This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

HCET: Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Record for Predicting Depression

Yiwen Meng, William Speier, *Member*, Michael Ong and Corey W. Arnold, *Member, IEEE*

Abstract— **Recent developments in machine learning algorithms have enabled models to exhibit impressive performance in healthcare tasks using electronic health record (EHR) data. However, the heterogeneous nature and sparsity of EHR data remains challenging. In this work, we present a model that utilizes heterogeneous data and addresses sparsity by representing diagnoses, procedures, and medication codes with temporal Hierarchical Clinical Embeddings combined with Topic modeling (HCET) on clinical notes. HCET aggregates various categories of EHR data and learns inherent structure based on hospital visits for an individual patient. We demonstrate the potential of the approach in the task of predicting depression at various time points prior to a clinical diagnosis. We found that HCET outperformed all baseline methods with a highest improvement of 0.07 in precision-recall area under the curve (PRAUC). Furthermore, applying attention weights across EHR data modalities significantly improved the performance as well as the model's interpretability by revealing the relative weight for each data modality. Our results demonstrate the model's ability to utilize heterogeneous EHR information to predict depression, which may have future implications for screening and early detection.**

***Index Terms*—Clinical decision support, deep learning, electronic health record, depression, temporal representation and reasoning.**

I. Introduction

With the rapid development of deep learning algorithms and widespread use of healthcare data sets, many models have presented state-of-the-art performance using patients' electronic health records (EHRs) for diagnostic tasks [1], disease detection [2], and risk prediction [3]. EHRs have been broadly adopted for documenting a patient's medical history [4]. They are composed of data from various sources, including diagnoses, procedures, medications, clinical notes, and laboratory results, which contribute to their high dimensionality and heterogeneity. Frequently, models built on EHR data have limited the number of data categories used [5], [6]. Few studies have attempted to use data from a broad set of categories as data heterogeneity remains a technical barrier for utilizing all types of EHR data in one model. As a consequence, there is an ongoing effort to construct a single model that is able to aggregate data from different data modalities. An additional complication is that EHR data includes temporal information from different patient visits, with each visit producing data from various sources.

Depression is one of the leading causes of disability worldwide [7]. Many depressed patients seek treatment from primary care providers, as 15% of primary care patients screen positive for depression, which makes improvement in the quality of depression care in primary care settings vital [8]**.** Despite the high prevalence and cost of depression, a previous meta-analysis found that the screening process for patients at high risk of depression only produced a true positive rate of 50% [9]. Ensuring that screening targets high-risk individuals minimizes the workload for primary care providers, who do not have enough time to do all relevant preventive health care screening [10]**.** To address this problem, studies have utilized LASSO logistic regression [11], random forests [12], support vector machines (SVM) [13] for predicting depression. While these methods are able to handle some data modalities, they do not model the EHR's heterogeneous structure, thus presenting an opportunity for new techniques.

To construct a predictive model with high accuracy for prediction of depression and mitigate the heterogeneity and sparsity of EHR data, we propose Hierarchical Clinical Embedding with Topic modeling (HCET), which aggregates diagnoses, procedure codes, medications, and demographic information together with topic modeling of clinical notes. Inspired by [5], HCET builds a hierarchical structure on different categories of EHR data with various embedding levels, while preserving the data's sequential nature. In this way, it learns the inherent interaction between EHR data from various sources within each visit and across multiple visits for an individual patient. This study points to a potential method for targeting depression screening among individuals in a single health system who have conditions that are associated with high risk for depression**.** Depression is often not evaluated in primary care settings. This approach could help in clinical practice by identifying individuals potentially at risk for developing depression within a specific time interval who should be screened (and potentially treated) for depression.

This work was supported by the National Heart, Lung, and Blood Institute (NIH/NHLBI R01HL141773).

Y. Meng, is with the Computational Diagnostic Lab, the Department of Bioengineering, at the University of California Los Angeles, 924 Westwood Blvd, Suite 420, CA 90024 USA (e-mail: lanyexiaosa@ucla.edu)

W. Speier is with the Computational Diagnostic Lab, the Department of Radiology at the University of California Los Angeles, 924 Westwood Blvd, Suite 420, CA 90024 USA (e-mail: speier@ucla.edu).

M. Ong is with the Department of Medicine at the University of California Los Angeles, 924 Westwood Blvd, Suite 420, CA 90024 USA (e-mail:mong@mednetucla.edu).

C.W. Arnold is with the Computational Diagnostic Lab, the Department of Bioengineering, the Department of Radiology and the Department of Pathology at the University of California Los Angeles, 924 Westwood Blvd, Suite 420, CA 90024 USA (e-mail:cwarnold@ucla.edu)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

II. Related Work

Temporal models based on RNN or LSTM have been applied on medical data, particularly for using EHR data to predict future diagnoses [1], [14]. [15] added an attention mechanism to an RNN to predict heart failure, which improved the model's interpretability for predicting the time of an event. Bai et. al also focused on improving their model's interpretability using self-attention, but only applied it on diagnosis and procedure codes [16]. [5] focused on learning the inner structure of an EHR by constructing a multiple level embedding with a bottom up hierarchy of diagnosis level, visit level, and patient level. However, these models did not apply on wide range of EHR data sources. [17] was able to predict clinical interventions from a deep neural network using lab results and demographics, but with a smaller feature dimension of 34 in total. Thus, this method did not resolve the data heterogeneity and sparsity issues for EHR data.

