IMPROVEMENTS ON MACHINE LEARNING METHODS FOR COPD PREDICTION

Yuncong Cui | yuncongc@usc.edu

AGENDA

- ❑ Introduction
- ❑ Relevant Work
- ❑ X-Epoch
- ❑ PCA Dimension Reduction
- ❑ Missing Data
- ❑ Summary

INTRODUCTION

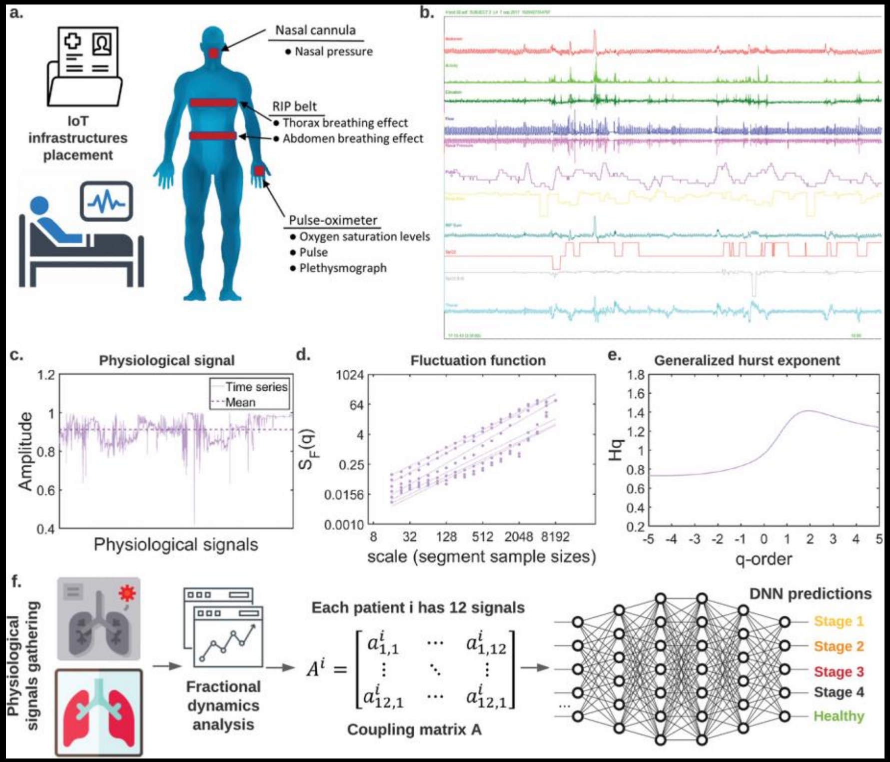
- **WHY TO DO**

Early identification of people at risk of developing COPD (Chronic Obstructive Pulmonary Disease) is crucial for implementing preventive strategies.

- **WHAT TO DO**

We aimed to systematically improve the performance of models that predicted development of COPD.

