Research Journal of Computer Systems and Engineering

ISSN: 2230-8571, 2230-8563

Volume 02 Issue 02 - 2021 (July to December)

Page 45:50

Dimensionality Reduction Based Diabetes Detection Using Feature Selection and Machine Learning Architectures

Adlin D Steffy
Physics and Chemistry
Nesamony Memorial Christian College, Marthandam, affiliated to Manonmanium Sundaranar University, Tirunelveli
steffy909090@gmail.com

Article History

Received: 15 July 2021
Revised: 20 September 2021
Accepted: 22 November 2021

Abstract

One of the illnesses that is sweeping the globe like an epidemic is diabetes. Every generation, including children, teenagers, young adults, and elderly, is seen to be affected by it. The study raised difficulties regarding the necessity to establish a connection between the primary causes of diabetes development. This paper compares pre-processing accuracies of various dimensionality reduction models. Here the proposed technique use two datasets, one is normal diabetic dataset and another is heart disease dataset. Both dataset has been pre-processed using dimensionality reduction (DR). In proposed work, DR process is divided into two stages: unsupervised DR and supervised DR. Prior to processing, improved DR unsupervised principle component analysis was performed. The two datasets were then combined.
Keywords: Diabetes, data pre-processing, dimensionality reduction, improved principle component analysis, unsupervised DR, supervised DR.

1 Introduction:

Diabetes is a varied group of disorders that can ultimately lead to an increase in blood glucose levels and a deficiency in urine glucose [1]. It is not an inherited illness. Despite attempting to address issues using historical or prior examples, ML is a subset of AI [2]. Machine learning, as contrast to AI applications, entails discovering hidden patterns in data as well as then using those patterns to categorise or forecast a problem-related event [3]. Dimensionality reduction is a potent way to deal with scaling the statistics back up. A decrease in dimensionality is sought while maintaining estimates between several over-the-top dimensional vectors using this technology [4].

2 Related works:

For instance, Cao and Chong [5] represented a comparison study of PCA, XPCA, and ICA as feature extraction for SVM; [6] compared PCA, KPCA, and ICA for SVM classification; [7] built an empirical study of dimensionality reduction in SVM using PCA, KPCA, and ICA; and [8] represented a comparative study of PCA, XPCA, and ICA [9] as feature extraction for SVM [10].

3 Research methodology:

Since a decade ago, the number of persons with diabetes has dramatically increased. Key factor contributing to rise of diabetes is current human behaviour. There are three basic sorts of errors that can occur in the present medical diagnosing process.

1. False-negative testing is when a patient's test results indicate that they do not have diabetes but, in fact, they already have the disease.

2. falsely positive in nature. In this case, despite test results indicating otherwise, the patient is not actually diabetic.

3. The third type is unclassifiable, which occurs when a system is unable to identify a certain scenario. The reason for this is that an unclassified form of prediction for a certain patient may result from insufficient knowledge extraction from historical data. The following is the suggested architecture for detecting diabetes:

