Reproducibility Project for CS598 DL4H in Spring 2022

Zewei Long and Keyuan Huang
{zeweil2, keyuanh2}@illinois.edu

Group ID: 142
Paper ID: 141, Difficulty: Hard
Presentation link: https://youtu.be/NM7pi_e3zsw
Code link: https://github.com/zewei-long/CS598-FP-HiTANet

1 Introduction

This paper introduced a new model, HiTANet (Luo et al., 2020), in order to predict the disease progression with the assumption that disease progression is non-stationary. There are two main contributions of this paper. First, the authors designed a time-aware Transformer to capture the time information. Instead of using a monotonic time decay function, this model includes time information into the visit embeddings. Second, the authors proposed a time-aware attention mechanism to identify the key visit steps in the visit history.

2 Scope of reproducibility

In the original paper, the authors proposed a time-aware attention mechanism and concluded that the hierarchical structure, which embeds local and global time information, help the HiTANet model improve the performance of prediction task.

The authors conduct an ablation study and claim that after removing time embedding from the local level and global level, the performance of the model decreased dramatically, compared to the original HiTANet model. However, the decrease in performance after moving the analysis component is not significant based on the original result in Table 1. We can observe that the best variants only have a performance decrease of 0.01 in most data sets and measures. In the Kidney disease data set, the F1 score is even higher than the full model.

Table 1: Average Performance of HiTANet's Variants in Original Paper.

Model	COPD		HF		Kidney	
	F1	Auc	F1	Auc	F1	Auc
HiTANet	**0.637**	**0.752**	**0.645**	**0.750**	0.702	**0.792**
HiTANet-LT	0.624	0.742	0.633	0.740	**0.707**	0.789
HiTANet-GT	0.589	0.718	0.599	0.718	0.699	0.786
HiTANet-GLT	0.547	0.694	0.616	0.730	0.661	0.761

Therefore, we want to conduct an ablation study on the effectiveness of the time-aware mechanism. Specifically, we will look into the architecture of the transformer and validate the performance of time-aware visit embedding and time-aware attention mechanism.

2.1 Addressed claims from the original paper

We will test the following important claims in HiTANet model:

- The performance of HiTANet decreases after removing time embedding from local-level visit analysis component.
- The performance of HiTANet drops significantly without modeling time information with time-aware key vector in the global attention construction.

3 Methodology

To reproduce the claims from the paper, we firstly give a brief introduction about the model descriptions (Section 3.1), data descriptions (Section 3.2), hyperparameters (Section 3.3), implementation details (Section 3.4), and computational requirements (Section 3.5).

3.1 Model descriptions

In the original paper, it proposed the HiTANet model, which consists of three major components: (1) a time-aware Transformer and local attention weight for local level visit analysis; (2) a time-aware key-query attention for global level comprehensive analysis; (3) a dynamic attention fusion to predict. The author formulate the risk prediction task as a binary classification problem to predict whether a patient has a specific disease. Thus, the objective function of risk prediction is the average

of cross-entropy:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{P}|}\sum_{p=1}^{|\mathcal{P}|}\left(\mathrm{y}_p^\top \log\left(\mathrm{y}'_p\right) + (1-\mathrm{y}_p)^\top \log\left(1-\mathrm{y}'_p\right)\right) \tag{1}$$

For the transformer, the dimension size of attention embedding as 64, the multi-head number as 4, the size of middle feed-forward network as 1024, and the total number of the parameters is 3067781.

3.2 Data descriptions

The author adopts three disease cohorts extracted from a real-world EHR database: COPD, Heart Failure, and Kidney Disease. The data contains the patient visit history of a disease as shown in Figure 1, which consists of three parts: (1) the ICD codes for each visit; (2) the time interval between visits; (3) the label that indicates whether the patient suffers the disease. We follow the same data structure as the original paper.

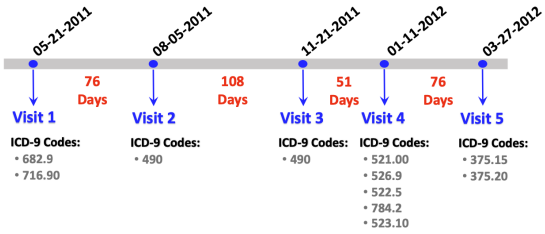

Figure 1: An example of time-ordered patient EHR data that includes five visits. Each visit records a set of diagnosis codes

Table 2: Dataset Details

Dataset	COPD	Heart Failure	Sample-hf
Case (Positive)	354	836	480
Control (Negative)	1416	3342	683
Avg visits per patient	4.6836	4.3597	7.0630
Avg codes per visit	13.4650	13.8670	2.8794
Unique ICD-9 codes	2121	2717	1834

During our reproduction, we first asked the author about the access to the original data, but unfortunately, it's not an open-source dataset. However, the existing code provides some heart failure data samples of the original data. The sample data contains about 1/10 of the original data set with the same structure as we described above, so we believed it was suitable for our preliminary study. Therefore, we started our experiment with the sample dataset that was provided with the code.

