Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

Feature Selection Based Diabetic Detection Using Ensemble Dimensionality Reduction with Machine Learning

S. Sutha[1], N. Gnanambigai[2], P. Dinadayalan[3] , A. Vettriselvi[4]

[1]Research Scholar, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; Email:ssutha1612@gmail.com

[2]Department of Computer Science, Indira Gandhi College of Arts and Science, Puducherry - 605009,India, Email: dgnanambigai@hotmail.com

[3]Department of Computer Sci, KanchiMamunivar Centre for Postgraduate Studies, Puducherry - 605008, India, Email: pdinadayalan@hotmail.com

[4]Research Scholar, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; Email: puppyselvi1@gmail.com

Abstract:
Diabetes is one of the diseases in the world that is spreading like epidemics. Every generation, ranging from girls, teenagers, young people and the elderly, is seen to suffer from this. In terms of organ failure such as liver, kidney, heart, stomach, prolonged impact can cause worse effects and can lead to death. It is commonly associated with retinopathy and neuropathy disorders. There are primarily two types of diabetes: type 1 and type 2.Diabetes diagnosis or prediction is carried out by various techniques of data mining, such as correlation, grouping, clustering and pre-processing of data. The research led to similar open problems to recognize the need for a relationship between the key factors contributing to the development of diabetes. This is possible through the mining patterns contained in the dataset between the independent and the dependent variable. This paper compares the accuracy of pre-processing of different models of dimensionality reduction. The process of DR is classified into two stages in the proposed work, one is unsupervised DR and other is supervised DR. before the processing using unsupervised DR improved principle component analysis has been done. Then the two dataset has been merged. On whole the aim for the proposed technique is to reduce the dimensionality which could improve the accuracy in feature selection. The simulation results gives the improved accuracy in pre-processing part.
Keywords: Diabetes, data pre-processing, dimensionality reduction, improved principle component analysis, unsupervised DR, supervised DR.

INTRODUCTION

Diabetes is not a genetic condition, but a heterogeneous collection of conditions that could potentially lead to a blood glucose boom and a loss of glucose in the urine. Typically, diabetes is the product of biology, diet and the climate. Eating a dangerous weight loss diet plays a part in the creation of diabetes because of being overweight. The tiny blood vessels in your frame will damage the extra sugar in the blood. Diabetes symptoms are blurry, inventive, prescient, intense appetite, unusual weight loss, common urination, and thirsty. Glucose, blood pressure, pores and skin texture, cortisol, age are criteria included within the information set to locate the diabetes in this article. Healthcare sectors produce large amounts

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

of statistics units. The data set out here are a compilation of diabetes patient details from hospitals[1].

Big data analytics is the processing in which the information units are analyzed and the secret information exhibited. Big data is a future research field for which academics and IT societies attract significant exposure. The volume of data produced and processed has increased within a brief period of time in the modern world. Consequently, this exponentially rising data rate has created several difficulties. To explore the roots of big data systems and their present developments, we use structuralism and functionalism paradigms in this article. Big data is a mixture of multiple granular data types. A part of artificial intelligence is machine learning, although it attempts to solve problems based on historical or previous examples[4]. Machine learning, unlike artificial intelligence software, entails learning secret patterns within the data (data mining) and then using the patterns to identify or forecast an issue-related event[5]. Intelligent machines simply rely on information to preserve their functionality, and such knowledge is provided through machine learning. Machine learning algorithms are basically integrated into computers and data sources, such that knowledge and information are extracted and fed into the system for systems to be handled more rapidly and effectively. In addition, these machine learning methods are also useful for strategies to minimize dimensionality. Parallel processing data analysis algorithms are useful in this approach for time confinement applications. One of the key tasks with regard to duplicate data dimensionality reduction applications is both time, speed and performance. In addition, different applications are presented, but changes are needed.

Reduction of dimensionality is a strong way of operating for the numbers restored by scaling. It is an approach that attempts to amplify various dimensional vectors to a reduced dimensional field over the pinnacle while retaining projections between them. In view of the expected increases, the techniques of AI and certainty mining might not be persuasive in attitude to the scourge of dimensionality for high dimensional realities, and the consistency and adequacy of the investigation may deteriorate rapidly. For 1) Visualization: The Size Decrease is used for mapping of high-dimensional dimensions to 2D or 3D. 2) Data Compression: Efficiently collecting and healing. Three) Reduction of noise: Positive impact on demand accuracy. Dimensionality procedures are carried out to remedy excessive dimensional numbers, as in exquisite microarray assessment of clarity, description of subject data, with burdens on a primary comprehensive range of skills, with various unessential and dreary tasks and late research findings, sparkle off overabundance relying primarily on knowledge willpower. The definition of lower measurement can be referred to as seeks after 1) from a mathematical revelation aspect of view, the unmistakable evidence of a reduced leisure schedule of characteristics that can be farsighted with factors may be useful. 2) The preparation, though, would definitely improve the representation time with the sum of functions with those having to know computations. 3) However, noisy or unimportant tasks may have a similar effect on the request as perceptive capacities, so that they can have an effect on accuracy in reverse.

A tool called Predictive Analysis involves a range of algorithms for machine learning, tools for data processing and mathematical models that use present and historical information to find insights and forecast future events. Important decisions can be drawn and forecasts can

be made by applying predictive technology to healthcare results. Deep learning and regression approaches can be used to do statistical analysis. The goal of predictive analytics is to identify the condition for the highest possible specificity, enhance health safety, maximize services and improve clinical results[24-25]. Machine learning is considered to be one of the most significant aspects of artificial intelligence that facilitates the development of computer systems capable of acquiring information from interactions without any need for programming. In order to eradicate human efforts by encouraging automation with minimal defects, machine learning is perceived to be a desperate need in today's case. Lab measures, such as fasting blood glucose and oral glucose resistance, are an existing tool for diabetes diagnosis. This approach is time-consuming, though [6].

Section II-gives a literature summary of the work performed earlier on diabetes prediction and machine learning algorithm taxonomy. Section III-discusses the proposed model for diabetes prediction. For the proposed model to be discussed, section IV gives experiment findings. Conclusion of Section V and Sources.

Related works:

A few examinations/correlations have been played out that expand assortment and extraction techniques for highlights for AI calculations; like the creator in [7] who analyzed PCA, KPCA and ICA for SVM grouping, in [8] who set up an experimental examination of dimensionality decrease in help vector machines utilizing PCA, KPCA and ICA, in [9] a relative investigation of PCA, XPCA and ICA as help vector machine work extraction was introduced, the author[10] who introduced a near investigation of highlight extraction ideal models dependent on neural organizations and who researched include choice procedures dependent on ICA and PCA for face acknowledgment in[11]. Also, an assortment of studies have been effectively applied to LR by PCA, such as[12] utilizing principle parts to appraise LR with high-dimensional multi-collinear outcomes, and[13] utilizing the PCA-LR model to develop the monetary evaluation model of a recorded firm. Also, a few examinations have been performed utilizing KPCA with LR, for example, applying KPCA with LR for the characterization of quality articulation data[14]; and building a nonlinear connection investigation dependent on KPCA with LR[15]. Free Component Analysis (ICA) is an exceptionally normal procedure that has shown viability in detachment of visually impaired causes, extraction of highlights and unaided acknowledgment; as of late, scientists have to a great extent examined it[16]. Past examinations have shown that in both directed and unaided characterizations, ICA can fill in as a proficient strategy for highlight extraction to upgrade grouping productivity. ICA has been applied to numerous classifiers; the creator [17] applied ICA to the arrangement of Naive Bayes and [18] addressed a face acknowledgment examination utilizing ICA and vector machine help. To the most amazing aspect our experience, LR didn't utilize ICA. As of late, the most un-square relapse (LSR) model with minimized structure for multi characterization utilizing e-hauling procedure has been impressively studied[19], which can normally be summed up for multi-class work assortment.

