Received: 24 September 2021 | Revised: 30 October 2021 | Accepted: 30 October 2021
DOI: 10.37917/ijeee.18.1.1 Vol. 18| Issue 1| June 2022

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

IJEEE University of Basrh College of Engineering

Face Recognition System Against Adversarial Attack Using Convolutional Neural Network

Ansam Kadhim, Salah Al-Darraji
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Iraq

Correspondence
*Ansam Kadhim
Computer Science Department,
College of Education for Pure Sciences,
University of Basrah, Basrah, Iraq
Email: pgs2181@uobasrah.edu.iq

Abstract
Face recognition is the technology that verifies or recognizes faces from images, videos, or real-time streams. It can be used in security or employee attendance systems. Face recognition systems may encounter some attacks that reduce their ability to recognize faces properly. So, many noisy images mixed with original ones lead to confusion in the results. Various attacks that exploit this weakness affect the face recognition systems such as Fast Gradient Sign Method (FGSM), Deep Fool, and Projected Gradient Descent (PGD). This paper proposes a method to protect the face recognition system against these attacks by distorting images through different attacks, then training the recognition deep network model, specifically Convolutional Neural Network (CNN), using the original and distorted images. Diverse experiments have been conducted using combinations of original and distorted images to test the effectiveness of the system. The system showed an accuracy of 93% using FGSM attack, 97% using deep fool, and 95% using PGD.

KEYWORDS: Face Recognition, Convolutional Neural Network, Adversarial Attacks.

I. INTRODUCTION

Face Recognition (FR) is a technique that is used to recognize faces in images and videos using various algorithms. As the face is the most important identification part in the body of a human, it is useful in many fields for people's identification, such as in airports for security issues. Therefore, face recognition is necessary for such applications. Many factors affect the clarity of the face such as resolution, illumination, and facial expressions. Noise also plays a negative role in faking faces. The technology of face recognition tries to remove noises from faces to get higher accuracy. Consequently, it discovers the original images to enhance the results using suitable algorithms for this purpose.

Hence, it is reasonable that most companies use this technology to get to know their staff and avoid strangers, especially, if the number of employees is high. This technology is related to computer programming and gives full information about a person rapidly. The technology of face recognition was used in most popular regions, for example, 98 countries use this technology. So, the defense techniques by using layers in the CNN algorithm were used to increase adversarial training [1, 2, 3, 4, and 5] on dataset training. This helps them to get results as soon as possible, especially, in the airport, passengers, traveling, working, medicine, and security issues. FR technology is considered one of the most important methods that deal with images of faces for different people. This technology is compatible to discover any noises in faces [6] using CNN algorithm. Therefore, it decreases noises from most images of faces for the train and test dataset.

It is not possible to tell the difference between a real face and an image of a face and cannot be easily recognized by machine learning algorithms. Therefore, biometric sensors can improve recognition accuracy. The advantage of this technology is to enhance security and social environments. It can be used in online banking and medical records for Personal Identification. However, the CNN algorithm is used for this purpose in many areas. In some applications, where the recognition accuracy is required to be high, some factors may affect the recognition, such as intentionally added noise. When the noise was added, the classification of the input image was wrong, as explained by Szegedy et al. [7]. To recognize faces, different systems can be used, and these systems are required to eliminate noise in faces.

This paper aims at identifying the problem of noisy faces (unclear faces) and adversarial images. FR technique is a difficult process, especially if the images are blurry or unclear. Therefore, FR with a suitable algorithm can be used to add some improvements to the faces to recognize them. Some problems related to attacking the FR system were noticed, such as generating adversarial perturbations images using different attacks such as FGSM, Deep Fool and PGD [8]. To eliminate these distortions, it is proposed to use an algorithm suitable for this purpose that includes obtaining high recognition accuracy.

The contributions of this paper are focused on recognizing faces if they are distorted using dataset regeneration. This is done by adding distorted images with three attacks such as FGSM [9], Deep fool [10 and 11], and PGD [12]. Therefore, the proposed system will be powerful enough to be robust against these three attacks. Hence, with difficult situations or environmental problems, the FR system will successfully cope with these difficult situations.

