



This is a repository copy of *A dimension-reduction based multilayer perception method for supporting the medical decision making*.

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/153665/>

Version: Accepted Version

Article:

Lee, S.-J., Tseng, C.-H., Lin, G.T.-R. et al. (4 more authors) (2020) A dimension-reduction based multilayer perception method for supporting the medical decision making. *Pattern Recognition Letters*, 131. pp. 15-22. ISSN 0167-8655

<https://doi.org/10.1016/j.patrec.2019.11.026>

Article available under the terms of the CC-BY-NC-ND licence
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

A Dimension-Reduction Based Multilayer Perception Method for Supporting the Medical Decision Making

Shin-Jye Lee^a, Ching-Hsun Tseng^a, G. T. -R. Lin^a, Yun Yang^{b*}, Po Yang^c, Khan Muhammad^d, Hari Mohan Pandey^e

^a Institute of Management of Technology, National Chiao Tung University, Hsinchu 300, Taiwan

^b School of Software, Yunnan University, Kunming, China

^c Department of Computer Science, Sheffield University, Sheffield, UK

^d Department of Software, Sejong University, Seoul 143-747, Republic of Korea

^e Department of Computer Science, Edge Hill University, UK

ARTICLE INFO

Article history:

Received

Received in revised form

Accepted

Available online

Keywords: *Deep Learning, Multilayer Perceptron, Weight Initialization, Medical Decision Support*

ABSTRACT

Due to the rapid development of Medical IoT recently, how to effectively apply these huge amounts of IoT data to enhance the reliability of the clinical decision making has become an increasing issue in the medical field. These data usually comprise high-complicated features with tremendous volume, and it implies that the simple inference models may less powerful to be practiced. In deep learning, multilayer perceptron (MLP) is a kind of feed-forward artificial neural network, and it is one of the high-performance methods about stochastic scheme, fitness approximation, and regression analysis. To process these high uncertain data, the proposed work based on MLP structure in particular integrates the boosting scheme and dimension-reduction process. In this proposed work, the advanced ReLU-based activation function is used. Also, the weight initialization is applied to improve the stable prediction and convergence. After the improved dimension-reduction process is introduced, the proposed method can effectively learn the hidden information from the reformative data and the precise labels also can be recognized by stacking a small amount of neural network layers with paying few extra cost. The proposed work shows a possible path of embedding dimension reduction in deep learning structure with minor price. In addition to the prediction issue, the proposed method can also be applied to assess risk and forecast trend among different information systems.

2015 Elsevier Ltd. All rights reserved.

1. Introduction

As for processing either high-dimensional or high-uncertain data[1], Multilayer Perceptron (MLP) can be regarded as an effective model because of its compact structure and adaptive mechanism. Especially for Medical IoT (MIoT) based data, these data usually comprise high-complicated features with tremendous volume, such as high-related relation among features or different bias among relations[2]. To address the problem above, a reliable method is essentially required to be developed based on more efficient and robust performance, and it can also avoid the disadvantage brought by the less-powerful feature selection under the problem of unbalanced data as much as possible[3]. Especially for processing huge amounts of similar features, it is tough for an inspector to accurately distinguish the index and then verify it only depending on a few indexes. Traditional methods include

ensemble learning [4] [5] [6] [7] and semi-supervised learning [8] [9] can achieve a satisfactory performance, while a very huge computational cost usually occurs in processing these types of data, and it particularly appears in processing the data generated by MIoT-based systems. On the basis of MLP structure, the objective of the proposed method is thus designed to decrease the computing cost in each layer but still keep a reliable performance or even better accuracy compared with some traditional methods. Meanwhile, under the insufficient medical resource, this feature can also bring a good effort to the medical staff making the therapy decision in a rush time because of the critical situation of patients. Further, the proposed method can process different clinical diagnosis works, such as brain micro wave and picture distinguishing on tumors. Also, the proposed method can be illustrated in the Fig. 1. As shown in the Fig. 1, the proposed method comprises four phases, including MIOT-Based Health

* Corresponding author. e-mail: yangyan19@hotmail.com

Care Systems, Data Preprocessing, Modified Multilayer Perceptron Networks with LDA-Based Input Layer, and Decision Making. In the Phase 1, the raw data are collected from a variety of health care systems (IoT and Non-IoT), and these data are picked as their compactness without missing values. In the Phase 2, the data are preprocessed to address the problem of imbalanced data, to reinforce their completeness. In the Phase 3, the data are processed by the improved MLP model with the technique of dimension reduction (LDA-based input layer). In the Phase 4, the final value is outputted for the purpose of decision making.

The organization of this paper is arranged as follows. The related work is presented in the Section 2. Then, the mechanism of the proposed method, Dimension-Reduction based Multilayer Perceptron (DR-MLP), is shown in the Section 3. In the Section 4, the simulation experiments and corresponding results are discussed. Finally, the summary is given in the Section 5.