Several previous studies have focused on semantic representation of clinical notes. Gligorijevic et. al proposed a model with attention to process clinical text with several hand crafted features for chronic disease prediction [18]. [11] was able to include diagnosis codes, demographic information, and clinical notes for predicting a future diagnosis of depression. However, their approach processed unstructured clinical text using a medical ontology for medical term extraction. In addition, it ignored temporal information by building a logistic regression classifier, which is a non-temporal model. [6] first applied topic modeling to parse clinical notes and combined it with other data modalities to input into an autoencoder as a feature extractor, while building a random forest classifier for future disease prediction. This method is a two-stage model, which incurs additional complexity to optimize compared to one end-to-end model. Our model aims to achieve better semantic representation of clinical notes and aggregates them with other EHR data to improve predicting diagnosis of depression. In addition, hierarchical embedding was built to reveal latent connection between various EHR data source to resolve data heterogeneity and sparsity issues.

III. Data Description

To capture a spectrum of clinical complexity for our analyses, we selected patients based on three primary diagnoses: myocardial infarction (MI), breast cancer, and liver cirrhosis. Generally, MI represents the least complexity, with acute onset, resolution, and straight-forward treatment. Breast cancer is increasingly complicated in terms of diagnoses and treatment options. Finally, a patient with liver cirrhosis may have many sequelae, generating a complex EHR representation. Patients for this project were identified from our EHR in accordance with an IRB (#14-000204) approved protocol. Each patient visit had EHR data types consisting of diagnosis codes in International Classification of Disease, ninth revision (ICD-9) format, procedure codes in Current Procedural Terminology (CPT) format, medication lists, demographic information, and clinical notes. All patient records coded with ICD-9 values for MI, breast cancer, or liver cirrhosis from 2006-2013 were included. In this data set, demographics were limited to the patient's gender and age at the time of each visit. Initially, there were 45,208 patients and after the preprocessing and patient including criteria in section III. D, 10,148 patients were included in the analysis. Table I shows statistics of the dataset. Note that there some patients have more than one primary diagnosis.

TABLE I
Statistics of EHR dataset

# of patients with MI	2,943 (1,280 depressed)
# of patients with breast cancer	5,568 (1,960 depressed)
# of patients with liver cirrhosis	2,218 (772 depressed)
Gender	Male (27.46%), Female(72.54%)
Age	68.78 ± 15.46, min: 18, max 98

A. Identifying Diagnosis of Depression

Because patients in this dataset were identified retrospectively and were not suspected for depression, common methods for identifying and assessing severity of depression such as Patient Health Questionnaire (PHQ-9) scores [19] were not available. Instead, depression onset was identified by three methods:

- depression related ICD-9 code [11]
- inclusion of an antidepressant drug in a patient's medication list
- appearance of an antidepressant drug in clinical notes (from https://www.whocc.no/atc_ddd_index/?code=N06A)

The earliest time stamp of an occurrence of any of these events was defined as the time of diagnosis with depression. In total, 3,047 patients out of the total 10,148 were identified as depressed. The diagnosis time of depressed for each patient occurred after the primary diagnosis.

IV. Methods

ICD-9 codes, CPT codes, medication lists, and patient's gender can all be considered as categorical variables while ages are numerical. Therefore, an intuitive approach is to encode these features in a multi-hot vector, where each row corresponds to a specific code or data element. Each row has a binary value, where 1 indicates have this item and 0 for not during one visit. ICD-9 codes are up to five digits long with three digits before a decimal point and two digits after, resulting in 9,285 unique code in our data set. In order to reduce the dimensionality of the feature vector, ICD-9 codes were grouped by the three numbers before the decimal point, as was previously done in [14]. Detailed descriptions of dimensionality reduction techniques for ICD-9, CPT, and medication lists are presented in section IV.C, definition of HCET.

Embedding is a technique that has been widely adopted in NLP to project long and sparse feature vectors into a dense lower dimensional space [20]. This approach efficiently reduces the size of a model's parameters as well as decreases training time. Recent models [5], [14], [21] have utilized embedding to process categorical data in EHRs, which we have adopted in the current model. The full definition is shown in section IV.C.

A. Topic Modeling of Clinical Notes

Latent Dirichlet allocation (LDA) is an unsupervised learning method to encode text by assigning words to underlying topics (semantic themes). Briefly, a topic is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

represented as a multinomial distribution over the unique words in a corpus, and a document is represented as a multinomial distribution over all topics. LDA is able to generate topics automatically from a corpus, providing generalized information. Recent work has applied topic modeling on clinical notes [22]–[25]. We chose to model clinical notes with 100 topics, each one with five words with highest probability to represent the semantic mean of the clinical notes, thus generating a 100-feature vector representation of the document in semantic topic space. Topic vectors was dichotomize using each topic's average value as a threshold among our data. For patients with multiple clinical reports in the six-month time window, probabilities were averaged first to reach on feature vector and then dichotomized using the same method.

B. Baseline Models

Traditional machine learning algorithms generally ignore temporal and sequential correlation among features by aggregating them over a time window for a patient. As mentioned in the first paragraph of section IV, the feature vector for each patient is a multi-hot vector which concatenated all five EHR data modalities over multiple visits. In order to leave out the bias for more frequent codes, each row of vector is 1 when this code shows in any of the visits. As a compensation factor for temporal information, the number of records in ICD-9, CPT, medication lists, and clinical notes are added as addition factors to capture the of frequency of patients visits of records. 10-fold cross validation was adopted for each model. In addition, patients in the test set were separated by their primary diagnosis and the results were compared for three primary diagnosis individually.