The remainder of this work is structured in the following manner. The review of the literature is presented in Section II. Section III provides an overview of the adversarial attacks. Section IV discusses the techniques that were suggested. The face database for this study is described in Section V. Section VI contains the results of the experiments in more detail. In Section VII, an explanation of the results and their implications were provided.

II. Literature Review

The FR technology has been used to enhance results by using the CNN algorithm to study the faces of people. It is widely used on smartphones and in other forms of technology, such as robotics. However, the algorithm is related to mathematical results suitable for this purpose. However, the results can be enhanced. Therefore, many methods are used for this purpose to increase accuracy, such as FGSM using MNSIT and CIFAR-10 databases.

Gael, Agarwal, et al. [13] showed the use of the filter to generate noise in the manner of agnostic for the network. Therefore, they suggested a defense layer that helps to protect against enemy attacks such as FGSM. Three databases (MNIST, CIFAR-10, and PaSC) were used to get high results. Therefore, efficiency is improved by using this defense layer without more mathematical work.

Carlini, Nicholas, et al. [14] proposed the classification of images for detection by changing small parts. Neural networks perform machine-learning tasks. The inputs such as X face adversarial examples. Therefore, it is difficult to use neural networks, especially in security fields. Distance is considered a very important thing for getting high activity. So, defense distillation was proposed to increase the robustness of the network. Another proposal was to use symbols.

Papernot, Nicolas, et al. [15] suggested some noises such as attacks on neural network, which contain two layers. Their layers are useful to remove (decrease) noises from images. Also, the distance between layers is necessary to recognize the faces clearly by FR technology. Hence, the distance should be simple between images to clearly recognize the face.

Szegedy, Christian, et al. [7] suggested that attack examples were generated by the L-BFGS box. This was done by using L2 distance by finding different images x which is similar to x under the distance of L2.

Deb, Zhang, et al. [16] proposed making the face have noise in some areas (set of pixels). They showed the dangers of adversarial examples in image classification. Hence, CNN might classify images wrongly when the pixels have any noise.

Cisse, Moustapha, et al. [17] proposed that on many tasks, the accuracy of neural networks in tasks is comparable to that of humans, particularly in perception, but the robustness of inputs to change is limited during testing. They suggested that by changing the structure of engineering to move attack examples from one network to another. However, a transferable attack example leads to the creation of a security threat to the production system as well as giving information about the lack of robustness of neural networks.

Dubey, Abhimanyu, et al. [18] proposed several adversarial attacks after adversarial examples which were discovered first. Therefore, changing the image by using oscillations with a scale of L2 or $\ell\infty$ norm leads to changing the predictions of the model. Hence, PGD is related to the gradient-sign method which considers a strong attack. So, the force against adversarial attacks can be increased by using defensive distillation.

In adversarial attacks, by using deep learning models resistant, Aleksander Madry et al. [12] suggested that there are weaknesses in deep learning that facilitate such adversarial attacks. Also, that is implemented values of loss on databases of MINST and CIFAR-10. Hence, the loss developed during 20 runs of PGD.

Xue, Jingsong et al. [11] proposed the face recognition neural network deceiving method that is based on the Deep Fool algorithm. FaceNet is used to generate adversarial samples. Table 1 shows the comparison of the related works.

III. Background

This section explores potential adversarial attacks on facial recognition systems. Face recognition systems are vulnerable to a variety of attacks. Three different types of attacks were mentioned in this paper: Fast Gradient Signed Method (FGSM), Deep Fool, and Projected Gradient Descent (PGD). There are several types of attacks, but these three types were particularly used in this paper because they are the most widely used and most well-known attacks.

A. *Fast Gradient Signed Method (FGSM) attack*

FGSM is called the Fast Gradient Signed Method because it computes the gradients of a loss function (for example, mean-squared error, or categorical cross-entropy) and then utilizes the sign of the gradients to create a new image (i.e., the adversarial image) that minimizes loss. In order to provide the same kind of noise that exists in the gradient, the FGSM method (forward-looking stochastic gradient descent) is used. The magnitude of the noise is scaled by the epsilon constant, and epsilon is typically limited to be a small integer to prevent excessive floating-point arithmetic. Another advantage of the FGSM is that it is a white-box attack, meaning that it is designed to target the specific network structure.

