Risk Prediction with Electronic Health Records: A Deep Learning Approach

Yu Cheng* Fei Wang† Ping Zhang* Jianying Hu*

Abstract

The recent years have witnessed a surge of interests in data analytics with patient *Electronic Health Records* (EHR). Data-driven healthcare, which aims at effective utilization of big medical data, representing the collective learning in treating hundreds of millions of patients, to provide the best and most personalized care, is believed to be one of the most promising directions for transforming healthcare. EHR is one of the major carriers for make this data-driven healthcare revolution successful. There are many challenges on working directly with EHR, such as temporality, sparsity, noisiness, bias, etc. Thus effective feature extraction, or phenotyping from patient EHRs is a key step before any further applications. In this paper, we propose a deep learning approach for phenotyping from patient EHRs. We first represent the EHRs for every patient as a temporal matrix with time on one dimension and event on the other dimension. Then we build a four-layer convolutional neural network model for extracting phenotypes and perform prediction. The first layer is composed of those EHR matrices. The second layer is a one-side convolution layer that can extract phenotypes from the first layer. The third layer is a max pooling layer introducing sparsity on the detected phenotypes, so that only those significant phenotypes will remain. The fourth layer is a fully connected softmax prediction layer. In order to incorporate the temporal smoothness of the patient EHR, we also investigated three different temporal fusion mechanisms in the model: early fusion, late fusion and slow fusion. Finally the proposed model is validated on a real world EHR data warehouse under the specific scenario of predictive modeling of chronic diseases.

1 Introduction

The global health care systems are rapidly adopting Electronic Health Records (EHR), which are systematic collections of longitudinal patient health information (e.g., diagnosis, medication, lab tests, procedures, etc.) generated by one or more encounters in any care delivery setting [1]. This will dramatically increase the quantity of clinical data that are available electronically. As a result, data driven healthcare, defined as usage of those available big medical data to provide the best and most personalized care, is becoming to be one of the major trends to the success of revolutionize healthcare industry [2,3]. As patient EHRs is the major carrier for conducting data-driven healthcare research, understanding the information contained in EHRs is crucial.

There are quite a few works in recent years on data analytics with patient EHRs. For example, Wang *et al.* [4] presented a multilinear sparse logistic regression for risk prediction with patient EHRs. Zhang *et al.* [5] proposed a similarity based approach for personalized treatment recommendation. A composite distance metric learning algorithm is presented in [6] for patient similarity evaluation by integrating the patient population from different clinical sites without sacrificing privacy. One key aspect to the success of those medical applications is extracting effective features from patient EHRs, which is usually referred to as *electronic phenotyping* in medical informatics [7,8]. Although some computational models have been proposed for EHR based electronic phenotyping recently (e.g., a matrix based method [9] and a tensor based algorithm [10]), a lot of challenges still remain, to list a few:

- *High-Dimensionality.* There is a large amount of distinct medical events (e.g., there are more than 14,000 different diagnosis codes in terms of International Classification of Diseases - 9th Version (ICD-9)) in patient EHRs. Those events also interact with each other.
- *Temporality.* The patient EHRs evolve over time. The sequentiality of the medical events reveal important information on impending patient disease conditions.
- *Sparsity.* The EHR data are extremely sparse, as illustrated in Fig.1. The data are largely missing

*Healthcare Analytics Research, IBM T.J. Watson Research Center

†Computer Science and Engineering, University of Connecticut

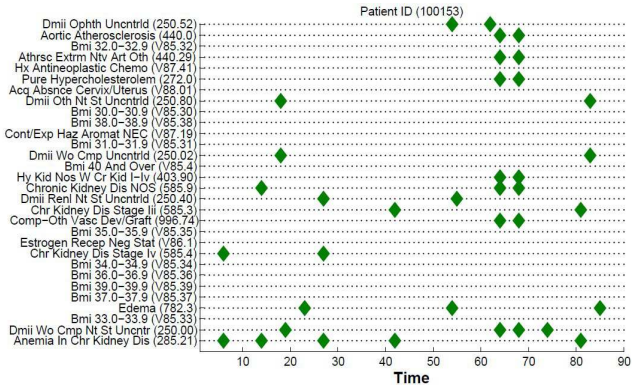


Figure 1: A real world example of the patient's EHR. The horizontal axis represents the number of days since the patient has records. The vertical axis corresponds to different diagnosis in terms of ICD-9 codes. A green diamond indicates the corresponding code is diagnosed for this patient at the corresponding day.

in several ways, such as recording mistake, patient relocation, lack of visits, etc.

