Generative Poisoning Attack on Face Biometric System

[1]Barjinder Kaur, [2]Manvjeet Kaur
[1]M.Tech Scholar, [2]Assistant Professor
[1,2]Department of CSE, Punjab Engineering College(Deemed to be university),Chandigarh, India
94kbarjinder@gmail.com,manvjeet@pec.ac.in

Abstract: Biometric security is the major concern nowadays as through biometrics an individual is being authenticated or identified using his or her biometric trait. The security level needs to be enhanced more so that an individual is authenticated securely. Biometric system is prone to various types of attacks. Poisoning attack is one of the major attacks as it manipulates the training database and that database is being tested on the model. So the model wrongly learns the infected samples as genuine one. In this work, Generative adversarial network has been established to perform poisoning attack on face biometric system where the generator network generates fake imagesand analyses how the generator and discriminator models plays a major role in fooling each other. The performances of generator and discriminator models have been analyzed by computing the accuracy and loss of training and testing databases. Gradient descent optimization algorithm has been used to counter the attack through back-propagation method.

Keywords: Biometric Security, Adversarial Machine Learning, Generative adversarial network: Generator and Discriminator, Poisoning Attack.

1. Introduction

Biometric recognition system is used to make an individual identify his/her identity after getting verified or recognized by the system. Many organizations had deployed biometric system in various applications and hence leave behind a major issue called security. Security and privacy of an individual’s identity is a major concern. The demand of biometric relies on the performance of the system robustness, low error rate and secured from attacks. Much research work has been done to secure the biometric system, where various level of attacks(Level 1 to Level 8) have been identified and various countermeasures has been proposed like liveness detection, challenge/Response, multimodal biometric, template encryption against these attacks [1].

As years passed techniques have been changed with advancement of latest technologies, the increased variability and sophistication of attack threats ,in response to the growing complexity and amount of vulnerable attack points in security systems has favored the adoption of machine learning and pattern recognition techniques to detect variants of known and never before seen attacks. The important use of Machine learning is because of varieties of large available databases, Computational processing which is cheaper and more powerful, affordable data storage.

With the advancement of Machine learning, it still suffered issues in adversarial settings for developing a secure system such as determining vulnerabilities during learning, how to detect attacks &estimate(evaluate) their impact on the attacked system and finding ways to develop countermeasures against the attacks[2]. To improve these conditions various research works has been done and still is in ongoing process.

Various potential attacks in machine learning have been identified as[2][3] [4]:-

1. Poisoning attack (Causative attack):- The attacker tries to manipulate the training database.
2. Evasion Attack (Exploratory attack):- The attacker tries to manipulate the testing database.
3. Data Reconstruction: - At feature space, it has the capability of reconstructing the data which looks similar as the original data, and becomes difficult to distinguish between original and fake. Countering this attack with deep neural networks has aimed at more power in recognizing the data.
4. Privacy Violation: -Compromising the confidential data resulting in violating the privacy of a user.
5. Integrity Violation: -The genuine person is misclassified as fake and vice-versa violating the integrity of the system. The data has been manipulated by the adversary thus frustrating the user for not being recognizable by the system.
6. Availability Violation: -The authorized person is not able to access the system when neededsuch as denial of service attack.

Security Arms Races follows two categories for evaluating the impact of system security before and after attack [5][6].

1. Reactive arms: Firstly the attacker will analyse the whole system, find out the loop holes and then perform an attack. After that the Security designer scrutinizes the attack and will develop countermeasures.

2. Proactive arms: Here the system designer will perform both tasks such as firstly the designer will develop a model and then execute the attack. Secondly, he will inspect the attack's impact and developing countermeasure depending upon the severity of the attack.
   We are considering the proactive arm approach throughout our work. The proactive arm approach anticipates new security vulnerabilities and forecast future attacks. Hence, it is advisable to use proactive approach for a system so as to maintain the security.