TABLE 1
EXPLAINS A SUMMARY OF THE RELATED WORKS

Authors	FGSM	PGD	Deep Fool	Method
(Deb, Zhang, et al., 2019)	✓	✓		The AdvFaces, an automated adversarial face synthesis method that learns to generate minimal perturbations in the salient facial regions via GAN.
(Dubey, Abhimanyu, et al., 2019)		✓		The image is transmitted by adversarial perturbations away from the image manifold. The aim is to return the image to a manifold before classification.
(Aleksander, Madry et al., 2017)	✓	✓		To address this problem, the aggressive robustness of neural networks was studied through a strong optimization lens.
(Xue, Jingsong et al., 2019)	✓		✓	The face recognition neural network deceiving method that is based on the Deep Fool algorithm is proposed.
(Gael, Agarwal, et al., 2020)	✓	✓	✓	Showed using the filter to generate noise in the manner of agnostic for the network.
Our Method	✓	✓	✓	The work focuses on recognizing faces if they are distorted using the regeneration of the dataset. This is done by adding distorted images with three attacks such as FGSM, Deep Fool, and PGD.

Goodfellow and Colleagues [9] use the term "consequence" to describe the Attack FGSM. In other words, the technique uses the loss gradient to modify the input data to maximize the loss. Also, this technique manipulates the input data to maximize the loss gradient while factoring in the loss function's gradient. In this way, an adversarial example is an instance in which tiny, deliberate feature perturbations lead a machine-learning model to produce an incorrect prediction. As a result, numeric vectors are accepted as inputs by machine learning algorithms. An adversarial attack is defined as the deliberate design of input in such a manner that it causes the model to provide the incorrect output. It is a significant issue in the field of Artificial Intelligence (AI) security to harness this sensitivity and use it in order to alter an algorithm's behavior. The evading attack thus requires less control over the disturbance than before. As a result, the original image is no longer recognizable due to the disturbance. There is still a possibility that you will not be identified as the subject of the Vitim images. The number of iterations increases with the number of recognitions. It is helpful for impersonation attacks, but it is not good for avoiding attacks. Face classification is a classification issue, and this is the problem that deep learning is trying to solve. As a result, the FGSM technique is suggested as an efficient way for generating adversarial samples that may be used to deceive the classifier and mislead it. Table 2 illustrates three different kinds of granularity of perturbations: 0.001, 0.01, and 0.1.

B. *Deep Fool attack*

In adversarial attack techniques, the Deep Fool is a well-known method that employs images in a variety of places. It identifies for the first time the sample oscillations and the model oscillations mirrors that correspond to them. This allows it to calculate the deep classifier's noise on large-scale data sets that include adversarial cases, which is very useful in machine learning [10].

TABLE 2
FGSM ATTACK WITH VARIOUS VALUES OF GRANULARITY OF PERTURBATIONS

Step	Dodging			Impersonation		
	€ =0.001	€ =0.01	€ =0.1	€ =0.001	€ =0.01	€ =0.1
1	81.62%	11.54%	1.74%	97.37%	28.95%	85.1%
5	83.40%	55.38%	46.29%	98.72%	50.42%	69.7%
10	88.43%	49.45%	44.26%	99.22%	57.21%	41.0%

In a recent study, researchers showed that this method is unsustainable when data are subjected to hostile modifications. Despite the fact that deep neural networks have shown remarkable performance in classification tasks, these attacks on the technique revealed many weaknesses in the system. As a result, these algorithms have the potential to enhance results by filtering out noise in the images. In other words, machine-learning models are capable of producing particular misclassifications based on some different kinds of sample data. While deep networks have been shown to be very successful at classification tasks, they are often misled by little and undetectable changes in the data sets they are trained on. It is shown in this case that adversarial cases in deep learning models are adequate to reveal our blind spots. Additionally, numeric vectors are accepted as inputs by machine learning algorithms. As a consequence, the Deep fool technique computes perturbations that fool deep networks in a brief span, allowing for the quantification of their resilience.