- *Irregularity.* Due to the complexity of patient diseases, there exists high variabilities among the EHRs of different patients, even with the same disease.

- *Bias.* The above challenges, including systematic errors, can result in significant bias when health record data are used naively for clinical research.

To overcome those challenges, we propose a deep learning approach for extract meaningful features, or phenotypes, from patient EHRs in this paper. As a prerequisite, we first convert the EHRs of every patient into a binary sparse matrix as in [9, 11], where the horizontal dimension is time, the vertical dimension is medical events. The (i, j)-th entry in the EHR matrix of a specific patient is 1 if the i-th event is observed at time stamp j in his/her EHR. Our model is adapted from the regular Convolutional Neural Network (CNN) model, which is a biologically-inspired variant of multi-layer perceptrons [12]. There are four layers in our model. The first layer is the basic patient EHR matrices. The second layer is a one-side convolution layer, where the features are obtained through one-side convolution on the first layer. The third layer is a max pooling layer for introducing sparsity on the learned features. The fourth layer is a fully connected layer linking with softmax classifiers for prediction. Considering the temporal continuity of the patient EHR, we also investigated different temporal fusion strategies in our model. Finally we validate the effectiveness of the proposed model on a real world EHR data warehouse under two specific clinical scenarios: early prediction of Congestive Heart Failure (CHF) and Chronic Obstructive Pulmonary Disease (COPD). The results show that those features learned from our model can not only produce better prediction performance, but also make clinical sense.

It is worthwhile to highlight the following aspects of the proposed model:

- To the best of our knowledge, this is the first attempt on applying deep learning technologies in analyzing discrete patient EHR data.

- Our model naturally explores the temporal characteristics of the patient EHR and incorporates them into the feature learning process.

- Different from the CNN methods applied in analyzing image and video data, the convolution operator on the second layer is only performed on the time dimension of the patient EHR matrices. Because it does not make clinical sense for convolve over the medical events.

- Three different temporal fusion strategies: early fusion, late fusion and slow fusion are investigated in our model for leveraging the temporal smoothness of EHR into the learning process.

- The effectiveness of our model is validated on real world EHR data warehouses.

2 Related Work

In this section we briefly review the existing work that is closely related to the research proposed in this paper. One is patient phenotyping from their EHRs. The other is the algorithmic research on deep learning and their applications on extracting effective temporal features.

2.1 Electronic Phenotyping Electronic phenotyping refers to the problem of extracting effective phenotypes from longitudinal patient EHRs. As pointed out by Hripsak *et al.* [7], this is a key step before we can perform any data-driven applications (e.g.,comparative effectiveness research [13], predictive modeling [14], etc.) with EHR, because there are many challenges working directly with raw EHR (such as the ones we listed in the introduction). In the following we will summarize the existing works according to the different representations of patient EHRs.

- *Vector Based Representation.* This method construct a vector for every patient. It dimensionality equals to the number of distinct events appeared

in the EHR, and the value on each dimension is the summary statistics (e.g., sum, average, max, min, etc.) of the corresponding medical event in a specific time period. With vector based representation, each phenotype is usually assumed to be a linear combination of those raw medical events and the combination coefficients are obtained by some optimization procedure [15]. The limitation of this representation is that it ignores the temporal relationships among those events.

- *Tensor Based Representation.* This method constructs a EHR tensor for every patient. Every mode of the tensor indicates a specific type of medical entity (e.g., patients, medications or diagnosis). The entry values will be the summary co-occurrence statistics of the different events of the corresponding dimensions. Ho *et al.* [10, 16] proposed a nonnegative tensor factorization based approach for phenotype extraction from those EHR tensors. This method explored the interactions among different medical entities. The limitation is that they did still not take event temporal relationships into consideration.

- *Sequence Based Representation.* This method construct a EHR sequence for every patient according to the time stamp of each event. Then frequent pattern mining approaches can be applied to identify temporal patterns as phenotypes [17, 18]. One problem is that because of the high variability among patient EHRs, this approach usually returns a huge number of patterns (which is also referred to as the "pattern explosion" phenomenon). It is very difficult to judge which phenotype is clinically useful.