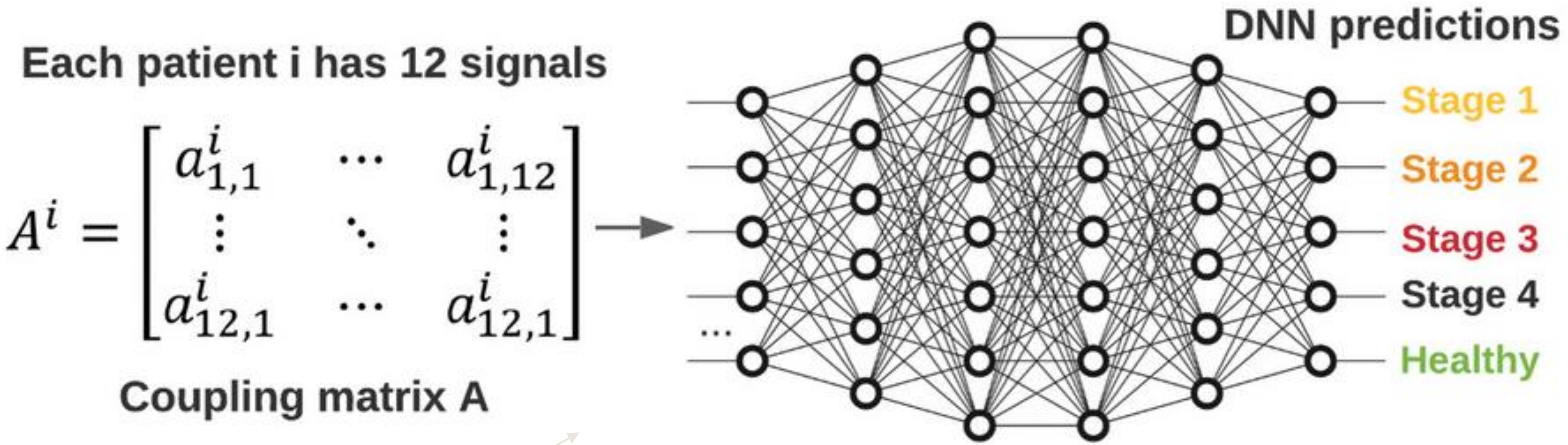
RELEVANT WORK



COPD Stage Prediction Process (Chenzhong Yin, 2023)

- 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 the last part, how to train the neural network from the coupling matrix more efficiently.

MAIN PROCESS



PCA Dimension Reduction

Missing Data

X-Epoch

X-EPOCH

- **Idea:**

  Data Labels: 0 1 2 3 4.

  Round the training results to get a specific type.