Lasso Previous work has applied Lasso for predicting depression [11], which is compared in the analysis. Lasso uses L1 regularization which brings sparsity to select the more correlated features for the task.

SVM SVM is also compared in the experiment as it been utilized to predict depression previously [13]. Here we used RBF kernel and five-fold cross validation with grid search to fine tune the regularization term.

MLP Multilayer perceptron (MLP) Two layers of MLP with a tanh activation function and 256 nodes is also compared here, following the implementation from previous studies [5], [15].

RF Nevertheless, ensemble methods like random forests (RF) [26] and gradient boost regression trees (GBRT) [27] have produced competitive results in disease detection and outcome prediction for healthcare. These models also compute the significance factor for each feature, which provides valuable information on feature selection as well as dimension reduction. Therefore, RF was adopted as a baseline model in comparison with HCET. The hyper parameters were chosen using grid search with five-fold cross validation on the training set.

VAE+RF [6] proposed pretraining autoencoder as the feature extractor for EHR and using RF for classification from the extracted features. This method is also compared.

MiME* The MiME model demonstrated state-of-the-art performance in predicting heart failure onset [5]. It consists of a temporal model using GRUs that learn the temporal character of disease progression with external knowledge of linked relation between ICD-9 codes and associated CPT codes and medication lists during each visit. The MiME model required removal of visits that did not include diagnosis codes to make sure diagnosis codes were present to input the model. Since there was no direct linked relationship between ICD-9 codes, CPT codes, and medication lists in our EHR data, these three features were processed in the same level instead of the two level structure proposed in MiME. In addition, there are many cases where procedure codes or medications are present in the EHR without associated diagnoses. Therefore, we revised the MiME model by removing this layer while keeping the remaining structure and parameter values consistent, denoted MiME*. The performance of this modified model was compared to our HCET model.

$$L_{aux} = -\lambda_{aux}\sum_{t}^{T}\Big(\sum_{i}^{|v^{(t)}|} CE\left(d_i^{(t)}, \hat{d}_i^{(t)}\right) + \sum_{i}^{\left|M_i^{(t)}\right|} CE\left(m_{i,j}^{(t)}, \hat{m}_{i,j}^{(t)}\right) \quad (1)$$

As shown above, MiME defined Eq. (1) to compute the auxiliary loss, where $d_i^{(t)}$ denoted the diagnosis code in t^{th} visit. Thus, calculating auxiliary loss required diagnosis codes present in each visit, which and this is not applicable to our dataset. On the other hand, we highly focus on the prediction accuracy of depression but not on other diseases or symptoms. Furthermore, the average increase after implementing this component was less than 0.01 from their reported results, so the auxiliary loss defined in MiME was not adopted in this study.

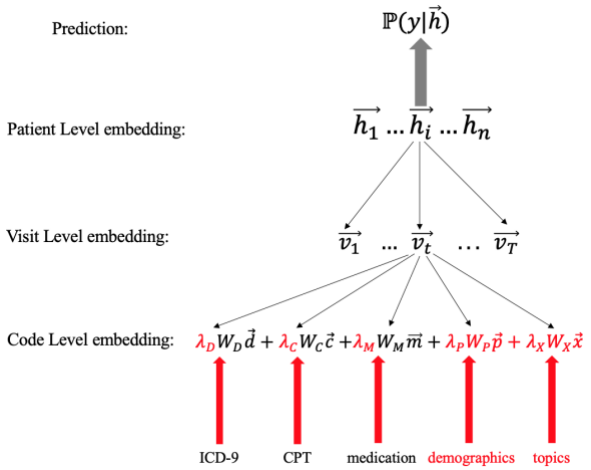

Fig. 1. Illustration of for EHR data. There are three levels of embedding: patient level, visit level and code level. λ denotes the attention weight for each embedding. The full explanation of symbols is described in Table II. The red color shows the new components added from MiME.

C. Definition of HCET

Fig. 1 illustrates the hierarchical structure of HCET. The ultimate goal of the model is to predict the probability of a chronic disease for patient i given the feature embedding representing a sequence of visits, $\mathbb{P}(y_i|\overrightarrow{\mathrm{h_i}})$. While the model is designed to be generalizable, we focus here on the prediction of depression, y_i. $\overrightarrow{\mathrm{h_i}}$ stands for the patient level embedding of a patient's EHR, and each patient has multiple hospital visits from $\overrightarrow{v_1}$ to $\overrightarrow{v_t}$, which compose the visit level embedding. During one visit $\overrightarrow{v_t}$, the code level embedding $\overrightarrow{e_t}$ is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

ensemble of multiple ICD-9 and CPT codes, medications, demographic information, and topic features extracted from associated clinical notes. Since there are five categories of EHR data, we built individual embedding for each first and aggregated them together.

TABLE II
NOTATION USED IN THE FORMULATION OF HCET

Notation	Definition
D	Unique set of ICD-9 codes
C	Unique set of CPT codes
M	Unique set of medications
X	Set of 100 topic features
P	Demographic information
λ_j	Attention weight for one data modality, j ∈ (*D,C,M,X,P*)
$\overrightarrow{e_t} \in \mathbb{R}^z$	Vector representation of summed EHR data at the t-th visit
$\overrightarrow{v_t} \in \mathbb{R}^z$	Vector representation of $t \in [1 \dots T]$ visit EHR data for a patient
$\overrightarrow{\mathrm{h_i}} \in \mathbb{R}^z$	Vector representation of EHR data for patient number i

The dimension of embedding z is the same for associated vectors due to the residual connection used in HCET.

