Model Inversion Attack: Analysis under Gray-box Scenario on Deep Learning based Face Recognition System

Mahdi Khosravy*, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi
Media Integrated Communication Laboratory, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
[e-mails: dr.mahdi.khosravy@ieee.org, k-nakamura@comm.eng.osaka-u.ac.jp, hirose@nanase.comm.eng.osaka-u.ac.jp, naoko@comm.eng.osaka-u.ac.jp, babaguchi@comm.eng.osaka-u.ac.jp]
*Corresponding author: Mahdi Khosravy

Received August 24, 2020; revised October 26, 2020; accepted November 12, 2020; published March 31, 2021

Abstract

In a wide range of ML applications, the training data contains privacy-sensitive information that should be kept secure. Training the ML systems by privacy-sensitive data makes the ML model inherent to the data. As the structure of the model has been fine-tuned by training data, the model can be abused for accessing the data by the estimation in a reverse process called model inversion attack (MIA). Although, MIA has been applied to shallow neural network models of recognizers in literature and its threat in privacy violation has been approved, in the case of a deep learning (DL) model, its efficiency was under question. It was due to the complexity of a DL model structure, big number of DL model parameters, the huge size of training data, big number of registered users to a DL model and thereof big number of class labels. This research work first analyses the possibility of MIA on a deep learning model of a recognition system, namely a face recognizer. Second, despite the conventional MIA under the white box scenario of having partial access to the users' non-sensitive information in addition to the model structure, the MIA is implemented on a deep face recognition system by just having the model structure and parameters but not any user information. In this aspect, it is under a semi-white box scenario or in other words a gray-box scenario. The experimental results in targeting five registered users of a CNN-based face recognition system approve the possibility of regeneration of users' face images even for a deep model by MIA under a gray box scenario. Although, for some images the evaluation recognition score is low and the generated images are not easily recognizable, but for some other images the score is high and facial features of the targeted identities are observable. The objective and subjective evaluations demonstrate that privacy cyber-attack by MIA on a deep recognition system not only is feasible but also is a serious threat with increasing alert state in the future as there is considerable potential for integration more advanced ML techniques to MIA.

Keywords: Model Inversion Attack, Deep Learning, Face Recognition System, Media Clone

http://doi.org/10.3837/tiis.2021.03.015 **ISSN : 1976-7277**

1. Introduction

Deep Learning (DL) [1, 2] has drawn the attention of research and development centers of Machine Learning (ML) due to its great ability to learn the features of huge databases, thereafter recognizing the data categories despite the vast diversity of data. Amongst the DL techniques, Convolutional Neural Networks (CNN) [3, 4] have been having the greatest influence on practical machine learning. Due to high capability of DL, a variety of complicated tasks have been carried out by the help of DL-based techniques like face classification by using deep belief networks [5], human sentiment classification [6], human consumption behavior forcasting [7], human activity recognition [8], hand gesture recognition [9], plannar pressure segmentation [10], driving behavior recognition [11], finally even in the case of recent COVID-19 detection from chest X-ray images [12-14], etc. One of the CNN-based systems widely used is deep face recognition system which due to its implementation on cloud level it can be targeted by privacy cyber-attackers. ML systems in general and DL systems in the special case is very much inherent to the data which has been used for their training. It means, although the training data is not available while the systems are in use, since the system structure and parameters are well-tuned by the training data, they can be traced through the model structure. This kind of cyber-attack access to the training data has become feasible on a shallow network by the methodology called Model Inversion Attack (MIA) [15]. MIA reversely goes through the model from its output and by using the class labels of a corresponding registered identity to the system, tries to estimate its corresponding input to the system. The estimated data can be a generated version of privacy-sensitive data for possible malicious use on the web. Therefore, MIA may violate the user privacy. An example of a privacy attack is an attack on a face recognition system wherein the attacker may generate the face image of one of the registered identities to the face recognition system and gets access to his/her face image for malicious purposes. **Fig. 1** illustrates the general concept of privacy cyber-attack on a face recognition system.

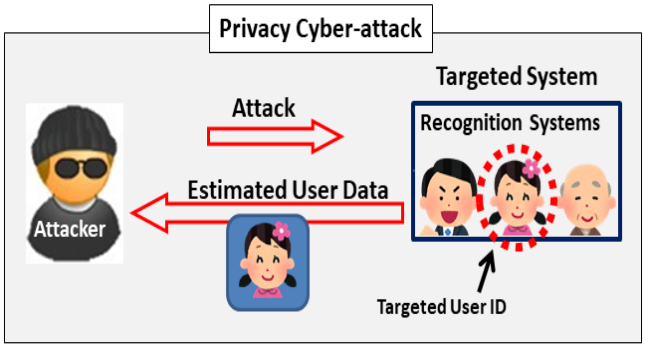

Fig. 1. Privacy cyber-attack to a face recognition system

Ref. [15] demonstrates the feasibility of MIA on a shallow ML system in a white-box scenario where there is partial access to the users' information, model structure, and parameters. In the case a DL-based recognition system, the feasibility of such privacy attack is under question due to deep learning complexities in model structure, a vast number of parameters, and huge and diverse training data sets as in the case of a deep face recognizer the training data may include millions of face images, and thousands of class labels. This research work targets a deep CNN-based face recognition system user identities by MIA under a scenario without any access to any information of the users but having access to the model

structure namely a semi-white scenario or as called here, a '*gray-box scenario*' of MIA.

Section 2 presents a review of the related MIA works, and subsequently, Section 3 illustrates the MIA under the gray-box scenario. Section 4 describes the experimental setup for both deep face recognizers and MIA, and analyses the results. Section 5 presents the future scope of the research, and finally Section 6 gives concluding remarks on gray-box MIA on a deep face recognition system.

2. Related Works

Due to the widespread application of ML systems on the cloud level, they are vulnerable to different types of attacks and malicious access to the users' information. By the presented categories of the attacks as (i) causative and (ii) exploratory attacks presented by Huang et. al [16], MIA is considered as an exploratory attack. Despite a causative attack, it does not interfere with the training of the ML system, but it explores the ML system structure and follows the trace of training data in the model structure for accessing the user information. ML model is very usable for the attacker in an exploratory attack as stealing the model has been suggested in research work on the attack by Ref. [17] which constructs a copy of the model. The copy of the model can be used for a variety of unauthorized applications like accessing sensitive information, applying unlimited trials on the model without any fee, etc. MIA is an exploratory attack by having the ML model.

MIA was initially introduced by Fredrikson et al. [18]. In their work, the model structure and parameters were used to estimate the user information corresponding to a class label from the available partial information of the same user. The partial information of the user is claimed in their work that can be available due to lack of sensitivity while the sensitive information is not available, as in the case of a face recognition system, the parts of the face like eye, nose, lips are privacy sensitive and the rest are non-sensitive in term of privacy. Initially, they targeted a linear regression model [18]. Thereafter, they implemented the MIA on a neural networks and decision tree [15]. Although MIA was in use, it was without a formal theoretical until the work presented by Wu et al. [19] wherein they presented a methodology for precise detection of the MIA on a system by a game-based technique. Their technique is inspired by the two-worlds concept in cryptography.

Another scenario is the black-box scenario wherein the attacker neither has any access to the machine learning model structure, nor any information of the user is provided. Such a scenario has been argued as the most difficult situation that an adversary can face [20]. This scenario has been left as an open problem. Aïvodji et al. [21] have introduced a generative adversarial model inversion under the black-box scenario, and they have achieved considerable results even against a deep model. Although their approach does not have any information regarding the model system, it is based on try and error based on assumptions of target model structure as well for the data structure. As their methodology involves training a GAN, and it is based on a sequence of assumptions for both the model and original data characteristics, it requires a costly process to perform. Their work has the merit of being one of very few works under the black-box scenario.