After a few rounds, the scientists discovered that Deep Fool converges to an oscillation vector that deceives the classifier, thereby fooling it (i.e., fewer than three). The oscillation vector is also more accurate than that of other current models, which is another plus. In contrast, "the Fast Gradient Sign" generates a perturbation image with a greater normal, while this method generates minimal adversarial perturbations. Because of this, with Deep Fool, it is recommended to create adversarial samples that are capable of deceiving current-generation classifiers.

C. Projected Gradient Descent (PGD) attack

PGD "Projected Gradient Descent" is a white-box attack, which implies the attacker has direct access to the model gradients during the attack. As a result of this attack, the attacker acquired a copy of the weights associated with your model. Then, the PGD attack is almost identical to the BIM (Basic Iterative Method) and IFGSM (Iterative–Fast Gradient Signed Method) attacks. After that, the BIM executes FGSM with reduced step size and restricts the updated adversarial sample to a range that the algorithm can handle in T iterations. PGD, on the other hand, initializes the example at a random location inside the ball of interest (specified by the L norm) and then performs random restarts, while BIM initializes the example at the starting position (which is decided by the L norm).

Consequently, the rapid gradient sign technique demonstrates its efficacy by creating an adversarial example using the neural network's gradients. A source image is utilized to generate a new image that is as similar to the original as feasible [12], which is then used to mitigate the loss.

IV. THE PROPOSED METHOD

This proposed system is capable of recognizing faces even if there are adversarial images in databases. It distorts images using three kinds of attacks: FGSM, Deep fool, and PGD. The system is composed of four steps: pre-processing, generation of adversarial images, feature extraction, and building the classifier. So, Figure 1 explains the proposed system diagram.

A. Pre-processing

Pre-processing is a crucial step that proceeds any recognition system. It can affect the accuracy of the system tremendously. However, the variety in size and distance of faces in images may lead to poor recognition. So, the LFW database contains a set of classes for one person in different situations. The original image size before pre-processing was (250 x 250) pixels. In this work, pre-processing includes face detection, cropping, and resizing. As indicated in Figure 2, the technique of image editing is used.

Before training the images, only faces are cropped. Its size before being cropped is (250x250) pixels, whereas after cropping, the size of the images became (146x146) pixels. Figure 3 shows the process of cropping and resizing the image to make all images equal. Figure 4 shows images after processing.

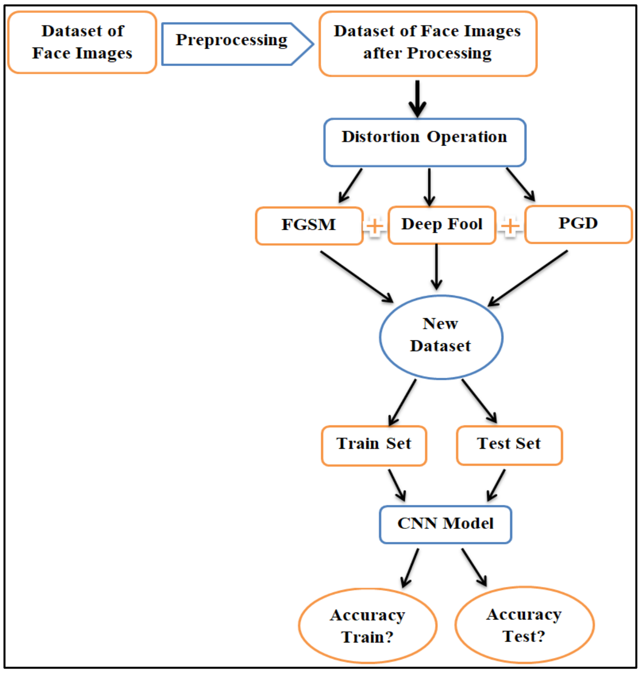

Fig. 1: The Structure of the proposed System.

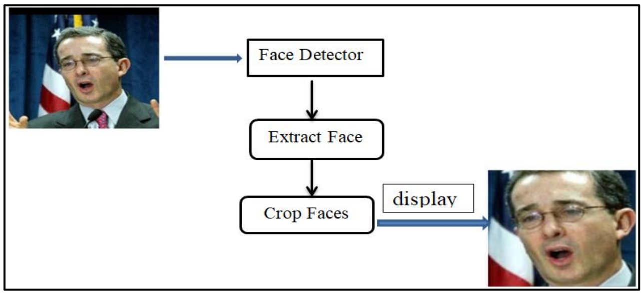

Fig. 2: Part of the image by cropping.