Table II shows the full list of notation and corresponding definitions of symbols used in HCET. $\vec{d}$ is a multi-hot binary vector with dimension of $\mathbb{R}^{D\times1}$, where each column corresponds to whether a specific ICD-9 code was assigned in the t_{th} visit. A similar approach applies to $\vec{c} \in \mathbb{R}^{C\times1}$, $\vec{m} \in \mathbb{R}^{M\times1}$, and $\vec{p} \in \mathbb{R}^{2\times1}$, which are the vector representations for CPT, medication, and demographic information, respectively. As described before, topic features are vector representations in topic space, which represent the distribution of topic occurrences in the document. In order to match the embedding size of other data types, a threshold was defined to dichotomies each topic word, which was computed by the average probability of each topic value across all patients. The threshold value for 100 topic words are described in the results section. Thus, $\vec{x} \in \mathbb{R}^{100\times1}$ is a multi-hot binary vector representation of topic features, where each column denotes the unique 100 topics for the t_{th} visit.

Eq. (2), Eq. (3) and Eq. (4) describe mathematical formulation of HCET in the top-down view, denoting the *Patient level*, *Visit level*, and *Code level* embeddings, respectively.

$$\overrightarrow{\mathrm{h_i}} = f(\overrightarrow{v_1} \dots \overrightarrow{v_t} \dots \overrightarrow{v_T}) \tag{2}$$

Eq. (2) shows the method to process temporal information of various visit level embeddings to compute a patient level embedding, where f stands for the function to input visit information in a sequential order. As mentioned before, RNNs, LSTMs, and GRUs have been widely used to fulfill this task. Since RNNs often encounters the vanishing gradient problem and better performance has been shown for a GRU over an LSTM in previous work [5], we used a GRU in the current model.

$$\overrightarrow{v_t} = \alpha(W_e \overrightarrow{e_t}) + \overrightarrow{e_t} \tag{3}$$

In Eq. (3), visit level embedding is generated by first performing a matrix transformation with weight $\boldsymbol{W}_e \in \mathbb{R}^{z\times z}$, followed by a non-linear ReLU transformation function α, where z is the embedding size. We omitted the bias term $\overrightarrow{b_t}$ here to formulate residual connection [28].

$$\overrightarrow{e_t} = \beta(F) + F \tag{4}$$

$$F = W_D \vec{d} + W_C \vec{c} + W_M \vec{m} + W_P \vec{p} + W_X \vec{x} \tag{5}$$

Eqs. (4) and (5) define the code level embedding by summing individual embeddings from five EHR data sources with a non-linear transformation function β. As in equation (3), we use a ReLu for β. The $W_D \in \mathbb{R}^{z\times D}$, $W_C \in \mathbb{R}^{z\times C}$, $W_M \in \mathbb{R}^{z\times M}$, $W_P \in \mathbb{R}^{z\times P}$ and $W_X \in \mathbb{R}^{z\times X}$ represent the weight matrices for transforming the feature vectors of ICD-9 codes, CPT codes, medication lists, demographics, and topic features with high and varied dimensionality into a latent space with the same lower dimension. For example, the diagnosis vector $\vec{d} \in \mathbb{R}^{D\times1}$, after multiplied with weight matrix, $W_D\vec{d}$ results in a vector of dimension $\mathbb{R}^{z\times1}$. Therefore all vectors can sum up as in Eq. (5). In the same manner to Eq. (3), all of the corresponding biased terms were omitted to denote residual connection. Finally, binary cross entropy was used as the loss function.

$$\sum \lambda_j = 1 \tag{6}$$

$$F' = \lambda_D W_D \vec{d} + \lambda_C W_C \vec{c} + \lambda_M W_M \vec{m} + \lambda_P W_P \vec{p} + \lambda_X W_X \vec{x} \tag{7}$$

In order to investigate the importance of each data modality in this prediction task as well as improve interpretability of HCET, attention weights λ_j were defined for each modality, where the sum of all weights equal to one, as shown in Eq. (6). A weighted sum code level embedding F' was input into HCET, indicated by Eq. (7), which substituted F in Eqs. (4) and (5). After training, attention weights reveal the importance for each feature type in prediction tasks.

D. Predicting Depression at Different Decision Points

Previous studies [1], [2], [5], [21] have used the data from the entire EHR for future disease prediction. This method could add bias for patients with longer medical histories. It also gives equal weight to old data that likely is not as useful as more recent data. As predicting the future risk of a disease in a prospective setting is an ongoing task, the time window of a patient's EHR is highly varied. Therefore, as a similar approach to [11], we defined four decision points in advance of the diagnosis of depression: two weeks, three months, six

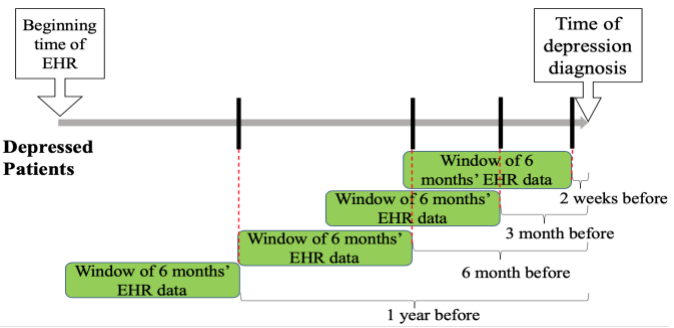

Fig. 2. Illustration of prediction at different time windows in advance of diagnosis of depression. The beginning time of EHR is defined by the timestamp of the primary diagnosis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

months, and one year. Fig.2 illustrates the four six-month time windows for using EHR data to predict depression diagnosis. For non-depressed patients, the last time step of the EHR was substituted for the diagnosis time.