The investigation of comparable examinations gives discoveries on various information bases of medical care, where various methodologies and systems have been utilized to perform investigation and figures. Utilizing renditions of information mining procedures, AI

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

calculations or even a combination of these strategies, diverse forecast models have been created and applied by various scientists. Jayanthi et al (2021) applied a strategy for investigation of diabetic information utilizing Hadoop and Map Reduce method [20]. This technique conjectures the kind of diabetes and the related dangers too. The strategy depends on Hadoop and is prudent for each organization in the medical care field. [21] To explore secret varieties in the diabetes dataset, utilized an order strategy. In that model, Naive Bayes and Decision Trees were utilized. The effectiveness of the two calculations was looked at and the adequacy of the two calculations was viewed therefore. [22] For the arrangement of dress deformities, Habib and Rokonuzzaman [23] utilized CPN. Utilizing the CPN model, they focused on grouping material imperfections. They additionally led research on the interrelationship between plan boundaries and CPN model proficiency.

Research methodology:

For a decade, there has been a dramatic rise in the number of people suffering from diabetes. The primary factor behind the rise in diabetes is the modern human lifestyle. There may be three distinct types of errors in the modern medical diagnosis process.

1. The false-negative types. Here, the person is actually diabetic but the test shows he isn't.
2. The second one is false-positive types. Here, the person is actually not a diabetic, but the test shows he is.
3. And the third type is unclassifiable. Here a given case cannot be diagnosed by a system. This occurs because a given patient can be predicted in an unclassified form due to inadequate information extraction from past data.

In reality, however, the patient must predict that he or she is either in the diabetic or non-diabetic group. Such diagnostic errors can result in unnecessary treatment or no treatment at all if needed. To keep away from or diminish the degree of such an impact, it is imperative to make a framework utilizing AI calculations and data-mining procedures, which will give precise outcomes and lessen human endeavors. The proposed diabetes diagnosis architecture is determined with an overview is indicated in the Figure.1.in which the Overall Architecture ICA -PLSDA(Partial Least square with Linear Discriminant Analysis) as follows:

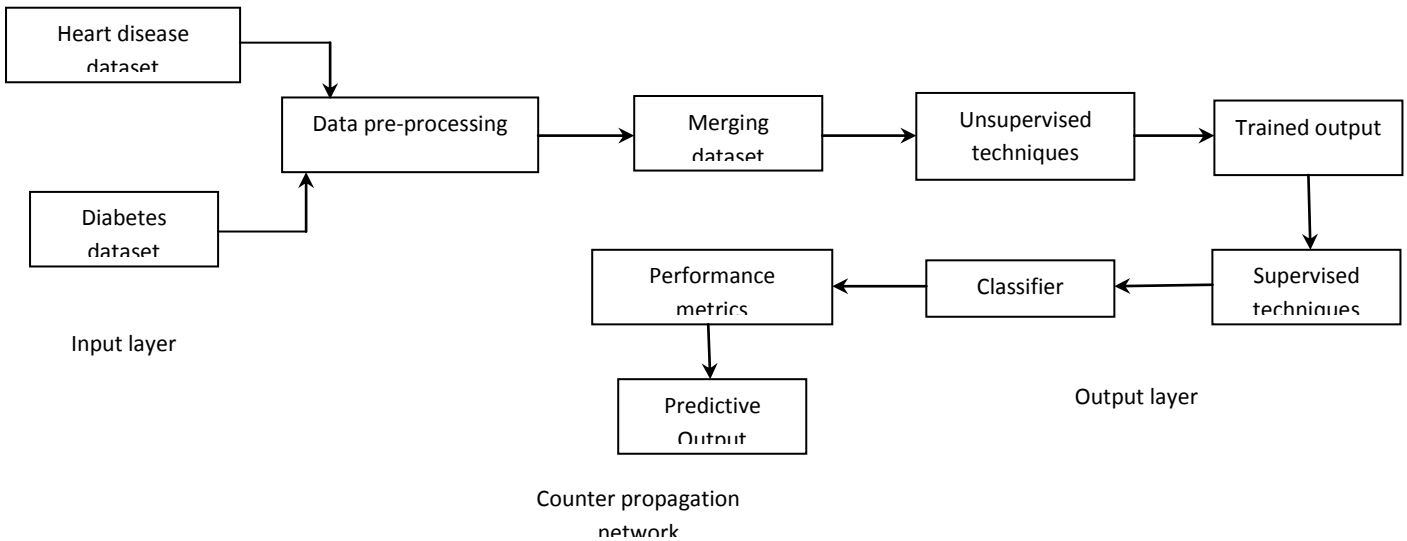

Figure 1- Overall Architecture ICA -PLSDA

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

Dataset Description:

To be evaluated by the UCI system, the data collection is derived from the Pima Indians Dataset Database (PIDD). The statistics collection includes several unbiased factors, along with glucose, blood pressure, skin thickness, BMI, etc. The record collection is trained to get the same end result, and it is equally verified. 800 records and 10 attributes are included in this Diabetes dataset.

Diabetes detection using Dimensionality Reduction (DR):

There are different approaches that can be used in the pre-processing technique for DR. For the following advantages it offers, the DR is considered:

1) With the decrease in the measure of dimensions, the memory accessible to save the information decays.

2) Reduced training or computation times are needed for fewer dimensions.

3) If data has several dimensions, most of the algorithms in the process of feature extraction do not cope well.

4) The multi-collinearity between different data features is well dealt with by DR techniques and redundancy between the features is eliminated.

5) Finally, the lower dimensions of the data aid to be visualized.

Unsupervised DR:

Independent component analysis:

The information dataset has been joined with the dataset for coronary illness in this interaction. For information handling, IPCA is a generally present day strategy for insights and calculation. ICA started from the sign handling society, where it was made as an amazing strategy to recognize dazzle sources. The basic ICA model for change of highlights can be composed as: s_t=ux_t. Where xt is n×p lattice, st is n×p network is the new free assessed vectors for grouping purposes, u is alluded to as the n×n de-blending framework and is utilized to discover an altogether new genuinely autonomous non-Gaussian heading coordinate framework, with the main IC course being the most non-Gaussian. The calculation works iteratively and the most non-Gaussian way is first portrayed. Fixed on this heading, which is free from the first bearing and so on, the second, more non-Gaussian bearing is found. For n×p dimensional information vectors, it decides up to n×p dimensional free vectors, so the element vectors addressing the first information should be determined for n×p dimensional information vectors from the information. Factual freedom implies that the capacity of the thickness of the joint likelihood of the variable st is equivalent to the result of the elements of the minor thickness parts. The xt measurement has been diminished by utilizing PCA to get the brightening and hence lessens the measure of st to be determined. The quantity of PCs selected in this paper is equal to the number used on the PCA. The fixed point approximation is done to conclude the shift lattice and individual parts after the knowledge brightening measure is completed. Popular knowledge is considered a calculation of the dependency between arbitrary variables. Augmenting their negentropy is equivalent to limiting the data between the segments that is proportional. Negentropy can be communicated approximately as follows in the quick ICA:

$$J_G(s_{t(i)}) \approx \left[E\{G(s_{t(i)})\} - E\{G(V)\}\right]^2 \qquad [1]$$

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

V - Gaussian variable with the mean of 0 and unit fluctuation, where G is any non-quadratic power for all intents and purposes, and $s_{t(i)}$ is an n-dimensional vector involving one of the u-framework lines. There are several attributes that can be used, much like G. Substituting in $s_{t(i)} = u_i^T x_t$ obtaining the following optimization problem:

$$\text{Maximize } J_G(s_{t(i)}) \approx \left[E\{G(s_{t(i)})\} - E\{G(V)\}\right]^2 \quad [2]$$

$$\text{Subject to } E\{(\mu_i^T x)^2\} = 1 \;\; i = 1,2,....n \quad [3]$$

By settling this advancement problem via the Quick ICA estimation, one new free factor can be calculated. In light of this, the whole reduced free factor network s_t can be determined. We utilized Quick ICA with slant in this paper to acquire the individual parts. The IPCA yield joins the PCA and is then marked utilizing PLS for this yield. This yield is then gone into the U-guide and grouped once more. The grouped yield is at long last gathered. Beneath, the extraction highlights and portrayal of the pictures have been examined.