In this work, the various potential attacks on machine learning has been examined and then proposed a generative method of generating fake images of training database to perform poisoning attack (causative attack).The proposed method is inspired by the model of: *Generative Adversarial Network (GAN)*[10]. The two main modules of GAN are- Generator Neural Network and Discriminator Neural Network as mentioned in Figure 1 and explained below:

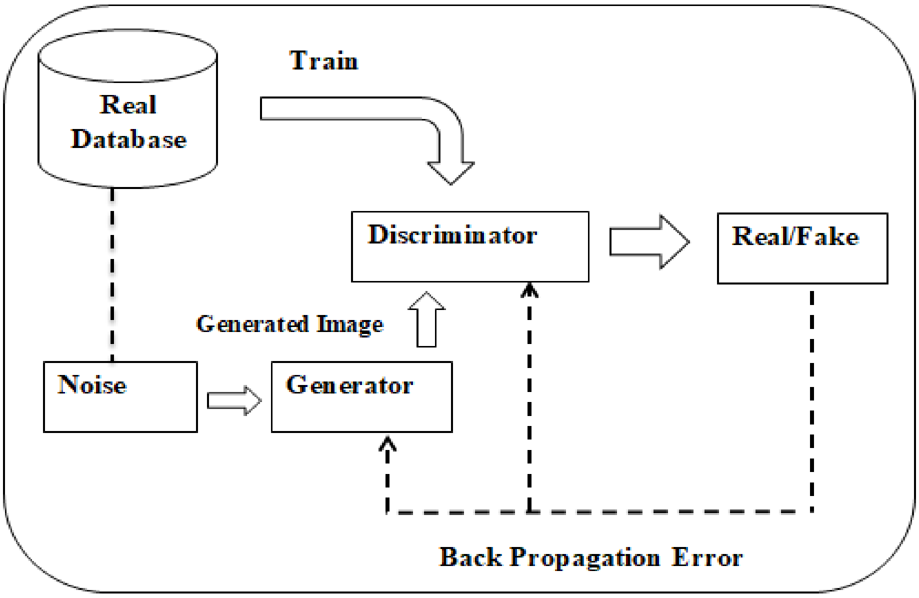


Figure 1: Generative Adversarial Network

a) The Generator Network takes arandom sample and generates a pattern of information which is being fake by the generator by adding noise to the images. The generated sample is then sent to the discriminator Network.
b) The work of Discriminator Network is to take the sample form the real database and from the generated sample, which then foresee whether the information is genuine or fake. The discriminator thenproduces the result in the form of 0(Fake) or 1(Genuine) and gives the feedback results back to both the network through back-propagation.

The attacker follows **adversary model** to simulate an generative poisoning attack which depicts what are the goals of an intruder, what kind of knowledge does he/she possess and what are the capabilities he hold to attack a system as shown in figure 2[5][3].

i. Adversary goal:-Attacker's goal is to generate fakesamples on random data from training database violating the security (privacy, integrity, availability) to misclassify the results (genuine/fake).
ii. Adversary Knowledge:-Depending upon limited or perfect knowledge, the attacker is able to perform an attack. In this scenario, attacker having perfect knowledge of components/algorithms such as training database, learning algorithm, to simulate the poisoning attack.
iii. Adversary Capability: - The capability of an attacker is to manipulate the data from the training database on any number of samples.

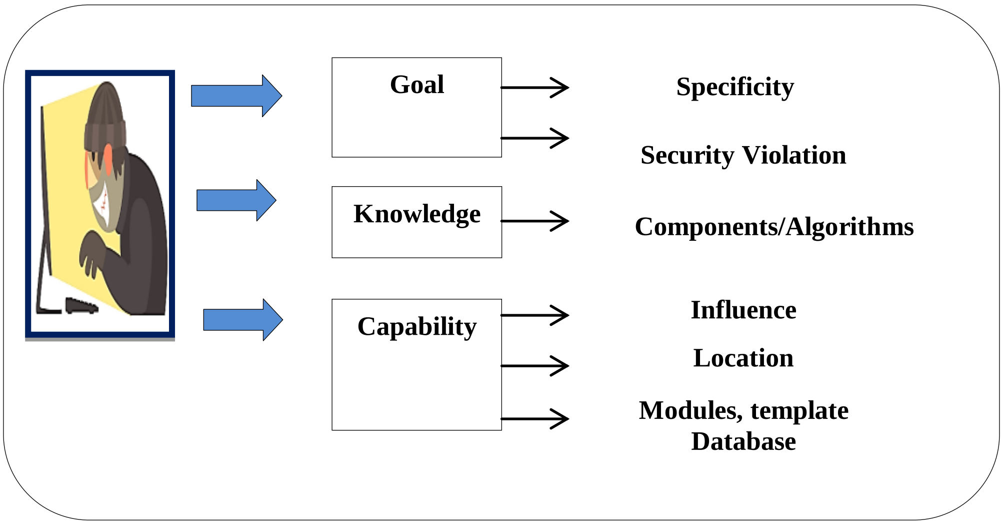


Figure 2: Adversary Model

Key contributions of research work are:

- To examinethe various potential attacks on machine learning and then proposed a generative method of generating fake images of training database to perform poisoning attack (causative attack).The proposed method is inspired by the model of: *Generative Adversarial Network (GAN).*
- The performance analysis of the proposed method is evaluated by performing experiments on live database(face) and ORL face database[11] under different parameters such as loss and accuracy.

