# Project Draft

Team C3: Bo Fan, Surya Garikipati

## Introduction

## Scope of reproducibility

## Methodology

### Model Description

To reproduce the work in “Distilling Knowledge from publicly available online EMR data to Emerging Epidemic for Prognosis”, three models (teacher, student, target) were employed to enable the transfer learning process shown in the paper. All models were constructed with three parts: feature extraction layer, self-attention layer and prediction layer. The feature extraction layer is response for capturing the underlying patterns and representation of the input data. Its output is set of feature embeddings that represent the learned patterns for each medical feature. The self-attention layer allows model to capture relationship between different feature step and time steps by dynamically adjusting model’s focus. The prediction layer is responsible for generating the final output of the model, giving the prediction value length of stay (LOS).

**Teacher model** is the model that learns the relative feature patterns from the large non-target source datasets (PhysioNet dataset). It was trained on PhysioNet dataset for all features in the PhysioNet dataset. The detailed architecture of the teacher model was printed out as shown in Figure 1 (a). As shown in the figure, the feature extraction is mainly done by and .

**Student model**. The student model is an immediate model that borrow extracted feature weights learned from teacher model and modified for further updating for target dataset(TJH). The architecture of the student model is similar to the teacher model as shown in Figure 1(b), but with less feature extractions in the feature extraction layer to match the number of feature to capture in target dataset. As shown in the architecture, student model has along with . The student model was trained on partial source dataset (PhysioNet) that only contains the features that also relative in the target dataset (TJH).

**Transfer from Teacher to student model**. The student model is initialized with the weights from the teacher model, but only remain the weights for features that related to TJH dataset. Then the distillation mechanism proposed in the work was used: the distillation loss () was calculated as the KL-divergence between the outputs of feature extraction layer by teacher and student models, defined by is the feature representation as the output of the extraction layers. Regular prediction MSE loss is also calculated by student model. By minimizing total loss , the student model was trained.

**Target model**. Target model is the model which predicts outcome on the target dataset (TJH). Its architecture is the same with student model but is further tuned on the TJH training dataset upon the trained student model.

**Transfer from student model to target model**. To maintain the capacity of trained student model, the weights of shared GRUs from student model was used in the target model and randomly initializing other parameters in the target model as the target baseline model. Then the target baseline model was trained on TJH training data and evaluated on TJH test data. To reproduce the results shown in the paper, fine tuning is necessary and need to be done.

A screenshot of a computer program

Description automatically generated

There are also some baselined models used to compare the result with the distillation model proposed: GRU[], StageNet[], ConCare[], T-LSTM[], AMT[], AttbiGRU[], TimeNet[].

### Data description

There were 3 main datasets used in the paper: COVID-10 Target Dataset from Tongji Hospital(TJH) [], PhysioNet Source Dataset [], and COVID-19 Target Dataset from HM Hospital (HMH) [].

**PhysioNet Source Dataset**. This is the data that teacher and student models were trained on. It contains 34 clinical features, 8 vital sign features and 26 lab features. As mentioned in the paper, there is 40336 unique patients’ record as .psv files recorded. The label in this dataset this the Sepsis-3 clinical criteria. The teacher models were trained on the full dataset while student model were trained on the partial dataset that contains only the features related to the target TJH dataset.

**TJH target dataset**. This is the most dataset we worked on in the reproducibly work, since all main results were based on the prediction of LOS of TJH dataset. TJH dataset has two files training and test file. Training TJH dataset has 6120 record instance, 374 unique PATIENT\_ID, and 81 features, while test dataset has 757 record instance, 110 unique PATIENT\_ID and 8 features.

**TJH X dataset post processing**. This is one of the main parts we worked on so far. Since only the raw data from TJH paper [] are available, we developed post-process method to generate time-series dataset for training and testing according to the description mentioned in Distcare paper. First, we only selected the related features according to test data, then we calculated the target LOS as shown above using [“Admission time”, “Discharge time”, “outcome”] columns, where 0 for live and 1 for dead in “outcome”. Then we group by “PATIENT\_ID” and sort it by “RE\_DATE” (the data recorded this instance). By defining the sequence length as the largest length of RE\_DATE for all patient and padding the sequence with zeros, we created a sequence of feature vectors in the shape ( for each patient. Hence, we need to modify the GRU layer as GRU(8, 32, batch size).

**TJH ground truth y dataset post processing**. The related 8 features shown in the test dataset were used as X data and the LOS data were calculated according to the definition mentioned in the paper as

### Computational implementation

**Data processing implementation:** codes were implemented by us, according to the description in the paper and the operations from homeworks. The processing was done mainly using pandas Dataframe in notebook file.

**Model implementation**: codes were implemented and modified based on the papers’ GitHub [], we made modifications such as usage of model and hyperparameter tuning so that we focus on the reproducing papers results. The modified codes were implemented in notebook.

**Experiments implementation**: Experiments were implemented by us. Codes were implemented to reproduce the results shown in the paper.

### Code

Please check the code files in the our Georgia Tech GitHub

<https://github.com/BoFanDY/BD4H-C3-project>

The codes included are:

**“distCare\_TJH\_data\_postprocess.ipynb”**: this file was developed to clean and transfer TJH Covid-19 data to generate X, y data for model training

**“main.ipnb”:** this is the main file to run, it import data, train model, test on test dataset and calculate metrics values and plots for results.

## Results

*Ps: only preliminary results are enough. You can just provide one evaluation metric of your developed model without any hyper-parameter tuning. You even don’t need to run all epochs on the full dataset, instead, show grader that your code could work.*

So far, we obtained the xx result without doing fine tuning on the target TJH dataset. The MSE obtained on the TJH test data are shown in the Figure 2. Figure 2(a) shows the MSE and MAE obtained by current model on test TJH data and Figure2(b) shows the prediction performance on TJH dataset under different training data volume to compare with the Figure 4 shown in the paper.

A white rectangular object with black text

Description automatically generated

## Discussion

*Ps: Discuss your current results and propose the continued optimization plan.*

### Discussion of current results

From Figure 2(a), we can see that the MSE and MAE results obtained from our current models are xxx.

From Figure 2(b), we can see that the MSE under different training data volume are xx.

### Future plans

The result we obtained now are not fully reproduce the results shown in the paper. To reproduce the results shown in the paper, we have future plan

Fine tuning the DistCare model.

Generate Confusion matrix.

Add