- *Temporal Matrix Based Representation.* This approach represents the patient EHRs as temporal matrices with one dimension corresponding to time and the other dimension corresponding to medical events. Zhou *et al.* [9] proposed a phenotyping method by grouping medical events with similar temporal trends together. However, they did not consider the temporal relationships across different events. Wang *et al.* [11] proposed a convolutional matrix factorization approach to detect shift-invariant patterns across patient EHR matrices, but they cannot determine the optimal pattern lengths and need to enumerate all possible values.

The method proposed in this paper is based on temporal matrix representation. With the smart CNN structure our algorithm can identify important phenotypes and weigh them automatically in the prediction phase. Also the temporal fusion scheme effectively compensates the different phenotypes with different window lengths.

2.2 Deep Learning for Feature Engineering

Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations. In the past few years, deep learning models have achieved remarkable results in computer vision [19] and speech recognition [20] applications. Convolutional Neural Networks (CNN) is one of the classic deep learning models.

CNN is a neural network that can make use of the internal structure of data (e.g., the 2D structure of image data) and utilize multiple layers with convolution filters applied to local features [21], wherein each computation unit responds to a small region of input data. Originally invented for computer vision, CNN models have subsequently been shown to be effective in other domains as well, such as search query retrieval [22] and word embedding learning [23]. In text mining, since the work on token-level applications by Collobert et al. [24], CNN has been used in systems for product feature mining [25], document classification [26], sentence modeling [27], and many others [28].

Traditional CNN can only handle static contents (e.g., images and documents). In our case, the patient EHR is longitudinal because the patient condition evolves over time. Therefore in order to apply CNN for analyzing patient EHR, we need to incorporate the rich temporal information. There are some works proposed to capture the temporal information in dynamic scenarios, such action recognition [29, 30] and object localization [31] from video sequences. Those methods typically use separate stacked video frames as input to the network, they try to combine those stacked videos by different fusion mechanisms on different layers of the CNN architecture [32, 33].

3 The Proposed Models

We build our model based on the temporal matrix representation of patient EHR as in [11]. Specifically, we model the EHR record as an longitudinal event matrix, where the horizontal dimension corresponds to the time stamps and vertical dimension corresponds to the event values. The (i, j)-th entry of an EHR matrix is 1 if the i-th event is observed at the j-th time stamp for the corresponding patient. However, different from images and videos, the standard CNN model can not be directly applied to this event matrix representation due to the following reasons.

- We cannot do regular convolutions on EHR matri-

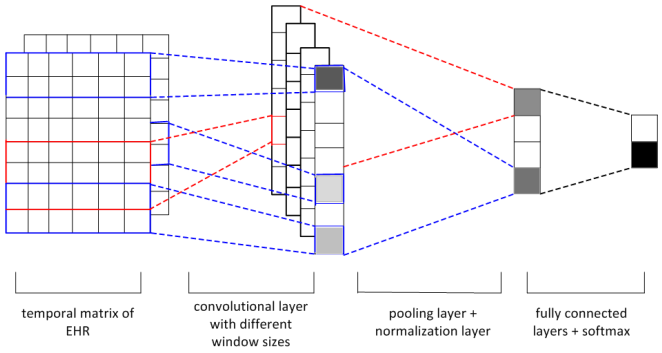


Figure 2: The basic model architecture for an example of an EHR data.

ces, as the convolution over events is not meaningful.

- Not like videos where each image frame has exactly the same dimensions, the EHR matrices for different patients have different sizes on the time dimension.
- We need to explore the temporal smoothness over the patient EHRs.

In the following, we will present the details of our proposed approach trying to address the above challenges. We will first introduce the basic CNN model, and then an advanced CNN architecture with temporal fusion.