2. Related Work

In [1] the author presented a broader and more practical view of biometric system attack vectors, placing them in the environment of a risk-based systems approach to security and outlining defenses.In [2] the authors studied about the machine learning techniques against an adversarial enemy and classified various attack against machine learning algorithms. They even addressed the capabilities & knowledge of an adversary’s to simulate an attack and explored the vulnerabilities.The adversarial machine learning attacks in intelligent and adaptive biometric systems was discussed in [3].They categorized various types of attacks against learning algorithms such as causative (poisoning), exploratory (evasion) and performed on application of adaptive biometric system and introduced the novel prevention method called template sanitization to counter the poisoning attack.The different types of attacks on machine learning techniques and systems were explained by the authors in [4] and mentioneda variety of defenses against those attacks to defend the question “Can machine learning be secure?”.In [5] the authorsdescribed the pattern recognition systems attacks in adversarial settings and explained the design issues of pattern recognition systems. They introduced guidelines to improve the designs against well-crafted attacks in applications like spam and malware detection. They also mentioned about the pro-active and reactive attacks and proposed various techniques to improve the system security design.The pattern classification systems under attacks in adversarial applications like spam filtering, network intrusion detection has been analyzed in [6], due to the exploitation of pattern classifiers the performance of systems was affected and highlighted that security of pattern classifier’s evaluated at design phase can lead to a better design choices i.e. proactive arm design phase. In [7] the authors performed poisoning attack on SVM; they inserted crafted malicious samples into training data which lead to increase the SVM’s test error. To perform poisoning attack they used gradient ascent strategy in which the gradients are calculated on the properties of SVM optimization solutions. The calculated gradients are optimized by training the network using gradient backpropagation which helped out in our proposed work.In [8] adaptive biometric recognition system was introduced due to factors such as aging, environmental conditions. This mechanism of adaptability leads attackers to compromise the stored templates, denying access to the clients and successfully misleading a simple PCA-based face verification system that performs self-update policy. PCA-based face verification systemone template per client as the templates are computed by averaging the n enrolled imaged and that template is referred as centroid. During verification the centroid is updated, if the threshold value of submitted sample and centroid is exceeded than a predefined threshold due to template self-update policy.In [9] the authors used the existing adaptive biometric recognition system. The update policy is exploited by the attacker by presenting fake biometric samples to the sensor. They have shown that the attack is possible in the case of multiple templates per client in PCA based method.In [10] the authors performed poisoning attack on traditional gradient method and on the proposed generative method. The proposed generative method used the concept of generative adversarial network to simulate the poisoning attack where the generator used an auto-encoder to generate the poisoned or fake images. The performance analyses of both the methods are based on loss and accuracy of the model. With the newly introduced generative attack method, the generation attack speed up the poisoned data generation rate (up to 239.38*) but lacked in model attack effectiveness (accuracy) as compared with gradient method.They even proposed an algorithm called Loss-based poisoning attack detection algorithm for detecting the occurrence of poisoning attack by checking the loss of the objective model.

3. Proposed Methodology

In the proposed scheme, firstly the face samples are being captured and resized in to size 92*112 and fed in to the neural network for training of the data. Specific patterns are being learned by the model for recognizing an individual. The patterns generated are very unique for every individual for their identity. The Generator module from GAN generates fakeimages. The Classification (Discriminator)module’s task is to identify the fakeness or genuiness of an image. The discriminator is so powerful in finding out the differences between the original and the fake images. However the generator is also too strong in fooling the discriminator to misclassify the result. Loss measures the performance of a classification model whose output is a probability value between 0 and 1.Accuracy of a model is calculated after the model parameters are learned and fixed and the test samples are fed in to model. Then the percentage of misclassification is calculated. The more accuracy, better the model is. The efficiency of model increases as the loss decreases.Following Figure 3 depicts the overall processing flow of the proposed methodology. For generation of fake images, Generative Adversarial network (GAN) is used to perform poisoning attack on training database.