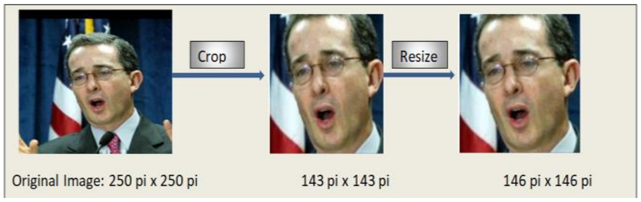

Fig. 3: Cropping and resizing of images with changing of size.

Fig. 4: A Sample of some images after processing.

B. *Generation of adversarial images*

Some improvements can be shown in adversarial attacks. There are also some advantages to the goal they achieve. Therefore, each attack represents the basics of the real world. However, here, various first strategies for producing adversarial situations are discussed.

1) *Fast Gradient Sign Method (FGSM) attack*

This method works by using special networks such as a neural network. This leads to creating an example for adversarial images. Also, this method uses the gradients of the loss in input images. So, it creates a new image that improves the technique of utilizing loss gradients in the input image. It also creates a new image to increase the loss. For the FGSM attack, the attack step size parameter is fixed to 0, 0.01, 0.1, and 0.15. However, it is explained by the following equation (1):

$$adv\_x = x + \epsilon \times \text{sign}\ (\nabla\ x\ J\ (\theta, x, y)) \qquad (1)$$

Figure 5 shows the input image when the epsilon value is zero respectively.

Fig. 5: Input image when epsilons = [0].

Running of FGSM attack to create disturbances (oscillations) used to distort original images. Figure 6 shows the addition of noise to the original image.

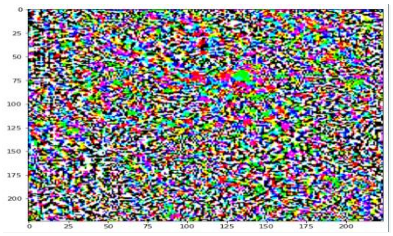

Fig. 6: Image noise.

Different values of epsilon can be used. Then, the output of the image can be noticed. The value of epsilon is (0.01, 0.1, and 0.15). Figure 7 shows adversarial images using three values of epsilon.

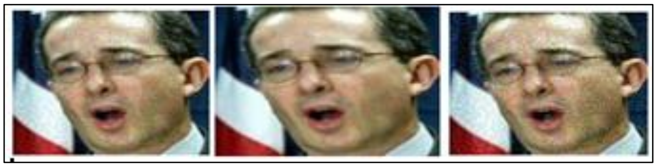

Fig. 7: Shows FGSM attack with various values of granularity of perturbations.

2) *Deep Fool attack*

Deep Fool is a simple algorithm used to find adversarial oscillation images in deep networks. Researchers proposed that the Deep Fool algorithm is used to compute adversarial examples that would noise modern classifiers. Figure 8 shows distorting the image of the deep fool attack method.

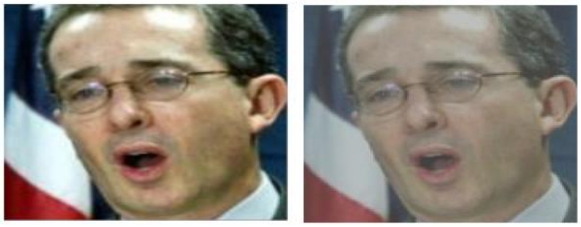

Fig. 8: Shows original and adversarial image using FGSM attack.

3) *Projected Gradient Descent (PGD) attack*

This kind of attack model works with many pixels of images. Each pixel can be distorted by at most epsilon = 0.8 of its initial value. All pixels can independently jam, so this is an endless attack. The test set should be configured as a one-row matrix for each example and each row has a flat matrix of (146 x 146) pixels. Hence, the overall dimensions are 10,000 rows and 21,316 columns. Each pixel must be in the range of [0, 1]. While the PGD attack parameter is fixed to 8.0. Figure 9 shows distorting the image of the PGD attack method.