In order to test the effect of temporal information and data size on model performance, previous work [5] used varying maximum lengths (visits) of the EHR. This resulted in a different number of patients in each of the four experiments as the number of visits was not consistent across patients. In our approach, we kept the number of patients consistent through the four predicting tasks, which revealed the temporal nature of prediction as the time to diagnosis varies. In this case, patients who had at least one of ICD-9, CPT, medication and topic feature in all four time windows were included in experiments. After processing data based on this method, 10,148 patients were selected, where 3,747 were diagnosed with depression. Basic statistics of the data are shown in Table III.

Ablation study Three feature sets were generated to compare the contribution to prediction of depression for demographics and topic features. There were applied to HCET: all data types (ICD-9 codes, CPT codes, medication lists, demographics, and topic features); ICD-9 codes, CPT codes, medication lists and topic features; ICD-9 codes, CPT codes, and medication lists. This ablation study was only applied to HCET while all baseline models used all data types as input.

TABLE III
STATISTICS OF DATA INPUT FOR HCET

Total # of patients	10,148 (Depressed:3,747; Non-depressed:6,401)
Total # of visits	294,941
Avg. # of visits	29.06
# of unique codes	D:1391, C:6927, M: 4181
# of demographics per visit	2 (Age, Gender)
# of topics per visit	100
Max / Avg. # of ICD-9 codes per visit	69 / 1.74
Max / Avg. # of CPT codes per visit	106 / 3.23
Max / Avg. # of medication per visit	14 / 0.09
Max / Avg. # of topics per visit	30 / 1.87

E. Training Details

All models were implemented in TensorFlow 1.12 and trained on a work station equipped with Intel Xeon E3-1245, 32 GB RAM and two NVIDIA Ti 1060 GPUs. Adam [29] was selected as the optimizer, with the same learning rate of $1e^{-3}$ as [5] for HCET. The number of parameters is 2.5M, which mainly depends on the size of embedding matrices. Reported results are averaged over 10 random data splits: training 70%, validation 10% and test 20%. Models were trained with the minibatch of 50 patients for a total of 2,000 iterations to guarantee convergence. The validation set was evaluated at every 100 iterations for early stopping. The vanishing gradient problem was avoided by using skip connections. To address over fitting, L2 regularization with coefficient $1e^{-4}$ was

TABLE IV
COMPARISON OF PREDICTION PERFORMANCE FOR DIFFERENT MODELS

Prediction window	Two weeks		Three months		Six months		One year	
Models	ROCAUC	PRAUC	ROCAUC	PRAUC	ROCAUC	PRAUC	ROCAUC	PRAUC
Lasso (codes+demo+topic)	0.66 (0.01)	0.55 (0.02)	0.65 (0.02)	0.52 (0.03)	0.63 (0.02)	0.51 (0.03)	0.63 (0.02)	0.50 (0.03)
SVM (codes+demo+topic)	0.72 (0.02)	0.62 (0.03)	0.69 (0.01)	0.59 (0.02)	0.68 (0.0176)	0.57 (0.02)	0.68 (0.02)	0.57 (0.03)
MLP (codes+demo+topic)	0.72 (0.01)	0.64 (0.01)	0.70 (0.02)	0.60 (0.02)	0.69 (0.02)	0.58 (0.02)	0.68 (0.02)	0.57 (0.02)
RF (codes+demo+topic)	0.76 (0.02)	0.67 (0.03)	0.73 (0.02)	0.62 (0.03)	0.70 (0.02)	0.59 (0.02)	0.69 (0.02)	0.58 (0.03)
VAE+RF (codes+demo+topic)	0.76 (0.02)	0.67 (0.02)	0.74 (0.01)	0.64 (0.02)	0.71 (0.03)	0.60 (0.02)	0.69 (0.01)	0.60 (0.02)
MiME* (codes)	0.76 (0.01)	0.67 (0.02)	0.74 (0.01)	0.64 (0.02)	0.72 (0.02)	0.61 (0.01)	0.70 (0.01)	0.61 (0.01)
HCET (codes+demo)	0.76 (0.01)	0.68 (0.01)	0.75 (0.02)	0.65 (0.02)	0.73 (0.02)	0.62 (0.01)	0.71 (0.01)	0.61 (0.01)
HCET (codes+demo+topic)	**0.81** † **(0.01)**	0.73 † (0.02)	0.80 † (0.02)	**0.71** † **(0.02)**	0.78 † (0.01)	0.68 † (0.02)	0.75 † (0.01)	0.66 † (0.02)
HCET + attention (codes+demo+topic)	0.81 (0.01)	**0.73 (0.01)**	**0.80 (0.01)**	0.70 (0.02)	**0.79** (0.01)**	**0.69 (0.01)**	**0.78** (0.01)**	**0.67 (0.01)**

Codes denote data from ICD-9, CPT, and medication lists, while *demo* stands for demographic information. Values in parenthesis refer to standard deviations across randomizations and bold values denotes the highest in each column. † indicates the value is significantly better than MiME* ($p<0.05$) while ** denotes the value is significantly better than no attention ($p<0.05$).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

chosen for the two HCET models instead of using dropout. The embedding size z was set as 200 and the number of nodes for the GRU was set at 256. The source code of HCET is available at https://github.com/lanyexiaosa/hcet.