3.1 Basic Model The basic model architecture, shown in Figure 2, is a slight variant of the CNN architecture of [24]. Each event matrix of length t is represented $\mathbf{X}$ and $\mathbf{X} \in \mathbb{R}^{d \times t}$. Let $\mathbf{x}_i \in \mathbb{R}^d$ be the d-dimensional event vector corresponding to the i-th event items. In general, let $\mathbf{x}_{i:i+j}$ refer to the concatenation of items $x_i, x_{i+1}, \cdot, x_{i+j}$. A one-side convolution operation involves a filter $w \in \mathbb{R}^{d \times h}$, which is applied to a window of h event features to produce a new feature. For example, a feature c_i is generated from a window of events $x_{i:i+h-1}$ by $c_i = f(\mathbf{w} \centerdot x_{i:i+h-1} + b)$, where $b \in \mathbb{R}$ is a bias term and f is a non-linear function such as rectification (ReLU), tangent (Tanh). This filter is applied to each possible window of features in the event matrix $\{x_{1:h}, x_{2:h+1}, ..., x_{n-h+1:n}\}$ to produce a feature map $\mathbf{c} = [c_1, c_2, ..., c_{n-h+1}]$, where $\mathbf{c} \in \mathbb{R}^{n-h+1}$. We then apply a mean pooling over the feature map and take the average value $\hat{c} = max\{\mathbf{c}\}$. The idea is to capture the most important feature one with the highest value for each feature map. This pooling scheme naturally can deal with variable time stamp lengths of the EMR records. The final layer is a connected layer with dense connections and a softmax classifier.

3.2 Temporal Fusion CNN Unlike images and documents which can be viewed as still, EMR data vary widely in temporal extent and the temporal connectivity is also important for the prediction. In this part we treat every data sample as a bag of short, fixed-sized sub-frames. Since each sub-frame contains several contiguous intervals in time, we can extend the connectivity of the model in time dimension to learn temporal features. Following [29, 30], we describe three broad connectivity pattern categories below.

The three proposed models are based on the fusing information across temporal domain: the fusion can be done early in the network by modifying the first layer convolution filters to extend in time, or it can be done late by placing two separate single-frame networks and fusing their outputs later in the processing:

Single-frame: This architecture views the EMR record as a static matrix and is the one we propose in the basic model. A slightly different component is the normalization layer added here. We use a single-frame architecture to understand the contribution of static appearance to the classification accuracy.

Temporal Early Fusion: The Early Fusion extension combines information across an entire time window immediately on the basic event feature level. This is implemented by modifying the filters on the first convolution layer in the single-frame model by extending them to be on the number of sub-frames k.

Temporal Late Fusion: The Late Fusion model performs the fusion on the fully connected layer. It first places several separate single-frame networks (5 sub-frames as shown in Figure 3) and then merges the these streams in the fully connected layer. In this setting, patterns existing in each sub-frame can easily be detected.

Temporal Slow Fusion: The Slow Fusion model is a balanced mix between the two approaches that slowly fuses temporal information throughout the network such that higher layers get access to progressively more global information in temporal dimensions. This is implemented by extending the connectivity of all convolution layers in time and the fully connected layer can compute global pattern characteristics by comparing outputs of all layers.

As a summary, all the three temporal fusion models try to bring the temporal conductivities into the CNN model. The late fusion model can well capture local

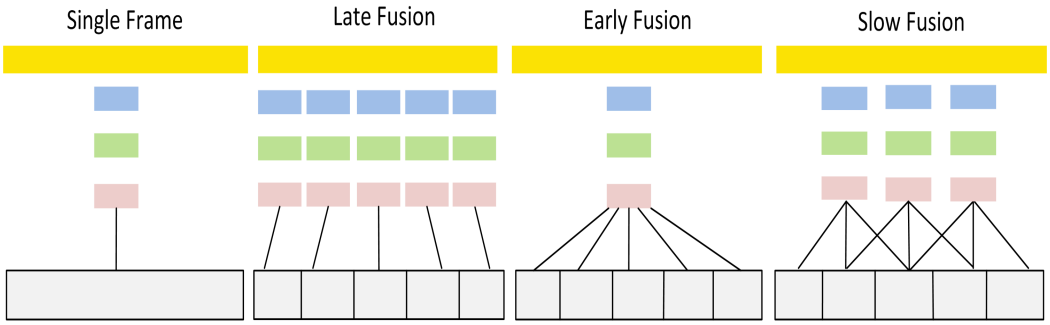


Figure 3: Explored approaches for fusing information over temporal dimension through the network. Yellow, red, green and blue boxes indicate fully connected, convolution, normalization and pooling layers respectively. In the Slow Fusion model, the depicted columns share parameters.

information in each sub-frame. The early fusion model tries to capture global patterns over the data. And theoretically, has very similar properties as the single frame model, in the one-side convolution setting. While slow fusion is a balanced mix between early and late ones, thus can achieve to capture both local and global temporal information. We will show the results over all the methods in the experimental section.