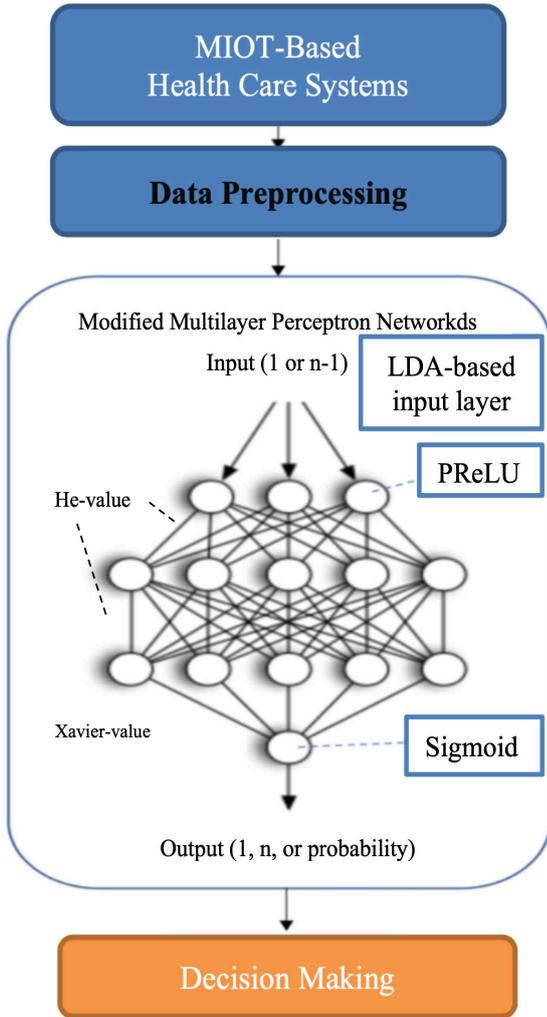


Figure 1. A graphical abstract of the proposed work and its context

2. Related works

2.1. Linear Discriminant Analysis (LDA)

In the Industry 4.0 era, the raw data are produced exponentially by a variety of industrial systems, and these data are usually either high-dimensional or high-complicated. As these raw data may carry a range of similar features, it is difficult to recognize each feature clearly. Only exam the potential meaning of each feature and then remove redundant features is possibly the reasonable way to make the features clear. However, this solution is less efficient and may lose useful information in processing high-related features, so the method of mapping the high-dimensional data the

lower dimension provides another solution to address such high-cost computational process. In accordance with this function, the algorithm of Linear Discriminant Analysis (LDA) [10] [11] can properly squeeze all dimensional signals (or vectors) into one dimension if the label type is two. In other words, the number of output dimensions is obtained by $n-1$ label types. Linear Discriminant Analysis, called Fisher Linear Discriminant, is one of linear classification, and its purpose aims to classify the data in an easy manner. Basically, the LDA method is a supervised learning method, and its distribution is Gaussian density distribution. If there are two classes in the data space, the LDA method maps these two classes into one dimension, in which data assigned in the same class keep as close as possible and data assigned in different classes keep as far as possible. In the light of the past works, the LDA method cannot only make convergence faster but also maintain the same accuracy. It somewhat avoids the risk of selecting feature improperly. Besides, as the process LDA is applied by the linear transformation, the output will be intuitional and low-cost. Compared with other feature selection methods, such as PCA and KPCA, the LDA method usually brings more effective performance to the precise prediction, and which can be observed from the experiments in the Section 4.

2.2. Multilayer Perceptron (MLP)

On the basis of MLP-based structure, the simplest structure is called “single-layer perceptron” [12], which is only one layer between the input layer and output layer. Its original objective aims to address the problem of linear separability, which the data is linearly separable, and it is only applied to classify binomial labels. Initially, a similar network structure was described by Warren McCulloch and Walter Pitts in the 1940s, and at present the highly computational performance can develop more complicated MLP mechanism. The MLP is also known as the initial deep structure, and it can structure a more efficient model on separating non-linear problem. Through stacking multiple layers of perceptron, the MLP indeed produces a good performance and saves the computational cost. Also, embedding suitable activated neuron after each layer and combining these neurons together can amplify the effect and avoid being affected by abnormal information, and this mechanism brings a concrete foundation to the development of deep learning. Moreover, several evidences and competition results support that deep layers of perceptron can attain a better performance than most traditional methods. In addition, the MLP can effectively address the XOR problem by decreasing the number of neurons in one layer and stacking multiple layers in a structure, which cannot be settled by using single layer perceptron[13] [14] [15].

As for the development of deep learning, its application in image classification provides a good effort in image denoising [16], image compression [17], image painting [18], and super-resolution [19], and other advanced works contributed to different image researches [20] [21] [22] [23] [24] [25] [26] [27]. In addition, it has been applied to traffic prediction [28] [29] [30]. The above works have already showed a possible way of applying image classification to solve the practical problems in the real world based on the deep structure of neural network. Also, some works [31] [32] show that using batch normalization [33] can help to maintain the gradient effect and prevent overfitting problem by gradient vanishing. Furthermore, introducing cross-channel correlations and incorporations as new combination of features and spatial attention, respectively, shows that processing spatial structure [34] [35] [36] [37] [38] can be helpful to enhance the performance. Although there are a series of good solutions to address current problems in deep structure of neural network, these problems can be relieved slightly instead of being eradicated.