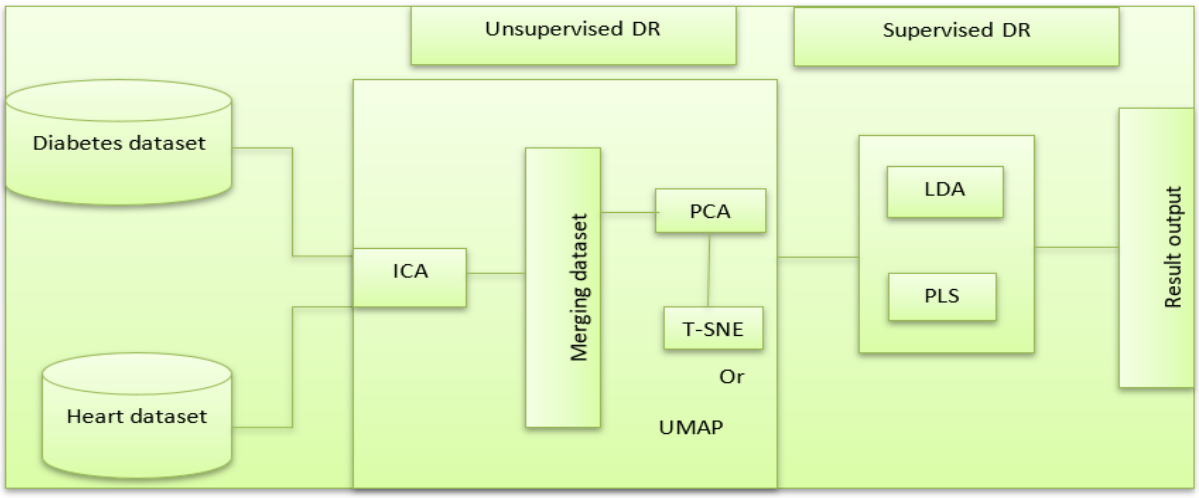

Figure 1: Overall architecture

By maximising the covariance between y and t1 while adhering to the restriction of k w1 k=1, PLS1 finds first latent component t1 = Xw1. The following objective function corresponds to eq. (1):

$$w1 \ = \ arg \ max \ wT \ w = 1 \ (Cov(Xw, y)) \ (1)$$

The Lagrange multiplier approach can be used to quickly solve the maximising problem in Equation (2).

$$w1 \ = \ X \ T \ y / \ k \ X \ T \ y \ k$$

$$E1 \ = \ X \ - \ t1p \ T \ 1 \ f1 \ = \ y \ - \ t1q \ T \ 1$$

$$p \ T \ 1 \ = \ (t \ T \ 1 \ t1) \ - 1 \ t \ T \ 1 \ X \ q \ T \ 1 \ = \ (t \ T \ 1 \ t1) \ - 1 \ t \ T \ 1 \ y$$

$$\boldsymbol{X = TP^T + E}$$

$Y = U\boldsymbol{Q}^T + \boldsymbol{F}.$

$$wk \ = \ E \ T \ k - 1 \ fk - 1 / \ k \ E \ T \ k - 1 \ fk - 1 \ k$$

$$tk = Ek - 1wk$$

$$p\,T\,k = (t\,T\,k\,tk)\ -1\,t\,T\,k\,Ek - 1$$

$$q\,T\,k = (t\,T\,k\,tk)\ -1\,t\,T\,k\,fk - 1$$

$$Ek = Ek - 1 - tkp\,T\,k \qquad (2)$$

$$fk = fk - 1 - tkq\,T\,k$$

$$V = W(P\,TW)\ -1\ ,\ (2)$$

4 Performance Analysis:

On Windows 10, we run the simulations using the PyTorch deep learning framework. The PC used for the tests has an AMD Ryzen 5 1600X Six-Core Processor and an 8GB GeForce GTX 1070Ti GPU. Programming is carried out using the Python language.

Table.1.The Overall Outcome of Performance Measure of ICA,PCA,UMAP -PLSDA

Dataset	Techniques	Accuracy Score(%)	Precision(%)	Sensitivity(%)	Specificity(%)	F1-Score(%)	AUC-Score(%)
Diabetes and Heart Disease Dataset	ICA (PLSDA)	84.49	89.57	74.64	92.73	81.43	83.68
Diabetes and Heart Disease(Merge)	**PCA(PLSDA)**	**95.05**	**88.00**	**91.67**	**96.10**	**89.80**	**93.89**
Diabetes and Heart Disease(Merge)	UMAP(PLSDA)	71.67	69.12	68.12	74.55	68.62	71.31

From data set on diabetes and heart disease, accuracy analysis is used to identify diabetes, and combined data value is used to assess patient's diabetes across board. Analysis shows that PCA with PLSDA has highest accuracy compared to other dimensionality reduction techniques.