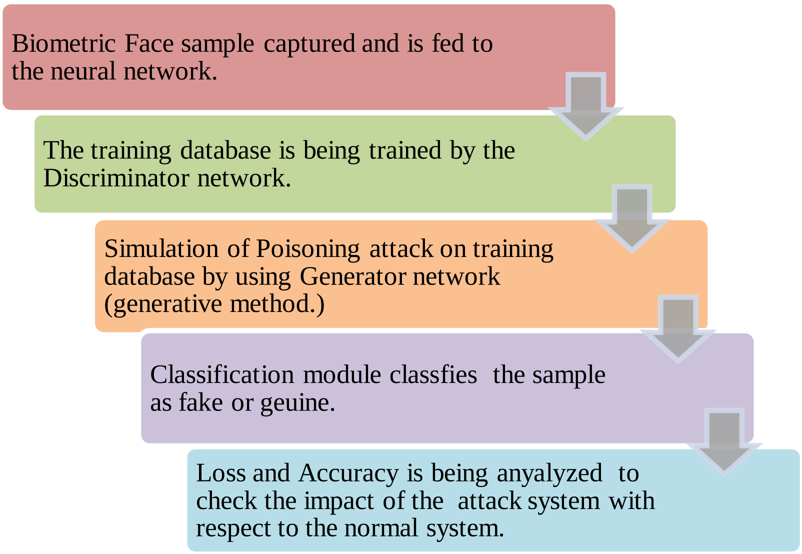

Biometric Face sample captured and is fed to the neural network.
The training database is being trained by the Discriminator network.
Simulation of Poisoning attack on training database by using Generator network (generative method.)
Classification module classfies the sample as fake or geuine.
Loss and Accuracy is being anyalyzed to check the impact of the attack system with respect to the normal system.

Figure 3: Overall processing flow of the proposed method.

The architecture of generative and discriminator networks are explained below, used in generative adversarial network:

a) Architecture of Discriminator Network:
Input the images in size of 92*112*3 which holds the raw pixel on an image where width is 92, height is 112 and the 3 represent the color channels [RGB].
Two Convolutional layers with 2*2 filter size(window size) and a stride of 2.

ReLU Activation layer is used after every convolutional layer.
Two Hidden layers are used with 2*2 filters and a stride of 2.
One regular fully connected layer is used, which calculates the output resulting in the form of [0,1] i.e. [fake, Genuine].

b) Architecture of Generative Network:
One regular fully connected layer with batch normalization and ReLU activation layer.
Two hidden De-convolutional layers with 2*2 filter size (window size) and a stride of 2.
ReLU activation layer is used after every de-convolutional layer.
One De-convolutional layer with 2*2 filter sizeto get the desired output shape.

The proposed methodology is based on three main categories i.e.

i. Training the discriminator network (Normal/Attack free system).
ii. The simulation of attack where generator generates fake images.
iii. Countering the attack where by increasing the number of epochs and gradient descent optimization technique.

To perform Poisoning attack on training database we use **Generative Adversarial Network** as show in Figure 4.