  e.g. 1.237 -> 1, 2.954 -> 3

- **Early Stopping Criticism**

  RMSE < 0.4

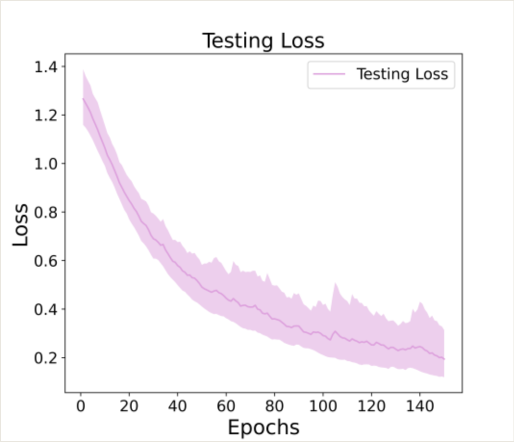
  Loss (Mean Squared Error) < 0.16

  Accuracy > 95%

500 Epochs on FDDLM

Average Loss: 0.089

Average Accuracy: 0.981

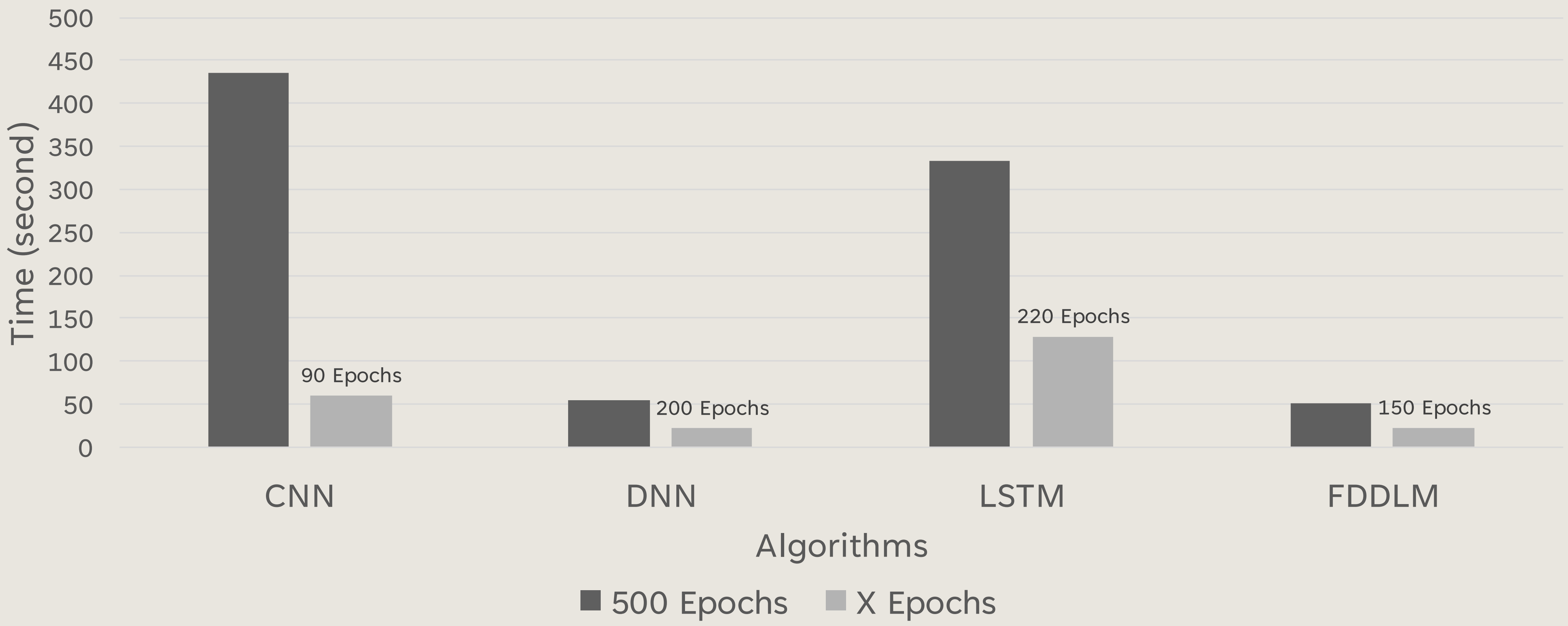


150 Epochs on FDDLM

Average Loss: 0.127

Average Accuracy: 0.959

X-EPOCH PERFORMANCE



Singular Value of Training Data Coupling Matirx

360.636	30.133	19.383	16.675	14.065	9.539
9.147	8.703	8.387	5.556	5.399	4.907
4.726	4.272	3.940	3.558	3.222	3.183
3.093	2.847	2.592	2.545	2.413	2.227
2.133	1.951	1.905	1.814	1.741	1.702

First 30 Singular Values

PCA DIMENSION REDUCTION

- **Idea:** 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.

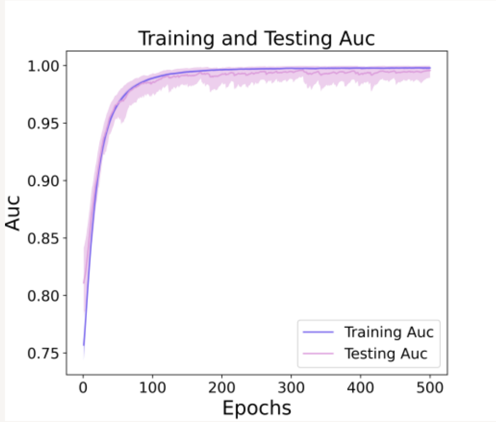
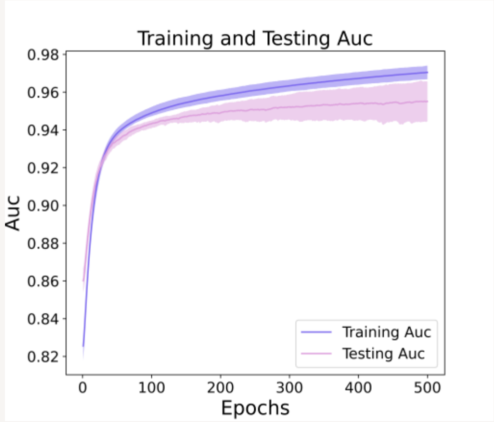
- **Step 1:** Take previous *k* singular values for

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

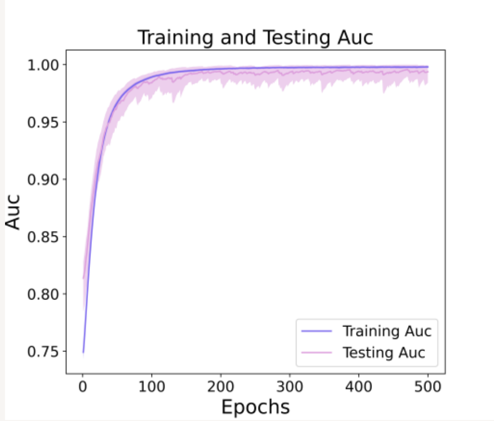
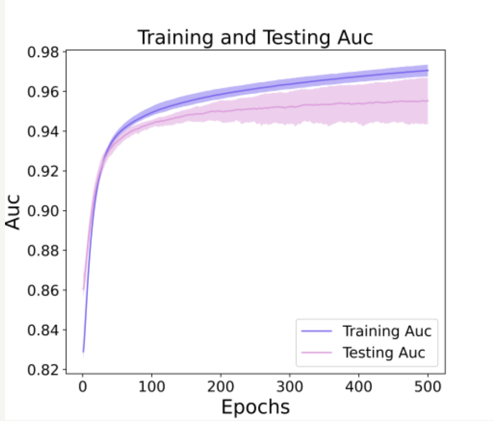
(choose percentage as 99%, *D* is the number of features)

- **Step 2:** Reduce the dimension from *D* to *k* on original matrix by PCA. Use **PCA** in Python to build models.

