Optimizations on Machine Learning Methods for COPD Prediction

*May, 2023

Yuncong Cui
yuncongc@usc.edu

***Abstract*—Early identification of people at risk of developing COPD (Chronic Obstructive Pulmonary Disease) is crucial for implementing preventive strategies. At the same time, machine learning can play an important role in predicting COPD, as it allows for the analysis of large and complex datasets, including medical records, environmental data, and genetic information. Scientists aim to systematically improve the performance of models that predicted the development of COPD. Current COPD prediction is In Fractional Dynamics Foster Deep Learning of COPD Stage Prediction (Chenzhong Yin, 2023), FDDLM (Fractional Dynamic Deep Learning Model), LSTM (Long Short-Term Memory), and DNN are introduced to solve COPD prediction problem. This project mainly focuses on how to train the neural network from the coupling matrix more efficiently.**

***Index Terms*—COPD Stage Prediction, Machining Learning Method, FDDLM, LSTM**

I. Introduction

Chronic obstructive pulmonary disease (COPD) is a chronic lung disease that causes airflow obstruction and makes it difficult to breathe. It is a progressive disease that affects millions of people worldwide and is a leading cause of morbidity and mortality. Early prediction of COPD allows for early intervention, which can slow the progression of the disease and improve the quality of life for people with COPD. This can include lifestyle changes, such as quitting smoking, as well as medical treatments. It is important to build fast and reliable COPD stage prediction algorithms. This project introduces PCA Dimension Reduction, Complementary for Missing Data, and X-Epoch these three methods to optimize the current models for COPD stage prediction.

II. Main Process

In Chenzhong's paper, he proposed a complete process for COPD stage prediction. In the data training part (Fig. 1), he used a neural network to predict the result from patients' health signals. A neural network is a computational model inspired by the structure and function of the human brain. It consists of interconnected artificial neurons, organized in layers, which process and transmit information in a non-linear, parallel manner. Neural networks are used to model complex relationships between signals of patients' health conditions and patients' COPD stages.

A. Dataset

In this dataset, each medical case consists of 12 signal records. All the medical records in this dataset are gathered from four Pulmonology Clinics in Western Romania (Chenzhong Yin, 2023). All signals will be transferred to a coupling matrix. This matrix is the input of the neural network.

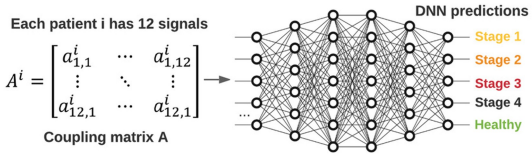


Fig. 1: The Details of Neural Network Training

B. Optimization on Dataset

The patients' dataset contains more than 144 numbers in a matrix. To extract the pattern of a matrix, Principal Component Analysis (PCA) is a widely-used statistical method for data analysis and dimensionality reduction. Its primary goal is to project high-dimensional data onto a lower-dimensional space by finding the most representative directions (principal components) within the data. The coupling matrix can be reduced to a low dimension after applied PCA on it. The time of neural network training will reduce with respect to the reduction in the scale of data.

The dataset contains signals generated by medical machines. The complementary for missing data is also insurance for machines that mistakenly missed signals due to both natural and human factors. The method for making up the missing numbers can make the algorithm robust.

C. Neural Network

Chenzhong (2023) applied CNN, DNN, LSTM and FDDLM to the neural network training. In the context of neural networks, an epoch refers to a single pass through the entire training dataset during the training process. The role of epochs in neural networks is to govern the number of times the model iterates over the entire dataset, updating its weights and biases to reduce the error between predicted outputs and actual outputs.

D. Optimization on Neural network

By iterating over the entire dataset multiple times (i.e., using multiple epochs), the neural network has more opportunities to learn from the data and fine-tune its weights and biases. However, it is essential to find an optimal number of epochs,

as too few may lead to underfitting (the model doesn't learn the underlying patterns well enough), while too many may cause overfitting (the model learns the noise in the data and performs poorly on unseen data).