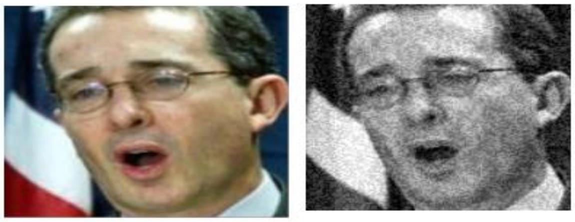

Fig. 9: Shows original and adversarial image using PGD attack.

C. *Feature Extraction*

Feature extraction is a process that divides and reduces a large collection of raw data into smaller, more manageable groupings. As a consequence, processing it will be more straightforward. As a result, CNN networks are responsible for extracting features from the images in the collection. The main distinguishing characteristic of these large data sets is their large number of variables. Thus, face feature extraction is the process of extracting individual facial component characteristics from a photograph of a human face, such as the eyes, nose, and mouth. Face feature extraction is critical for the beginning of processing methods such as face tracking, facial emotion detection, and face recognition. The proposed work uses CNN, which does not need a stand-alone feature extraction. It uses a learnable feature extractor as a part of the network, which extracts a lot of suitable features.

D. Classification

This paper uses the CNN network to deal with the original images and to distorted images of the dataset. It enhances the results with high accuracy. This network helps to reduce or remove perturbation from the original images. However, building a CNN structure consists of ten layers, made up of five “convolutional layers” and five “pooling layers” with “Fully-Connected (FC) layers”. After that, they extract all the features and use the Softmax classifier for recognition. Then, all the original and distorted images are trained with several methods of tests on the neural network. Thus, Figure 10 shows the structure of CNN and Figure 11 shows the general structure of CNN respectively. The classification process is composed of two steps, which are the training and prediction model for face images.

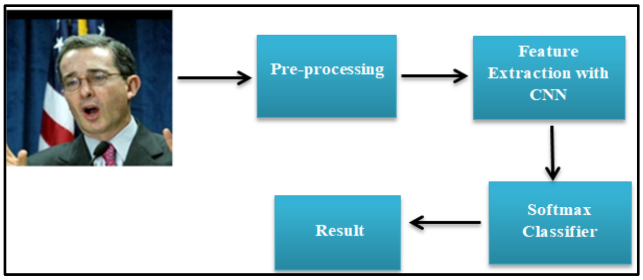

Fig. 10: Structure of the proposed convolutional neural network.

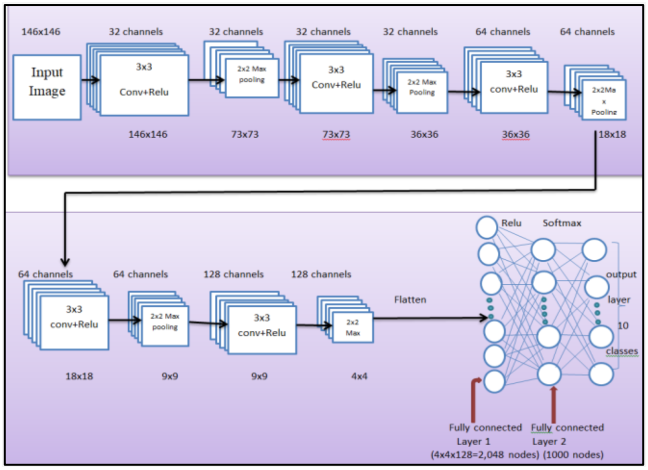

Fig. 11: The general structure of CNN of the original and distorting of the faces recognition system.

1) Training

The images are chosen from the LFW database and then the number of classes is determined, which is about 10 classes for males and females in different positions. The total number of original images is about 1,083 images. Then these images are divided randomly and manually into 70% (720 images) for training and 30% (363 images) for testing.

After that, all database images are distorted with three kinds of attacks; FGSM, Deep Fool, and PGD. Finally, the CNN model will be trained using original and distorted images set to get high accuracy of recognition which is strong to face any problems on the FR system.