The white-box scenario may not be interesting enough compared to the black-box scenario in the view of some researchers, but it is more practical and realistic. Because, first, assuming once MIA performs well under black-box scenario, it would be costly and time-consuming, and its process can be suspicious by the severs and detectable. Second, white-box or gray-box senario have their own value since having a hundred percent secretive model structure which

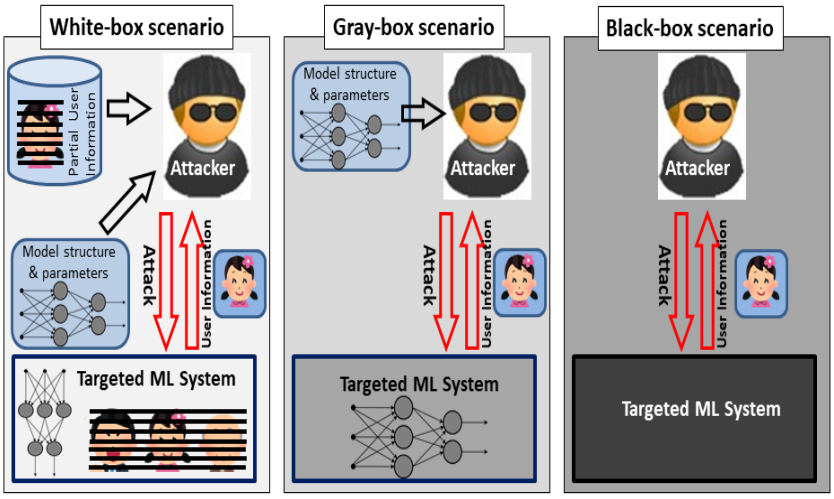

Fig. 2. From left to right: white-box scenario, gray-box scenario, and black-box scenario of MIA

requires a black-box scenario is not realistic and it would not be relied as a security solution as the Kerckhoff's principle [22] states:

"the system security should not rely on an unrealistic level of secrecy"

The first approach to a semi-white scenario of MIA was by Hidano et al. [23, 24] as they proposed a framework wherein the non-sensitive user information was not necessary. Since their approach covers the white-box scenario of Fredrikson MIA, they called generalized paradigm of MIA. In the sense of performing under a gray-box scenario, our work presented in this paper is similar to Hidoano's work, but the present work is on a CNN-based deep model. Although the MIA has not been applied to a deep model, but there are some works using the deep generative models for MIA as Ref. [25] wherein initially, a generative adversarial network (GAN) is trained by using the non-sensitive face parts of the targeted user, then the trained GAN generates the full face of the targeted user. The main drawback of their technique is the need for an enormous volume of non-sensitive data for training GAN. Finally, Khosravy et al suggest the seed image generated by a pretrained GAN model for initialization of MIA on deep face recognizer [26]. **Fig. 2** illustrates the black-box, white-box and gray-box scenarios of MIA. **Table 1** gives a brief review to the above mentieond related works.

Table 1. Related works to MIA

Literature	MIA contribution
[15]	MIA in privacy violation in pharmacogenetics by Fredrikson et al.
[18]	Initial introduciton of MIA by Fredrikson et al.
[19]	Stealing machine learning models via prediction apis by Tramèr et al.
[20]	Discussion on black-box scenario
[21]	Introduced a generative adversarial model inversion under the black-box scenario by Aïvodji et al.
[23, 24]	Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes by Hidano et al.
[25]	Using generative models in MIA on deep neural networks by Zhang et al.
[26]	Generating and implementing optimum seed image by pre-trained GAN for initialization of MIA process on DL based recognition system by Khosravy et al.

3. Gray-box MIA on Deep Face Recognizer

This section represents the implementation of MIA on a deep face recognition system in a gray-box scenario. MIA aims to generate a face image of a targeted user identity registered to the face recognizers. As the scenario of MIA implementation is gray, the information of the recognition model related to its structure and parameters are already available to the attacker, but there is not any access to any type of user information, except their class labels. Thue, the MIA should generate the face image of the targeted user identity just through its corresponding confidence information and the model structure and parameters and there is not any initializing available image with partial information of the user for beginning the process. Thereof the process should be initialized by a seed image. The seed image acquired in this work is an average of an already available face image database without any connection the system under attack. **Fig. 3** illustrates the MIA on a recognition system under the gray-box scenario as formally explained in the sequence.

This work applies the MIA on a face recognizer R whose structure and parameters are setup by CNN model of deep learning. The input to the system R is a face image x with dimensions matching with the input layer of the system model. As the system R receive the input x, it generates a corresponding output which determines to which of n user identities registered to the system it belongs. The corresponding class label is determined at the output by a vector of length n as follows:

$$\begin{aligned}\vec{y} &= R(x) \\ &= [y_1 ... y_2] \in [0,1]^n,\end{aligned} \tag{1}$$

wherein each element of y is a confidence score corresponding to a user identity. The result of recognition will be the one shows the highest confidence. As their indice corresponds to the indices of the registered users, mathematically the detected index of recognized user identity is as follows:

$$i^* = \max_{y_i \in R(x)} y_i, \tag{2}$$

As MIA under-work scenario has access to the class labels of the targeted identities, it acquires a corresponding one-hot vector as an approximation for the desired output of the system for the corresponding targeted identity of index i as follows:

$$R(x_i) \approx \vec{\tilde{y}}_i \in \{0,1\}^n, \tag{3}$$

$$\vec{\tilde{y}}_i(j) = 1, \quad \text{if} \quad j = i \tag{4}$$

$$\vec{\tilde{y}}_i(j) = 0, \quad \text{if} \quad j \neq i \tag{5}$$

wherein the corresponding element to the targeted user identity is 1 and the rest are 0. As x is a true image of the targeted identity, MIA target is to generate a clone image of same identity as $\hat{x}$. This generation in MIA is iterative starting from the seed image. By having the output of the system model $R(\hat{x})$ and $\vec{\tilde{y}}$ the approximate of the desired output a loss function is made as a function of the difference between them. The difference can be a least square

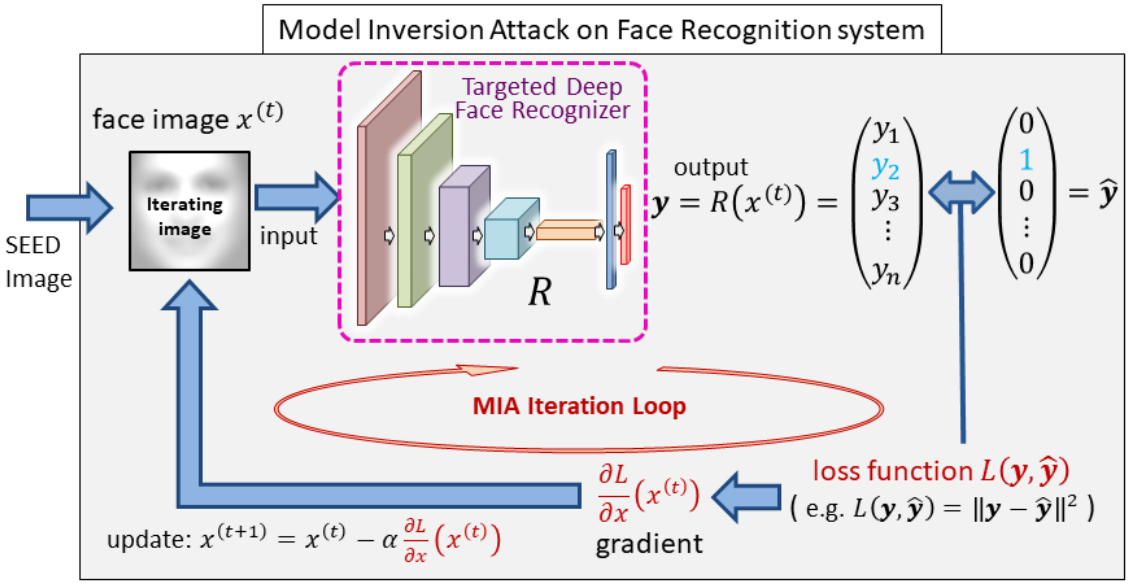

Fig. 3. Model Inversion Attack (MIA) on a face recognition system under gray-box scenario