3.3 Regularization and Learning Regularization: For regularization we employ dropout on the fully connected layer with a constraint on l_2-norms of the weight vectors [34]. Dropout prevents co-adaptation of hidden units by randomly dropping out i.e., setting to zero a proportion p of the hidden units during forward-back propagation. That is, given the penultimate layer $\mathbf{z} = [\hat{c}_1, \hat{c}_2, ..., \hat{c}_{n-h+1}]$, for output unit y in forward propagation, dropout uses

$$y = \mathbf{w} \cdot (\mathbf{z} \circ \mathbf{r}) \tag{3.1}$$

where $\circ$ is the element-wise multiplication operator and $\mathbf{r} \in \mathbb{R}^m$ is a "masking" vector of Bernoulli random variables with probability p of being 1. Gradients are back-propagated only through the unmasked units.

Learning: We use Downpour Stochastic Gradient Descent [35] to optimize our models across a computing cluster. The number of replicas for each model varies between 10 and 50 and every model is further split across 4 to 32 partitions. We use mini-batches of 32 examples, momentum of 0.9 and weight decay of 0.0005. All models are initialized with learning rates of 0.001 and this value is further reduced by hand whenever the validation error stops improving.

4 Evaluation

In this part, we studied the effectiveness of our proposed approaches on a real world EHR data warehouse

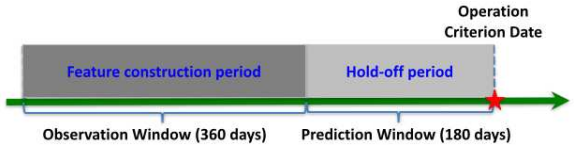


Figure 4: Experimental setting of early prediction of the chronic disease onset risk.

including the records of 319,650 patients over 4 years. We use the diagnosis information in terms of the first three digits of ICD-9 to construct the EHR sequences. We will study the problem of early prediction of the onset risk of chronic diseases. The detailed setting is illustrated in Fig.4. For each disease, our domain expert first help us identify a set of case patients who are confirmed with the disease according to the medical diagnosis guidelines, and then a set of group matched controls is collected according to patient demographics and clinical characteristics. For every patient, we set an operation criterion date, which is the chronic disease confirmation date for case patients, the last day in our database for control patients. We then trace back from the operation criterion date, hold off the records with in the prediction window, and use the records in observation window for analysis.

The two specific disease studied in our experiments are Congestive Heart Failure (CHF) and Chronic Obstructive Pulmonary Disease (COPD). The CHF patient cohort consists of 1127 cases and 3850 controls. The COPD patient cohort includes 477 cases and 2385 controls. For both diseases, we set the prediction window to be 180 days and utilize all the records available in our database before the prediction window to train our proposed model. That is, we use all the historical patient records to predict their onset risk of CHF or COPD half year later.

We implemented the following models to extract patient features (a.k.a. phenotypes): 1) Basic CNN model (BS-CNN); 2) Temporal Early Fusion CNN (EF-CNN); 3) Temporal Late Fusion CNN (LF-CNN); 4) Temporal Slow Fusion CNN (SF-CNN); 5) Combined Features (Combined): the combination of aggregated feature and the feature learned from the SF-CNN model, as additional features for predicting the diagnoses. For comparison purpose, we also report the performance of logistic regression using the aggregated clinical features (i.e., the vector based representation with value on each dimension indicating the frequency of the corresponding feature).

For all the experiments we set the models hyper-parameters in the following way (1) filter windows(h) from 2 to 8 with 105 feature maps each; (2) dropout rate p of 0.5; (3) the number of sub-frames k of 5; and (4) mini-batch size of 50. These values except the sub-frames number k were chosen via a grid search and 10 fold cross validation is used, i.e., every time we randomly select 10% of the training data as the cross validation set. Training is done through stochastic gradient descent over shuffled mini-batches [36].

We then randomly split the samples into different ratio of training and testing, ranging from 60% to 90% and train the classifier on the training data. We repeat the random splitting 50 times independently, and report the average performance. For the sake of fairness, the splitting is the same for all methods in each iteration.