2.3. Parametric Rectified Linear Unit

Parametric Rectified Linear Unit (PReLU) [39] is one of the modified ReLU activation models, and it includes Leaky ReLU (LReLU) and ReLU. As for the definition of ReLU, once the input value is less than 0, the input parameter will be set as 0. If the input parameter is close enough to 0, the PReLU will turn to Leaky ReLU (LReLU). Even though the input parameter of LReLU is almost to 0, it can be observed that the performance of LReLU is better than ReLU in most cases. In other words, as the input parameter of PReLU is larger than that of LReLU and ReLU, this mechanism effectively maintains the activation of following neurons by taking some minor information into account. Besides, as the input parameter can adapt to different inputs, which means the parameter is more flexible and dynamic, PReLU can learn from past information and output a more adaptive result to the next layer.

2.4. Xavier-value and He-value

In the process of parameters initialization, Xavier-value [40] and He-value[39] are two of popular skills to maintain neurons activity before tanh-based and ReLU-based activation functions. Before considering the performance of Xavier-value, the activation value drops after the first layer as increasing the number of layers of deep structure so the following neurons can barely dedicate their ability and fade into useless. This phenomenon is more severe in a highly deep structure of neural network especially, as its mechanism is always regarded as a black box modelling. To relief the problem above, the Xavier-value can be applied to the last layer, so it can confirm that information learned from the previous layer can be sent to the next layer without much distortion.

3. A Dimension-Reduction MLP Based Deep Learning Method

As for enhancing the performance of MLP, applying a suitable deep structure of perceptron to enhance the prediction performance is usually an effective way. Due to its simple structure and flexible modules, the number of neurons can be adjusted dynamically, and the result of prediction can be used to support situation analysis. In order to decrease the complexity of data in advance, the procedure of the proposed method begins with LDA method. As the computational cost will increase exponentially when the structure of neural network is ramped up with numerous layers and neurons, the LDA method is applied to reduce the dimension of data to $n-1$ dimension, which is the number of object category. Here, the process of selecting features is not required, as certain features are hard to find the connection with labels. Also, avoiding losing information because of discarding potential features unknowingly is another consideration, as it might contain useful messages. Therefore, the proposed method introduces the LDA method to do the dimensional reduction and turn the input layer to few neurons. After the process, the convergence is speeded up in first five epochs and reached to stable accuracy rapidly in the end. Moreover, the he-value and Xavier-value are applied as the parameter initialization among layers, to make converge faster in a reliable way. In addition to make a quick converge, these skills are also able to maintain the activity of each neuron. Further, the PReLU is applied as the activation function of the proposed DR_MLP model, and this will consider weakness information, that the activated neuron will be more flexible and even effective to increase the effect of rapidly converging and the activity of neurons. Moreover, this structure is either reliable or robust and

which can be proved on following data test. Further, the steps of the proposed method are described as follows and illustrated in the Fig. 2:

- (1) Step 1: Data preprocessing in training data and then load which to the input layer of model.
- (2) Step 2: The LDA method is applied to map the original data to the $n-1$ dimension in the input layer, and then transmits these values to the next layer.
- (3) Step 3: The PReLU is applied as the activation function to process the value of former layer than output these values to the next layer.
- (4) Step 4: The he-value is applied as the parameter initialization during the path of propagating value.
- (5) Step 5: Repeating the step 3 to the step 4 in each layer then sending the final value to output layer.
- (6) Step 6: The Xavier-value is applied to the last layer as the parameter initialization during the path of propagating value.
- (7) Step 7: The sigmoid function is applied as the value of the last layer.

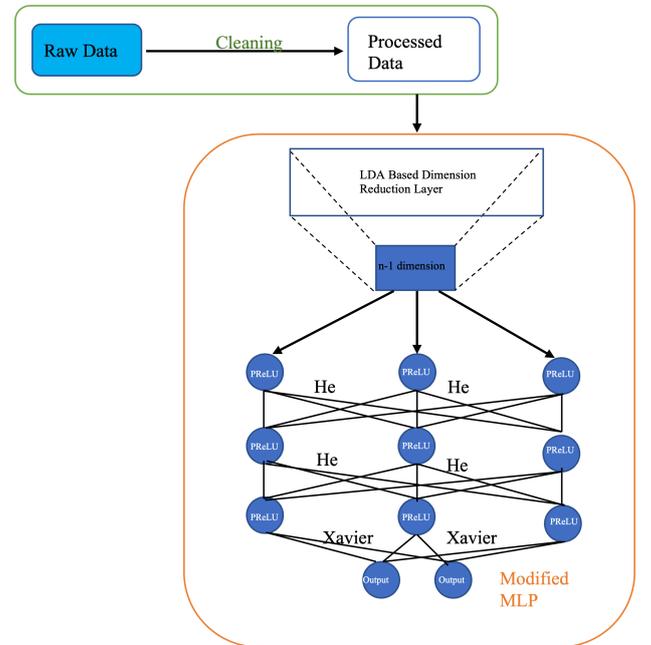


Figure 2. The structure of the proposed method. After a data preprocessed, it will experienced dimension reduction by being inputted in the first layer in our model with $n-1$ dimension. Then, among the feed forward paths, the He-value initializations is equipped before each ReLU-based activation function. In the last layer, the Xavier-value will be added before a tanh-based activation function.

3.1. LDA Based Dimension Reduction Layer

In order to decrease the dimension of data, linear projecting can efficiently achieve this task. Also, to make the process more fluent, taking the linear dimension reduction as the input layer is a good process, so the neurons can linearly transform the input value to few neurons in the next layer. To achieve this, each feature will be produced and presented as the center of each dimension and feature. Also, the matrix of m_i can be represented as follows:

$$\text{Assuming the input data are } x_{m,n} = \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{c,1} & \dots & x_{c,n} \end{bmatrix}, \text{ where } m$$

is the type of category and n is feature.