measure of their difference or an Infomax measure; $\mathrm{L}(R(\hat{x}), \vec{\tilde{y}})$. Having the loss function MIA performs in the frame of the following optimization problem:

$$\min_{\hat{x}} L(R(\hat{x}), \vec{y}) \qquad \hat{x} \in \mathcal{X} \tag{6}$$

where $\mathcal{X}$ is the image space. The optimization problem of (6) has contineous nature and its solvable through a gradient descent iterative process:

$$\hat{x}(i+1) = \hat{x}(i) - \alpha \frac{\partial}{\partial \hat{x}} L(\hat{x}(i)), \tag{7}$$

where α is updating rate, and i in this equation is iteration step. Since attacker has the structure of the model and its parameters, $\frac{\partial}{\partial \hat{x}} L(\hat{x})$ is obtainable as given below:

$$\frac{\partial}{\partial \hat{x}} L(\hat{x}) = \frac{\partial}{\partial \vec{y}} L(\vec{y}) \frac{\partial}{\partial \hat{x}} \vec{y}(\hat{x}) \tag{8}$$

$$= \frac{\partial}{\partial \vec{y}} L(\vec{y}) \frac{\partial}{\partial \hat{x}} R(\hat{x}) \tag{9}$$

Due to adjustability of loss function L by the attacker its gradient with respect to the output of the model is available to the attacker too. Also, since the attacker has accessed the model of the system R, can obtain the $\frac{\partial}{\partial \hat{x}} R(\hat{x})$ too. The process of Eq. (7) iteratively modifies $\hat{x}$ until it converge to an estimation of a clone of x. Mathematically,

$$\hat{x} = \lim_{i \to \infty} \hat{x}(i) \tag{10}$$

The seed image $\hat{x}(0)$ initializes the iteration process. In order to keep close to the space of face images as a subspace in the image space, the suggested MIA approach takes an average of the a face image data as the seed image.

4. Experimental Setup

This section illustrates the setup of the experiment of MIA under gray-box scenario on CNN-based face recognition system. The setup is explained in three parts as below:

- Deep Face recognition training setup
- Evaluation setup of face recogntion systems
- Gray-box MIA setup

Each of them has been explained in the sequence. For the training the face recognition systems as well as for making the seed image to initialize the MIA, we have deployed the face database of VGGFace2 [27]. It should be noted that different sections of the same database has been deployed for different systems used in the experiment as well as for making a seed image. VGGFace2 is a rich database of face images with above hree millions of images. VGGFace2 charachterics are as (i) age diversity, (ii) gesture diversity, (iii) race diversity, (iv) variety of lighting condition, etc.

4.1 Deep Face Recognition Model Structure

The deep learning models of face recongnition systems acquired in this reseach work copmrises the following structure:

- Four convolutional layers followed by max pooling layers, namely as L1, L2, L3, and L4. The kernel size in these layers is respecitvely 5×5, 5×5, 3×3 and 3×3. The number of channels in these layers are respectily 8, 16, 32 and 64, and each channel is for the layers L1 to L4 are respectively 128×128, 64×64, 32×32 and 16×16.
- Three fully connected layers, namely as L5, L6 and L7. The two layers of L5 and L6 are of 1024 neurons. The length of the L7 is equal to the number of class labels 2041 in the case of the targeted recongnition system and 2041+5 in the case of the evaluation recongtion system. The 5 more are related the targeted identity which are in common between two recognition systems.

Fig. 4 illustrates the the structure of the under work deep face recognition systems.

4.2 Deep Face Recognition System Training Setup

In this experiment two deep CNN-based face recognition systems has been used; target system R_T and the evaluation system R_E. The first face recognition system R_T is targeted by MIA. Already it includes the targeted identities which means the targeted identities are involved in the system class labels. The second deep face recognition system R_E is used for the evaluation of the generated cloned face images by MIA. Indeed both systems involved the targeted identities, and R_E as a separate system is used for testing the recognizability of the generated clone images apart from the targeted model that is R_T. For the training of R_T and R_E, the first and second quarters of VGGFace2 have been used, respectively. Theses two quarters are without any intersection except the training images of the targeted identities. As the VGGFace2 includes three millions of images, for training of each face recognition system more than $800{,}000$ face images are used which covers more than 2000 identities for each.

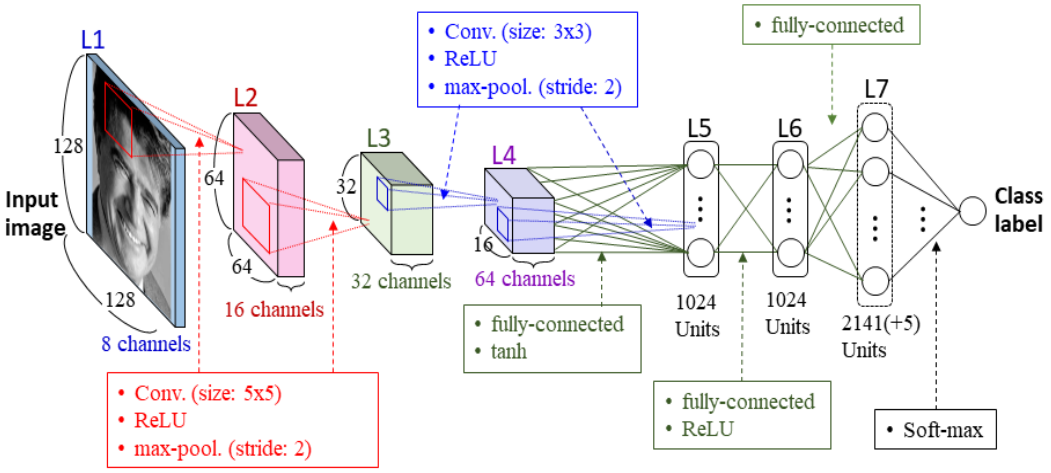

Fig. 4. Deep learning based face recognition model structure used in this research work for Model Inversion Attack under Gray Scenario

4.3 Evaluation Setup of Face Recognition Systems

For evaluation of efficiency of MIA on deep face recognition system, two CNN-based face recognition systems of the same structure have been required as R_T and R_E; targeted and evaluation face recognizers. But as mentioned in Section 4.1, more than 2000 identities are registered to each of them. They are trained by different set of images belonging to different user identities which results in despite of using the same structure but having different parameters. The targeted user identities are the only common part of their training data set, which comparing to 2000 identities, having five common identity does not bring to them any considerable similarity in parameters of the model. After targeting one face recognizer, the generated clone images are evaluated by the other one, since already the same targeted identities are registered to the evaluation system too, and if the clone image is mainly representing its corresponding identity rather than being affected by the system which is used for its generation, it should equivalently perform on another recognition system which the

same identity is registered too. On other hand, since already the clone image has gone through iteration process using the target model, its efficiency score by the target model is naturally high and its evaluation by the same model will result in a same good score due to the model not exactly due to be representative of the targeted identity. For, evaluation, we have targeted five identities which are the only common ones registered to both models. Some of the face images of these targeted identities have been depicted in **Fig. 5**.