Table.1.The Outcome of the Performance Measure of ICA -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

The success indicators of the PLSDA classification for the Dataset as seen in Table.1 above (Diabetes and Heart Disease Dataset). The implementation architecture is seen in Figure 2.

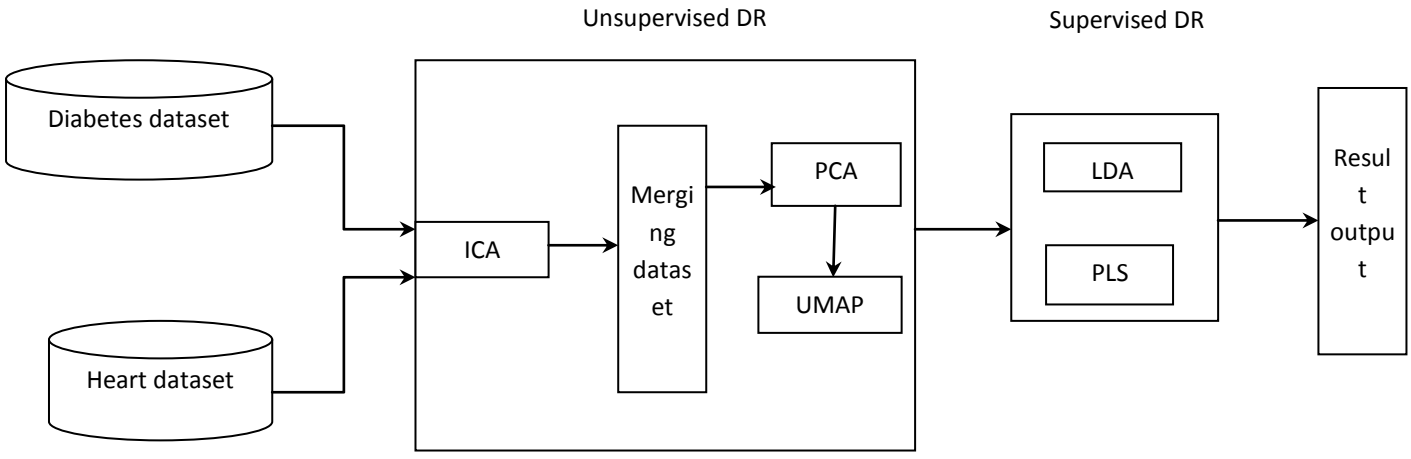

Figure-2 Implementation Architecture of ICA -PLSDA

Principal component analysis:

In a dataset, the PCA converts the attributes into new functions called primary components. A linear combination of the dataset's initial variables is the key component. The primary elements are grouped such that the first primary element is the one that reflects the biggest disparity in the results. These are not, however, correlated to the first main data variable. Centered on the notion of eigenvalues and eigenvectors, the initial matrix is decomposed into its components by Singular Value Decomposition (SVD), and this is used to extract

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

redundant features. This transformation is described in such a way that the first principal variable has the greatest possible variance. The mathematical concept of PCA is as follows: for a given input central x_t where t = 1, 2,…, p, $x_t \in R^n$and $\sum_{t=1}^{p} x_t = 0$, usually $n < p$, so that x is $n \times p$ input matrix. PCA transforms this vector space into new vectors by solving the Eigen values problem given in

$$\lambda_i \mu_i = c\mu_i \; i = 1,2,...,n \quad [4]$$

Where $\lambda_i \geq 0$ is one of the n Eigen values of the $n \times p$ covariance matrix c, where $c = (\frac{1}{p}(\sum_{t=1}^{p} x_t x_t^T))$, and $\mu_i$is the Eigenvector whose corresponding to the Eigen value λ_i, then the PCs $y_t$are calculated as the orthogonal transformation of x_t;

$$y_t(i) = \mu_{1i}x_1 + \mu_{2i}x_2 + \ldots\ldots\ldots + \mu_{ni}x_n = \mu_i^T x_t \quad i = 1,2,...,n \quad [5]$$

The primary $y_t(i)$variable with the highest Eigen value reflects the largest variance vector in the data collection. Only by using the first few Eigenvectors ordered in descending order by their corresponding Eigen values; by using only the first few Eigenvectors ordered in decreasing order by their corresponding Eigen values, it is possible to reduce the initial vector space PCs to new PC vectors k < n.

Table.2.The Outcome of the Performance Measure of PCA -PLSDA

Dataset (Merge)	Techniques	Accuracy Score(%)	Precision (%)	Sensitivity(%)	Specificity (%)	F1-Score(%)	AUC-Score(%)
Diabetes and Heart Disease	PCA(PLSDA)	95.05	88.00	91.67	96.10	89.80	93.89

Table 2 above. Demonstrates the output metrics of the PLSDA classification with the Merge Dataset (Diabetes and Heart Disease Dataset).

The patterns are recognized in non-linear forms by t-SNE. In order to transform data points into a lower-dimensional data representation, local approaches (The mapping of adjacent points in the low-dimensional data environment on the manifold to adjacent points) and global approaches (Geometry preservation at all scales, i.e. mapping neighboring points to neighboring points on a manifold and mapping far away points to faraway points in the low-dimensional data setting.) are used. It assesses the probability similarity of points in high dimensional representation and low dimensional representation. Basically, in high or low dimensional expanses between points, it assesses the Euclidean distances and transforms these distances to conditional probabilities to represent similarities. The t-SNE mechanism has many drawbacks, however, such as lack of wide-scale information, longer execution times, inability to represent larger datasets, etc. Where the dataset is not large, t-SNE operates effectively because there is non-linear dependence between the data characteristics.In order to minimize elevated loading times, GPU-accelerated tSNE implementations (such as Barnes-Hut method, RAPIDs, etc.) are commonly used. In the series of experiments in this work, however, only the CPU simulations of it's DR techniques were performed.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

Uniform Manifold Approximation and Projection (UMAP):

UMAP is a nation DR approach that, as opposed to t-SNE, retains much of the local and global structure. The k-Nearest Neighbor Theory is used. The time between high-dimensional feature extraction points and projects on lower-dimensional feature extraction is defined in the lower dimensional setting and Stochastic Gradient Descent (SGD) is used to minimize the distance. In addition, it has the following benefits, such as managing massive datasets, faster turnaround time as compared to t-SNE, and protection of local structure and global data structure.

Table.3.The Outcome of the Performance Measure of UMAP -PLSDA

Dataset (Merge)	Techniques	Accuracy Score	Precision	Sensitivity	Specificity	F1-Score	AUC-Score
Diabetes and Heart Disease	UMAP(PLSDA)	71.67	69.12	68.12	74.55	68.62	71.31

The above Table.3. shows the performance measures of the classification PLSDA with the Merge Dataset (Diabetes and Heart Disease Dataset).