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^c x_m \quad (1)$$

where $m_i = \begin{bmatrix} \mu_{i, \text{feature}_1} \\ \vdots \\ \mu_{i, \text{feature}_n} \end{bmatrix}$, $i \in \{1 \dots n\}$ is the mean vector which also presents the category i 's μ_m .

Using the mean vector m_i , the inner-categories' scatter matrix S_w can be represented below:

$$S_w = \sum_{x \in D_i}^c S_i \quad (2)$$

where $S_i = \sum_{x \in D_i}^c (x - m_i)(x - m_i)^T$

Also, the between-categories' scatter matrix S_B can be represented below:

$$S_B = \sum_{x \in D_i}^c (m_i - m)(m_i - m)^T \quad (3)$$

where m is whole mean, which includes whole categories' value.

Accordingly, multiplying inner-categories' scatter matrix S_w^{-1} and between-categories' scatter matrix S_B and then find the eigen values of $S_w^{-1}S_B$. Therefore, it might get the $w \in \mathbb{R}^{n \times c'}$, where m follows the rule of less than $c - 1$ number of categories. Lastly, following below to produce projecting result of new data dimension:

$$x' = xW \quad (4)$$

Further, the algorithm extends to the whole training data as follows:

$$X' = XW \quad (5)$$

Since $X' \in \mathbb{R}^{n \times c'}$, it will be used as the new input data, which is a dimension-reduction data transmitting to the next layer to learn.

Also, the LDA Based Dimension Reduction Layer can be represented below:

$$Z^{(2)} = W^{(1)}X^{(1)} \quad (6)$$

$$X' = LDA(Z^{(2)}) \quad (7)$$

Moreover, an apparent computing cost dropping has been applied in the simulation experiment.

3.2. Using PReLU as Activation Function

Although the conventional activation function, ReLU, brings a good effect to the performance of MLP structure, the minor values are ignored based on its mechanism. To avoid this disadvantage and keep the all values as much as possible, the proposed method applies PReLU as the activation function to provide a better performance in proposed MLP model, which can be represented as follows:

$$Z^{(2)} = W^{(2)}[X']^T \quad (8)$$

$$A^{(3)} = \phi_{PReLU}(Z^{(2)}) \quad (9)$$

Then, keep repeating the process of basic MLP step to the next iteration.

3.3. Embedding Xavier-value and He-value

Effectively maintain the activity of each neuron can confirm that the learning ability of the whole MLP structure is stable as much as possible, and furthermore the problem of vanish gradient is also a critical issue in deep learning works. To address this problem, the introducing initialization of Xavier-value and he-value literally spray a bright light on this dilemma. Though the PReLU is applied as the activation function in the MLP model, the problem of gradient vanishing still can be relieved by applying the Xavier-value. Further, the formula can be represented as follows:

$$Z^{(2)} = W^{(2)}[X']^T \quad (10)$$

where $W^{(2)} \sim N(\mu = 0, S = \sqrt{\frac{2}{n}})$

$$A^{(3)} = \phi_{PReLU}(Z^{(2)}) \quad (11)$$

Assume the layer three is the output layer. Then, before passing the sigmoid function, each weight initialization using Xavier-value will be represented as follows:

$$Z^{(3)} = W^{(3)}[A^{(3)}]^T \quad (12)$$

where $W^{(3)} \sim N(\mu = 0, S = \sqrt{\frac{1}{n}})$

$$A^{(4)} = \sigma_{PReLU}(Z^{(3)}) \quad (13)$$

After outputting the final value, whether it is binomial or multiple problems, the proposed DM-MLP model can achieve a good performance without too much computational cost. Further, the simulation experiment will examine the reliability of the proposed DM-MLP model in the following section.

4. Experiments

In the simulation experiment, the datasets are from UCI standard machine learning library. Among these datasets, the selected datasets can be divided into three different types: (1) one with large volume (adult), (2) one with standard size and attributes (Breast Cancer) (3) one with a myriad of features (Sonar, Parkinson, Waveform, and Wisconsin-breast), (4) and others with both big volume and great feature (connect-4, truck). Furthermore, the information of each dataset is recorded in the Table 1.

Table 1. Each Dataset's Profile

Dataset	Data size	Attribute	Class
Adult	32,561	14	2
Sonar	208	60	2
Parkinson	197	23	2
Waveform	5,000	40	3
Connect-4	67,557	42	3
Truck	60,000	171	2
Brain Wave	12,812	15	2
Wisconsin-breast	286	26	2
Breast Cancer	699	9	2

To evaluate the reliability of the proposed DM-MLP model, the accuracy and F1 score are applied as the evaluation criteria. In F1 score, it combines precision and recall which are shown as follows:

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

where TP is true positives, FP represents false positives, and FN represents false negatives. These can be viewed in the following Table.

Table 2. Fusion Matrix

Name	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Therefore, the F1 score can be present under like:

$$F1 = 2 * \frac{Precision * Recall}{Precision * Recall + 2TP + FN + FP} \quad (16)$$

In that the F1 score is already combining the information of both Precision and Recall, the F1 score is chosen as the evaluation criterion to the experiment results.