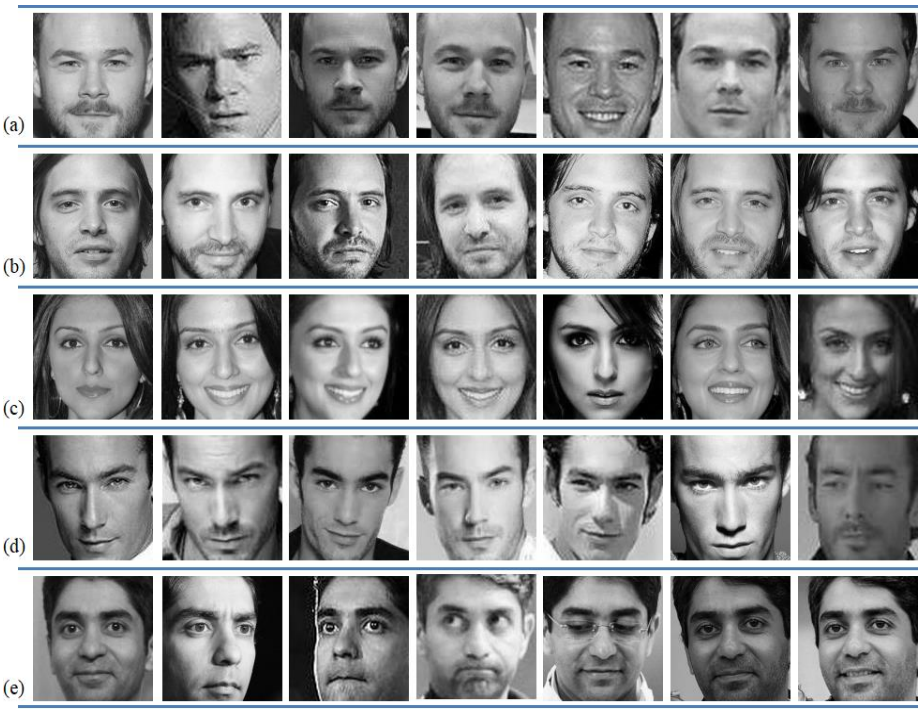

Fig. 5. The five personal registered IDs targeted in this research work namely as (a) ID-1, (b) ID-2, (c) ID-3, (d) ID-4, and (e) ID-5

4.4 Gray-box MIA Setup

The setup of the proposed gray-box MIA comprises two parts (i) setting the updating rate α of the iterations, (ii) preparing a proper seed image for initialization. Taking α very small results in very slow MIA process and even sometimes without any considerable change in the seed image, and taking the α as a big value does not mean having a fast process of convergence to a clone image. But it can results in very high level noisy results that the resultant image not even be recognizable as a face image. Therefore, we took α as a moderate value to be a trade-off between speed and accuracy. As we observed in the case of our practice $\alpha = 0.05$ is a proper choice. For the seed image we have used a number of images taken from the second half of the VGGFace2 database which is used in training of neither the targeted nor the evaluation system. In our setup the seed image is made of average of 256 images.

5. Results and Discussion

After having setting up the experiment, having available both the targeted deep face recognition system and the evaluation one, as well as having the seed image required for MIA prepared and ready, we have targeted five registered user identities and generated their

corresponding clone images by the proposed gray-box scenario MIA. In order to have a more rigorous evaluation for each of the targeted identities hundred times MIA was applied and hundred clone images were generated. **Fig. 6-10** respectively show twenty five generated clone images for each of the identities ID-1, ID-2, ID-3, ID-4, and ID-5 depicted in **Fig. 4**.

As can be seen in all the above-mentioned figures there is a high level of noise and not a crystal clarity in generated clone images which considering the condition of gray-scenario and the deep model of the system, it could be expected. Seeing the good side of the results, by a quick view to all the generated images for all the five targeted identities, the following advantages of the proposed technique is observable:

- The similarity of the clone images generated for one targeted identity to each other, as it can be said all of them belongs to the same target.
- The general difference between the images generated for one targeted identity from the ones generated for the other one.
- Similarity of the features of the generated clone images to the original image as some common features are observable. It is in the way if an observer has already well seen the face images of all the five targets can distinguish the generated clone images to which one blongs. This is called recognizability and similarity between the clone and target real identity.

As can be seen in **Fig. 5**, clearly the face features of nose, eyes, eyebrow of the targeted image are observable in all the depicted clone images. Similarly, the same features are strongly observable in the case of **Fig. 9**. In the case of **Fig. 7**, the similarity is much less observable, and the generated image show much less consistency with the targeted identity, and finally in the case of **Fig. 8** and **Fig. 6** the similarity of features is limited to the nose and partially in eyes. In addition to this visual analysis, we have applied some objective and subjective numerical evaluations which are presented in the sequence.

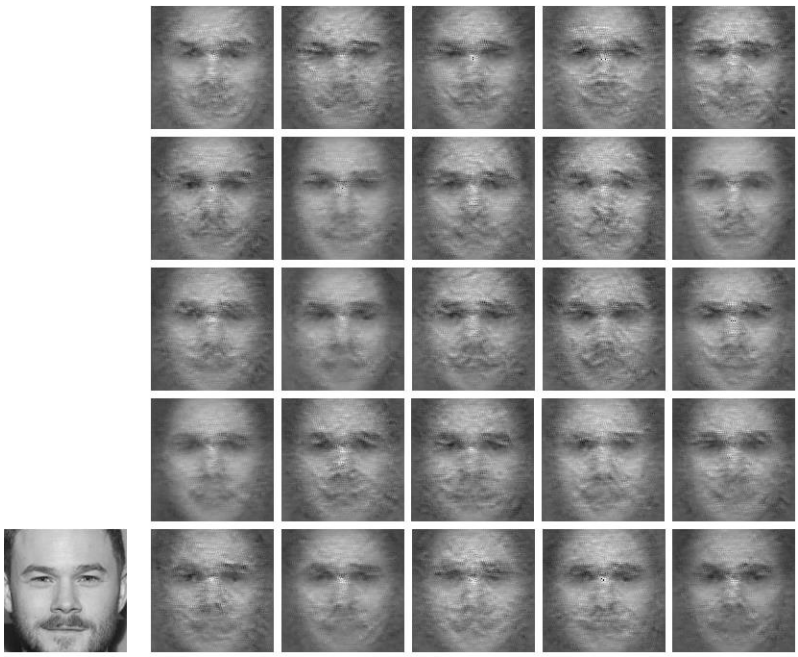

Fig. 6. A real face image of ID-1 (bottom-left) and some of the generated clone images by MIA for ID-1