2) Prediction Model for Face Images

Recognizing and confirming individuals from an image of their face is a computer vision job known as face recognition. Although various open-source implementations and pre-trained models for Google's net facial recognition system about the face, face image predictions are being used in increasing numbers by facial analysis apps, due to the fact that the technology has a wide range of applications. However, although the existing models are still lacking in accuracy, they are hindered by the vast variety of face images that exist (such as differences in lighting, poses, and angles). It is necessary to follow such a process in order to use these models in real-world situations.

For the accurate prediction of a collection of face images, an improved deep learning structure based on the combination of attention and residual convolutional networks was presented.

Using multitasking learning, the accuracy of face prediction may be enhanced by adding predicted faces to the feature embedding of the face classifier, which can then be used to further train the model. When our proposed model was trained, an image of a well-known individual and a frequently used dataset were used, and the results were very remarkable. Observing our trained model's attention maps, it can be seen that it has learned to be aware of different facial areas over time. Image prediction is accomplished via the usage of the CNN. The methods used for image processing are shown in the following Figure 12.

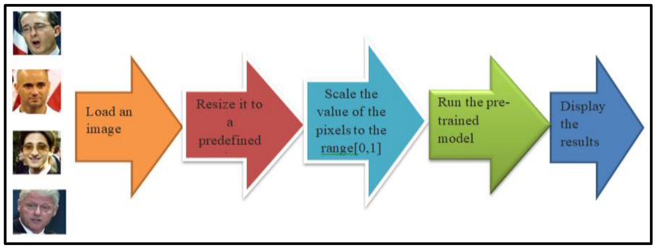
Load an image
Resize it to a predefined
Scale the value of the pixels to the range[0,1]
Run the pre-trained model
Display the results

Fig. 12: The image prediction process.

V. Experimental and Results

Experiments were performed to illustrate the suggested method's efficacy of recognizing faces in spite of different attack generations on the CNN model.

A. Setup

Databases: LFW (Labeled Faces Wild) [19] is used for the tests. Face images are in the LFW database for the purpose of studying the issue of unlimited face recognition. To run the tests, the LFW database will be used. However, the method's performance is assessed using the LFW database. The database has more than 13,000 images of faces. Each portrait is labeled with the subject's name. The images are available in two sizes (250 by 250) pixels with variations in emotion, posture, time, and gender. However, not all classes were used because not all classes have a large number of images, some of them with only one or two images, which is not sufficient to recognize faces. Therefore, 10 classes were selected, in which each person has at least 50 images with several shots and different angles and directions.

Furthermore, they can only be seen by the Viola-Jones face detector. The LFW dataset was next processed, and Figure 13 displays a sample of the post-processing results.

Fig. 13: Some Face Images of Different Subjects of the LFW Database.

To set up our experiments, our system is implemented using Python 3.7 language and then using the programs of the Anaconda Navigator and Spyder 4.1.4 environment on an Intel (R) 2.20 GHz Core (TM) i7-8750H CPU with 12.0 GB of RAM running Windows 10. So, the test is conducted on five experiments, and each experiment includes, in the train set and test set, a different number of images. Table 3 illustrates the number of images in the train set and test set for each experiment.

B. *Results*

In all experiments, the CNN is trained with images generated using three types of attacks. The performance of the proposed CNN is evaluated to recognize the difference between original images and adversarial images with the use of a Rmsprop optimizer for perturbation optimization. However, the first experiment is by training the CNN model using the original image set. Therefore, the accuracy obtained for the testing set is approximately 95%.

While the second experiment is by training the original and distorted images using the FGSM, Deep Fool and, PGD attack with different cases. Table 4 illustrates the accuracy of face recognition for several situations.

- **Case 1**: when training the CNN model using original image and testing with original and distorted images.
- **Case 2**: when training the CNN model using original and distorted images and testing with original images.
- **Case 3**: when training the CNN model using original and distorted images and testing with original and distorted images. However, the proportions are varied with respect to accuracy.

As for the final merge experiment, training is done in three cases:

- **Case 1**: when training the CNN model using original and testing with original images and all distorted images of three types of attacks are FGSM, Deep Fool, and PGD to obtain a medium accuracy.
- **Case 2**: when training the CNN model using original and all distorted images of three types of attacks are FGSM, Deep Fool, and PGD and testing with original images, and the accuracy was high.
- **Case 3**: when training the CNN model using original and all distorted images of three types of attacks are FGSM, Deep Fool, and PGD and testing with original and all distorted images and the accuracy is close to the accuracy of the original images.

