**Please find below the referee reports. Based on these and the corresponding recommendation of the associate editor, I have to inform you that your paper AutoGAN-based Dimension Reduction for Privacy Preservation with manuscript number: NEUCOM-D-19-00791 in its present form cannot be accepted for publication in Neurocomputing.**

**However, I would very much like to invite you to revise your paper, seriously taking into account the comments of the reviewers, and to resubmit your revised version by Jun 25, 2019 (mm/dd/yy). Any revision received after that may be treated as a new submission.**

**Reviewer #1: The authors proposed an input-privacy-preserving technique that reduces input images into lower-dimensional signals.  
The difficulty of image recovery is enhanced by training the encoder ("Generator" in this manuscript)  
to produce low-dimensional signal that cannot be decoded by another neural network ("Reconstruct" in Figure 3).  
  
Their idea is reasonable to some extent, but it is difficult for me to regard it as a serious privacy-preserving mechanism,  
mainly because the security is validated only heuristically by the re-constructor and discriminator used in the training phase.  
Since the capacity of re-constructor is finite, it may possible to decode reduced inputs by a more powerful re-constructor which can be easily prepared.  
My concern is enhanced by the fact that I could not find the precise information on the neural network architectures used in the study.  
  
In summary, I do not think the results are meaningless, but the current manuscript is difficult to recommend  
for publication in Neurocomputing journal.**

**“Since the capacity of re-constructor is finite, it may possible to decode reduced inputs by a more powerful re-constructor which can be easily prepared.”**

* Thank you for the insightful comment. To the best of our knowledge, in the case of using non-linear methods to reduce number of data dimension especially neural network, the well-known method that could be used to recover the original data is using an auto-encoder (*cite auto-encoder*). The architecture of an auto-encoder mainly includes two parts, encoder and decoder. The encoder structures vary and could be fully connected network, convolutional network. The decoder structure usually is inverted structure of encoder.

Instead of using typical fully-connected neural network, we use Deep Convolutional Neural Networks which are more effective for images. In this revision, we added more experiments using most current powerful convolution networks (i.e., deep convolutional network, VGG16, VGG19) (*cite very deep convolutional network for large-scale image recognition*) for encoders and their inverted structures for decoders so that we simulated powerful adversaries who aims to reconstruct our original data.

“

**My concern is enhanced by the fact that I could not find the precise information on the neural network architectures used in the study.**

**“**

To be clearer, we also add more information of neural network structures in the experiment section as the table below.

“”

**Reviewer #2: This paper proposed a dimension reduction-based method for privacy preservation. However, there are some problems.  
1. The organizational structure of the article is unreasonable, section 2, section 3 and section 4 should be introduced in the introduction and related work. I hope it can improve readability.**

* Sec 3 (preliminary) move to sec 2
* Change sec 2 to be related work (was literature review)
* Change sec 4 title “Problem” : We already introduce in introduction, this is a formal presentation of the problem

**2. Please briefly describe the content in Figure 1, which will be more helpful for readers to understand.**

* Thank you for your comment. It was our mistake not citing figure 1 in a proper position (Sec 4.1) while explaining it. In the revision, we cite the figure and update (blue part) the description as follows.

“

*4.1. Problem statement*

*We introduce the problem through the practical scenario mentioned in Section 1. Figure 1 briefly describes the entire system in which staff members (n members) in a company request access to its data resources, such as websites and data servers through a face recognition access control system. For example, if member n requests to webserver 2, the local device takes a facial photo of the member by an attached camera, locally transforms it into lower dimension and send to an authentication server. The authentication server then obtains the reduced dimensional features to determine his/her access eligibility without the clear face of the requesting member. We consider that the system has three levels of privilege (i.e., single level, four-level, eight-level) corresponding to three groups of members. We assume the authentication server is semi-honest (it obeys the work procedure but might be used to infer personal information). Once the server is compromised, an adversary in the authentication center can reconstruct the face features to achieve plain-text face images and determine members' identity.*

”

**3. Please explain what is the operation of E[] in equation (1)?**

**We add this explanation**

*“where E[] is the expectation, epsilon >= 0 0, x0 = F(x), ^x = R(x0), and R(.) is the Recon-struction Function.”*