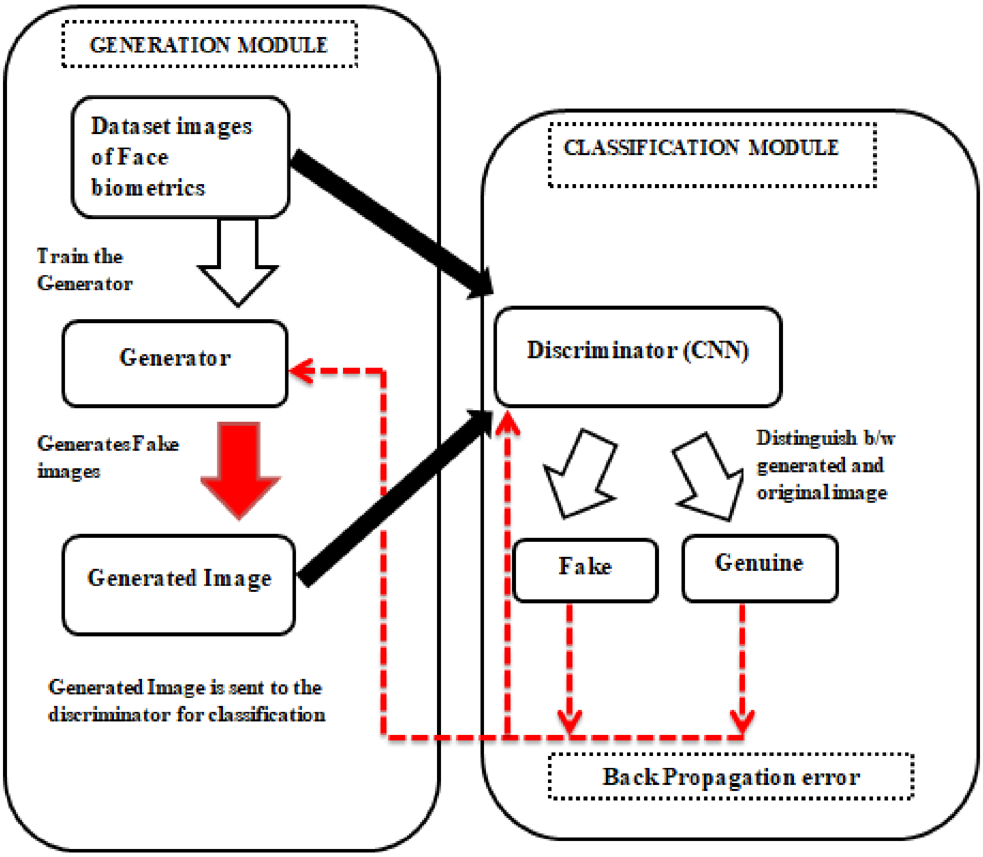


Figure 4: The architecture of the generative adversarial network for face biometric system.

Step1. Define generator and discriminator network structure as explained above.
Step2. ***Case1: Normal System***

Discriminator: Biometric face Samples of size 92*112 pixels of the training database are being trained by the discriminator network for n number of epochs and calculates the loss and accuracy of the training database. In this case, where only the discriminator plays a role of predicting the samples as genuine and fake according to the input images being trained .This case performs forward pass only i.e. the output of discriminator goes to the generator but doesn’t update as no back propagation error is send, as in the normal system which is free from an attack.

Step3. ***Case 2: Attack System***

Generator: In this case the following steps are followed for simulating an attack.

a. The generator take inputs from training database and tries to manipulate the data thus generating fake images which look as similar to the original database which is then trained on discriminator.
b. The discriminator should declare the image as fake.

Step4. The output of the discriminator network is sent back to the generator through the back propagation error and the generator is trained with the result of discriminator.

Step5. If the generator finds that he failed to fool the discriminator, and then the generator needs to be trained again with the database and output of the discriminator, so as to fool the discriminator with the fake image.

Step6. Repeat steps 3 to 5 until both the networks reach an equilibrium point.

Step7. Check if the generator successfully generates fake image then stop the procedure of training the networks.

The mathematical model of the generative adversarial network of face biometric system is expressed using equations (a) and (b).

$$\mathrm{Min}_G \mathrm{Max}_D F(D,G) \text{ ……... (a)}$$

$$F\ (D,G) = E_{\mathbf{x\varepsilon db}}[\log D(x)] + E_{\mathbf{x\varepsilon G}}[\log\ (1\text{-}D(G(x))]\text{……… (b)}$$

where notations are described in Table1.

Table1:-Notations described

Terms	Description
E	Expectation
db	Training Database
x	Sample from db
G	Generator Network
D	Discriminator Network.

In the function F(D,G) the first term ($E_{x\varepsilon db}$[log D(x)]) represents the entropy that the sample (x) from training database (db) passes through the Discriminator Network (D(x)). The Discriminator Network tries to maximize this to 1. The second term ($E_{x\varepsilon G}$ [log (1-D(G(x))]) represents the entropy that the sample (x) passes through the Generator Network (G(x)), which then generates a fake sample which is further passed through the discriminator to identify the fakeness. In this term, discriminator tries to maximize it to 0 (i.e. the log probability that the fake sample produced from Generator Network is fake and is equal to 0). So, the discriminator is trying to maximize the function F. However, the task of Generator Network is exactly opposite, i.e. it tries to minimize the function F so that the differentiation between real and fake image is minimum.