Besides, the author also mentioned that the MIMIC-III dataset can be used for model training and evaluation after preprocessing. After investigating the MIMIC-III dataset, we found that the only necessary tables are ADMISSIONS and DIAGNOSES_ICD tables in this dataset. Therefore, after the preliminary study, we started working on extracting the structured data from ADMISSIONS and DIAGNOSES_ICD tables. The ADMISSIONS table contains the information of every admission, including time, patient ID, admission type, etc. The DIAGNOSES_ICD table contains the ICD9_CODE of each corresponding admission and patient. By combining these two tables, we had enough information that we need to reproduce the paper. The information of this data can be found in table 2. However, the most challenging part is the data processing. Even though we had the data needed, we had to reconstruct the data so that it matched the original data structure. We spent most of our time with this and eventually completed this task. We were able to extract and process the data on heart failure and COPD disease from the MIMIC-III dataset. There is only a small number of kidney disease data samples in the MIMIC-III dataset, and it might not be useful in the training process. Therefore, we decided not to use the kidney disease cohort in our experiment.

3.3 Hyperparameters

In our reproduction, we set the hyperparameters from the paper implementation details and the existing code. Table 3 shows the hyperparameters of the proposed framework.

Table 3: Hyperparameters for HiTANet.

Symbol	Value	Definition
l_r	1e-4	the learning rate
b	50	training batch size
d	0.5	dropout rate
n	20	the number of epoch
$visit\_size$	256	the size of input embedding
$hidden\_size$	256	the size of hidden layer

3.4 Implementation

We implement some reproduction based on the existing code from the author and will develop other baselines based on the implementation details in the paper. Here is the link to our GitHub report and codes: https://github.com/zewei-long/CS598-FP-HiTANet.

Table 4: Average Performance on Three Disease Prediction Tasks.

Dataset	Model	Acc	Pre	Recall	F1	Auc
COPD	HiTANet	0.922 ±0.019	0.924±0.047	0.759±0.119	0.825±0.059	0.868±0.051
	HiTANet-C1	**0.950±0.010**	**0.930±0.029**	**0.866±0.067**	**0.894±0.027**	**0.922±0.029**
	HiTANet-C2	0.936±0.022	0.907±0.069	0.841±0.105	0.866±0.049	0.904±0.044
Heart Failure	HiTANet	0.903±0.025	0.914±0.053	0.690±0.149	0.768±0.084	0.832±0.066
	HiTANet-C1	**0.937±0.015**	**0.901±0.064**	**0.854±0.092**	**0.871±0.034**	**0.909±0.036**
	HiTANet-C2	0.916±0.013	0.882±0.064	0.782±0.102	0.821±0.040	0.872±0.039
HF Sample	HiTANet	0.770±0.033	**0.840±0.020**	0.881±0.069	0.853±0.028	**0.645±0.036**
	HiTANet-C1	**0.777±0.021**	0.799±0.027	**0.951±0.045**	**0.867±0.014**	0.583±0.059
	HiTANet-C2	0.773±0.028	0.830±0.020	0.888±0.065	0.856±0.025	0.644±0.039

3.5 Computational requirements

we simulate the model environment on a commodity machine with 1 Intel E5-2650 v4 CPU and 1 NVIDIA 2080Ti GPU. In our experiment for sample data, it takes 1 minute to train the data. In the future, it might take more time to train and test the full dataset.

4 Results

In this section, we reproduce the main experiment of the full HiTANet model, and conduct the ablation study on two important time-aware component in HiTANet. Although the ablation study is implemented on three data sets, there is evidence to support or refute the claims clearly. In summary, Section 4.1 is a sample study in the original data set; Section 4.2 validates claim 1 by removing the time-aware visit embedding; Section 4.3 validates claim 2 by removing the time-aware global attention; Section 4.4 validates the model performance on different data sets.

4.1 Sample Study in Original Data Set

To begin with, We create the model based on the implementation details discussed in the paper and existing source code. Then we are able to train the model with provided heart failure sample data. Similar to the original paper, we used accuracy, precision, recall, F1, and Auc scores in the evaluation. We performed 10 random runs and reported the mean scores for testing performance. Table 5 shows results from our test and the original paper.

Note that because the training data sets are different, there is a significant difference between the resulting scores of these two models. We noticed that several of our testing scores (Precision, recall, F1, and Auc) are higher than the initial result. It might be due to the small size of data that we used. Because we don't have access to the original data, we will conduct our other experiments on both the sample data and extracted data on MIMIC-III data.

Table 5: Our Result on Heart Failure (HF) Data Sample and Result on Full HF Data on Original Paper.

Model	Acc	Pre	Recall	F1	Auc
Our Implementation	**0.772**	**0.823**	**0.897**	**0.857**	**0.630**
Original Result	0.823	0.724	0.587	0.647	0.564

4.2 Claim 1: Removing Time-aware Visit Embedding

In this section, we build HiTANet-C1 from the original model by removing the time-aware visit embedding from the encoder. Based on Table4, we can observe that HiTANet-C1 receives a significantly high performance (up to 0.105) on all measures in COPD and Heart Failure data sets. This illustrates that the local time-aware mechanism is unnecessary and decreases the performance of the transformer on these datasets.