**4. Please explain which method is used to project original data to low dimensional space?**

* An auto-encoder (Sec 3.1) with feedback from other components (i.e., Classifier, discriminator, constrain) was used to project data into lower dimensional space. In our work, the Generator and Re-constructor form an auto-encoder in which the lower dimensional represents are extracted from the middle layer of the auto-encoder (also output of Generator). Figure 1 show the architecture of an Auto-encoder.
* “*To tackle the problem, we propose a deep learning framework for transforming face images to lower dimension vectors which are hard to be clearly reconstructed. We leverage the structure of an auto-encoder \cite{Baldi2012} which contains encoder and decoder (in this work, we called them generator and re-constructor) in order to reduce data dimension. More specifically, the lower dimensional represents are extracted from the middle layer of the auto-encoder (output of the generator). The dimensionally-reduced data can be sent to the authentication server as an authentication request. We consider an adversarial as a re-constructor implemented by the decoder of the Auto-encoder. To prevent fully reconstructing images, the framework utilizes the discriminator in GAN \cite{Goodfellow2014} to direct reconstructed data to a desired target distribution. In this work, the target distribution is sampled from Gaussian distribution and the mean is the average of the whole training data. After the transformation projects the data into a lower dimension domain, the re-constructor can only partially reconstruct the data. Therefore, the adversary might not be able to recognize an individual's identity. To maintain data utility, we also use feedback from a classifier while training the auto-encoder as a part of its loss function. The entire framework can enlarge the distance between original data and its reconstruction to preserve individual privacy and retain significant data information. The dimensionally-reduced transformation model is extracted from the framework and provided to clients for reducing their facial image dimensions. The classification model will be used in an authentication center that classifies whether an authentication request is valid to have access \{1\} or not \{0\}.”*

**5. Please explain how to estimate the range of values of <epsilon> ?**

* In this study, the distance is defined as the L2 norm distance between the original data and reconstructed data. Since the reconstructed data is expected to expose less individual information and the driver the reconstructed images will be close to the mean of training dataset. Therefore, the epsilon can be estimated base on the expectation of the distance between the original data and the mean of training data. (will give more detail in this revision based on this answer)

**6. There are few methods for comparing experiments, no convincing. More recent methods should be used in the comparison experiments. In addition, computation complexity or time consumption of the methods should also be given. I hope this will help to improve the quality of the paper.**

* In this revision, besides comparison with the most similar work (GAP), we also perform more experiments for comparison with well-known methods (i.e., differential privacy and random projection).
* In addition, detail information of components’ structures also included in a table **(section …)**.

**7. What is the advantage of AutoGAN-DRP in terms of privacy protection compared to GAP?  The advice I would like to give is that the authors could give more details to explain the advantages of this method further.**

* GAP’s goal is hiding private labels and using a classifier to evaluate the privacy, our method is visually protecting the images themselves by reducing image dimension.

(A re-constructor plays a role of an aggressive adversary who aims to recovering the images.)

**8. Please rearrange the chart in paper, I hope it can help the paper to improve the readability.**

* Small format changes.

**9. In section 6.2, there is a spelling mistake, please correct it.**

* Thank you for your comment. We fixed the mistake.

“*We use an constrained optimization method to put a constraint on the*”->

*“We use a constrained optimization method to put a constraint on the”*

**10. In the paper, some formulas have no serial numbers, please correct it.**

* Thank you for the comment. In this revision we add serial numbers for all formulas.

**11. The number of references is insufficient, please add the corresponding references, I hope this will help to improve the quality of the paper.**  
 -> Adding more references.

**Reviewer #3: Overall, this paper proposed a GAN-based dimension reduction method for privacy preservation. The motivation is practical while the method is well presented. However the novelty is still limited and the authors need to explain more clearer about their contribution.**  
  
**Main issues:  
1. The main optimization formulation needs more explanation, especially from the perspective of theoretical analysis.**

* Formula 2:

**2. Not enough comparison in the experiments section, especially evaluation standard is not clear.**

* To have more comprehensive present, in this revision we add more experiments in linear cases (i.e., PCA, LDA).
* Regarding privacy evaluation, we consider the distance between the reconstructed images using an aggressive re-constructor and the original images. The distance in this case are defined as L\_2 norm distance. The high distance implies high level of privacy. For utility, the evaluation bases on the classification accuracy.

**3. Overall, the paper is not well written, please propose a comprehensive revision.**

Summary of changes

1. Sec 3 (preliminary) move to sec 2

Change sec 2 to be subsection (related work) inside “literature review” section

1. Explaining and make the scenario clearer in “problem statement”.