CHF: The results for CHF prediction is summarized in Table 1, measured by AUC over 10-fold cross-validation. As can be seen from the table, our methods consistently and significantly outperform the feature-based baseline. Particularly, the SF-CNN achieve the best performance over all the methods. Measured by AUC, SF-CNN increases the prediction accuracy by 1.5% when 60% training data is used, and 5.2% when 90%. The basic CNN model (BS-CNN) and EF-CNN have comparable performance as their architectures are similar. The prediction accuracy of LF-CNN is slightly better than BS-CNN and EF-CNN. One possible reason might be LF-CNN can capture some local temporal patterns, which is discriminative for the classification. It is notable that with the number of training data increases, the performances gain of all proposed model also increase, which show the CNN-based models would be benefited if the training data is large.

COPD: The predictive performance on COPD is given in Table 2. Similar trend can be observed as on CHF cohort. SF-CNN also outperforms others and improve the predication AUC by 5.3% when 90% training data is used, over the baseline. However, on this dataset, the performance of LF-CNN is not as good as on CHF Cohort. This probably because ESRD Cohort is a smaller dataset than ESRD Cohort and LF-CNN suffers more from over-fitting since it has more parameters than the other model. When the training set is reduced to 60%, LF-CNN even perform worse than EF-CNN and BS-CNN.

As a summary, the experimental results have demonstrated the effectiveness of the CNN-based model on real clinical data. The temporal fusion framework over CNN can significant improvements on predictive performance, which showed that incorporating temporal connectivity can boost the performance. On the other hand, LF-CNN generally achieves a better performance and outperform EF-LNN, which shows that the local patterns are also important for classification.

However, treating the EHR data in the temporal fusion framework is not an easy way and sometime this may fail (like LF-CNN on COPD). The fusion type, the number of sub-frames, as well as the over-fitting problem are all important. We will lead some discussion in the following Section. Nevertheless, all models will gain benefits if the training data is large, which suggests that collecting more real data is important.

5 Discussions

In this section we discuss the effectiveness of the CNN-based model in a qualitative way. We argue that the proposed framework can produce meaningful phenotypes by fully taking advantage of the higher-order temporal event relationships. We will also compare those phenotypes resultant from different strategies.

Limited by the model architecture, it is not easy to visualize the feature maps with deconvolutional nerual networks [37]. Thus we exploit a method by observing the activity of neurons. Recall that the output/activation of the neurons (after pooling) serve as features in the top layer, and the top layer assigns weights to the features. We record the neurons whose output received the highest weights in the top layer for the negative and positive class respectively. In this way, we can find the regions appearing in the training set that highly activate the corresponding neurons. Via sliding window cut (minimal and maximal window size), we can obtain several top ranked regions (patterns).

In particular, for both scenarios, some event conditions were selected as being relevant to the progress of CHF/COPD based on the guidance of our medical advisors. Figure 5(a) shows the top-5 learned patterns by different models on CHF Cohort, with different window sizes. For LF-CNN, most of examples that highly active model's top-layer neurons are within small window, which confirm that LF-CNN tends to capture local patterns. For EF-CNN, the window size of most of the

Table 1: Prediction AUC and standard deviation on the CHF cohort with different ratio of training data.

Method	60%	70%	80%	90%
Baseline	0.5317 ± 0.090	0.5992 ± 0.078	0.6593 ± 0.049	0.7156 ± 0.044
BS-CNN	0.5346 ± 0.107	0.6133 ± 0.092	0.6754 ± 0.071	0.7388 ± 0.048
EF-CNN	0.5389 ± 0.102	0.6195 ± 0.094	0.6797 ± 0.071	0.7402 ± 0.047
LF-CNN	0.5414 ± 0.113	0.6232 ± 0.095	0.6815 ± 0.068	0.7569 ± 0.049
SF-CNN	0.5462 ± 0.101	0.6309 ± 0.088	0.6963 ± 0.061	0.7675 ± 0.045
Combined	0.5405 ± 0.103	0.6038 ± 0.084	0.6779 ± 0.061	0.7355 ± 0.048

Table 2: Prediction AUC and standard deviation on the COPD cohort with different ratio of training data.