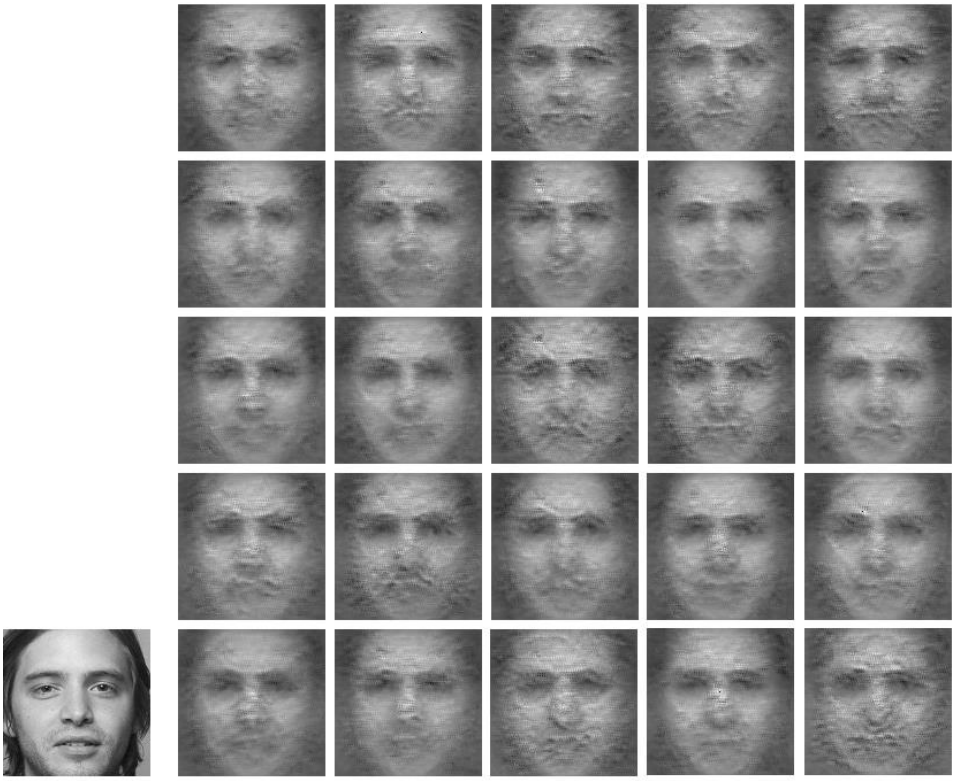

Fig. 7. A real face image of ID-2 (bottom-left) and some of the generated clone images by MIA for ID-2

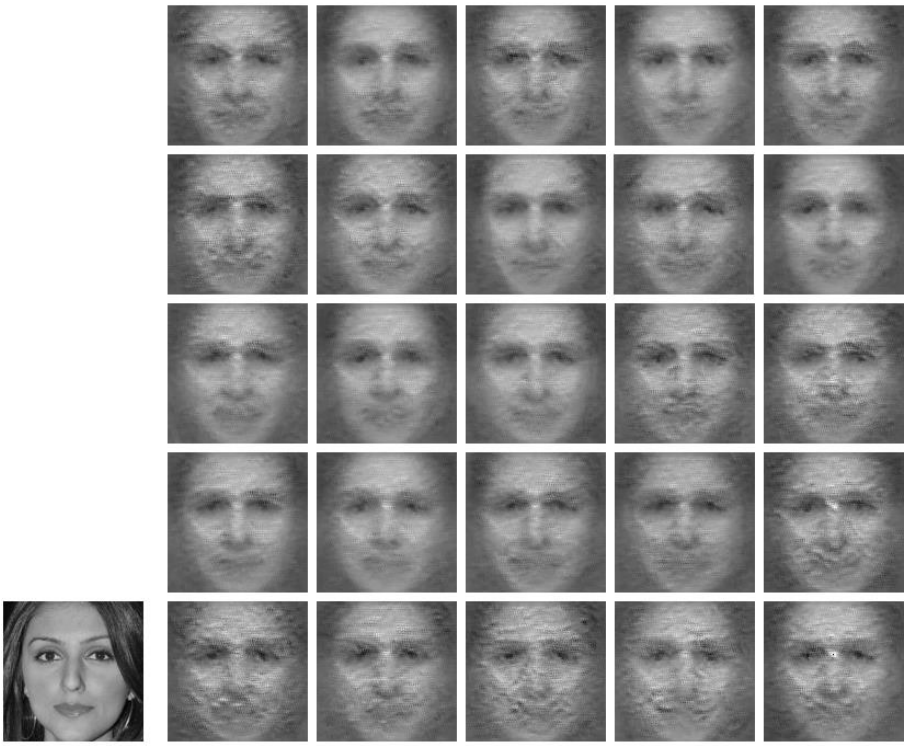

Fig. 8. A real face image of ID-3 (bottom-left) some of the generated clone images by MIA for ID-3

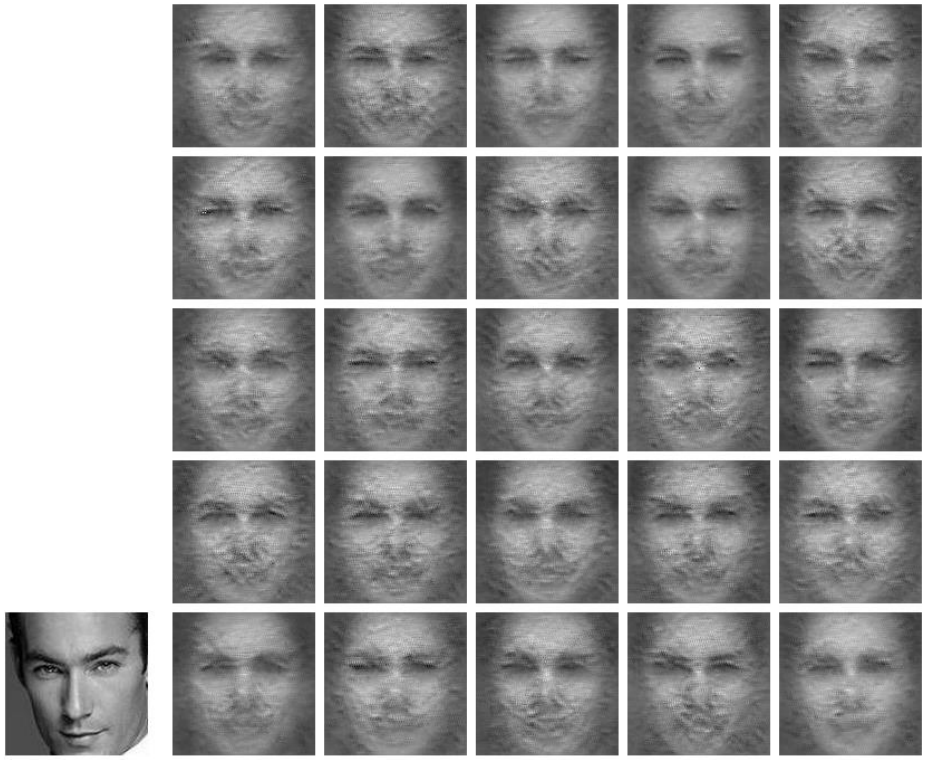

Fig. 9. A real face image of ID-4 (bottom-left) and some of the generated clone images by MIA for ID-4