4.1. Experiment Design

The design of the simulation experiment is conducted into three types: (1) applying the LDA based dimension-reduction neuron to squeeze the multiple dimensions to less dimensions, (2) applying the novel MLP boosting skills, including PReLU, he-value, and Xavier-value, to improve the overall performance, (3) integrating all the mentioned methods to the proposed DR-MLP method, and then make a complete comparisons. In the above simulation experiments, the number of parameters of each model is remain the same, including neurons, layers, learning rate, and epochs.

4.2. Applying LDA Based Dimension Reduction to Squeeze the Data's Dimension

The dimension reduction of data is usually applied as the dimension of data is so large that easily meets the problem of curse of dimension. In some cases, the certain technique of dimension selection picking up few features can distinguish part of classes, but this kind of technique is not easy to perform well in high-dimensional data, especially to the either high-dimensional or high-complicated data. To solve this dilemma and avoid picking up bias features, linearly projecting the entire dimension to the fewer dimensions is one of available options to complete this work. Therefore, the following table shows the features of each data after the process of dimension reduction. Basically, all the dimension of data is combined into one dimension, because the number of categories among data is various. Moreover, the comparisons also include the number of attributes, which of parameters, and the corresponding computation cost in each experiment, seeing Table 3.

As shown in the Table 3, compared with the original model, the number of parameters in the LDA based DR-MLP model of each dataset is decreased. Except for Sonar, Wisconsin-breast, and Breast Cancer dataset, other datasets can improve their computation efficient by reducing the potential low-impact or redundant parameters over 10%. Among these improvements, medical dataset, and Parkinson, has a big flip because its features basically will be transformed to one-hot encoding after data preprocessing. This situation is normal, especially to MIoT-based data. Through introducing the proposed structure, certain potential low-impact or redundant parameters can be reduced.

We have also analyzed two state-of-the-art techniques for breast segmentation [41] and breast abnormality detection [42]. Authors in [41] propose a two-stage approach for segmentation by combining the Firefly Algorithm (FA) assisted Kapur's thresholding and Hidden Markov Random Field (HMRF). We have validated their approach on our dataset and found their approach to provide better breast segmentation results. Authors in [41] proposed symmetry examination, social group optimization, and Kapur's entropy (SGO+KE) thresholding, and fuzzy C-means (FCM) to extract and estimate the breast abnormality known as the DCIS using the RGB BTI. We have used the symmetry analysis approach proposed by the authors in [42], and we were able to segregate the BTI into normal, moderate, and abnormal.

4.3. DR-modified MLP's Performance

The performance of DR-modified MLP model and the corresponding comparison is stated in the Table 4. In the Table 4 and Fig. 3, the DR- modified MLP model can bring a good performance to four of nine datasets, including Sonar, Parkinson, Waveform, and Breast Cancer. As for the Adult data, both the

original MLP model and DR-modified MLP model get the same full performance, so it is impossible to observe a further progressing but it is still seeing that the DR-modified MLP model gets the same precise performance as original MLP model. Furthermore, the comparison also presents the performance of DR-modified MLP model with other combination of only using LDA in MLP model (LDA based DR-MLP test) and applying weight initialization MLP model (PReLU + He-value and Xavier-value

Table 3. Model's Parameters toward Each Dataset

Dataset	Original MLP		LDA based DR-MLP	
	Attribute	Model's Parameter	Attribute	Model's Parameter
Adult	14	30,701	1	20,501 (-33.2%)
Sonar	60	2,064,001	1	2,005,001 (-2.9%)
Parkinson	23	116,001	1	40,701 (-64.9%)
Waveform	40	103,503	2	92,103 (-11%)
Connect-4	42	66,203	2	41,403 (-37.5%)
Truck	171	57,601	1	40,701 (-29.3%)
Brain Wave	33	571	1	251 (-29.4%)
Wisconsin-breast	11	21,501	1	20,501 (-4.7%)
Breast Cancer	43	44,901	1	40,701 (-9.4%)

*The values in the brackets are the increasing ratio toward using LDA-based DR-MLP compared with the original MLP setting.

Table 4. DR-Modified MLP's Accuracy Performance

Testing Accuracy	Original MLP	DR-modified MLP	Increasing Ratio
Adult	100.0%	100.0%	0.0%
Sonar	83.2%	85.7%	2.9%
Parkinson	89.4%	100.0%	11.0%
Waveform	84.5%	86.7%	2.6%
Connect-4	83.5%	75.5%	-9.9%
Truck	99.2%	98.4%	-0.8%
Brain wave	91.9%	70.5%	-23.3%
Wisconsin-breast	100.0%	97.8%	-2.2%
Breast Cancer	67.5%	74.0%	10.1%

test). Through observing the F1 score in the Table 5 and Fig. 4, almost each dataset has been improved except for Braine wave and Wisconsin-breast. In this regard, the proposed method indeed can enhance the accuracy in most data types. As for the Adult data, both the original MLP model and DR-modified MLP model get the same full grade in F1 score. Comparably, the comparison also presents the performance of DR-modified MLP model with other combination of only using LDA in MLP model (LDA based DR-MLP test) and applying weight initialization MLP model (PReLU + He-value and Xavier-value test). Through observed the result in the Table 6, the proposed method can slightly save the computational cost in medical dataset with a reliable accuracy, including Parkinson, Brain Wave and Breast-cancer. In the Fig. 5, it can be observed that the overfitting problem can be blessed by the proposed method. Thus, the DR- modified MLP model not only provides a positive sway in both accuracy and F1 score but also possesses the robustness on the basis of avoiding the overfitting problem.