However, the performance of full HiTANet is close to (some even higher than) the HiTANet-C1 on the HF sample dataset. To the best of our knowledge, the performance gain on HiTANet-C1 might come from the difference in data sets. The COPD and Heart Failure data sets, which were extracted from MIMIC-III, have lower average visits per patient and higher average codes per visit compared with the HF sample, which is part of the original data set. These properties will make the time-aware mechanism meaningless. Besides, the model performance of HiTANet in the HF sample data sets is worse than the original paper as well. This is because the HF sample is too short to train a complex model, which will easily lead to overfitting. At the

same time, the simplified model e.g. HiTANet-C1 will have a better performance with less overfitting. Therefore, although our experiment illustrates the performance decrease with time-aware visit embedding, it's unconvincing to refute the claim in the paper.

4.3 Claim 2: Removing Time-aware Global Attention

In this section, we build HiTANet-C2 from the original model by removing the time-aware global attention. Based on Table4, we can observe that HiTANet-C2 outperforms the original model in all the measures except for the precision on COPD and Heart Failure datasets. However, the performance of HiTANet-C2 is worse than the full model on HF sample dataset, which is similar to HiTANet-C1. In this case, we can attribute the performance difference between the original paper and our result to the difference in datasets (lower average visits per patient, higher average codes per visit, and limited patient cases). In this case, although our experiment suggests that the time-aware global attention doesn't improve the model performance, it's unconvincing to refute the claim in the paper.

4.4 Additional Result: Performance Evaluation on Different Data Sets

Just like the results that we show in previous sections, we conducted the experiment on the new datasets (COPD and Heart Failure), which is extracted from the MIMIC-III dataset. The result shows that the data properties will have a significant influence on the model performance, and the time-aware component is unnecessary for the proposed HiTANet model in some datasets.

5 Discussion

In general, we conduct parts of the experiment especially the ablation study successfully on other extracted datasets. However, we cannot get comparable results and lack the necessary information to refute or support the claims of the paper. The strengths of our approach are that we reproduce the main experiment of HiTANet and design the ablation study on time-aware components. The weakness is that the datasets we used have different distributions on many key properties and the total number of records is limited as well. This leads to the inaccuracy of model outputs and unconvinced support or rebuttal of the claims.

5.1 What was easy

First, the architecture of the model is easy to understand. The authors provide a detailed graph about the model architecture in the paper and discuss every major component of it clearly. The illustration helped us grasp the big picture while reading the paper. Second, the math described in this paper is easy to follow. The authors provide thorough explanation while writing the math notation. The authors not only give the detailed formula, but also mention the meaning and dimension of each variable. For example, when the authors discuss the time information embedding, in addition to the actual formula, they also provide the meaning and size of each variables. This helped us a lot to understand and implement the model. Third, during coding process, we got lots of help from the code that provided with the paper. Even though there is no documentation, the code is clean and structure, which makes it easy to follow. Besides, the authors follow the variable naming conventions, and make the code clearly communicate its intent to the reader. Overall, the thorough explanation of the model and well-structured source code help us accomplish the paper reproduction.

5.2 What was difficult

The most challenging task that we encountered was the data processing. The original dataset is not publicly available, so we needed to obtain the data from MIMIC-III dataset and do the preprocessing on our own. After studying the data structure of the paper and the data of the MIMIC-III, we realized that all the data that we needed would be from ADMISSION and DIAGNOSES_ICD tables in the dataset. However, since the data tables have a different structure from the one that used in the paper, it's necessary to do data processing in order to pass the data to the model. It turned out that this part was the most time-consuming during our entire reproduction. Even though we fully understood the data structure of two tables, the actual coding portion remained challenging because it involved data reformatting and encoding. Besides, the structure was clear as a graph but can be confusing in actual model representation. As a result, we spent the majority of the time working on data cleaning and processing. Once we completed this task, the rest of our experiment became easy and straightforward.

5.3 Recommendations for reproducibility

We think the reproducibility is excellent: the implementation details are stated in the original paper and the public code is available on GitHub. However, the biggest difficulty comes from the closed-source datasets: (1) we spend a lot of time to get the access to MIMIC-III dataset and extract the data sets about admission history and disease prediction. (2) the extracted data sets have a very different properties on many attributes and lead to different model performance. We hope the original authors or others who work in this area for improving reproducibility can make the original data be public for the future researchers.

6 Communication with original authors

The data used in the original project is not publicly accessible, so we contacted the author about the data access. We were told that we could use the data samples provided in the github repo to do our preliminary study. Then we could also preprocess the MIMIC-III dataset to do actual model training and evaluation.

References

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, page 647–656, New York, NY, USA. Association for Computing Machinery.