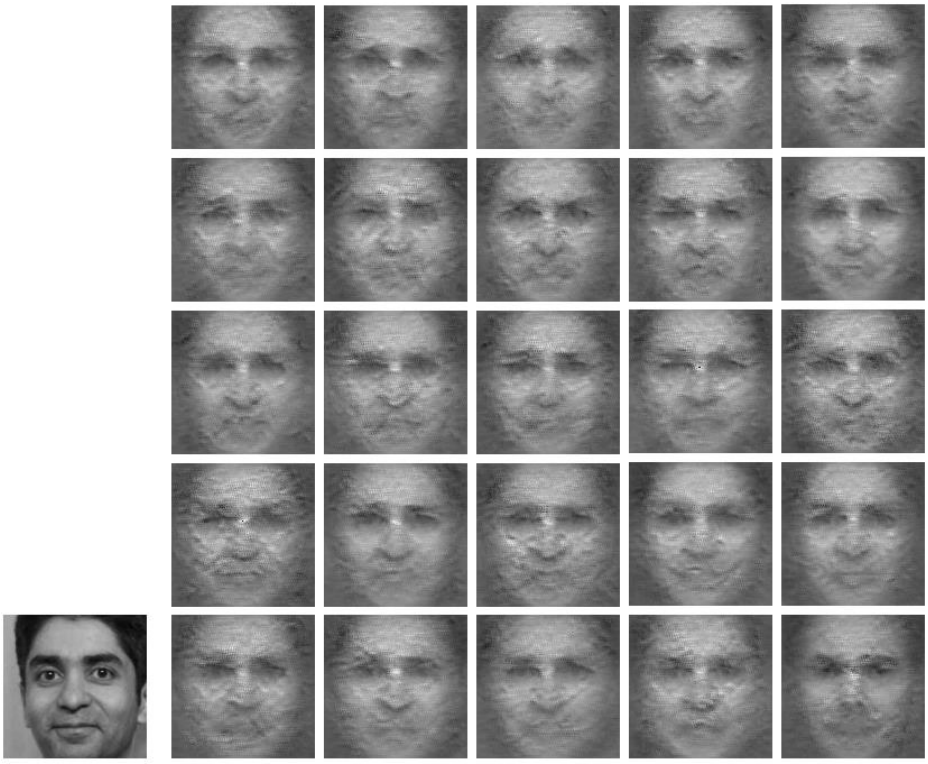

Fig. 10. A real face image of ID-5 (bottom-left) and some of the related generated clone images by MIA

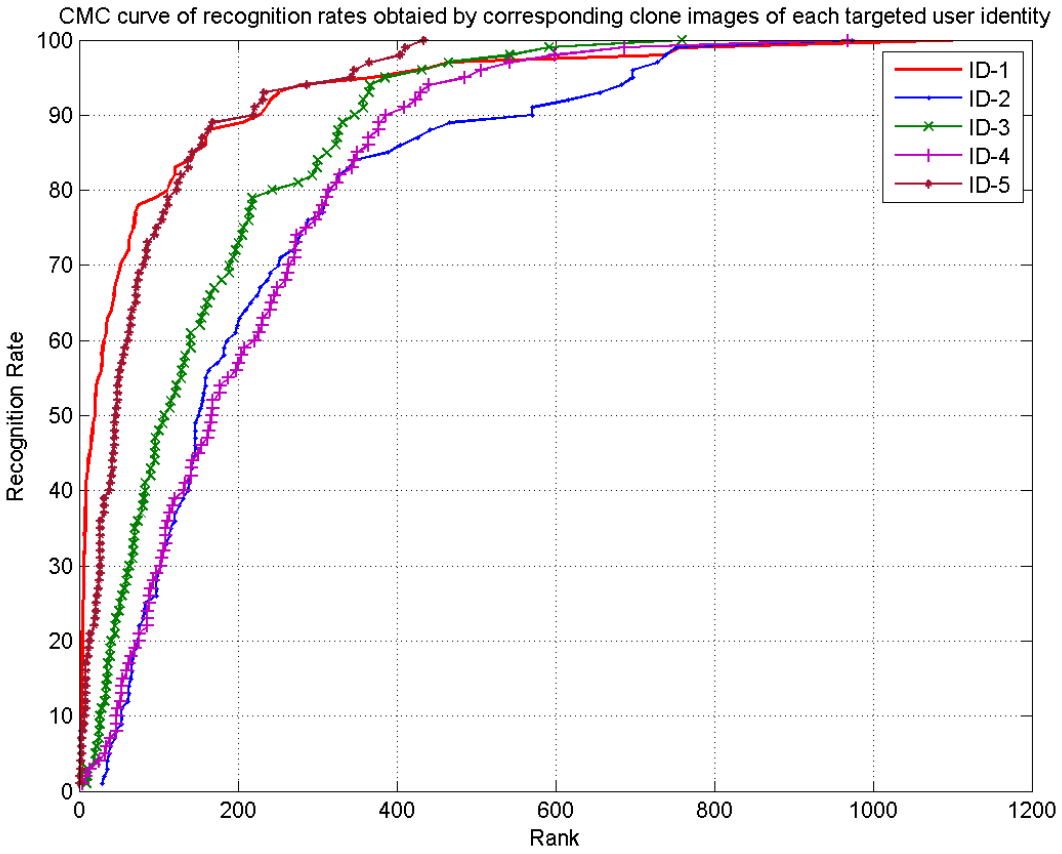

Fig. 11. CMC curve of recognition rates obtained by the corresponding generated clone images for each targeted user identities ID-1, ID-2, ID-3, ID-4, and ID-5

As a robust analysis on the generated clone images for each of the targeted identities, as a hundred images for each of the are generated, the images are fed to the evaluation face recognition system and rank and score according to the training information of the evaluation model is given to each of them. For each hundred clone images corresponding to one of the targeted identities, a cumulative matching characteristic curve (CMC) is generated. The CMC cure is an accurate measure of the recognition precision by the generated clone for a user ID. As the recognition rate is higher the CMC curve will resemble a more rectangular shape with a sharper corner on up left of the curve and the curves go up with a sharper ramp. As can be seen in **Fig. 11**, in general, the CMC curves shows an acceptable level of recognition rate for the generated clone images. In the case of ID-1 and ID-5 as we expected from the visual observation the CMC curves also show a higher level of recognizability. Similarly, we can see a lower rate of recognizability by the MIA generated clone images for the ID-2 and ID-4 which matches the visual observations. What CMC curves convey is acceptable and even high recognizability by the generated clone images for some identities.

To have another visual evaluation besides the available scores and ranks for the generated clones, we have demonstrated the highest rank clone images for each ID-1 to ID-5 as suggested by the evaluation face recognition system score and rank and depicted besides its target identity face image in **Fig. 12**. Besides, **Table 2** demonstrates the recognition score by the target face recognition system, the recognition score and the rank by the evaluation face

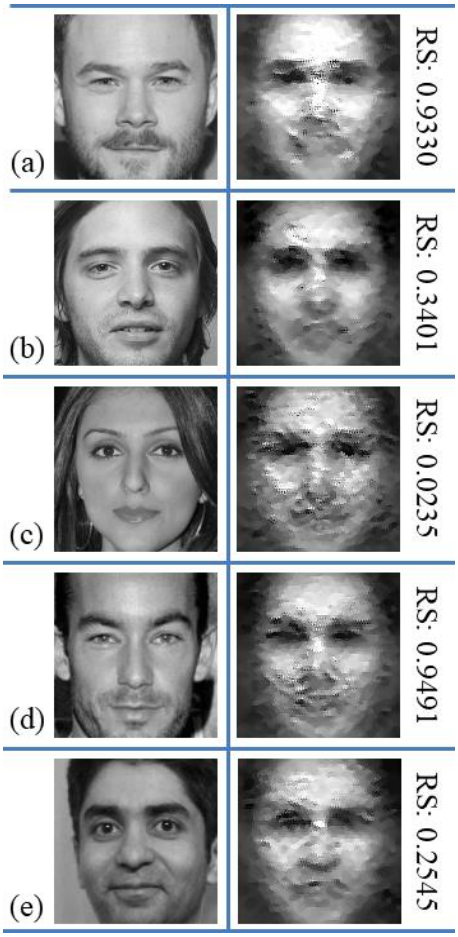

Fig. 12. The highest rank clone images for (a) ID-1, (b) ID-2, (c) ID-3, (d) ID-4, and (e) ID-5 according to the score and rank given by the evaluation face recognition system

recognition system for each of the clone images depicted in **Fig. 12**. As can be seen in the special cases where the results are cherry-picked, the similarity between the clones and the targeted identities is considerable as in the case of ID-1 and ID-4 the similarity is very high. However, in some cases, even the well-chosen image does not show a good similarity as in the case of ID-3. What we can conclude from **Fig. 12** is that a secondary evaluative system can assist a secondary choice amongst the generated clone image for a highly effective clone. This is a potential for the future focus of this research for further improvement of the results of MIA.