The separate losses for generative and discriminator networks are defined by the min-max game according to the equation (a) and (b) i.e.

$$\mathrm{Min}_G \mathrm{Max}_D F\ (D,G) = E_{\mathbf{x\varepsilon db}}[\log D(x)] + E_{\mathbf{x\varepsilon G}}[\log\ (1\text{-}D(G(x))]$$

Algorithm1: - Training Generative Adversarial Network

```
n =number of images
For number of training iterations do
    For every n images in training database($x^{(1)}$… $x^{(n)}$) should be trained by the discriminator network.
        For every fake n images ($x^{(1)}$… $x^{(n)}$) generated from the generator network.
        Update the discriminator network by increasing the gradients:
            $\frac{1}{n}\sum_{i=0}^{n}\left[\log D(x^i) + \log\left(1 - D\left(G(x^i)\right)\right)\right]$
End for
For every fake n images ($x^{(1)}$… $x^{(n)}$) generated from the generator network.
        Update the generator network by decreasing the gradients:
```

$$\frac{1}{n}\sum_{i=0}^{n}\left[\log\left(1-D\left(G(x^{i})\right)\right)\right]$$

End for

4. Experiment Results

Performance analysis is done on the basis of loss and Accuracy of the proposed model which is evaluated on the basis of following criteria:

1. Training the discriminator network(Attack free system).
2. The simulation of attack where generator generates fake images
3. Countering the attack by increasing the number of epochs and gradient descent optimization technique.

For these threecriteria's the results are evaluated at epoch 10, epoch 25 and epoch 50. In addition to this, graphs have been plotted between Training Loss vs. Testing Loss and Training Accuracy vs. Testing Accuracy.

A. Database Collection

The experiment analysis is performed on combination of two different database with same pixel resolution i.e. 92*112 pixels. In table 4, a description of each database along with parameters is mentioned. The First database is considered from AT&T laboratories [11]: ORL face database and the second database is considered as live database of face captured at college using Canon E750 D DSLR camera.

Table 2: Statistics of database

PARAMETERS	**DATABASE**	
	Live Database	**ORL face Database**
Number of users	96	40
Number of samples per user	10	10
Total Number of samples	960	400
Total number of samples in combined database	1360	
Total number of samples for Training Database	1090	
Total number of samples for Testing Database	270	

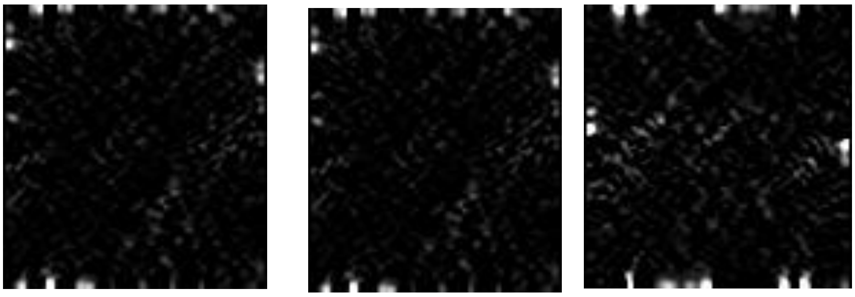
B. The Effectiveness of Poisoning attack

Case 1: Normal System

The evaluation of the proposed method has been calculated on the two basis: Normal system (Attack free system) and the Attack system (Generative Poisoning Attack).The Generative Adversarial Network (GAN) consists discriminator and generator networks, where the discriminator network's original recognition accuracy is 100% for all three epochs 10, 25 and 50 which indicates that the face biometric system is free from attack and the discriminator network correctly recognizes the genuine and fake samples.

Case 2: Attack System

The generator network manipulates the training database images by adding noise to the images and confuses the discriminator to misclassify the results which indicates that the system is attacked and the training accuracy has been degraded to 0.8844% in epoch 10, 0.8760% in epoch 25 &0.9000% in epoch 50.The Fake images are mentioned in figure 5(a-c) generated by the generator network.



(a) (b) (c)

Figure 5: Fake images of training database.

During the attack, a sudden increase of loss and breakdown of accuracy of the training database occurred which indicated a very heavy loss to the model. Figure6 demonstrates the loss of training and testing databases at different number of epochs 10, 25 and 50.