Counter Propagation Network (CPN)

It depends on the mixture of the details, severe, and yield layers, multilayer feed forward organization. There is a CPN model in Staroutstar. Input yield knowledge preparation is carried out by a three-layer neural organization, that is, generating yield in the reaction to a data vector dependent on competitive learning. In CPN, the relation between the input layer and the serious layer is instar construction, and the outstar arrangement is aligned with the serious and yield layer. Two-stage planning steps are used in the Counter Propagation Network. The input vector is packed based on Euclidean distance in the first stage and the organization's appearance is enhanced using direct geography. We evaluated BMN for the Euclidean distance between the info and weight vector. In stage II, by changing the loads from extreme layer to yield layer, the optimal reaction is obtained. Let x=[$x1$ $x2$ ⋯xn] and y=[$y*$ 1 $y*$ 2 ⋯$y*m$] be input and wanted vector, separately, let Vij be weight of information layer and serious layer, where 1≤i≤n and 1≤j≤p, and let wjk be the load between serious layer and yield layer, where 1≤k≤m. Euclidean distance between input vector x and weight vector Vij is

$$Dj = n \sum i=1 (xi - \mathrm{V}ij) \text{ [6]},$$

where j = 1, 2, . . . ,p.

The architecture of CPN is shown in Figure 3.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

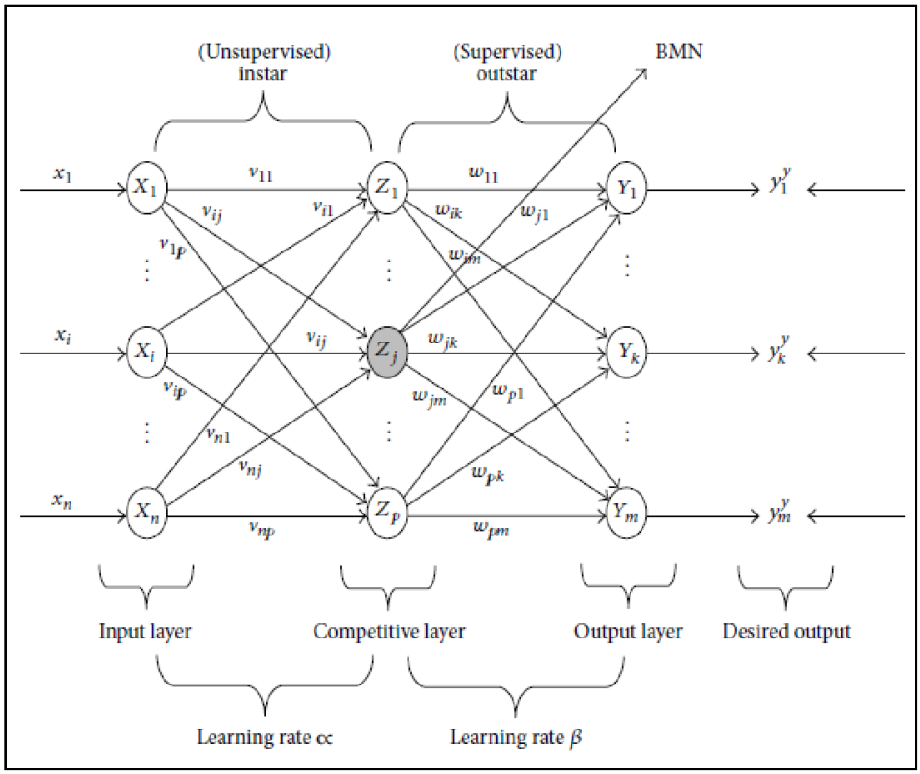

Figure.3. Architecture of Counter Propagation Network

Supervised DR:

Linear Discriminant Analysis:

Projecting the original data matrix into a space with a lesser dimension is the goal of the LDA approach. Three actions needed to complete this. Finding the inter-class variance(i.e. the distance between the average of different classes) is the first step. Calculating within-class variance b/w the samples and the mean of each class, is the second step. Maximizing the variance b/w and minimizing the variance within classes to build a tiny dimensional space is the third step.

LDA is the most popular method in reducing dimensions in supervised classification problems. It is used in groups to model distinctions, i.e. to distinguish two or more classes. It is used to project the characteristics into a lower dimension space in higher dimensional space.

We have two classes, for instance, and we need to efficiently distinguish them. Classes may have different features. As shown in the figure below, using just a single feature to distinguish them can result in some overlap. So, for proper classification, we will keep increasing the number of features.

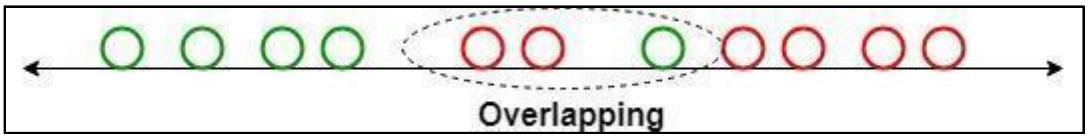

Figure.4.Diagramatic Representation of the Merging Dataset.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

It comprises your data's statistical properties, determined for all classesFor any input vector (x), for any class, this is the variance and the mean of that variable. The means and the covariance matrix are calculated with several variables over the multivariate Gaussian.

LDA uses above statistical properties calculated from input data to make assumptions. The Figure.4.is the pictorial representation of the Merging Dataset of which the variables are combined to specify the required dataset.

Some simpler assumptions about your data are made by LDA:

1. As we use Gaussian as input, when plotted, each variable forms a bell like curve.
2. Variance of every attribute are same, i.e., each variable's value varies by the same value on average around the mean.

The Variance and mean of every class is predicted by the LDA model is from those assumptions made from the input data. For univariate with two classes, it's enough to assume like this.

The average value(mu) for every input(x) from every class(k) is calculated by:

$$mu\_k = (1 / n\_k) * sum(x) \quad [7]$$

If mu_k for class k is the mean value of x, n_k is the number of instances of class k. The variance is measured for both groups as the average square deviation of the average value.

$$sigma^2 = 1 / (n-K) * sum((x - mu)^2) \quad [8]$$

For all inputs (x), where sigma^2 is the variance, n is the number of instances, K is the number of groups, and mu is the x input mean.
The LDA makes predictions by calculating the probability that each class belongs to a new set of inputs. The class that gets the highest likelihood is the output class, and a prediction is made.
The formula uses the Bayes Theorem to predict the chances. The Bayes formula can be used to calculate the likelihood of the output class (k) using the probability of each class and the probability of the data belonging to each class, provided the input (x) is given:

$$P(Y=x|X=x) = (PI\_k * f\_k(x)) / sum(PI\_l * fl(x)) \quad [9]$$

Where the simple probability observed for each class (k) in your training data corresponds to PI kk (e.g. 0.5 for a 50-50 split in a two class problem). This is considered the prior likelihood in Bayes' Theorem.

$$PI\_k = n\_k/n \quad [10]$$

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

The f(x) above is the approximate probability that x belongs to the class. For f, a distribution characteristic of Gaussian is used (x). By using the following equation, we end up plugging the Gaussian into the above equation by simplifying it. This is called a discriminatory function, and by the performance classification, the class is calculated as having the greatest value (y):

$$D_k(x) = x * (mu_k/sigma^2) – (mu_k^2/(2*sigma^2)) + ln(PI_k) \quad [11]$$

From your results, both mu_k, sigma^2 and PI_kare determined, with input x given for class k. The discriminating attribute of D_k is (x).