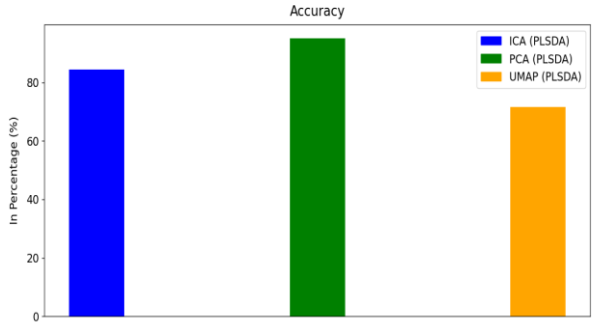

Figure.2. Comparison of Accuracy for various DR techniques

Contents of table 1 are graphically represented in figure 2. While the UMAP technique, which had an accuracy value of approximately 71.6%, had the worst result. In addition, ICA method gains greater accuracy when compared to prior one, which had an accuracy of roughly 84.49%. Finally, the PCA

technique achieves maximum Accuracy value of 95.05% after merged dataset, outperforming performance of other models.

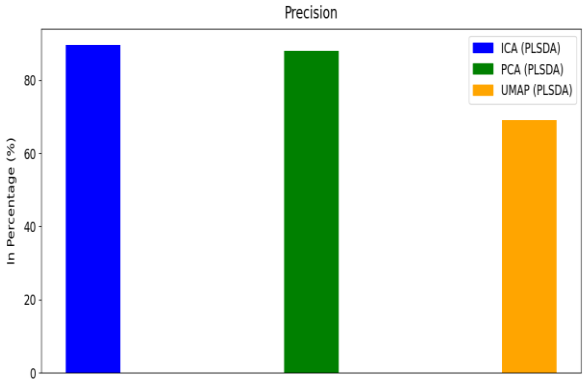

Figure.3. Comparison of Precision for various DR techniques

Contents of table 1 are graphically represented in figure 3. While the UMAP technique, which had a Precision value of approximately 69.12%, had worst result. The PCA method simultaneously gains greater Precision compared to prior one by roughly 88.00%. Finally, by achieving a maximum Accuracy value of 89.57%, the ICA approach performs more effectively than the performance of the other models.

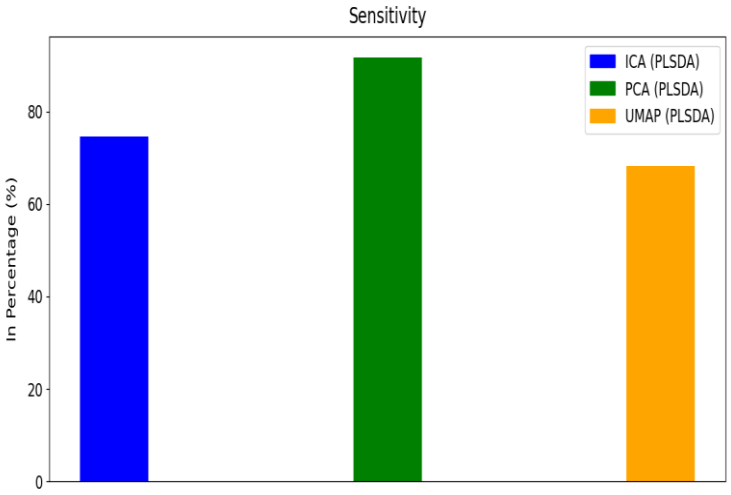

Figure.4. Comparison of Sensitivity for various DR techniques

Graphic representation of Table.2.contents is Figure.4. While the UMAP technique, which had a Sensitivity value of approximately 68.12%, had worst result. The ICA method simultaneously gains increased sensitivity compared to prior one, which was roughly 74.64%. Finally, PCA method achieves the greatest Sensitivity value of 89.57% with the combined dataset, outperforming the performance of the other models.

