M-SEQ: Early Detection of Anxiety and Depression via Temporal Orders of Diagnoses in Electronic Health Data

Jinghe Zhang†, Haoyi Xiong†, Yu Huang‡, Hao Wu†, Kevin Leach‡, Laura E. Barnes†
† Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA
‡ Department of Computer Science, University of Virginia, Charlottesville, VA
{jz4kg, hx6d, yh3cf, hw4tm, kjl2y, lbarnes}@virginia.edu

***Abstract*—According to a 2014 Spring American College Health Association Survey, almost 50% of college students reported feeling things were hopeless and that it was difficult to function within the last 12 months. More than 80% reported feeling overwhelmed and exhausted by their responsibilities. This critical subpopulation of Americans is facing significant levels of mental health disorders, challenging colleges to provide accessible and high quality behavioral health care. However, psychiatric disorders are frequently unrecognized in primary care settings, posing physical, emotional, economic, and social burdens to patients and others.**

Towards the goal of earlier identification and treatment of mental health disorders, this paper proposes *M-SEQ*, an early detection framework for anxiety/depression using electronic health data from primary care visit sequences. Specifically, compared to existing methods that predict a future disease state using frequency of diagnoses in a patient's medical history, we hypothesize that future disease might also be correlated with the temporal orders of diagnoses. Thus, *M-SEQ* first discovers a set of diagnosis codes that are discriminative of anxiety/depression, and then extracts each diagnosis pair from each patient's health record to represent the temporal orders of diagnoses. Further, it incorporates the extracted temporal order information with the existing representation to predict whether a patient is at risk of anxiety/depression. We evaluate *M-SEQ* using the electronic health record (EHR) data of 213,112 college students from 10 schools participating in the College Health Surveillance Network (CHSN) from January 1, 2011 through December 31, 2014. The experimental results shows that our framework can detect a future diagnosis of anxiety and depression based on the primary care visit data up to 3 months in advance, with approximately 1%–4.5% higher accuracy, compared to baseline methods using frequency of diagnoses.

***Keywords*-predictive models, early detection, anxiety/depression, temporal order, electronic health data**

I. Introduction

Psychiatric disorders in the college student population have increased in frequency and severity in the United States with 18.6% of adults suffering from at least one active mental health disorder [1]. Anxiety and mood disorders occur more frequently with incidence rates of 18.2% and 9.6% among adults, respectively [2]. College students responding to the Spring 2014 American College Health Association's National College Health Assessment reported feeling things were hopeless (46%), felt overwhelming anxiety (54%), and 86% reported feeling overwhelmed by all they had to do [3]. Suicide is also the leading cause of death [3]. Early detection and treatment of depression or anxiety disorders are crucial to preventing more severe outcomes. Psychiatric disorders are frequently unrecognized in primary care settings. As the most accessible institute for students' health care, student health centers hold the best opportunity to improve the early detection and intervention of anxiety/depression disorders.

Despite significant resources being devoted to the provision of mental health services at most colleges and universities in the United States, there have been no efforts towards data-driven methodologies for early detection of mental health disorders in primary care settings. Temporal data from a patient's electronic health record could be used to aid clinicians in the identification of patients at high risk of a mental health disorders potentially leading to more timely treatment and better health outcomes. This work represents the first attempt in college health to develop predictive models for anxiety and depression based on primary care data.

A. Assessment of Mental Health Disorders

In order to detect mental health disorders for a particular patient, a variety of predictive models utilizing heterogeneous medical data have been studied [4]–[9]. Questionnaire-based assessments are commonly used to detect mental health disorders. In some of the studies [4]–[6], researchers designed a specific questionnaire, a structured interview, or evaluation targeting a certain mental health disorder to collect the patients' behavioral information. In other cases, researchers used standard psychological measures, such as PHQ-9 [10], also called psychological screening, to assess patients' risk of suffering from mental health disorders. However, psychological questionnaires are not generally applicable in primary care and the evaluation data is not widely accessible. This minimizes possibility of early detection of mental health disorders. In contrast, Electric Health Record (EHR) data has a higher accessibility to clinicians and researchers and holds comprehensive information of patients medical history especially within the primary care setting. Thus, this data also provides a promising opportunity for the early detection of mental health disorders due to its accessibility and standardized use and features.