Typically, the number of epochs is chosen based on the performance of the model on a separate validation dataset. Monitoring the performance on the validation set can help to identify when the model starts to overfit and allows for early stopping or other regularization techniques to prevent overfitting.

III. Optimization Approaches

This part covers the formula and description of three main methods for optimizing the model, and them with the performance on the Loss and AUC curve.

A. PCA Dimension Reduction

Zhang (2016) proposed a complete solution to compress big data by PCA. PCA (Principal Component Analysis) is used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still retains most of the information. PCA in this algorithm consists of two steps.

The first step is to take the previous k singular values for

$$\sum_{i=2}^{k} \sigma_i^2 > 0.99 \times \sum_{i=2}^{D} \sigma_i^2 \tag{1}$$

The parameter D is the dimension of the coupling matrix A (Fig. 1). The parameter σ_i means the singular values of the coupling matrix A. In PCA Dimension reduction, choosing singular values with percentages large than 0.92 is enough to demonstrate most patterns of a matrix.

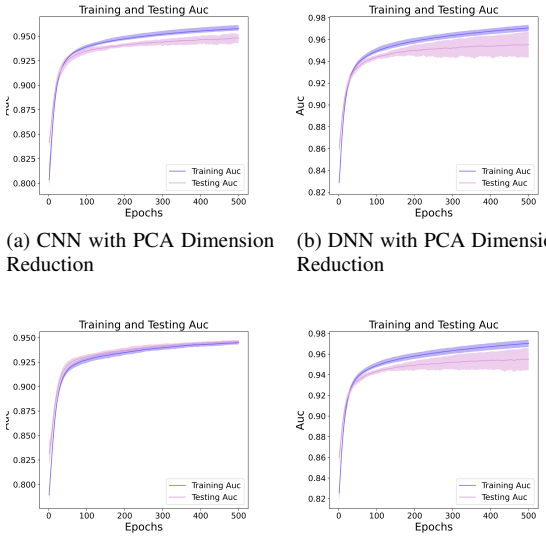
The second step is to reduce the dimension from D to k on the original matrix by PCA. The neural network uses the matrix after dimension reduction as the input to get the result.

	DNN	DNN with PCA
Testing AUC	0.995	0.957
Testing Accuracy	0.975	0.771
Testing Loss	0.130	0.581
Running Time	59.21s	19.32s

TABLE I: Training Performance on DNN with PCA Dimension Reduction

	FDDLM	FDDLM with PCA
Testing AUC	0.997	0.955
Testing Accuracy	0.982	0.771
Testing Loss	0.089	0.592
Running Time	56.81s	37.56s

TABLE II: Training Performance on FDDLM with PCA Dimension Reduction



(a) CNN with PCA Dimension Reduction

(b) DNN with PCA Dimension Reduction

(c) LSTM with PCA Dimension Reduction

(d) FDDLM with PCA Dimension Reduction

Fig. 2: AUC Curve of Training Results with PCA Dimension Reduction

	LSTM	LSTM with PCA
Testing AUC	0.996	0.957
Testing Accuracy	0.974	0.745
Testing Loss	0.124	0.636
Running Time	332.41s	758.84s

TABLE III: Training Performance on LSTM with PCA Dimension Reduction

A characteristic of the current scheme is that it encourages higher speed, but causes lower accuracy. In general, this is a good practice for neural networks. From Table. I - III, the FDDLM and DNN have a great improvement in the running time; however, LSTM does not demonstrate a decrease in running time but a great increase instead. For all three algorithms above, a negative impact has been added to accuracy. The model can take more risks with well running time if the accuracy can have endurance.

B. Complementary for Missing Data

Medical equipment and human factors could cause errors in the data record. This method is designed to fix the missing numbers in the coupling matrix. The experiment will randomly remove one-tenth of the data from the matrix and then try to fill the matrix in three different ways to prepare complete input data for the neural network.