PLS (Partial Least Square):

PLS is a collection of methods for modeling relationships using latent variables between blocks of observed variables. The fundamental theory of PLS is that a system or method that is guided by a few latent (not necessarily observable or measured) variables produces the observed data. Therefore, PLS attempts to distinguish the original predictor variables (latent components) that have strong covariance with the response variables by uncorrelated linear transformations. PLS predicts the variables of the y response and reconstructs the initial matrix X simultaneously on the basis of these latent components. Let matrix T = [t1, . . . ,tK] ∈ R n×K represents the n observations of the K components which are usually denoted as latent variables (LV) or scores. The relationship between T and X is defined as:

$$T = XV \quad [12]$$

where$V = [v1, \ldots, vK] \in R\, p \times K$ is the matrix of projection weights. The projection weights V are determined by PLS by optimizing the covariance between the reaction and latent components. X and y are decomposed on the basis of these latent elements as:

$$X = T\,PT + E \quad [13]$$

$$y = T\,QT + f \quad [14]$$

where$P = [p1, \ldots, pK] \in R\, p \times K \; and \; Q = [q1, \ldots, qK] \in R\, 1 \times K$ are denoted as loadings of X and y respectively. Typically, ordinary least squares are determined by P and Q (OLS). E and f are residuals of X and y, respectively. Response values are determined by the latent variables being decomposed by X and y, not by X. (not direct at least). It is believed that this model would be more accurate than OLS if the latent variables coincide with the real underlying structure of the original data. By projecting X on the weights V of the elements, the key point of PLS is the construction. The classical criteria of PLS is to sequentially maximize the covariance between solution y and latent components. There are many variants of PLS approaches to solve this problem. By disregarding the miner contradictions of these algorithms, we view the most commonly used PLS approach: PLS1. The first latent variable t1 = Xw1 is determined by PLS1 by minimizing covariance between y and t1 under the constraint k w1 k = 1. The corresponding objective function is:

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

$$w1 \ = \ arg \ max \ wT \ w = 1 \ (Cov(Xw, y)) \ [15]$$

Using the Lagrange multiplier form, the maximization question of Equation (1) can be easily solved.

$$w1 \ = \ X \ T \ y / \ k \ X \ T \ y \ k \ [16]$$

To sequentially remove other latent components, we need to model the X and y residual information that previous latent variables did not define. Therefore, after the extraction of the score vector t1, PLS1 deflates matrices X and y by subtracting their rank-one approximations based on t1. The X and y matrices are deflated as:

$$E1 \ = \ X \ - \ t1p \ T \ 1 \ f1 \ = \ y \ - \ t1q \ T \ 1 \ [17]$$

where p1 and q1 are loading determined by OLS fitting:

$$p \ T \ 1 \ = \ (t \ T \ 1 \ t1) \ - 1 \ t \ T \ 1 \ X \ q \ T \ 1 \ = \ (t \ T \ 1 \ t1) \ - 1 \ t \ T \ 1 \ y \ [18]$$

As an iterative process, PLS1 builds other latent components in turn by using the residuals E1 and f1 as new X and y.

Models based on variables from l < m are the first type of regularization. This means that regularization is accomplished by fitting a model into a lower dimensional space. The theory of partial least squares (PLS) is to factor in both the solution matrix Y and the regression matrix X:

$$\boldsymbol{X} \ = \ \boldsymbol{T} \ \boldsymbol{P}^{T} \ + \boldsymbol{E}$$
$$\boldsymbol{Y} \ = \ \boldsymbol{U} \ \boldsymbol{Q}^{T} \ + \ \boldsymbol{F} \ [19]$$

where the covariance between T and U is maximized. X is an indicator matrix n x m. Y is a matrix with responses n x p. T and U are matrices of n x l, which are projections of X and Y, respectively. P and Q are, respectively, orthogonal matrices of m x l and p x l. E and F are additive noise terms that are considered to be spontaneously distributed independently.

$$wk \ = \ E \ T \ k - 1 \ fk - 1 / \ k \ E \ T \ k - 1 \ fk - 1 \ k \ [20]$$

$$tk \ = \ Ek - 1wk \ [21]$$

$$p \ T \ k \ = \ (t \ T \ k \ tk) \ - 1 \ t \ T \ k \ Ek - 1 \ [22]$$

$$q \ T \ k \ = \ (t \ T \ k \ tk) \ - 1 \ t \ T \ k \ fk - 1 \ [23]$$

$$Ek \ = \ Ek - 1 \ - \ tkp \ T \ k \ [24]$$

$$fk \ = \ fk - 1 \ - \ tkq \ T \ k \ [25]$$

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

For ease of speech, the matrices X and y are also denoted as E0 and y0, respectively. The number of components is a PLS parameter that the user can set or determine through a cross-validation scheme. In general, the maximum number of latent components is the rank of a matrix X that has non-zero covariance with y. PLS decreases the difficulty of original data analysis by creating a few fresh predictors, T, that are used to replace the vast number of original features. In contrast, the components of the PLS are typically more predictive than those obtained by other unmonitored strategies, such as PCA, obtained by optimizing the covariance between the components and the variables of the response. Classification methods and other approaches to mathematical analysis can be used after dimension reduction, based on these new predictors. The PLSDR solution uses the derived score vector T as the new representation of the original results, where T = X x V. The reciprocal orthogonality of the derived score vectors T, i.e., is evidently guaranteed by the PLS deflation scheme. TT = I. However, the projection vectors V are non-orthogonal. As we understand, the PLS deflation method maintains the reciprocal orthogonality of the extracted score vectors T.. By the arguments it can be seen that the weights $W = [w1, \ldots, wK] \in R\,p \times K$ are also orthogonal. Furthermore, the relation between V and W was demonstrated as:

V = W(P TW) -1, [26] from which, on the residual matrix Ek, we evade the iterative construction of latent components T, but directly relate T to X. Loading vectors P and Q are not, in general, orthogonal. From Equation [26], we can also see that the projection weights V are not orthogonal. We assume the orthogonality of projection vectors for the implementation of dimension reduction is more significant than that of score vectors. As a result, we are contemplating using orthogonal projection vectors W instead of non-orthogonal ones V

Performance Analysis

Table.4.The Overall Outcome of the 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

Tables 1,2 and 3, show the classification results of proposed PLSDA selected features that obtained using ICA,PCA and UMAP respectively.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

Performance Metrics:

Accuracy: This shows correctly classified instances percentage in course of classification. It is evaluated as

$$Accuracyrate = \frac{TruePositive + TrueNegative}{TotalInstances} * 100$$ [26]

Precision: It measure gives what proportion of data that transmit to the network, actually had intrusion. The predicted positives (Network predicted as intrusion is TP and FP) and the network actually having a intrusion are TP. This is used to measure the quality and exactness of the classifier as shown below:

$$Precision = \frac{Truepositive}{Truepositive + FalsePositive}$$ [27]

Recall:

Recall is the ratio Real Positives which are correct Predicted Positive and is defined as

$$Recall = \frac{TruePositive}{TruePositive + Falsenegative}$$ [28]

F1 Score: F1 Score is basically the mean value of precision and recall. Also statistical measure is used in F1 score to performance rate of individual classifier of FN and FP. Definition of precision is judgment of accuracy whereas recall is detecting the sample instance based on the attribute called faulty or non faulty.

$$F1 - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$ [29]

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. Accuracy performance analysis is used as a powerful measurement analysis in machine learning techniques with higher analytical values. Accuracy is a well-known performance metric for algorithm evaluation and comparison, and it has been regarded as an alternative measure to compare classifier performance over the entire range of class distributions. The estimated outcome is used to construct the Accuracy analysis after each for the unsupervised and supervised dimensionality selection methods. The Accuracy analysis is used to determine the diabetes from the Diabetes and heart disease data set and the merged data value is used to evaluate the diabetes of the patient from the whole data set. The dimensionality reduction method is applied as the priority measure to determine the features selection and it has been analysed that the PCA with PLSDA has the highest accuracy than the other dimensionality reduction methods. Table 4. shows some of the observation of Diabetes and Heart Disease Dataset, the outcome of the classifier has been estimated from the Diabetes and Heart Disease datasets

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

instances. Then, classifying the instances with the same observation is performed. Next, the performance measures of the various DR techniques of ICA,PCA and UMAP are calculated. Finally, these results are classified with that of the hybrid PLSDA. Table 4. shows the comparison of the performance in terms of Accuracy, precision, recall and F1-score. It has been analysed from the actual and predicted values from the objective of classes in confusion matrix and it is represented in percentage.