Method	60%	70%	80%	90%
Baseline	0.4536 ± 0.104	0.5738 ± 0.087	0.6324 ± 0.062	0.6624 ± 0.053
BS-CNN	0.4643 ± 0.101	0.5814 ± 0.085	0.6512 ± 0.065	0.7072 ± 0.058
EF-CNN	0.4625 ± 0.095	0.5854 ± 0.081	0.6533 ± 0.059	0.7083 ± 0.048
LF-CNN	0.4517 ± 0.112	0.5865 ± 0.091	0.6583 ± 0.073	0.7109 ± 0.067
SF-CNN	0.4749 ± 0.106	0.6086 ± 0.078	0.6735 ± 0.064	0.7388 ± 0.055
Combined	0.4572 ± 0.099	0.5815 ± 0.086	0.6523 ± 0.061	0.6924 ± 0.050

top-activated examples are longer since it can capture global patterns. SF-CNN can locate between those two. Similar trends can be observed in Figure 5(b) for COPD. The average length of patterns on COPD is shorter than that on CHF, which indicates that COPD data contains more discriminative local patterns.

By aggregating weights of the neurons assigned to each medical features, we can obtain the importance of each medical feature for prediction. In Table 3, we show a list of top ranked medical features for CHF and COPD predictions. From the table we can observe that the important features for CHF prediction are those those heart disease including heart failure itself, cardiac dysrhythmias, hypertension, etc. Some chronic renal and lung disorders are also there because they are common comorbidities of CHF. For COPD, the most important risk factors are of course chronic lung disorders including symptoms on respiratory system and chest, chronic airways obstruction and asthma. Diabetes and hypertensions are also observed because those chronic conditions are usually co-existing with each other.

6 Conclusion

We propose a deep learning framework for analyzing patient EHRs in this paper. Our framework is composed of four layers: input layer, one-side convolution layer, max-pooling layer and softmax prediction layer. Different temporal fusion mechanisms are also investigated to explore temporal smoothness of patient EHRs in the proposed framework. Finally we validate the effectiveness of the proposed model on both synthetic and real world data quantitatively and qualitatively.

One potential future work could be exploit parameters reducing framework [38, 39] to further light the CNN model to prevent over-fitting. Applying the current work to other domain [40] is also promising.

References

[1] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[2] "Data driven healthcare," *MIT Technology Review Business Report*, vol. 117, no. 5, pp. 1–19, 2014.

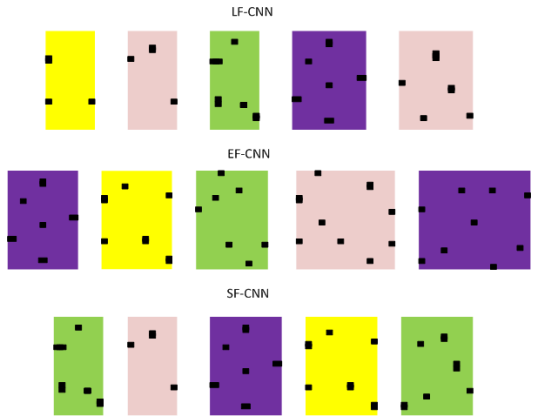
[3] L. B. Madsen, *Data-Driven Healthcare: How Analytics and BI are Transforming the Industry*. Wiley, 2014.

[4] F. Wang, P. Zhang, B. Qian, X. Wang, and I. Davidson, "Clinical risk prediction with multilinear sparse logistic regression," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 145–154.

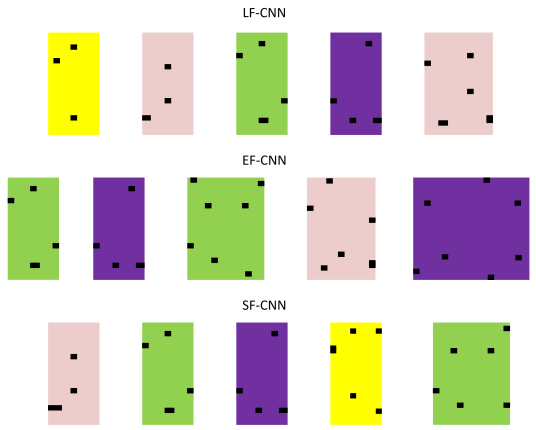
[5] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: Leveraging patient similarity and drug similarity analytics," *AMIA Joint Summits on Translational Science*, 2014.

[6] F. Wang, J. Sun, and S. Ebadollahi, "Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 1, pp. 54–69, 2012.

[7] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 117–121, 2013.



(a) CHF



(b) COPD

Figure 5: Top-5 examples of highly activate model's top-layer neurons for CHF and COPD.