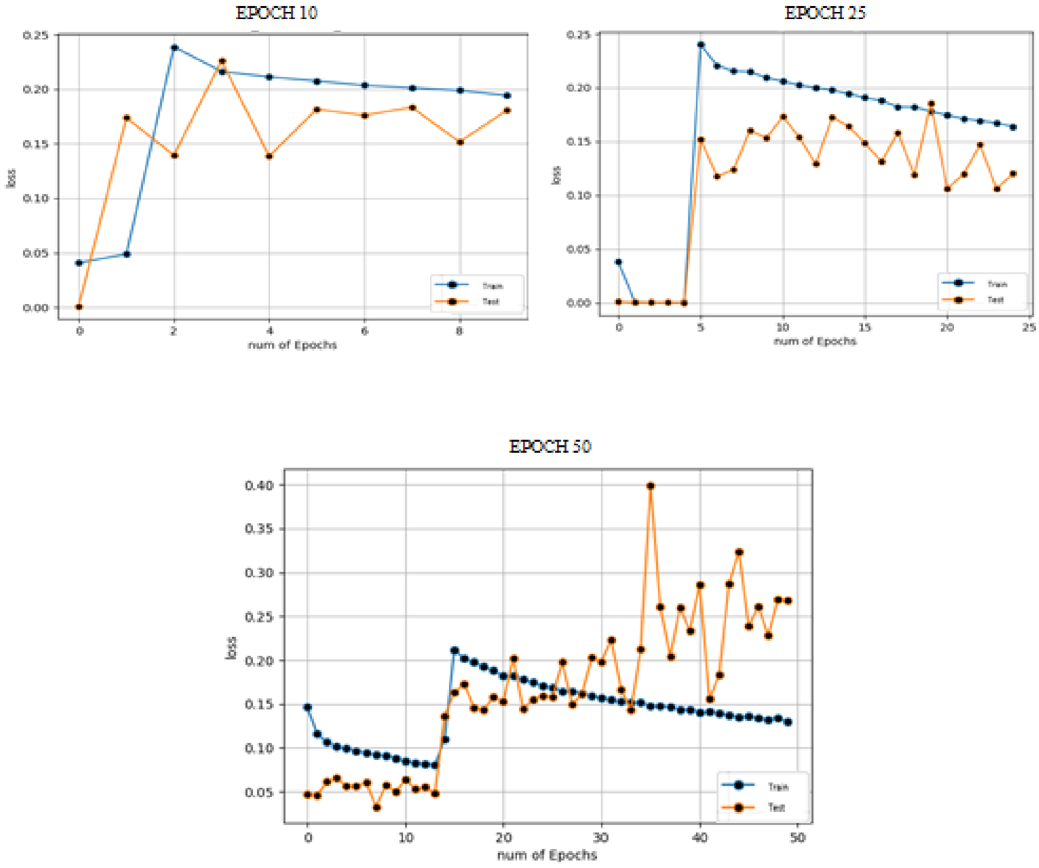


Figure 6: Detection of training and testing loss at different epochs (10, 25, and 50)

But as the epochs are increased, the data is being trained and validated simultaneously, giving feedback results to both the networks by back propagation error. **Back propagation** isbasically the **gradient descent optimization algorithm** which calculates the gradient of loss values and this result is fed back to both the networks for improvising their individual task.Figure 7 demonstrates the accuracy of training and testing databases at different number of epochs 10, 25 and 50. Thus this optimization algorithm counters the attack and makes the discriminator network too intelligent to recognize the fake samples and the accuracy increases as the optimization algorithm is activated.

Unexpectedly, as we increase the epochs the training and testing loss drops. At a certain period, the training loss continues to drop (as the network learns the data better) and the testing loss starts to rise resulting in overfitting problem evaluated in case of epoch 50 as seen in Figure 6.