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

B. Early Detection of Diseases based on EHR Data

Given (1) a disease as the prediction target (i.e., anxiety/depression in our study), (2) the EHR data of a large population with/out the target disease, and (3) the EHR data of the patient for whom we aim to predict, several models [11]–[14] have been used to predict whether the given patient would develop that disease in the near future. Most of these methods consist of following three unified steps:

- *Feature extraction*: Given the raw visit data from each patient's EHR, this step extracts some attributes and information that are relevant to the target disease and represents the extracted information as a data vector with a uniform structure (e.g., a numeric vector consisting of the results of a set of relevant diagnoses in past visits). Please note that each patient's data vector should be in the same structure and contain the necessary information for disease prediction;
- *Supervised learning*: Given each patient's data vector as well as his/her class label which identifies if the patient is diagnosed with the target disease, this step builds a predictive model using supervised learning algorithms (e.g., k nearest neighbor (kNN), support vector machine (SVM), or random forest (RF)), to compute the risk of developing the target disease as a function of past diagnoses represented by the data vector; and
- *New patient prediction*: Given the predictive model and the medical history of a new patient, we predict whether the patient will develop the target disease according to the data vector produced from one's medical history and the patterns learned in the predictive model.

Among the three steps discussed above, our research focuses primarily on improving the accuracy of existing predictive models through enhanced feature extraction, while pushing the frontiers of the state-of-the-art of supervised learning and disease prediction methods.

C. Feature Extraction from EHR Data for Mental Health Disorder Prediction

Given each patient's EHR data, which consists of the patient's demographic information and a sequence of past visits, existing methods first retrieve the diagnosis codes recorded during each visit [11], [12]. Then, the frequency of each diagnosis appearing in all past visits are counted, followed by further transformation on the frequency of each diagnosis into a vector of frequencies (e.g., $\langle 1, 0, ..., 3\rangle$, where 0 means the 2^{nd} diagnoses does not exist in all past visits). Patients might have differing numbers of visits, and each visit might consist of multiple diagnoses—aforementioned methods can be used to transform differing patients' visit sequences to a vector in a unified space $\mathbb{R}^D$, where D is the total number of diagnoses and vectors, can be handled by common machine learning algorithms. However, psychological studies and clinical research demonstrate that the mental health status of a patient can be observed through the change of some relevant diagnoses over time [15], [16]. Thus, we hypothesize that temporal orders of diagnoses in past visits are predictive of anxiety/depression. In order to extract temporal features for the early detection of anxiety/depression, we propose to create a vector space characterization of the temporal orders of diagnoses in patients' medical histories.

D. Proposed Research

In this work, we study the early detection of anxiety/depression among college students using EHR data to discriminate a patient's risk for anxiety/depression by incorporating the temporal orders of past diagnoses. We present a novel framework for the early detection of anxiety/depression by leveraging patient demographics and the temporal order and frequencies of the past diagnoses in patients' EHR data. We summarize the contributions of this work as follows:

- In this work, our goal is to improve the early detection for anxiety/depression by incorporating the temporal orders of past diagnoses in EHR data from non-mental health visits. To the best of our knowledge, this paper is the first study to address the problem of early detection on mental health disorders using *pairwise temporal order*, *diagnosis frequency* and *diagnoses transition frequency*.
- In order to improve the accuracy of existing predictive models, we propose *M-SEQ*, a predictive framework. To extract an appropriate set of features preserving the temporal order information, *M-SEQ* uses a three-step algorithm: (1) we select an optimal subset of diagnoses that are most relevant to the anxiety/depression according to the *maximal information gain criterion*, (2) transform the *pairwise order* between each two diagnoses into a vector space (i.e., $\mathbb{R}^{D\times D}$) so as to represent the temporal orders of diagnoses for each patient, and (3) compress $\mathbb{R}^{D\times D}$ into a small vector space $\mathbb{R}^k, k << D$), where k is the dimension of the most relevant transitions between pairwise diagnoses containing the temporal orders through *supervised dimension reduction*. After combining the *pairwise transition vector* and the *frequency vector* extracted from each patient's EHR data (e.g., $v_i \in \mathbb{R}^D$ and $v_i' \in \mathbb{R}^k \Rightarrow v_i'' \in \mathbb{R}^{D+k}$ for patient i), *M-SEQ* further leverages a set of alternative supervised learning algorithms (i.e., SVM, RF, and linear discriminant analysis (LDA)) to predict future diagnoses of anxiety/depression.
- We evaluate *M-SEQ* using the electronic health data from 10 universities participating in CHSN containing 213,112 patients from January 1, 2011 through December 31, 2014. The data contains both mental health and non-mental health diagnoses recorded in the primary care visits of college students to the student health centers. We present further comparison of the performance of *M-SEQ* with several baseline approaches that do not consider the temporal orders of diagnoses. The evaluation results demonstrate that the proposed method improves the accuracy of the predictive models for anxiety/depression compared to baselines.