V. Results

A. Comparison of Performance in Depression Prediction

Table IV displays the results from all baseline models and HCET with abalation analysis at four time points in advance of diagnosis in terms of receiver operating characteristic area under the curve (ROCAUC) and PRAUC. HCET using all EHR modalities with attention outperformed other models for every prediction window. Lasso generated the worst accuracy. There is no significant difference between results from RF and VAE+RF. There is conistent decrease of accuracy for each model as the prediction point moves further away from the time of diagnosis, where the number achieves the highest at window of two weeks.

Adding demographic information and topic features improved the performance for HCET, which demonstrates their significant contribution in predicting depression as well as emphasizes the advantage of building a model being able to aggregate EHR data from multiple sources. The values between MiME* and HCET(codes+demo) are similar, while the difference between HCET(codes+demo) and HCET(code+demo+topics) are relatively large. HCET with all types of EHR data achieved the highest accuracy at each prediction than all baseline models. It generated the highest mean ROCAUC of 0.81 when predicting two weeks prior to the diagnosis, and the value dropped to 0.7541 when predicting one year in advance. After applying attention weights to each embedding at the code level, the ROCAUC at six months and one year are significantly improved with p=0.04 and p=3e-5, respectively.

B. Model's performance for each primary diagnosis

Table V shows the results for each of three primary diagnosis in predictions windows of two weeks and one year in ROCAUC and PRAUC. Only HCET+attention was compared as it demonstrated the best performance in the previous ablation study on its own. HCET+attention also achieved the best performance for three primary diagnosis for two prediction windows. The low variance also indicated that it is more robust than other models. The ROCAUC for every model is quite similar even though the number of patients with breast cancer was substantially higher than the other diseases (Table I), which indicated no bias toward any primary diagnosis in the prediction. On the other hand, it is noticeable that the PRAUC for patients with myocardial infarction is relatively higher than other two.

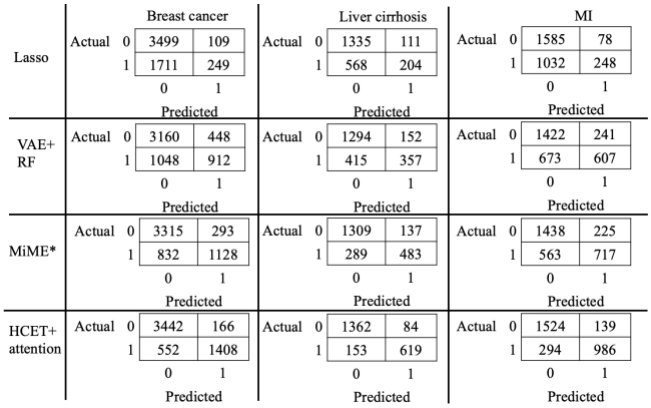

Fig. 3 Confusion matrix for patients separated by three primary diagnosis at a window of two weeks for four models. The numbers are aggregated together with 10-fold cross validation. Label 0 means non-depressed while 1 means depressed.

TABLE V
COMPARISON OF PREDICTION PERFORMANCE FOR THREE PRIMARY DIAGNOSIS

Prediction window	Two weeks						One year					
Diseases	Breast cancer		MI		Liver cirrhosis		Breast cancer		MI		Liver cirrhosis	
Models	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
Lasso	0.67 (0.02)	0.54 (0.03)	0.66 (0.02)	0.62 (0.03)	0.65 (0.02)	0.55 (0.02)	0.64 (0.03)	0.49 (0.03)	0.62 (0.02)	0.56 (0.04)	0.62 (0.04)	0.53 (0.02)
SVM	0.72 (0.02)	0.61 (0.03)	0.71 (0.02)	0.68 (0.03)	0.71 (0.02)	0.60 (0.03)	0.68 (0.03)	0.56 (0.03)	0.67 (0.02)	0.62 (0.03)	0.66 (0.02)	0.55 (0.02)
MLP	0.74 (0.02)	0.63 (0.02)	0.72 (0.02)	0.69 (0.02)	0.72 (0.02)	0.62 (0.02)	0.69 (0.01)	0.56 (0.02)	0.66 (0.01)	0.62 (0.02)	0.66 (0.02)	0.56 (0.02)
RF	0.76 (0.02)	0.66 (0.03)	0.74 (0.03)	0.71 (0.02)	0.75 (0.03)	0.65 (0.03)	0.70 (0.03)	0.57 (0.03)	0.67 (0.02)	0.63 (0.02)	0.67 (0.01)	0.57 (0.03)
VAE+RF	0.76 (0.02)	0.67 (0.02)	0.75 (0.02)	0.71 (0.01)	0.75 (0.02)	0.65 (0.02)	0.70 (0.02)	0.58 (0.03)	0.68 (0.01)	0.63 (0.01)	0.68 (0.02)	0.58 (0.02)
MiME*	0.77 (0.01)	0.67 (0.02)	0.75 (0.01)	0.70 (0.02)	0.76 (0.02)	0.67 (0.01)	0.71 (0.02)	0.61 (0.01)	0.69 (0.02)	0.64 (0.01)	0.70 (0.01)	0.61 (0.02)
HCET+ attention	**0.81 (0.01)**	**0.73 (0.01)**	**0.79 (0.01)**	**0.77 (0.01)**	**0.80 (0.01)**	**0.72 (0.01)**	**0.78 (0.01)**	**0.67 (0.01)**	**0.77 (0.01)**	**0.71 (0.01)**	**0.77 (0.01)**	**0.66 (0.01)**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3004072, IEEE Journal of Biomedical and Health Informatics