1) Average Feature Value: The most common approach to make up for the missing data is the mean of a feature.

$$A_{ij} = \frac{1}{m} \sum_{k=1}^{m} A_{kj} \tag{2}$$

The parameter m is the number of samples. The parameter k is which feature has missing data.

2) Average Feature Value: Most data in the coupling matrix concentrates from 0.45 to 0.55. The algorithm can use 0.5 to replace the missing numbers.

$$A_{ij} = 0.5 \tag{3}$$

3) Random Forest: Random Forest (Leo Breiman & Adele Cutler, 2001) is used to build a model for known data. Then the model could be used to predict the unknown data.

	CNN	DNN	LSTM	FDDLM
Original Dataset	0.992	0.975	0.974	0.981
Average Feature Value	0.992	0.967	0.974	0.919
Certain Value: 0.5	0.990	0.974	0.979	0.960
Random Forest	0.989	0.971	0.976	0.971

TABLE IV: The Accuracy in the Original Dataset and the Three Compensation Datasets

As can be seen from the data in the Table. IV, after filling in the missing data, all three methods are able to recover the entire matrix almost completely as well as maintain the prediction results of the original matrix. For Random Forest, it will build a model for every feature that has missing data. In the worst case, using Random Forest to fix the missing data will build n (n is the number of features) models. The model construction itself spends much time so the efficiency of this method is far smaller than using an average value or a certain value.

However, all these methods pay attention to the suitability of a single feature without analyzing the interactions within the full set of features.

C. X-Epoch

Chenzhong (2023) set the standard of the number of epochs as 500 in the code. X-Epoch approaches refer to strategies that aim to adaptively decide the number of epochs during the training process to save time without compromising accuracy. These methods can be particularly useful when training deep neural networks, as they often require a large number of epochs, leading to long training times. The idea of X-Epoch in this project is to specify another certain number X of epochs for each neural network to save the redundant times of epochs. For data labels only have 5 types from 0 to 4 (0 means health, 1 to 4 means the stage of COPD), the final results will round the training results to get a specific type (e.g. 1.237 will be rounded to 1; 2.954 will be rounded to 3). Therefore, errors that are less than 0.4 are acceptable. The iteration can stop when the error (RMSE) satisfies the demand:

$$\text{RMSE} < 0.4 \tag{4}$$

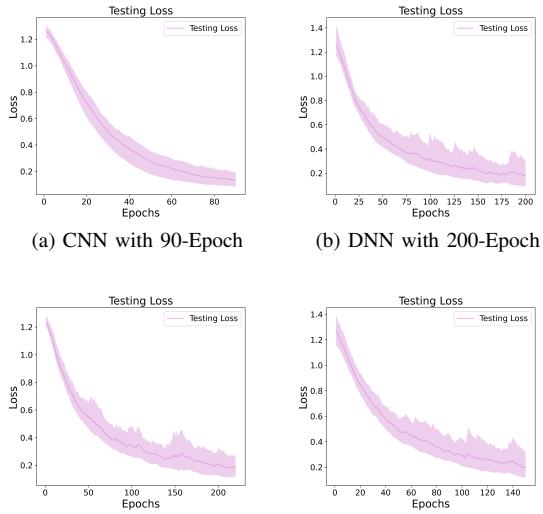
$$\text{Loss(MSE)} = \sqrt{\text{RMSE}} < 0.16 \tag{5}$$

$$\text{Accuracy} > 0.95 \tag{6}$$

Besides the Loss in neural network training, the accuracy should also be paid enough attention. Thus the lower bound of accuracy that allows the algorithm to stop the iteration is set to 0.95. Thus the iteration will stop when loss drops under 0.16 and accuracy increases up to 0.95.