The paper is structured as follows: Section II discusses the previous studies that have been done in the mental health

disorders and predictive modeling in healthcare. Section III describes the proposed methodology. Section IV describes the data used in this research, the experimental design, and the experimental results and analyses. Finally, the summary of this work, future work, and clinical context are discussed in Section V.

II. Related Work

Mental health disorders have been extensively studied by researchers in terms of the causes, prevention, and treatment, where the prevailing approach is cohort studies on behavioral and social information. In recent years, a few researchers studied the prediction of mental health disorders using machine learning approaches [17], [18]. In this section, we will discuss these works from two perspectives:

A. Predictive Models for Early Detection of Diseases

Predictive models have been applied to help with decision-making in many medical domains. These include the prediction of breast cancer, type II diabetes, cardiovascular disease, and mortality for critically-ill hospitalized adults to name a few [19]–[22]. According to the predictive models, a high-risk patient will be referred to intensive interventions and attention (e.g., screening and counseling) to prevent potential disease. These models have the potential to reduce the mortality rates and improve the quality of life of high-risk patients and control cost and complications for low-risk patients [23]. Notably, predictive models have become vital tools to assist with medical decision-making and can bring benefits to both healthcare providers and patients. Accurate prediction of mental health disorders can assist clinicians in identifying high-risk patients in an early stage ultimately leading to more timely diagnosis and treatment of mental health disorders.

Disease prediction can be treated as a classification problem in which many well-established classifiers are actively used in this arena. Researchers have attempted to predict depression severity to help personalize treatment for those patients. In [18], the features used for supervised learning include gender, ICD-9 codes, disease and drug ingredient terms, and average number of visits. A LASSO logistic regression model is trained on the feature vectors to predict potential depression. The prediction model is better at recognizing low-risk patients with a 90% specificity. Patients at high risk of depression are identified with a 25% sensitivity 12 months before the diagnosis and with a 50% sensitivity at the time of diagnosis [18].

B. Data Representation of Electronic Health Data

Electronic health data is heterogeneous and cannot be readily expressed in a unified vector space. Thus, an appropriate representation of those data is the cornerstone for further advancements in analytics and modeling. Poor representation of data lacking vital information can adversely affect predictive models. Usually, frequency and presence (or absence) are used as the representations for the categorical features of an instance, where presence or absence is coded as a binary variable [11], [18]. In [11], diagnosis of diabetes is predicted based on the past diagnoses information, medication and procedure orders, and lab tests from patients' medical records. For each patient, a feature vector is constructed based on the longitudinal health data. A feature value is generated by aggregating all the events of the same feature occurring in the pre-defined time window, where frequency is used for categorical features. While in a predictive model for the early detection of depression, the ICD-9 codes are converted to binary features where 1 indicates that the feature is present in the patient's medical history, and 0 otherwise [18].

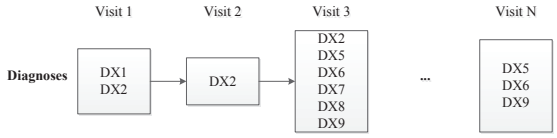


Fig. 1: An example of a patient profile in CHSN

However, the features extracted in the above works have omitted the temporal orders of events, which might contain important information for the early detection of a disease. Thus, we propose to generate a feature vector representation for each patient with both commonly used frequency information and temporal orders of clinical events in order to improve the early detection accuracy using the predictive models trained on those feature vectors.

III. *M-SEQ* Framework and Algorithms

In this paper, we propose *M-SEQ* for representing patient profiles in an EHR database to enable the early detection of anxiety/depression from primary care visit data. In this section, we define the problem and introduce the design of the *M-SEQ* framework.