Fig. 3 contains confusion matrices with patients separated in three primary diagnosis in the prediction window of two weeks from four models at the same threshold of 0.5, after probability calibration using isotonic regression [30]. VAE+RF is chosen here rather than SVM, MLP, RF and as it generated slightly higher results in previous analysis for non-temporal models. The numbers were aggregated from 10-fold cross validation. For each primary diagnosis, the distribution was imbalanced with a lower number of depressed patients. Lasso generated poor accuracy as it almost always predicted the negative class. VAE+RF slighted reduced false negative cases but the number of true negatives was worse than Lasso. MiME* both improved the numbers in true positives and true negatives while HCET+attention improved it further. The average precision and recall over three primary diagnosis from HCET+attention were 0.88 and 0.76, respectively.

C. Interpretation of Feature Importance from Attention Weights

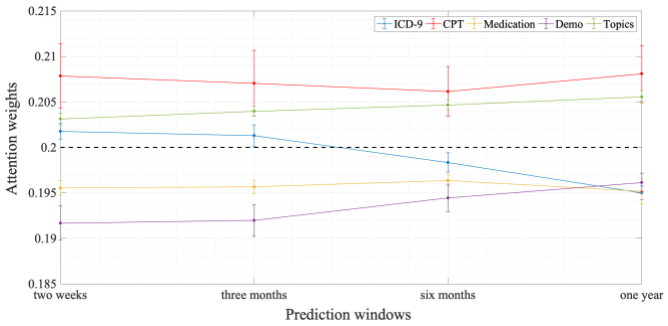

Fig. 4 Attention weights from every EHR data modalities in four prediction windows. Error bars denotes the standard deviation. The black dash line is at threshold of 1/5, which indicates constant weights in HCET models before.

As mentioned in the method section, one advantage of RF over majority of deep learning models is the ability to provide information on the importance factor of each feature contributing to classification [31]. However, there is a consistent effort to improve the interpretability of deep learning models like HCET. Fig. 4 shows the attention weights for each of EHR data modalities over four prediction windows According the result, medication and demo both are below the average value of 0.2 in four prediction windows. Attention for ICD-9 is above 0.2 in window of two weeks and three months but it drops in six months and one year. There is a consistent increase of attention for topics while the attention from CPT always ranks top.

VI. Discussion

Accoring to result in Table IV and V, Lasso generated the worst performance as it is a linear classifier which indicats that predicting depression from the EHR is a complicated task which requires more advanceted models. In addition, the Lasso method also provides sparsity of using more correlated features but the poor accuracy reveals that this task need to include more features than only the most correlated ones. There is no significant different between results from RF and VAE+RF which indicated the power of classfication maily depends on RF. Models starting from MiME* are all temporal models and they all achieved higher performance than non-temporal ones, which further confirms the advantage of using a temporal model over non-temporal methods in predicting chronic disease. Furthermore, the performance consistently declined for each model as the prediction window moved further away from the diagnosis time point, which agrees with our expectation that records closer to the diagnosis are more likely to contain relevant information and provide better predictions.

The improvement of HCET with attention over all baseline models demonstrated the advantage of utilizing temporal information and hierarchical embedding to aggregate more heterogenous EHR data modalities in the prediction of depression. In the original implementation of the MiME model [5], interactions between diagnosis codes with associated procedures and medication were explicitly modeled, but this linked relation was not available in our EHR, a situation that commonly applies to other medical systems. Meanwhile, MiME also has another limitation of ignoring data when no diagnosis code is present for each visit. Our results indicate that treating all EHR data types in one level of code embedding during each visit is a viable solution in this scenario while being able to include all data from each visit. Another adjustment in our model is the extension of embedding to process demographics and clinical notes, which further addresses the heterogeneity issue in EHR data. Furthermore, we applied attention weights on each data modality, which further improved our model's interpretability by showing the relative importance of each data modality.

The results presented in Table IV and V both demonstrate the contribution of topic features in temporal models for predicting depression. Future work may include more clinical notes with other EHR modalities in a single model when building machine learning models for healthcare tasks. In addition, the attention weights of topics were consistently above the average value, demonstrating their important contribution in our prediction task. Topic modeling methods, such as LDA, are one way of processing texts. They are based on the bag of words assumption, which may not be the ideal way to represent clinical text. Future studies could utilize more recent NLP tools, such as BERT [32] to process clinical notes, which could further improve the overall performance. On the other hand, our attention weights were not applied on individual visits and codes, so HCET did not learn the latent relation between them. Future work can improve this attention by applying BERT [32] to do representation learning using self-attention and a multi-head attention mechanism.

As mentioned in our methods section, the clinical standard for depression diagnosis is the PHQ-9 questionnaire, which is not routinely collected clinically. Instead, three criteria were used for determining depression diagnosis, which could have led to errors in our labels. Thus, future prospective studies could periodically administer PHQ-9 surveys, which may provide more precision in depression diagnosis. Temporal models could then be built to track the disease progression as well as early detection. There are other chronic diseases with high prevalence, such as hypertension, diabetes, and obesity,

which could provide more applications for the HCET model in future work. Finally, the EHR includes other data sources that are not currently included in the HCET model, such as laboratory results [17]. Future studies may extend the model to include these other data sources to further utilize the heterogeneity of EHR data.