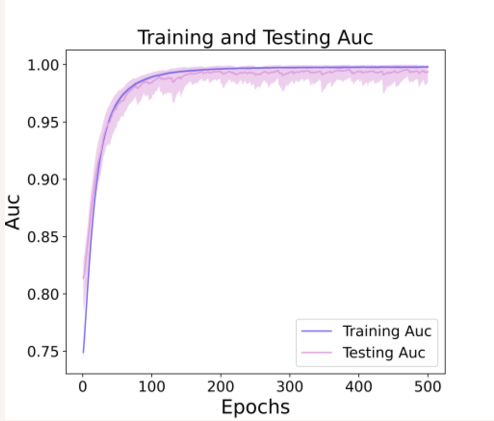
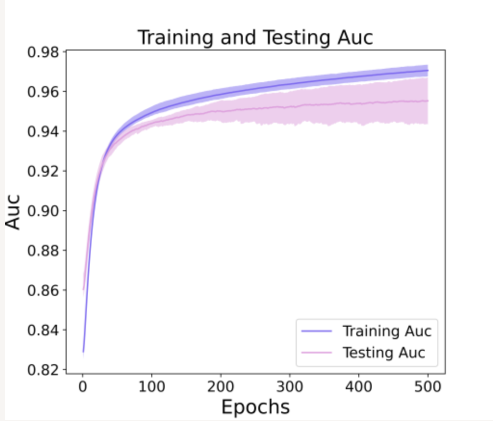
PCA DIMENSION REDUCTION PERFORMANCE

<table>
<tr><th></th><th>FDDLM</th><th>FDDLM with PCA</th></tr>
<tr><td>AUC Chart</td><td>
</td><td>
</td></tr>
<tr><td>Testing AUC</td><td>0.997</td><td>0.955</td></tr>
<tr><td>Testing Accuracy</td><td>0.982</td><td>0.771</td></tr>
<tr><td>Testing Loss</td><td>0.089</td><td>0.592</td></tr>
<tr><td>Running Time</td><td>56.81s</td><td>37.56s</td></tr>
</table>

PCA DIMENSION REDUCTION PERFORMANCE

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

PCA DIMENSION REDUCTION PERFORMANCE

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

MISSING DATA

What if some pixels are missing? E.g. medical equipment error.

Average Feature Value

We use the mean of a feature to replace the missing numbers.

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

(m is the number of samples)

Certain Value: 0.5

Most data in the coupling matrix concentrates from 0.45 to 0.55. We use 0.5 to replace the missing numbers.

$$A_{ij}=0.5$$

Random Forest

We use Random Forest to build a model for known data, then use the model to predict the unknown numbers.

Use **RandomForestRegressor** in Python to build models.

ACCURACY AFTER THE COMPLEMENTARY

	CNN	DNN	LSTM	FDDLM
Sample	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

Randomly delete 1/10 numbers in the coupling matrix.

SUMMARY

- X-Epoch

Choose an adaptive number of epochs helps save the time without causing a reduction to accuracy.

- PCA Dimension Reduction

We can choose to trade the time with the accuracy. PCA dimensionality reduction is a lossy compression of data. It depends what is the priority to this problem. But PCA does not perform well in LSTM.

- Missing Data

The performance of choosing average feature value and a certain value is basically the same. Random forest is not a good method for this problem, because we need to build too many models and it spends too much time.

REFERENCES

- Yin, Chenzhong, et al. "Fractional dynamics foster deep learning of COPD stage prediction." *Advanced Science* (2023): 2203485.

- Data and Code Source: https://github.com/chenzhoy/Fractional-dynamics-foster-deep-learning-of-COPDstage-prediction

THANK YOU

Yuncong Cui

yuncongc@usc.edu