Table 2. Evaluation scores and the rank of clone images in **Fig. 12**

Targeted Identity	Recognition score by the Targeted system	Recognition score by the Evaluation system	Rank by the Evaluation system
ID-1	0.9900	**0.9330**	1
ID-2	0.9844	0.3401	1
ID-3	0.9374	0.0235	12
ID-4	0.9791	**0.9491**	1
ID-5	0.9834	0.2545	1

Since the ultimate goal of an attacker on a face recognition system is reveling the identity of a registered user in a malicious way, the recognizability of the clone images by a human observer is of importance. To evaluate the clone images in this regard, we have applied a subjective evaluation to all generated images in two general evaluation by a group of human observers as (i) subjective similarity to the targeted identities, (ii) subjective recognition accuracy. To perform these evaluations, for each of them a quiz by involving clone images and targeted face images were designed and handed to more than fifty observers for grading the similarities of the clone image to their corresponding target as well for recognizing that each clone image belongs to which target identies. Interestingly, the total results averaged from all the quiz attendees show 33% accuracy in subjective recognition of the generated clone image and 45% similarity to their corresponding target identities (**Table 3**). Considering the simplicity of the deployed MIA which does not use any deep generative adversary, as well considering the complexity of the deep model under attack, the results indicate the increasing threat of the MIA even for the case of a deep recognition system. As in future, the deep generative models would be introduced to the same approach and as the search space would be transferred to the face image data space rather than the general data space of images, due to narrowing the search space the results would be more accurate, more precise, more natural and the most important more recognizable.

Table 3. Subjective evaluation Scores

Subjective naturalness	Subjective Similarity	Subjective Recognition Accuracy
20%	45%	33%

6. Future Scope

In the future work on the sequence of this research work, we will try to increase the MIA efficiency even more on a deep face recognition systems by integrating the deep generative models to the MIA loop. A well trained deep generative model can effectively transfer the search domain from the face image domain to the latent variable domain. Search in latent variable domain would be faster and more effective with high expectation of converging to an accurate generated by MIA process which is can be real privacy leakage of a registered user to the system. Another future direction of the current research work is to reduce the noise-like effect of MIA at each iteration by implementing pre-processing and post-processing techniques like non-linear morphological filters [28, 29] and image adaptation [30]. Also, another open aspect of this research work is initialization of the suggested MIA by a proper seed image. The seed image can be optimally generated by using meta-heuristic optimization techniques [31] like genetics algorithms and variants [32, 33], and the recent Mendelian evolutionary optimization [34-35]. Also in the case of future integration of the deep generative models to the MIA process, the structure and the length of the latent variable of the model can be optimized by the using the above mentioned optimization technique. This research orientation will lead a novel research orientation as evolutionary model inversion attack.

7. Conclusion

The privacy of the user information registered to a pattern recognition machine learning system that is used on the cloud level can be subjected to different types of cyber-attacks. One of them is the model inversion attack (MIA) which targets accessing the features of the

training data of the machine learning model. In a white-box scenario, MIA accesses this privacy-sensitive information via the model structure and parameters besides some partial information of the users. This research work applies MIA on a face recognition system which is structured and trained based on deep learning under a gray-box scenario wherein just the model structure and parameters are available for the attacker but not any user information. This scenario is called gray-box scenario since it is not as blind as a black-box scenario also not as open as a white-box scenario. Despite the depth of the model, as the objective and subjective evaluations of the generated clone images indicate, the proposed technique of MIA demonstrates a considerable level of efficiency in both terms of similarity of the generated clone images to the targeted identities and recognizability the targeted identities by their corresponding clones. In a nutshell, this research work promotes the alert state of MIA as an increasing threat of a privacy cyber-attack even for the case of a deep learning based face recognition system.

Acknowledgement

This work was supported by Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (JSPS KAKENHI) under Grant JP16H06302 and Grant JP17K00235.

References

[1] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp.1527-1554, 2006. Article (CrossRef Link)

[2] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3-4, pp. 197-387, 2004. Article (CrossRef Link)

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 60, no. 6, 2012. Article (CrossRef Link)

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708, 2017. Article (CrossRef Link)

[5] N. Bouchra, A. Aouatif, N. Mohammed, and H. Nabil, "Deep belief network and auto-encoder for face classification," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, pp. 22-29, 2019. Article (CrossRef Link)

[6] N. Y. Ali, G. Sarowar, L. Rahman, J. Chaki, N. Dey, and J. Tavares, "Adam Deep Learning with SOM for Human Sentiment Classification," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 10, no. 3, pp. 92-116, July 2019. Article (CrossRef Link)

[7] N. Dey, S. Fong, W. Song, and K. Cho, "Forecasting energy consumption from smart home sensor network by deep learning," in *Proc. of International Conference on Smart Trends for Information Technology and Computer Communications*, pp. 255-265, 2017. Article (CrossRef Link)

[8] K. K. Verma, B. M. Singh, H. L. Mandoria, and P. Chauhan, "Two-Stage Human Activity Recognition Using 2D-ConvNet," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 2, pp. 125-135, 2020. Article (CrossRef Link)

[9] R. Ahuja, D. Jain, D. Sachdeva, A. Garg, and C. Rajput, "Convolutional Neural Network Based American Sign Language Static Hand Gesture Recognition," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 10, no. 3, pp. 60-73, 2016. Article (CrossRef Link)

[10] D. Wang, Z. Li, N. Dey, A. S. Ashour, L. Moraru, R. S. Sherratt, and F. Shi, "Deep-segmentation of plantar pressure images incorporating fully convolutional neural networks," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 546-558, 2020. Article (CrossRef Link)