VII. Conclusion

We have developed a temporal deep learning model, HCET, which was able to integrate five types of EHR data during multiple visits for depression prediction. HCET consistently outperformed the baseline models tested, achieving an increase in PRAUC of 0.07 over the best baseline model. The results demonstrate the ability of HCET as an approach to deal with data heterogeneity and sparsity in modeling the EHR. Adding attention weights improved model's interpretability. In future work, HCET could possibly be used as the basis for constructing a screening tool by utilizing the models' predictions to intervene with individuals who have a higher risk of developing depression.

References

[1] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016, pp. 1–18.

[2] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Am. Med. Informatics Assoc.*, vol. 24, no. 2, pp. 361–370, 2017.

[3] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records : A deep learning approach," *J. Biomed. Inform.*, vol. 69, pp. 218–229, 2017.

[4] N. Menachemi and T. H. Collum, "Benefits and drawbacks of electronic health record systems," *Risk Manag. Healthc. Policy*, vol. 4, pp. 47–55, 2011.

[5] E. Choi, C. Xiao, W. Stewart, and J. Sun, "MiME : Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare," *Adv. Neural Inf. Process. Syst. 31 (NIPS 2018)*, no. Nips, 2018.

[6] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci. Rep.*, vol. 6, pp. 1–10, 2016.

[7] S. L. James *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017," *Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[8] M. K. Ong and L. V. Rubenstein, "Wishing Upon a STAR*D: The Promise of Ideal Depression Care by Primary Care Providers," *Psychiatr. Serv.*, vol. 60, no. 11, pp. 1460–1462, 2009.

[9] A. J. Mitchell, A. Vaze, S. Rao, and R. Infi, "Clinical diagnosis of depression in primary care : a meta-analysis," *Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[10] K. S. H. Yarnall, K. I. Pollak, T. Østbye, K. M. Krause, and J. L. Michener, "Primary care: Is there enough time for prevention?," *Am. J. Public Health*, vol. 93, no. 4, pp. 635–641, 2003.

[11] S. H. Huang, P. LePendu, S. V Iyer, M. Tai-Seale, D. Carrell, and N. H. Shah, "Toward personalizing treatment for depression: predicting diagnosis and severity," *J. Am. Med. Informatics Assoc.*, vol. 21, no. 6, pp. 1069–1075, Nov. 2014.

[12] H. Jin, S. Wu, and P. Di Capua, "Development of a Clinical Forecasting Model to Predict Comorbid Depression Among Diabetes Patients and an Application in Depression Screening Policy Making," *Prev. Chronic Dis.*, vol. 12, pp. 1–10, 2015.

[13] J. Zhang, H. Xiong, Y. Huang, H. Wu, K. Leach, and L. E. Barnes, "M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data," *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, pp. 2569–2577, 2015.

[14] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks.," *Proc. Mach. Learn. Healthc. 2016*, vol. 56, pp. 301–318, 2016.

[15] E. Choi., M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," *Adv. Neural Inf. Process. Syst. 29 (NIPS 2016)*, 2016.

[16] T. Bai, B. L. Egleston, S. Zhang, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 43–51, 2018.

[17] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical Intervention Prediction and Understanding using Deep Networks," *arXiv:1705.08498v1*, pp. 1–16, 2017.

[18] D. Gligorijevic *et al.*, "Deep attention model for triage of emergency department patients," *Proc. 2018 SIAM Int. Conf. Data Min.*, pp. 297–305, 2018.

[19] M. Kurt Kroenke, MD; Robert L. Spitzer, "The PHQ-9 : A New Depression Measure," *Psychiatr. Ann.*, vol. 32, no. 9, pp. 509–515, 2002.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," *Adv. Neural Inf. Process. Syst. 2013)*, pp. 3111–3119, 2013.

[21] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 787–795.

[22] C. W. Arnold, S. El-Saden, A. A. T. Bui, and R. K. Taira, "Clinical case-based retrieval using latent topic analysis," *AMIA Annu. Symp. Proc.*, pp. 26–30, 2010.

[23] C. W. Arnold and W. Speier, "A topic model of clinical reports," in *35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1031–1032.

[24] C. W. Arnold, A. Oh, S. Chen, and W. Speier, "Evaluating topic model interpretability from a primary care physician perspective," *Comput. Methods Programs Biomed.*, vol. 124, pp. 67–75, 2015.

[25] W. Speier, M. Ong, and C. Arnold, "Using phrases and document metadata to improve topic modeling of clinical reports," *J. Biomed. Inform.*, vol. 61, pp. 260–266, 2016.

[26] E. Hsich, E. Z. Gorodeski, E. H. Blackstone, H. Ishwaran, and M. S. Lauer, "Identifying important risk factors for survival in patient with systolic heart failure using random survival forests," *Circ. Cardiovasc. Qual. Outcomes*, vol. 4, no. 1, pp. 39–45, 2011.

[27] N. Limsopatham, C. Macdonald, and I. Ounis, "Learning to Combine Representations for Medical Records Search," *Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 833–836, 2013.

[28] J. S. Kaiming He, Xiangyu Zhang, Shaoqing Ren, "Deep Residual Learning for Image Recognition Kaiming," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015, pp. 1–15.

[30] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *34th International Conference on Machine Learning*, 2017.

[31] Y. Meng *et al.*, "A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease Using Activity Tracker Data," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 3, pp. 878–884, 2020.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:11810.04805*, 2018.