In addition to our network, the database was trained using VGG-16 and VGG-19 network; however, the results obtained were not very accurate compared to our model which was of high accuracy in distinguishing between faces in original and distorted images.

TABLE 3
NUMBER OF IMAGES IN EACH EXPERIMENT

Experiment	No. original image in train set	No. original image in the test set	No. image (original & distorted) in test set	No. image (original & distorted) in train set	No. image (original & distorted) in train set & test set
Clean	720	363	-	-	-
FGSM	720	363	1799	3590	5389
Deep fool	720	363	718	1436	2154
PGD	720	363	718	1436	2154
Merge	720	363	2517	5026	7543

TABLE 4
RECOGNITION FACE ACCURACY FOR SEVERAL OF SEVERAL SITUATIONS

Experiment	Train set (Original) & Test set (Original)	Train set (Original) & Test set (Original & Distorted)	Train set (Original & distorted) & Test (Original)	Train & Test (Original & Distorted)
Clean	95%	-	-	-
FGSM	-	54%	89%	93%
Deep fool	-	70%	94%	97%
PGD	-	94%	95%	95%
Merge	-	60%	93%	89%

VI. Conclusion

To distinguish between original and adversarial images, a CNN algorithm was used. Therefore, ten classes were used for ten people. That means each class for every person and everyone has many different captures of images with different positions. Thus, classes were divided into groups of train and test datasets. So, every person has two folders for both train and test images. However, the number of train images should be more than the images in the test to get clear results. Hence, the first method is by training original images on the CNN algorithm. So, the rate of recognition is "95%" between train images and test images.

For future work, The number of attacks can be increased to distort images such as One Pixel attack, Carlini & Wagner attacks (C&W), Visible Light-based attack (VLA), AdvHat attack, Face Friend-safe attack, etc. The number of classes can also be increased with new images.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] Mofeed T. Rashid, "Modeling Of Self-Organization Fish School System By Neural Network System," *Basrah Journal for Engineering Science*, vol. 15, no. 1, pp. 14-19, 2015.

[2] I. Goodfellow, Jonathon Shlens, Christian Szegedy, "Explaining and harnessing adversarial examples," 2014.

[3] R. Huang, B. Xu, D. Schuurmans, and C. J. Szepesvári, "Learning with a strong adversary," CoRR, 2015.

[4] Harini Kannan, Alexey Kurakin, Ian Goodfellow, "Adversarial logit pairing," CoRR, Vol. abs/1803.06373, 2018.

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial machine learning at scale," Computer Vision and Pattern Recognition, 2016.

[6] A. Bharati, R. Singh, M. Vatsa, K. W. Bowyer, and Security, "Detecting facial retouching using supervised deep learning," IEEE Transactions on Information Forensics and Security, vol. 11, no. 9, pp. 1903-1913, 2016.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, "Intriguing properties of neural networks," Computer Vision and Pattern Recognition, 2013.

[8] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Are image-agnostic universal adversarial perturbations for face recognition difficult to detect?," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1-7, 2018.

[9] Y. Liu, S. Mao, X. Mei, T. Yang, and X. Zhao, "Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 433-436, 2019.

[10] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574-2582, 2016.

[11] J. Xue, Y. Yang, and D. Jing, "Deceiving Face Recognition Neural Network with Samples Generated by Deepfool," in *Journal of Physics: Conference Series*, vol. 1302, no. 2, 2019.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv:1706.06083, 2017.

[13] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. K. Ratha, "DNDNet: Reconfiguring CNN for adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 22-23, 2020.

[14] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39-57, 2017.

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*, pp. 582-597, 2016.

[16] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1-10, 2020.

[17] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *International Conference on Machine Learning*, pp. 854-863, 2017.

[18] A. Dubey, L. Maaten, Z. Yalniz, Y. Li, and D. Mahajan, "Defense against adversarial images using web-scale nearest-neighbor search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8767-8776, 2019.

[19] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.