Table 5. DR-Modified MLP's F1-Score Performance

Test	Original MLP	DR-modified MLP	Increasing Ratio
Adult	100.0%	100.0%	0.0%
Sonar	83.2%	85.7%	3.0%
Parkinson	89.4%	100.0%	11.9%
Waveform	84.5%	86.7%	2.6%
Connect-4	83.5%	84.4%	1.1%
Truck	99.2%	99.9%	0.7%
Brain Wave	91.9%	70.4%	-23.4%
Wisconsin-breast	100.0%	97.8%	-2.2%
Breast Cancer	67.5%	71.8%	6.4%

Table 6. DR-modified MLP's computing time (Time/s)

Cost	Original MLP	DR-modified MLP	Increasing Ratio
Adult	49.342041	49.645839	7%
Sonar	22.066402	23.268208	7%
Parkinson	14.303724	13.199277	-7%
Waveform	16.964323	18.916078	13%
Connect-4	65.649524	67.42403	12%
Truck	3.809537	4.565453	5%
Brain Wave	5.196181	4.591883	-6%
Wisconsin-breast	2.219732	2.872996	13%
Breast Cancer	6.833997	6.618329	-2%

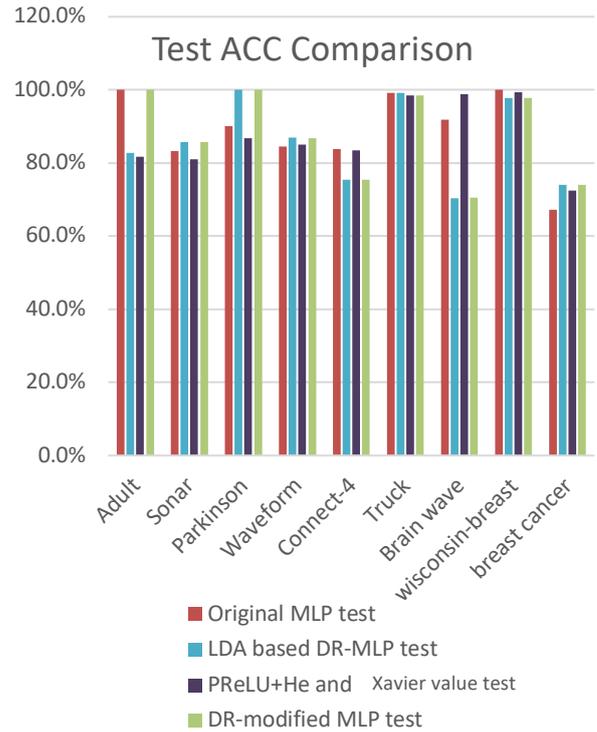


Figure 3. The overall accuracy comparison among original MLP, LDA based DR-MLP, PReLU+He and Xavier value MLP, and DR-modified MLP. Except of the dataset of connect-4, truck, brain-wave and breast cancer, there are four tasks can be classified clearer comared with typical MLP with same parameter and layers.

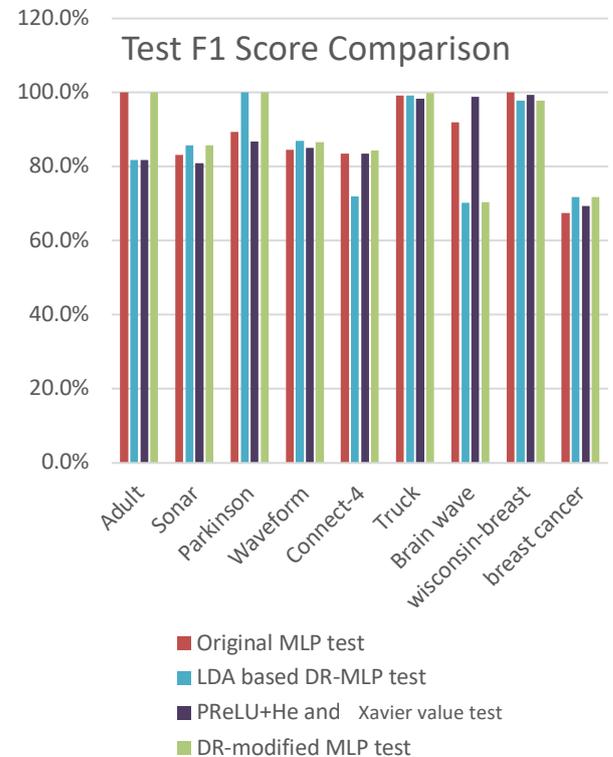


Figure 4. The overall F-score comparison among original MLP, LDA based DR-MLP, PReLU + He and Xavier value MLP, and DR-modified MLP. By using a comprehensive measure, f1-score, the performance of proposal model show a better result, apart from brain wave and Wisconsin-

breast datasets, although there are decreasing outcomes in connect-4 and truck in the index of accuracy.