A. Problem Formulation

Given (1) the training set for mental health disorder prediction in which each patient's record consists of a sequence of visits and each visit containing a set of diagnosis codes (shown in Figure 1), namely, s_i, for patient i, and (2) each patient's label (i.e., $+1/-1$) representing a mental health disorder (e.g., $l_i = +1$ indicating patient i has a diagnosis of depression or anxiety at any time in their medical history), our research problem is to find a method to transform each s_i to a fixed-length data vector v_i, where $v_i \in R^{D*}$ (D^* is not foreknown), so as to maximize the accuracy of any arbitrary classifier. That is:

$$\max \sum_{0 \le i < N} l_i * \mathit{classifier}(v_i), \tag{1}$$

where the function $\mathit{classifier} : R^{D*} \rightarrow \{+1/-1\}$ refers to an arbitrary binary classifier on top of vector space R^{D*}.

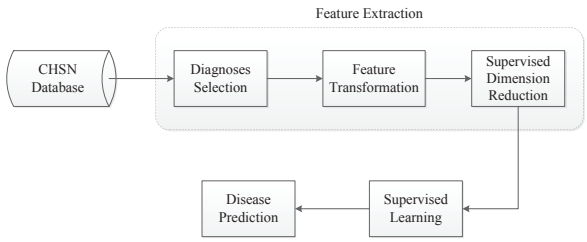


Fig. 2: The *M-SEQ* framework

B. M-SEQ *Framework*

In this section, we introduce the *M-SEQ* framework, illustrated in Figure 2. Our framework contains the following three components:

1) **Feature Extraction** — Given each patient's medical history and the label of each patient, this step transforms each patient's records into a data vector of fixed-length using supervised approaches.
2) **Supervised Learning** — Given the transformed data vector from Step 1 and the label of each patient, this step trains a predictive model based on the given data vectors and labels using supervised learning algorithms (i.e., SVM, LDA, and RF).
3) **New Patient Prediction** — Given a feature vector from Step 1, this step uses the predictive model (from Step 2) to predict the new patient's label. The outcome of this step is +1 or -1, which refers to whether the patient will develop (+1) a mental health disorder (-1 otherwise) in the future.

Our *feature extraction* method consists of following three steps.

C. Information Gain for Maximal Diagnosis Code Selection

Given the set of diagnosis codes in ICD-9 scheme used in EHRs, this step intends to select a small subset of codes that can predict the diagnosis of anxiety/depression of the patient, so as to reduce the complexity and noise. Patient health data from the CHSN database are processed into sequences of ICD-9 diagnosis codes. Thousands of ICD-9 codes are clustered into 283 categories according to the AHRQ Clinical Classification Software and expert opinions [24]. Using the more general categories reduces the noise issues resulting from using overly granular features resulting from heterogeneous coding practices at the schools. For example, the category of *abdominal pain* includes 789.00 (abdominal pain unspecified site), 789.01 (abdominal pain right upper quadrant), 789.60 (abdominal tenderness unspecified site), 789.61 (abdominal tenderness right upper quadrant), etc. Further, given these code groups, we select a subset of diagnoses, according to their ability to distinguish between patients with and without anxiety/depression. We first calculate the information gain [25] of each diagnosis and then utilize the top M diagnoses with the highest information gain as the features in our models.

Information gain is a common method for feature selection which measures the decrease in entropy when the feature is given versus when it is absent [26]. The expected information needed to classify a tuple $x \in \mathbb{D}$ is:

$$I(D) = -\sum_{i=1}^{C} p_i \log(p_i) \tag{2}$$

where C is the number of distinct classes. p_i is the probability that an arbitrary object in D is from Class i. We compute additional information needed to arrive at an exact classification after partitioning feature A with values $v_1, v_2, \ldots, v_n$ using:

$$I_A(D) = -\sum_{j=1}^{n} \frac{D_j}{D} \cdot I(D_j) \tag{3}$$

where $\frac{D_j}{D}$ is the fraction of the jth partition. Hence, the information gained from having feature A is:

$$\text{gain}(A) = I(D) - I_A(D) \tag{4}$$

D. Pairwise Order Transition Transformation

Each patient's visit sequence is transformed to a data vector of the diagnosis set with the highest information gain. To create a unified feature space for each patient, we first generate a vector of frequencies $A_1, \ldots, A_M$, where A_i refers to the count of the ith selected diagnosis in one's previous visits and M is the total number of the selected diagnoses. For example, there are 3 visits for patient x and 5 visits for patient y, $x = ((A_1, A_2), (A_1, A_2, A_3), (A_1, A_5))$ and $y = ((A_2), (A_2, A_3, A_4), (A_3, A_5), (A_2, A_5), (A_5))$, where $A_i \in (A_1, \ldots, A_5)$. The frequency vectors of patients x and y are:

