



SDI Review Form 1.6

Journal Name:	Current Journal of Applied Science and Technology
Manuscript Number:	Ms_CJAST_42730
Title of the Manuscript:	A Machine Learning Algorithm Based on Inverse Problems for Cyber Anomaly Detection
Type of the Article	Original Research Article

General guideline for Peer Review process:

This journal's peer review policy states that **NO** manuscript should be rejected only on the basis of '**lack of Novelty**', provided the manuscript is scientifically robust and technically sound. To know the complete guideline for Peer Review process, reviewers are requested to visit this link:

(<http://www.sciencedomain.org/page.php?id=sdi-general-editorial-policy#Peer-Review-Guideline>)



SDI Review Form 1.6

PART 1: Review Comments

	Reviewer's comment	Author's comment (if agreed with reviewer, correct the manuscript and highlight that part in the manuscript. It is mandatory that authors should write his/her feedback here)
Compulsory REVISION comments	1) The resolution of the figures is low and not enough for publication. In addition to this, I have seen bunch of distortion that might have happened through the PDF conversion. 2) The mathematical notations contain confusion. For instance, what is the difference between “A” and “A” as shown in eq. 1 and 2. If there is any difference, it should be noted. 3) The authors have used a Kaggle database with 284,807 exemplars and 31 features. They mentioned that these features are the results of Principal component analysis (PCA). Therefore, the real data have not been used here or they do not have access to that. It should be noted that PCA is only one way of dimension reduction. There are different methods including independent component analysis (ICA), singular value decomposition (SVD), factor analysis (FA), regularized feature selection, and ... that can be employed. The following paper can be added to the manuscript to point out the chance of various data reduction methods: - Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Schmaal, L., Goudriaan, A. E., & Meyer-Baese, A. (2017, June). An evolutionary approach for fMRI big data classification. In <i>Evolutionary Computation (CEC), 2017 IEEE Congress on</i> (pp. 1029-1036). IEEE. 4) The authors need to show why PCA is better that was used in this database. In addition to this, in the manuscript PCA just mentioned without any definition. The following reference might help: - Duda, R. O., Hart, P. E., & Stork, D. G. (2012). <i>Pattern classification</i>. John Wiley & Sons. 5) For the database, it is a good idea to visualize the correlation matrix, density matrix, and scatter matrix of the features. It would give the readers some idea how the feature selection was done. The following article can give you some idea about this part: - Tahmassebi, A., Gandomi, A. H., & Meyer-Baese, A. (2018). Stock Risk Assessment via Multi-Objective Genetic Programming. <i>Journal of Postdoctoral Research</i>, 6(3). 6) There is a typo regarding the name of TensorFlow package on line 132. Please do check the whole manuscript again for any possible typos. In addition to this, the authors have employed Scikit-Learn which requires the following reference: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. <i>Journal of machine learning research</i>, 12(Oct), 2825-2830. 7) On line 144, the authors explained that the data randomly splitted into train/test size. However, it should be noted that what was the size of split? Did the authors use stratified split? 8) Again on line 152, the authors mentioned SVM without any definition	Thank you for the opportunity to revise the manuscript. Below I summarize my responses to the comments. I believe that all of the comments have been addressed in a way that you would find satisfactory. I thank you very much for their detailed and thoughtful critiques. The manuscript has been greatly improved as a result of your effort. Our point-by-point responses to your comments are detailed here. Responses highlighted in yellow in the manuscript. 1) The resolution of the figures has been improved by including the original figures. Definitely, the pdf conversion creates the distortions. Now, I don't use the figures created through the PDF conversion. 2) It was a typo. I changed them to “A” instead of “A”. 3) We have added the reference you provided for readers to refer to other various data reduction methods. 4, 5, and 10). First of all, please know that I appreciate very much your <i>suggestions</i> and advices. These are great suggestions and I was gladly surprised to see this since my colleague and I are currently working on it as stated in the conclusion section of the paper. This work is by no means exhaustive of the method discussed here. We recognize that improvements have to be done both on the algorithm (different regularizers must be investigated) and the applications. More detailed statistical analysis is needed to improve the effectiveness of the model in general. Further, the answer to exactly how sensitive the method is to moderate amounts of noise should be investigated. Also, the answer to which data reduction methods in the preprocessing needs to be used should be investigated. We believe that, to enhance the performance, it is possible to modify the suggested method and devise a scheme using k-folds cross-validation on different number of folds, using ensemble models, etc. One has to make further analysis not only on how to improve the efficiency of the proposed model, but also on how to take advantage of parallel and distributed computing for detecting anomalies. With this enhancements, we hope to improve the model in the future work. Overall, these results motivate further research of additional studies of the proposed model to detect anomalies. As of now, we are working on all that's suggested, but I feel that it represents more of a second, follow-up paper. The goal of this study is thus to better understand how well the proposed model works. These suggestion are added to the conclusion part of the paper as a future work. Suggested references are added in the manuscript accordingly, as well. Thank you very much for your great suggestions. 6. Typo is fixed and the reference is added and highlighted. 7. This is another great point. I have updated the manuscript to answer this question. 80 percent of the dataset will be used to actually train the model, while the rest will be used to