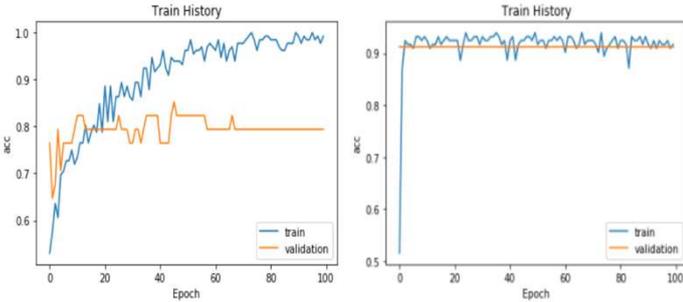
5. Conclusion

As considering the mechanism of the proposed method, the theoretical evidence is positive as a result of its logic reasoning. After examining the results of a series of experiments, the proposed method has certain advantages, including effectively improving the accuracy, F-measure, and computational cost. Especially for the dataset with huge amounts of features, the problem of curse of dimensionality is always unavoidable. Through applying the technique of dimension reduction, it does not only either relieve the problem of curse of dimensionality or save the computational cost, but it also obtains a much robust outcome by avoiding the overfitting problem. In addition, the aforementioned situations also imply that the possible of more robust prediction toward a high variety of real-world data is promising with high-sophisticated mechanism. Furthermore, through the weight initialization, advanced activation function, and dimension reduction, the training patterns usually converges in a soon time. Hence, the training epochs can be adaptively decreased, and which prevent the overfitting problem in advance.

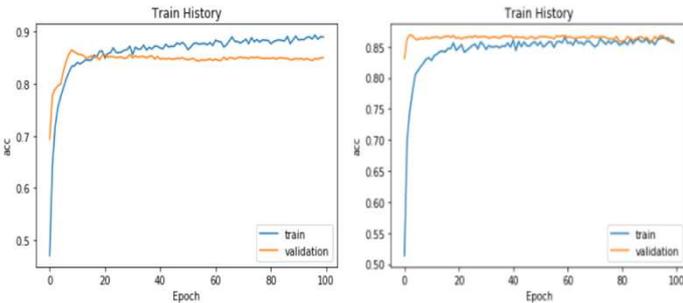
References

- [1] Q. Jiang, X. Jin, S.-J. Lee, S. Yao, A new similarity/distance measure between intuitionistic fuzzy sets based on the transformed isosceles triangles and its applications to pattern recognition, *Expert Systems with Applications*, 116 (2019) 439-453.
- [2] S.-J. Lee, Z. Xu, T. Li, Y. Yang, A novel bagging C4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making, *Journal of biomedical informatics*, 78 (2018) 144-155.
- [3] C.-H. Lai, S.-J. Lee, H.-L. Huang, A social recommendation method based on the integration of social relationship and product popularity, *International Journal of Human-Computer Studies*, 121 (2019) 42-57.
- [4] Y. Yang, J. Jiang, Adaptive Bi-weighting toward automatic initialization and model selection for HMM-based hybrid meta-clustering ensembles, *IEEE transactions on cybernetics*, 49 (2018) 1657-1668.
- [5] Y. Yang, X. Liu, Q. Ye, D. Tao, Ensemble Learning-Based Person Re-identification with Multiple Feature Representations, *Complexity*, 2018 (2018).
- [6] Y. Yang, J. Jiang, Bi-weighted ensemble via HMM-based approaches for temporal data clustering, *Pattern Recognition*, 76 (2018) 391-403.
- [7] Y. Yang, J. Jiang, Hybrid Sampling-Based Clustering Ensemble With Global and Local Constitutions, *IEEE Trans Neural Netw Learn Syst*, 27 (2017) 952-965.
- [8] Y. Yang, Z. Li, W. Wang, D. Tao, An adaptive semi-supervised clustering approach via multiple density-based information, *Neurocomputing*, 257 (2017) 193-205.
- [9] Y. Yang, X. Liu, A robust semi-supervised learning approach via mixture of label information, *Pattern Recognition Letters*, 68 (2015) 15-21.
- [10] R.A.J.A.o.e. Fisher, The use of multiple measurements in taxonomic problems, 7 (1936) 179-188.
- [11] C.R.J.J.o.t.R.S.S.S.B. Rao, The utilization of multiple measurements in problems of biological classification, 10 (1948) 159-203.
- [12] S. Raudys, Evolution and generalization of a single neuron: I. Single-layer perceptron as seven statistical classifiers, *Neural Netw*, 11 (1998) 283-296.
- [13] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, CORNELL AERONAUTICAL LAB INC BUFFALO NY1961.
- [14] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, *California Univ San Diego La Jolla Inst for Cognitive Science*1985.
- [15] G.J.M.o.c. Cybenko, signals, systems, Approximation by superpositions of a sigmoidal function, 2 (1989) 303-314.
- [16] J. Ballé, V. Laparra, E.P. Simoncelli, Density modeling of images using a generalized normalization transformation, *arXiv preprint arXiv:1511.06281*, (2015).