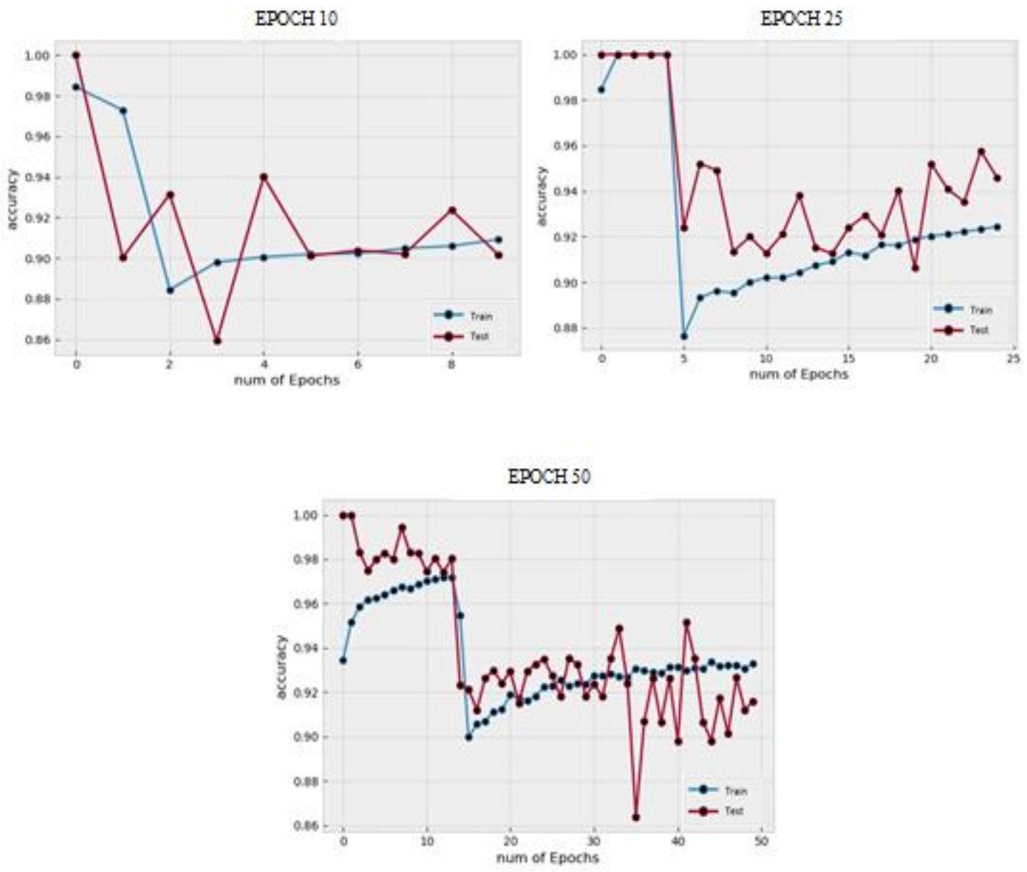


Figure 7: Detection of training and testingaccuracy at different epochs (10, 25, and 50)

5. CONCLUSION

In this work, Generative adversarial network is utilized to perform the poisoning attack on training database on face biometric samples and also utilize an attempt to improve the accuracy of the classification module. The fake images are being generated randomly on the training database and being trained by the discriminator network to improve the accuracy results as analyzed in the case of different epochs i.e. epoch 10, epoch 25 and epoch 50.The gradient descent optimization algorithm which calculates the gradient of loss values and this result is send back to both the networks for improvising their individual task, thus this optimization algorithm counters the attack and makes the discriminator network too intelligent to recognize the fake samples .Since the approach improved the results in some cases, but it still requires some more work toovercome the overfitting problem which occurred in epoch 50.

References

[1]. C. Roberts. "Biometric attack vectors and defences." *Computers & Security* 26, no. 1: 14-25(2007).

[2]. L.Huang, D. J. Anthony,N. Blaine, Benjamin IP Rubinstein, and J. D. Tygar. "Adversarial machine learning." In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43-58. ACM, 2011.

[3]. B. Biggio, G. Fumera, L. Didaci, P. Russu and F. Roli" Adversarial Biometric Recognition-A Review on biometric system security from the adversarial machine learning-perspective", IEEE Signal processing magazine,2015 IEEE

[4]. M. Barreno, B. Nelson, Russell Sears, Anthony D. Joseph, and J. Doug Tygar. "Can machine learning be secure?" In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pp. 16-25. ACM, 2006.

[5]. B. Biggio, G. Fumera, F. Roli, "Pattern Recognition systems under attack: Design Issues and research challenges" International Journal of Pattern Recognition and artificial intelligence, vol.28 no.7 2014.

[6]. B. Biggio,G. Fumera, and F. Roli. "Security evaluation of pattern classifiers under attack." *IEEE transactions on knowledge and data engineering* 26.4 (2014): 984-996

[7]. B. Biggio, B. Nelson, and P. Laskov"Poisoning Attack against SVM" ACM publisher: Omni press ICML'12 Proceedings of the 29th International Conference on International Conference on Machine Learning, 2012

[8]. B. Biggio, G. Fumera, L. Didaci, and F. Roli "Poisoning adaptive biometric systems" Publisher: Springer 417-425. 10.1007/978-3-642-34166-3_46 2012.

[9]. B. Biggio,G. Fumera, L. Didaci, and F. Roli "Poisoning attacks to compromise face templates." 2013 International Conference on Biometrics (ICB), IEEE, 2013.

[10]. C. Yang,Q.Wu, H. li, Y. Chen,"Generative Poisoning attack method against neural networks" *arXiv preprint arXiv: 1703.01340v1* (2017).

[11]. Olivetti Research Laboratory (ORL) database of faces, 2002. AT&T Laboratories Cambridge. www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.