SDI Review Form 1.6

of support vector machines. In addition to this, it should be noted that what type of kernel was used for SVM as the classifier.

9) The result that is presented in Figure 4 is completely confusing. As mentioned before, the size of training/testing set is crucial here. Based on my calculation, the testing set of Figure 4 is 30% of total 284000 instances. The results of AUC is great with an AUC of 91%. However, the number of False positives are high. Then, the authors under-sampled the data. Based on my calculation the testing set contains 296 instances. As mentioned before, the database contains 284000 instances. Do you think with this type of sampling you can generalize the importance of the model? The results are true, however they cannot be generalized. In this regards, at least the authors need to to show different number of samplings (at least 100 times) and then present the summary statistics including average values, standard deviations, median, and possible z-scores and p-value to make sure the sampling was right.

10) For the result presented in Figure 4, the authors can also use k-folds cross-validation using a grid-search on different number of folds to get the best results (least number of False Positives).

11) The authors employed SVM, Decision Tree, and Bayesian Network models as a comparative study. However, none of the presented models are ensemble models including random forest and gradient boosted trees such as XGBoost which is so popular these days. In addition to this, the size of the data is large enough to try some deep learning and neural network methods. The following articles can address some of these issues:

- Tahmassebi, A., Gandomi, A. H., Schulte, M. H., Goudriaan, A. E., Foo, S. Y., & Meyer-Baese, A. (2018). Optimized Naive-Bayes and Decision Tree Approaches for fMRI Smoking Cessation Classification. *Complexity*, 2018.
- Tahmassebi, A. (2018, May). ideeple: Deep learning in a flash. In *Disruptive Technologies in Information Sciences* (Vol. 10652, p. 106520S). International Society for Optics and Photonics.
- Tahmassebi, A., & Gandomi, A. H. (2018). Building energy consumption forecast using multi-objective genetic programming. *Measurement*, 118, 164-171.

11) Finally, the authors can visualize the comparative study results using some radar plot (spider plot).

evaluate the accuracy of this model.

8. Since the machine learning algorithms are studied extensively in the literature, for detailed definition and application of SVM, Bayesian Network, and Decision Trees in the context of intrusion detection we refer reader to the reference [2]. We used RBF kernel SVM in this study and this is updated in the manuscript as well. Since the **Support Vector Machine kernel** selection is important, as a future work, we may need to also check which kernel works best.

9. Total 284,807 rows of data are considered, but this data is fed into the preprocessing module as described in Figure 3 and missing values and junk values of rows are removed. After preprocessing, data set has 125974 rows. Data preprocessing is a very important and quite underestimated step in Machine Learning models. It provides cleaned and relevant datasets which then can be used in further steps. These changes are reflected in the manuscript under the Section 4 - Results and Discussions.

In future work, we also want to consider alternative preprocessing strategies to validate our results. Therefore, we ultimately seek to develop a robust and proven methodology of preprocessing.

11. Finally, we provided a spider plot for the comparative study results in Figure 11. The visualization results show that they are very much in line with each other.

Minor REVISION comments

Optional/General comments