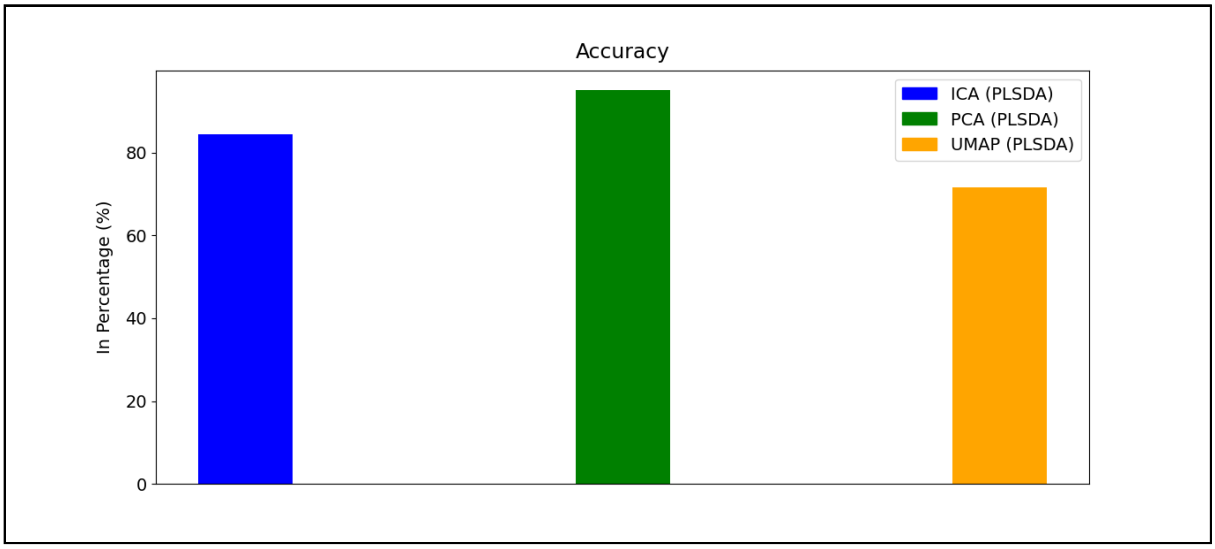

Figure.5. Comparison of Accuracy for various DR techniques

Figure.5. is the graphical representation of the table 4 contents. It shows the Accuracy comparison for Diabetes and Heart Disease Dataset among the DR techniques. As shown in this figure, the PCA achieves the maximum Accuracy percentage compared with the other techniques. Whereas, the UMAP approach obtained the worst performance by furnishing the Accuracy value of about 71.6 % . Simultaneously, the ICA model acquires more Accuracy compared to the previous one of about 84.49 % . Finally, the PCA technique operates more efficiently after the merged dataset compared with the other models performance by acquiring the maximum Accuracy value of 95.05% .

Figure.6. is the graphical representation of the table 4 contents. It shows the Precision comparison for Diabetes and Heart Disease Dataset among the DR techniques. As shown in this figure, the ICA achieves the maximum Precision percentage before merging the dataset

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

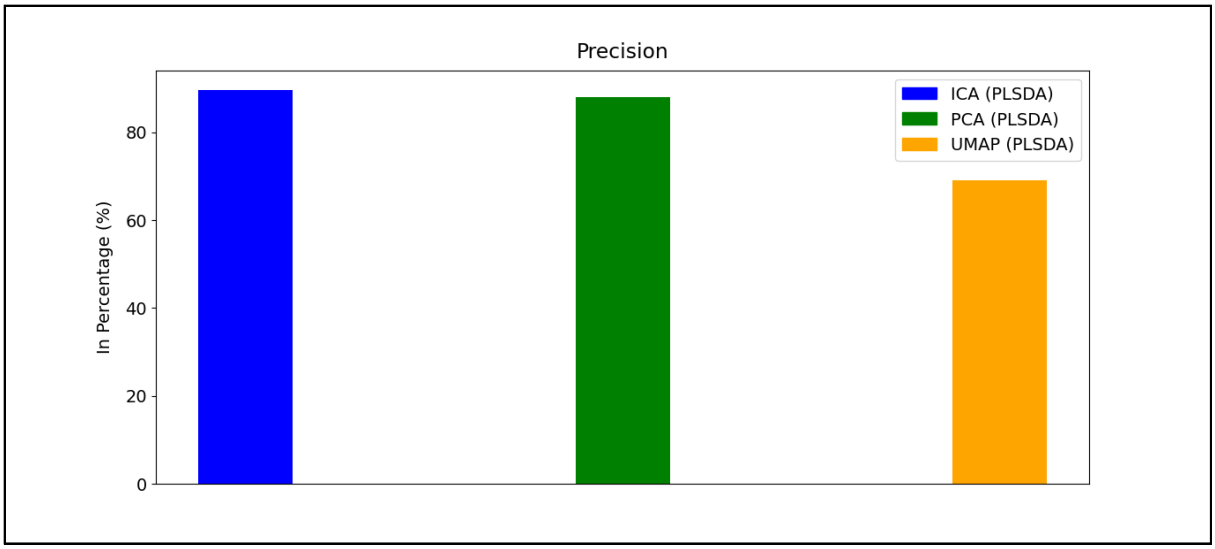

Figure.6. Comparison of Precision for various DR techniques

compared with the other techniques. Whereas, the UMAP approach obtained the worst performance by furnishing the Precision value of about 69.12 % . Simultaneously, the PCA model acquires more Precision compared to the previous one of about 88.00 % . Finally, the ICA technique operates more efficiently compared with the other models performance by acquiring the maximum Accuracy value of 89.57% .

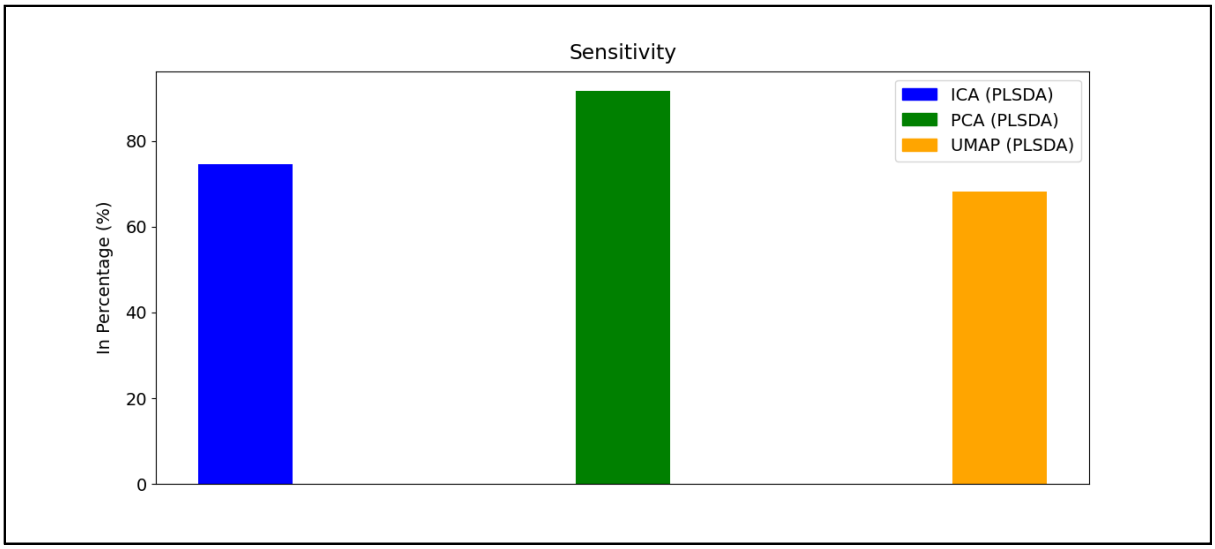

Figure.7. Comparison of Sensitivity for various DR techniques

Figure.7.is the graphical representation of the Table.4.contents. It shows the Sensitivity comparison for Diabetes and Heart Disease Dataset among the DR techniques. As shown in this figure, the PCA achieves the maximum Sensitivity percentage compared with the other techniques. Whereas, the UMAP approach obtained the worst performance by furnishing the Sensitivity value of about 68.12 % . Simultaneously, the ICA model acquires more Sensitivity compared to the previous one of about 74.64 % . Finally, the PCA technique operates more efficiently with the merged dataset compared with the other models performance by acquiring the maximum Sensitivity value of 89.57% .