[11] A. H. Ali, A. Atia, and M. S. M. Mostafa, "Recognizing driving behavior and road anomaly using smartphone sensors," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 8, no. 3, pp. 22-37, 2017. Article (CrossRef Link)
[12] F. A. Saiz, and I. Barandiaran, "COVID-19 Detection in Chest X-ray Images using a Deep Learning Approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 11-14, 2020. Article (CrossRef Link)
[13] G. R. Shinde, A. B. Kalamkar, P. N. Mahalle, N. Dey, J. Chaki, and A. E. Hassanien, "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art," *SN Computer Science*, vol. 1, no. 4, pp. 1-15, 2020. Article (CrossRef Link)
[14] S. Ahuja, B. K. Panigrahi, N. Dey, V. Rajinikanth, and T. K. Gandhi, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," 2020. Article (CrossRef Link)
[15] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322-1333, 2015. Article (CrossRef Link)
[16] L. Huang, A. D. Joseph, B. Nelson, B. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43-58, 2011. Article (CrossRef Link)
[17] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *Proc. of the 25th USENIX Security Symposium*, vol. 16, pp. 601-618, 2016. Article (CrossRef Link)
[18] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. of the 23rd USENIX Security Symposium*, vol. 14, pp. 17-32, 2014. Article (CrossRef Link)
[19] X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton, "A methodology for formalizing model-inversion attacks," in *Proc. of the 29th Computer Security Foundations Symposium (CSF)*, pp. 355-370, 2016. Article (CrossRef Link)
[20] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint* arXiv*:1611.03814*, 2016. Article (CrossRef Link)
[21] U. Aïvodji, S. Gambs, and T. Ther, "GAMIN: An Adversarial Approach to Black-Box Model Inversion," *arXiv preprint arXiv:1909.11835*, 2019. Article (CrossRef Link)
[22] A. Kerckhoffs, "La cryptographic militaire," *Journal des Sciences Militaries*, pp. 5-38, 1883. Article (CrossRef Link)
[23] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, "Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes," in *Proc. of the 15th Annual Conference on Privacy, Security and Trust (PST)*, pp. 115-11509, 2017. Article (CrossRef Link)
[24] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, "Model inversion attacks for online prediction systems: Without knowledge of non-sensitive attributes," *IEICE Transactions on Information and Systems*, vol. 101, no. 11, pp. 2665-2676, 2018. Article (CrossRef Link)
[25] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: generative model-inversion attacks against deep neural networks," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 250-258, 2020. Article (CrossRef Link)
[26] M. Khosravy, K. Nakamura, N. Nitta, and N. Babaguchi, "Deep Face Recognizer Privacy Attack: Model Inversion Initialization by a Deep Generative Adversarial Data Space Discriminator," in *Proc. of 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020. Article (CrossRef Link)
[27] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognizing faces across pose and age," in *Proc. of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67-74, 2018. Article (CrossRef Link)
[28] M. H. Sedaaghi, R. Daj, and M. Khosravi, "Mediated morphological filters," in *Proc. of 2001 International Conference on Image Processing*, vol. 3, pp. 692-695, 2001. Article (CrossRef Link)

[29] M. Khosravy, N. Gupta, N. Marina, I. K. Sethi, and M. R. Asharif, "Morphological filters: An inspiration from natural geometrical erosion and dilation," *Nature-inspired Computing and Optimization*, pp. 349-379, 2017. Article (CrossRef Link)
[30] M. Khosravy, N. Gupta, N. Marina, I. K. Sethi, and M. R. Asharif, "Perceptual adaptation of image based on Chevreul–Mach bands visual phenomenon," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 594-598, 2017. Article (CrossRef Link)
[31] M. Khosravy, N. Gupta, N. Patel, and T. Senjyu, "Frontier Applications of Nature Inspired Computation," *Springer*, 2020. Article (CrossRef Link)
[32] N. Gupta, N. Patel, B. N. Tiwari, and M. Khosravy, "Genetic algorithm based on enhanced selection and log-scaled mutation technique," in *Proc. of the Future Technologies Conference*, vol. 880, pp. 730-748, 2018. Article (CrossRef Link)
[33] G. Singh, N. Gupta, and M. Khosravy, "New crossover operators for real coded genetic algorithm (RCGA)," in *Proc. of 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pp. 135-140, 2015. Article (CrossRef Link)
[34] N. Gupta, M. Khosravy, O. P. Mahela, and N. Patel, "Plant biology-inspired genetic algorithm: superior efficiency to firefly optimizer," *Applications of Firefly Algorithm and its Variants*, pp. 193-219, 2020. Article (CrossRef Link)
[35] N. Gupta, M. Khosravy, N. Patel, N. Dey, and O. P. Mahela, "Mendelian evolutionary theory optimization algorithm," *Soft Computing*, vol. 24, pp. 14345-14390, 2020. Article (CrossRef Link)

Mahdi Khosravy received BSc. in Electrical Engineering (bio-electric) from Sahand University of Technology, Tabriz, Iran, and MSc. in Bio-electric Engineering from Beheshti University of Medical Studies, Tehran, Iran, respectively in 2002 and 2004. He received Ph.D. in Interdisciplinary Intelligent Systems from University of the Ryukyus, Okinawa, Japan, in 2010, where he was awarded by the head of University for his excellence in research activities. From September 2010 to 2017, he joined University of Information Science and Technology, Ohrid, Macedonia. From March 2018 to September 2019, he was a visiting associate professor in Electrical Engineering Department, Federal University of Juiz de Fora in Brazil and jointly, a research scholar in Electrical department, University of the Ryukyus, Okinawa, Japan. Currently, he is a specially appointed researcher in Media integrated Laboratories, graduate school of engineering, Osaka University, Japan. He is a member of IEEE.

Kazuaki Nakamura received the B.S. degree in Engineering from Kyoto University in 2005, and theM.S. and Ph.D. degrees in Informatics from Kyoto University in 2007 and 2011, respectively. He is currently an Assistant Professor at Graduate School of Engineering, Osaka University, from 2012. His research interests include image processing, image recognition, and video analysis. He is a member of IEEE, IEICE, IPSJ, and ITE.

Yuki Hirose received the B.E. and M.E. degrees in Engineering Science from Osaka University in 2018 and 2020, respectively. He is currently a doctoral course student in Graduate School of Engineering, Osaka University. His research interests include gait anonymization and face image generation. He is a student member of IEICE.

Naoko Nitta received the B.E., M.E., and Ph.D. degrees in Engineering Science from Osaka University, in 1998, 2000, and 2003, respectively. She is currently an Associate Professor in Graduate School of Engineering, Osaka University. From 2002 to 2004, she was a research fellow of the Japan Society for the Promotion of Science. From 2003 to 2004, she was a Visiting Scholar at Columbia University, New York. Her research interests are in the areas of video content analysis and image/audio processing.She received Best Paper Award of 2006 Pacific-Rim Conference on Multimedia (PCM2006). She is a member of IEEE, ACM, IEICE, and ITE.

Noboru Babaguchi received the B.E., M.E. and Ph.D. degrees in communication engineering from Osaka University, in 1979, 1981 and 1984, respectively. He is currently a Professor of the Department of Information and Communications Technology, Osaka University. From 1996 to 1997, he was a Visiting Scholar at the University of California, San Diego. His research interests include image/video analysis, multimedia computing and intelligent systems. Recently, he has been engaged in privacy protection for visual information, and security and fabrication of multimedia. He has published over 250 journal and conference papers and several textbooks. Dr. Babaguchi received Best Paper Award of PCM2006 and IAS2009. He is on the editorial board of Multimedia Tools and Applications, and New Generation Computing. He served as a Guest Editor of IEEE TIFS, Special Issue on Intelligent Video Surveillance for Public Security \& Personal Privacy.He also served as a workshop Co-chair of MIR2001, a Track Co-chair of IEEEICME2006, a General Co-chair of MMM2008, a Sub-Track Chair of APSIPAASC 2009, a General Co-Chair of ACM Multimedia 2012, a Track Co-Chair of ICPR2012, an Area Chair of IEEE ICME2013, and an Honorary Co-Chairof ACM ICMR2018. He also served on program committee of international conferences in multimedia related fields. He is a Fellow of IEICE, a Senior Member of IEEE, and a member of ACM, IPSJ, ITE and JSAI.