Diseases: CHF		
Weight	DxGrp	Description
0.101	428	Heart failure
0.083	427	Cardiac dysrhythmias
0.075	272	Diso. of lipoid metabolism
0.066	250	General sympt.
0.052	401	Essential hypertension
0.043	414	Other chronic ischemic heart disease
0.039	585	Chronic renal failure
0.035	785	Symptoms involving cardiovascular system
0.033	518	Other diseases of lung
0.032	493	Asthma
Diseases: COPD		
Weight	DxGrp	Description
0.122	513	Other diseases of lung
0.094	786	Symptoms involving respiratory system and other chest symptom
0.081	496	Chronic airways obstruction
0.070	493	Asthma
0.052	250	Diabetes mellitus
0.041	427	Cardiac dysrhythmias
0.039	401	Essential hypertension
0.038	588	General symp.
0.035	272	Diso. of lipoid metabolism
0.035	414	Other unspecified disorder of joint

Table 3: Top 10 medical features for predication and their diagnosis group codes for CHF and COPD cohort. In each disease, we provide the normalize the weights of the medical features and rank the features.

[8] J. Pathak, A. N. Kho, and J. C. Denny, "Electronic health records-driven phenotyping: challenges, recent advances, and perspectives," *Journal of the American Medical Informatics Association*, vol. 20, no. e2, pp. e206–e211, 2013.

[9] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 135–144.

[10] J. C. Ho, J. Ghosh, and J. Sun, "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 115–124.

[11] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 453–461.

[12] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," DTIC Document, Tech. Rep., 1961.

[13] B. J. Miriovsky, L. N. Shulman, and A. P. Abernethy, "Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care," *Journal of Clinical Oncology*, vol. 30, no. 34, pp. 4243–4248, 2012.

[14] R. Amarasingham, B. J. Moore, Y. P. Tabak, M. H. Drazner, C. A. Clark, S. Zhang, W. G. Reed, T. S. Swanson, Y. Ma, and E. A. Halm, "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data," *Medical care*, vol. 48, no. 11, pp. 981–988, 2010.

[15] X. Wang, F. Wang, J. Hu, and R. Sorrentino, "Exploring joint disease risk prediction," in *Proceedings of the Annual Symposium of American Medical Informatics Association (AMIA)*, 2014.

[16] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, "Limestone:

High-throughput candidate phenotype generation via tensor factorization," *Journal of biomedical informatics*, vol. 52, pp. 199–211, 2014.

[17] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of biomedical informatics*, vol. 48, pp. 148–159, 2014.

[18] A. Perer and F. Wang, "Frequence: interactive mining and visualization of temporal frequent event sequences," in *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014, pp. 153–162.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[20] A. Graves, A. rahman Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[22] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, ser. WWW Companion '14. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 373–374.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Nov. 2011.

[25] L. Xu, K. Liu, S. Lai, and J. Zhao, "Product feature mining: Semantic clues versus syntactic constituents," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014, pp. 336–346.

[26] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *CoRR*, vol. abs/1412.1058, 2014.

[27] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014, pp. 655–665.

[28] Y. Xie, P. Daga, Y. Cheng, K. Zhang, A. Agrawal, and A. Choudhary, "Reducing infrequent-token perplexity via variational corpora," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 609–615. [Online]. Available: http://www.aclweb.org/anthology/P15-2101

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classication with convolutional neural networks," in *Proceedings of International Computer Vision and Pattern Recognition (CVPR 2014)*, 2014.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 568–576.

[31] A. Bergamo, L. Bazzani, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," *CoRR*, vol. abs/1409.3964, 2014.

[32] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[33] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3021–3028.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, *et al.*, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[36] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[37] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning." in *ICCV*. IEEE, pp. 2018–2025.

[38] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. N. Choudhary, and S. Chang, "Fast neural networks with circulant projections," *CoRR*, vol. abs/1502.03436, 2015.

[39] Y. Cheng, F. X. Yu, R. Feris, S. Kumar, and S.-F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *International Conference on Computer Vision (ICCV)*, 2015.

[40] L. Liu, J. Tang, Y. Cheng, A. Agrawal, W.-k. Liao, and A. Choudhary, "Mining diabetes complication and treatment patterns for clinical decision support," in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, New York, NY, USA, pp. 279–288.