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

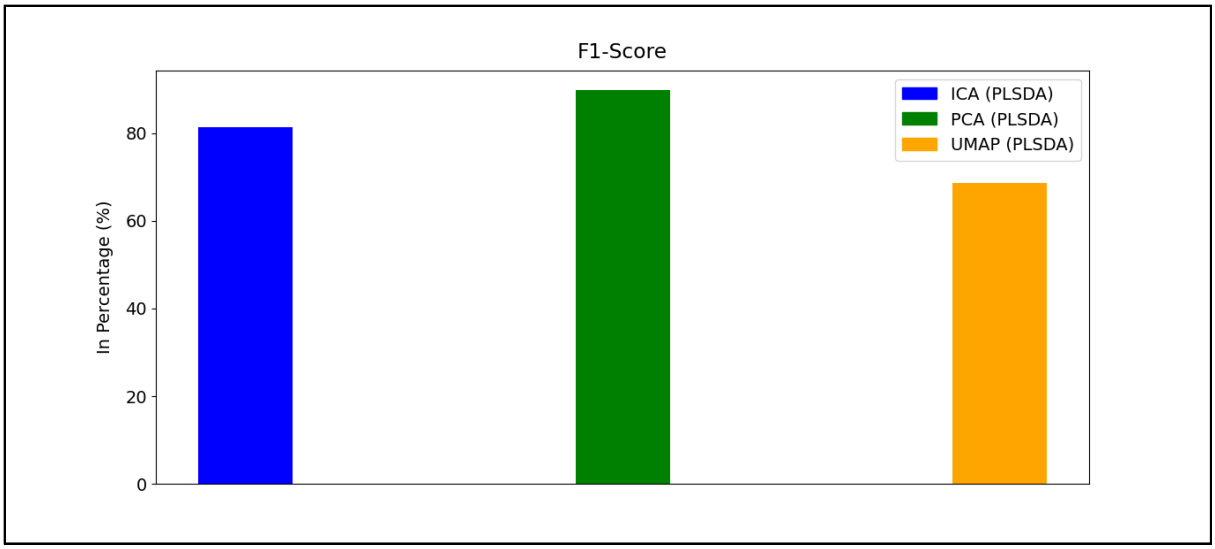

Figure.8. Comparison of F1-Score for various DR techniques

Figure.8.is the graphical representation of the Table.4.contents. It shows the F1-Score comparison for Diabetes and Heart Disease Dataset among the DR techniques. As shown in this figure, the PCA achieves the maximum F1-Score percentage compared with the other techniques. Whereas, the UMAP approach obtained the worst performance by furnishing the F1-Score value of about 68.72 % . Simultaneously, the ICA model acquires more F1-Score compared to the previous one of about 81.43 % . Finally, the PCA technique operates more efficiently with the merged dataset compared with the other models performance by acquiring the maximum F1-Score value of 89.80% .

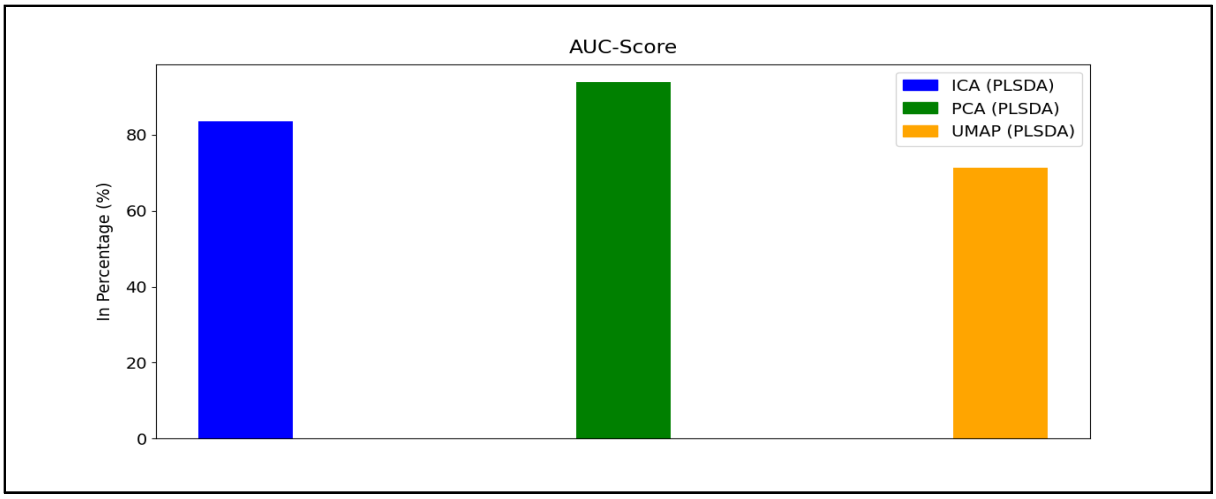

Figure.9. Comparison of AUC-Score for various DR techniques

Figure.9.is the graphical representation of the Table.4.contents. It shows the AUC-Score comparison for Diabetes and Heart Disease Dataset among the DR techniques. As shown in this figure, the PCA achieves the maximum AUC-Score percentage compared with the other techniques. Whereas, the UMAP approach obtained the worst performance by furnishing the AUC-Score value of about 71.31 % . Simultaneously, the ICA model acquires more AUC-Score compared to the previous one of about 83.68 % . Finally, the PCA technique operates more efficiently with the merged dataset compared with the other models performance by acquiring the maximum AUC-Score value of 93.89% .