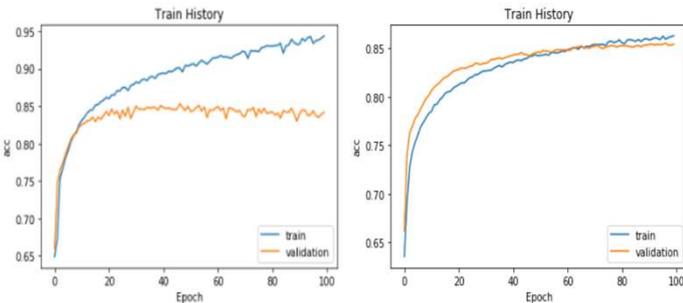
Sonar



Waveform



Connect-4



Breast-cancer

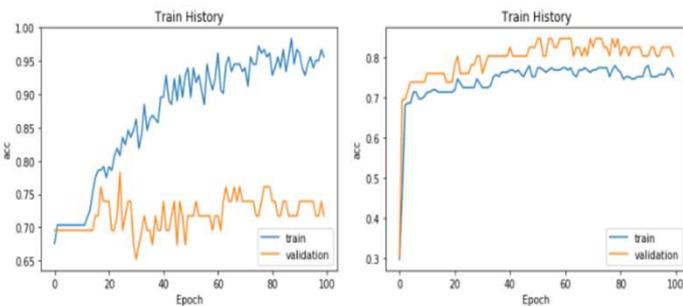


Figure 5. Sonar, Waveform, Connect-4, and Breast-cancer's train history on using DR-modified MLP. Original MLP's train histories are on the left and DR-modified MLP's are on the right. It is clear that the series of right picture has a better converging ability, which means they can output a stable prediction. However, the pictures in the left experience a high risk of overfitting. Although the left pictures still can offer a acceptable prediction in validation data, the

- [17] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, M. Covell, Full resolution image compression with recurrent neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*2017, pp. 5306-5314.
- [18] A.v.d. Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, *arXiv preprint arXiv:1601.06759*, (2016).
- [19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, Photo-realistic single image super-resolution using a generative adversarial network, *Proceedings of the IEEE conference on computer vision and pattern recognition*2017, pp. 4681-4690.
- [20] D.P. Kingma, M. Welling, Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114*, (2013).
- [21] J.T. Springenberg, Unsupervised and semi-supervised learning with categorical generative adversarial networks, *arXiv preprint arXiv:1511.06390*, (2015).
- [22] A. Odena, Semi-supervised learning with generative adversarial networks, *arXiv preprint arXiv:1606.01583*, (2016).
- [23] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, *Advances in neural information processing systems*2016, pp. 2234-2242.
- [24] C. Blundell, B. Uria, A. Pritzel, Y. Li, A. Ruderman, J.Z. Leibo, J. Rae, D. Wierstra, D. Hassabis, Model-free episodic control, *arXiv preprint arXiv:1606.04460*, (2016).
- [25] S.-J. Lee, T. Chen, L. Yu, C.-H. Lai, Image classification based on the boost convolutional neural network, *IEEE Access*, 6 (2018) 12755-12768.
- [26] Q. Jiang, X. Jin, S.-J. Lee, S. Yao, A novel multi-focus image fusion method based on stationary wavelet transform and local features of fuzzy sets, *IEEE Access*, 5 (2017) 20286-20302.
- [27] Q. Jiang, X. Jin, J. Hou, S.-J. Lee, S. Yao, Multi-sensor image fusion based on Interval Type-2 Fuzzy Sets and regional features in Nonsubsampled Shearlet Transform domain, *IEEE Sensors Journal*, 18 (2018) 2494-2505.
- [28] W. Huang, G. Song, H. Hong, K. Xie, Deep architecture for traffic flow prediction: deep belief networks with multitask learning, *IEEE Transactions on Intelligent Transportation Systems*, 15 (2014) 2191-2201.
- [29] Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic flow prediction with big data: a deep learning approach, *IEEE Transactions on Intelligent Transportation Systems*, 16 (2015) 865-873.
- [30] D. Agrawal, 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2010): November 2-5, 2010, San Jose, California, ACM2010.
- [31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, (2014).
- [32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*2015, pp. 1-9.
- [33] B. Normalization, Accelerating deep network training by reducing internal covariate shift, *CoRR*–2015.–Vol. abs/1502.03167.–URL: <http://arxiv.org/abs/1502.03167>, (2015).
- [34] F. Chollet, Xception: Deep learning with depthwise separable convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*2017, pp. 1251-1258.
- [35] M. Jaderberg, A. Vedaldi, A. Zisserman, Speeding up convolutional neural networks with low rank expansions, *arXiv preprint arXiv:1405.3866*, (2014).
- [36] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*2016, pp. 2874-2883.
- [37] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, *European Conference on Computer Vision*, Springer2016, pp. 483-499.
- [38] M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, *Advances in neural information processing systems*2015, pp. 2017-2025.
- [39] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*2015, pp. 1026-1034.
- [40] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the thirteenth international conference on artificial intelligence and statistics*2010, pp. 249-256.
- [41] N. Raja, V. Rajinikanth, S.L. Fernandes, S.C. Satapathy, Segmentation of breast thermal images using Kapur's entropy and hidden Markov random field, *Journal of Medical Imaging and Health Informatics*, 7 (2017) 1825-1829.
- [42] S.L. Fernandes, V. Rajinikanth, S. Kadry, A hybrid framework to evaluate breast abnormality using infrared thermal images, *IEEE Consumer Electronics Magazine*, 8 (2019) 31-36.