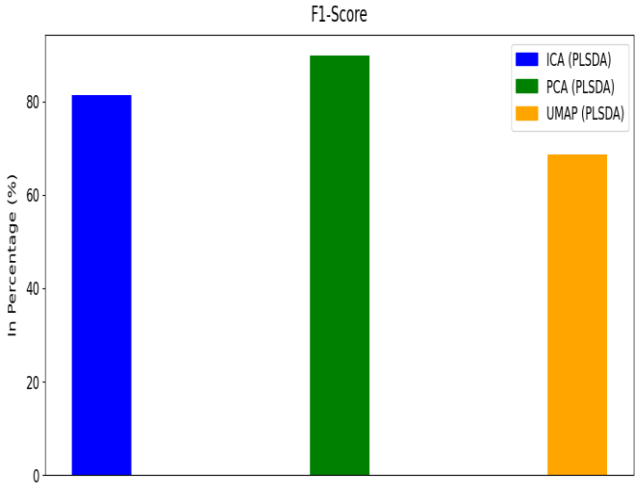

Figure.5. Comparison of F1-Score for various DR techniques

Graphic representation of Table.1's contents is Figure.5. It displays F1-Score comparison for the DR approaches for the Diabetes and Heart Disease Dataset. According to this figure, the PCA outperforms the other approaches in terms of F1-Score %. While the UMAP technique, which provided an F1-Score value of around 68.72%, had the worst result. In parallel, the ICA model gains a greater F1-Score than the previous one, which was roughly 81.43%. Finally, the PCA technique achieves the highest F1-Score value of 89.80% with the combined dataset, outperforming the performance of the other models.

5 Conclusion:

The goal of this study is to develop a predictive model for type 2 diabetes mellitus. Output of most popular methods for lowering simple dimensionality are compared in this research. While PCA is a traditional strategy, the comparison demonstrates that contemporary methods are still unable to exceed it. The characteristics in DR must be uncorrelated but not distinct as in naive bays classifier, therefore using ICA does not have any real relevance. This is supported empirically by the experiment, which shows that ICA performs less accurately than the other approaches. Finally, it should be noted that DR has shown to be a successful HD data set classifier that also performs well when applying techniques for feature selection.

Reference:

[1] Mateen, M., Wen, J., Song, S., & Huang, Z. (2018). Fundus image classification using VGG-19 architecture with PCA and SVD. *Symmetry*, *11*(1), 1.

[2] Chan, G. C., Kamble, R., Müller, H., Shah, S. A., Tang, T. B., & Mériaudeau, F. (2018, July). Fusing results of several deep learning architectures for automatic classification of normal and diabetic macular edema in optical coherence tomography. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 670-673). IEEE.

[3] Espadoto, Mateus, et al. "Towards a quantitative survey of dimension reduction techniques." *IEEE transactions on visualization and computer graphics* (2019).

[4] Abdulhammed, Razan, et al. "Features dimensionality reduction approaches for machine learning based network intrusion detection." *Electronics* 8.3 (2019): 322.

[5] Xu, Xinzheng, et al. "Review of classical dimensionality reduction and sample selection methods for large-scale data processing." *Neurocomputing* 328 (2019): 5-15.

[6] Becht, Etienne, et al. "Dimensionality reduction for visualizing single-cell data using UMAP." *Nature biotechnology* 37.1 (2019): 38-44.

[7] Zhang, Bing, et al. "Network intrusion detection method based on PCA and Bayes algorithm." *Security and Communication Networks* 2018 (2018).

[8] Esfahani, M. T., Ghaderi, M., & Kafiyeh, R. J. L. E. J. P. T. (2018). Classification of diabetic and normal fundus images using new deep learning method. *Leonardo Electron. J. Pract. Technol*, *17*(32), 233-248.

[9] Pavithra, M., Saruladha, K., & Sathyabama, K. (2019, March). GRU based deep learning model for prognosis prediction of disease progression. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 840-844). IEEE.

[10] Sun, Y. (2019). The neural network of one-dimensional convolution-an example of the diagnosis of diabetic retinopathy. *IEEE Access*, *7*, 69657-69666.