10.1109/IIHC55949.2022.10060230

Guarnera L, Giudice O, Guarnera F, Ortis A, Puglisi G, Paratore A (2022) The face deepfake detection challenge. *J. Imaging* 8(10): 263 https://doi.org/10.3390/jimaging8100263

Li B, Lima D (2021) Facial expression recognition via ResNet-50. *Int. J. Cogn. Comput. Eng* 2: 57-64 https://doi.org/10.1016/j.ijcce.2021.02.002

Khoo B, Phan RCW, Lim CH (2022) Deepfake attribution: On the source identification of artificially generated images. *Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov* 12(3): e1438 https://doi.org/10.1002/widm.1438

Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv*:1812.08685 https://doi.org/10.48550/arXiv.1812.08685

Man K, Chahl J (2022) A Review of Synthetic Image Data and Its Use in Computer Vision. *J. Imaging* 8(11): 310 https://doi.org/10.3390/jimaging8110310

Mirsky Y, Lee W (2021) The creation and detection of deepfakes: A survey. *ACM Comput. Surv (CSUR)* 54(1): 1-41 https://doi.org/10.1145/3425780

Meyer DW (2022) Find the Real: A Study of Individuals' Ability to Differentiate Between Authentic Human Faces and Artificial-Intelligence Generated Faces. *In International Conference on Human-Computer Interaction 655-662 Cham: Springer Nature Switzerland,* https://doi.org/10.1007/978-3-031-19682-9_83

Monkam G, Yan J (2023) Digital image forensic analyzer to detect AI-generated fake images. In 2023 8th International Conference on Automation, *Control Robot Eng (CACRE)*, 366-373 IEEE, DOI: 10.1109/CACRE58689.2023.10208613

Moshel ML, Robinson AK, Carlson TA, Grootswagers T (2022) Are you for real? Decoding realistic AI-generated faces from neural activity. *Vis Res* 199: 108079 https://doi.org/10.1016/j.visres.2022.108079

Pan Z, Ren Y, Zhang X (2021) Low-complexity fake face detection based on forensic similarity. *Multimed. Syst* 27: 353-361 https://doi.org/10.1007/s00530-021-00756-y

Partadiredja RA, Serrano CE, Ljubenkov D (2020) AI or human: the socio-ethical implications of AI-generated media content. *In 2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275*) 1-6 IEEE, DOI: 10.1109/CMI51275.2020.9322673

Pataranutaporn P, Danry V, Leong J, Punpongsanon P, Novy D, Maes P, Sra M (2021) AI-generated characters for supporting personalized learning and well-being. *Nat. Mac. Intell* 3(12): 1013-1022 https://doi.org/10.1038/s42256-021-00417-9

Quach LD, Quoc NP, Thi NH, Tran DC, Hassan MF (2020) Using surf to improve resnet-50 model for poultry disease recognition algorithm. *In 2020 International Conference on Computational Intelligence (ICCI),* 317-321 IEEE, DOI: 10.1109/ICCI51257.2020.9247698

Senapati B, Talburt JR, Naeem AB, Batthula VJR (2023) Transfer learning based models for food detection using ResNet-50. *In 2023 IEEE International Conference on Electro Information Technology (eIT)* 224-229 IEEE, DOI: 10.1109/eIT57321.2023.10187288

Shen B, RichardWebster B, O'Toole A, Bowyer K, Scheirer WJ (2021) A study of the human perception of synthetic faces. *In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* 1-8 IEEE, DOI: 10.1109/FG52635.2021.9667066

Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* 64: 131-148 https://doi.org/10.1016/j.inffus.2020.06.014

Verdoliva L (2020) Media forensics and deepfakes: an overview. *IEEE J. Sel. Top. Signal Process* 14(5): 910-932 DOI: 10.1109/JSTSP.2020.3002101

Voss JL, Cohen NJ (2017) Hippocampal-cortical contributions to strategic exploration during perceptual discrimination. Hippocampus 27(6): 642-652 https://doi.org/10.1002/hipo.2271

Wang, X, Guo H Hu S, Chang MC Lyu S (2022) Gan-generated faces detection: A survey and new perspectives. *arXiv preprint arXiv*:2202.07145 https://doi.org/10.48550/arXiv.2202.07145

Whittaker L, Kietzmann TC, Kietzmann J, Dabirian A (2020) All around me are synthetic faces": the mad world of AI-generated media. *IT Prof* 22(5): 90-99 DOI:10.1109/MITP.2020.2985492

Yang J, Xiao S, Li A, Lan G, Wang H (2021) Detecting fake images by identifying potential texture difference. *Future Gener Comput Syst* 125: 127-135 https://doi.org/10.1016/j.future.2021.06.043

Yegemberdiyeva G, Amirgaliyev B (2021) Study Of AI Generated And Real Face Perception. *In 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)* 1-6 IEEE, DOI: 10.1109/SIST50301.2021.9465908

Zhang X, Ghorbani AA (2020) An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manage* 57(2): 102025 https://doi.org/10.1016/j.ipm.2019.03.004