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

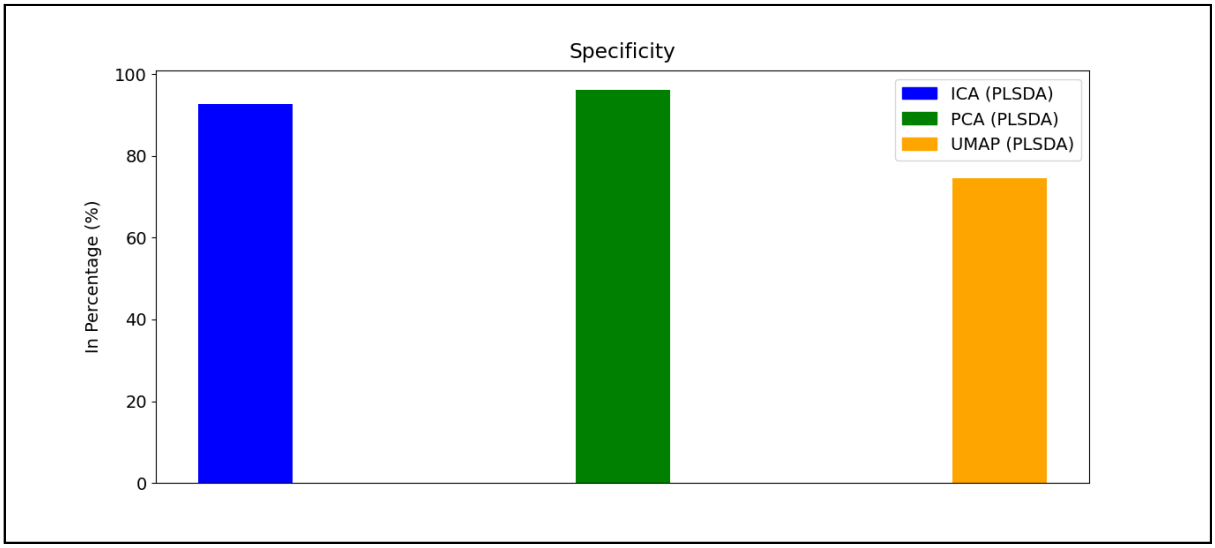

Figure.10. Comparison of Specificity for various DR techniques

Figure.10.is the graphical representation of the Table.4.contents. It shows the Specificity comparison for Diabetes and Heart Disease Dataset among the DR techniques. As shown in this figure, the PCA achieves the maximum Specificity percentage compared with the other techniques. Whereas, the UMAP approach obtained the worst performance by furnishing the Specificity value of about 74.55 % . Simultaneously, the ICA model acquires more Specificity compared to the previous one of about 92.73 % . Finally, the PCA technique operates more efficiently with the merged dataset compared with the other models performance by acquiring the maximum Specificity value of 96.10% .

Conclusion:

Identifying the risk of diabetes at its early stage is one of the global health challenges. This research aims to structure a system that predicts the likelihood of type 2 diabetes mellitus. This paper compares the output of the most common approaches to reducing simple dimensionality.Numerous statistical analyses and multiple performance metrics are included in the comparison, resulting in a clear and accurate conclusion. In terms of accuracy, sensitivity, specificity, F-score, precision, AUC and ROC analysis, the performance of these approaches is assessed.The comparison is performed by experiments conducted on different types/sizes of data sets, including data sets for heart and diabetes expression. This systematic analysis therefore offers a detailed insight into these widely used methods of selection for practical use with DR. The comparison shows that while PCA is a conventional approach, the modern techniques are still unable to outperform it.In DR, the characteristics must be uncorrelated but not separate as in the classifier of naı̈ve bays , so use of ICA does not give a significant meaning. This is empirically approved by the experiment, where, relative to the other methods, ICA appears to work less accurately.Finally, it can be mentioned that DR has proven to be an efficient high-dimensional data set classifier and also provides good efficiency when using the methods of selection of features.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5350 - 5369
Received 15 December 2020; Accepted 05 January 2021.

Reference:

1. Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." *2014 science and information conference*. IEEE, 2014.
2. Kavitha R.J., Avudaiyappan T., Jayasankar T., Selvi J.A.V. (2021) Industrial Internet of Things (IIoT) with Cloud Teleophthalmology-Based Age-Related Macular Degeneration (AMD) Disease Prediction Model. In: Gupta D., Hugo C. de Albuquerque V., Khanna A., Mehta P.L. (eds) Smart Sensors for Industrial Internet of Things. Internet of Things (Technology, Communications and Computing). Springer, Cham.pp.161-172 https://doi.org/10.1007/978-3-030-52624-5_11
3. S. Venkatraman, P. Muthusamy, Bhanuchander Balusa, T. Jayasankar,G. Kavithaa · K. R. Sekar,C. Bharatiraja "Time dependent anomaly detection system for smart environment using probabilistic timed automaton," *Journal of Ambient Intelligence and Humanized Computing (2020)*, https://doi.org/10.1007/s12652-020-02769-3
4. Espadoto, Mateus, et al. "Towards a quantitative survey of dimension reduction techniques." *IEEE transactions on visualization and computer graphics* (2019).
5. Abdulhammed, Razan, et al. "Features dimensionality reduction approaches for machine learning based network intrusion detection." *Electronics* 8.3 (2019): 322.
6. Xu, Xinzheng, et al. "Review of classical dimensionality reduction and sample selection methods for large-scale data processing." *Neurocomputing* 328 (2019): 5-15.
7. Becht, Etienne, et al. "Dimensionality reduction for visualizing single-cell data using UMAP." *Nature biotechnology* 37.1 (2019): 38-44.
8. Ali, Liaqat, et al. "LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine." *Neural Computing and Applications* (2020): 1-10.
9. Zhang, Bing, et al. "Network intrusion detection method based on PCA and Bayes algorithm." *Security and Communication Networks* 2018 (2018).
10. Ogbuanya, ChisomEzinne. "Improved Dimensionality Reduction of various Datasets using Novel Multiplicative Factoring Principal Component Analysis (MPCA)." *arXiv preprint arXiv:2009.12179* (2020).
11. Zhu, Tao, et al. "An online incremental orthogonal component analysis method for dimensionality reduction." *Neural Networks* 85 (2017): 33-50.
12. Zhu, Yani, Chaoyang Zhu, and Xiaoxin Li. "Improved principal component analysis and linear regression classification for face recognition." *Signal Processing* 145 (2018): 175-182.
13. Reddy, G. Thippa, et al. "Analysis of dimensionality reduction techniques on big data." *IEEE Access* 8 (2020): 54776-54788.
14. Xu, Xinzheng, et al. "Review of classical dimensionality reduction and sample selection methods for large-scale data processing." *Neurocomputing* 328 (2019): 5-15.
15. García-Gil, Diego, et al. "Principal components analysis random discretization ensemble for big data." *Knowledge-Based Systems* 150 (2018): 166-174.

16. Kaur, Devinder, et al. "Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: SDN perspective." *IEEE Transactions on Knowledge and Data Engineering* 30.10 (2018): 1985-1998.

17. Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." *International Journal of Engineering Research and Applications* 3.2 (2013): 1797-1801.

18. Salo, Fadi, Ali Bou Nassif, and Aleksander Essex. "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection." *Computer Networks* 148 (2019): 164-175.

19. Kumar Dewangan, Amit, and Pragati Agrawal. "Classification of diabetes mellitus using machine learning techniques." *International Journal of Engineering and Applied Sciences* 2.5 (2015): 257905

20. J. Jayanthi · E. Laxmi Lydia · N. Krishnaraj · T. Jayasankar · R. Lenin Babu · R. Adaline Suji," An effective deep learning features based integrated framework for iris detection and recognition," *J Ambient Intell Human Comput* (2020). https://doi.org/10.1007/s12652-020-02172-y

21. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." *arXiv preprint arXiv:1502.03774* (2015).

22. J.Jayanthi, T.Jayasankar, N.Krishnaraj, N.B.Prakash, A.Sagai Francis Britto, K.Vinoth Kumar, "An Intelligent Particle Swarm Optimization with Convolutional Neural Network for Diabetic Retinopathy Classification Model*," Journal of Medical Imaging and Health Informatics* (2020), Volume 11, Number 3, March 2021, pp. 803-809, https://doi.org/10.1166/jmihi.2021.3362

23. M. T. Habib and M. Rokonuzzaman, "An empirical method for optimization of counter propagation neural network classifier design for fabric defect inspection," International Journal of Intelligent Systems and Applications, vol. 6, no. 9, pp. 30–39, August 2014.

24. Mohamed Yacin Sikkandar, T. Jayasankar,· K. R. Kavitha, N. B. Prakash, Natteri M. Sudharsan,G. R. Hemalakshmi, "Three Factor Nonnegative Matrix Factorization based HE Stain Unmixing in Histopathological Images," *Journal of Ambient Intelligence and Humanized Computing (2020),* https://doi.org/10.1007/s12652-020-02265-8

25. A.Sheryl Oliver,M.Anuratha,M.Jean Justus,Kiranmai Bellam, T.Jayasankar, "An Efficient Coding Network Based Feature Extraction with Support Vector Machine Based Classification Model for CT Lung Images," *J. Med. Imaging Health Inf.* ,vol.10,no.11.pp.2628–2633(2020).