Fig. 3 shows that for the four neural networks, the algorithms will stop when the loss drops under 0.16 instead of the loss converges. Fig. 4 illustrates the running time comparison between the results before and after the algorithm restricts the number of epochs. The running time with X-Epoch could be greatly reduced to less than half of that with 500 epochs. Fig. 5 also proves that all the AUC curves converge before the neural network training stops.

These graphs reveal that using the X-Epoch method can significantly reduce training time while maintaining model performance. By setting a lower limit on the loss value to limit the number of iterations, training can be terminated early when the loss value reaches an acceptable range, avoiding unnecessary computations. Comparing the running times shows the advantages of the X-Epoch method in accelerating the training process. Finally, the convergence of the AUC curves shows that the model performance is maintained with the X-Epoch method and is not affected by the shortened training time.



(a) CNN with 90-Epoch (b) DNN with 200-Epoch

(c) LSTM with 220-Epoch (d) FDDLM with 150-Epoch

Fig. 3: Loss Curve of Training Results with X-Epoch

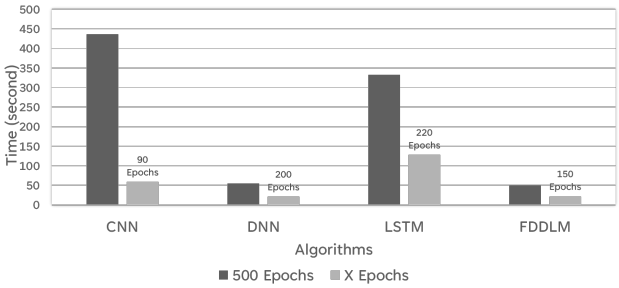
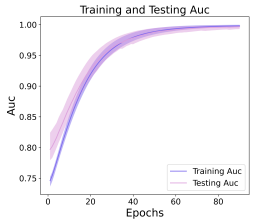
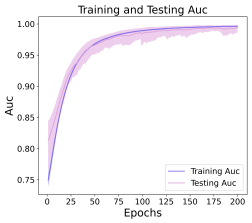


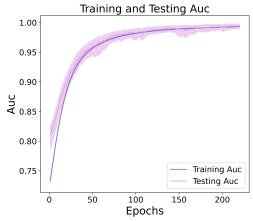
Fig. 4: The Running Time of Neural Network Training



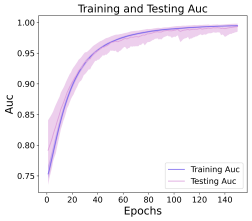
(a) CNN with 90-Epoch



(b) DNN with 200-Epoch



(c) LSTM with 220-Epoch



(d) FDDLM with 150-Epoch

Fig. 5: AUC Curve of Training Results with X-Epoch

IV. Conclusion

For PCA dimension reduction, in conclusion. The neural network can choose to trade the time with accuracy. PCA dimensionality reduction is a lossy compression of data. It depends what is the priority of this problem. But PCA does not perform well in LSTM. Future work could have two directions, the first is to figure out why the PCS is not suitable for LSTM, and the second is to find a better way to compress the data without reducing the accuracy.

To conclude the approaches for complementing the missing data, the performance of choosing an average feature value and a certain value is basically the same. Random forest is not a good method for this problem, because it needs to build too many models and it spends too much time. For future work, the relations between features could be added as a factor to complement the missing data.

X-Epoch, in summary, approaches can decide an adaptive number of epochs to help save time without causing a reduction in accuracy. By monitoring the performance of the model on the validation set, algorithms can stop training earlier to achieve more efficient model convergence.

References

[1] Breiman, Leo. ”Random forests.” Machine learning 45 (2001): 5-32.

[2] Yin, Chenzhong, et al. ”Fractional dynamics foster deep learning of COPD stage prediction.” Advanced Science 10.12 (2023): 2203485.

[3] Zhang, Tonglin, and Baijian Yang. ”Big data dimension reduction using PCA.” 2016 IEEE international conference on smart cloud (SmartCloud). IEEE, 2016.