TABLE I: Frequency vectors of patients x and y

	A_1	A_2	A_3	A_4	A_5
x	3	2	1	0	1
y	0	3	2	1	3

However, the frequency vectors do not contain any temporal information. Thus, we introduce the pairwise transitions into the vector space where the transition $A_{ij} (\in \{A_{11}, A_{12}, \ldots, A_{MM}\})$ refers to the number of co-occurrences of each two diagnoses A_i and A_j in two distinct visits (i.e., the p^{th} and q^{th} visit, and $|p - q| = 1, 2, 3, \ldots$). Table II presents the pairwise transition matrix for patient y.

TABLE II: Pairwise transition matrix of patient y

	A_1	A_2	A_3	A_4	A_5
A_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
A_2	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}
A_3	f_{31}	f_{32}	f_{33}	f_{34}	f_{35}
A_4	f_{41}	f_{42}	f_{43}	f_{44}	f_{45}
A_5	f_{51}	f_{52}	f_{53}	f_{54}	f_{55}

f_{ij}^{y} contains two components: the value of A_{ij} occurring in a patient's records and its frequency. If diagnosis A_i and A_j

occur in two nearby visits, it indicates a strong correlation between these two diagnoses, while A_i and A_j are less correlated if they occurred in two distant visits, respectively. Thus, a transition feature t_{ij} of patient y is defined as:

$$t_{ij}^y = \exp(-k \cdot |p - q|) \tag{5}$$

where t_{ij}^y is the transition from i to j for patient y. The parameter k is the decay factor characterizing the correlation between the two diagnoses in a transition (we assume the correlation between two diagnoses decreases if $|p-q|$ is large and k is the factor to control the decay speed. When k is 0, t_{ij}^y is 1 which is equivalent to merely considering the presence of diagnoses x_i and x_j in two distinct sequences. The larger k is set, the faster the transition decays. Hence, f_{ij}^y is defined as:

$$f_{ij}^y = \sum_{p=1}^{m-1} \sum_{q=p+1}^{m} t_{ij}^y \tag{6}$$

where m is the total number of distinct visits of patient y.

E. Supervised Transition Selection

Given the data matrix of transitions representing each patient's record, we convert the matrix into a smaller data vector by removing the noisy and redundant transition frequencies, considering the predictive power of each transition. First, the transition matrix is converted to a data vector of length M^2. The pairwise transition vector of patient y is $y' = (f_{11}^y, \ldots, f_{1M}^y, \ldots, f_{21}^y, \ldots, f_{M1}^y, \ldots, f_{MM}^y)$. Second, we further down-select features using the χ^2 test to remove redundant transitions. The χ^2 test is used to test the independence of an event and the occurrence of the class—if a feature is independent of the occurrence of the class, it cannot provide information to help with the classification [27]. Thus, the features statistically independent of the class outcome are removed from the transition set. The finalized vector to represent a patient consists of the *frequency vector* and the selected transitions. A patient is represented as $y' = (c_1^y x_1, \ldots, c_i^y x_i, \ldots, c_M^y x_M, f_{11}^y, \ldots, f_{ij}^y, \ldots, f_{MM}^y)$, where $(c_1^y x_1, \ldots, c_i^y x_i, \ldots, c_M^y x_M)$ is the *frequency vector* and $f_{11}^y, \ldots, f_{ij}^y, \ldots, f_{MM}^y$ are the selected transitions.

IV. Evaluation

In this section, we illustrate the experimental results for our work. We describe the data used in the experiments, followed by the details of the experimental design and results.

A. Data Description

This research used the de-identified electronic health records (EHR) data from 10 schools participating in the College Health Surveillance Network (CHSN) from January 1, 2011 through December 31, 2014. The CHSN database contains data from 31 student health centers across the US with over 1 million patients and 6 million visits [28]. However,the data from the 10 participating schools which upload both mental health and non-mental health visit data were used for this study. CHSN provides ICD-9 diagnostic codes and CPT procedural codes associated with a student health center encounter, as well as limited demographic information. The selected 10 schools include 263,947 enrolled students representing all geographic regions of the United States. The demography of enrolled students (sex, race/ethnicity, age, undergraduate/graduate status) closely matched the demography for the population of 108 Carnegie Research Universities/Very High classification.

The data utilized in this study consists of the primary care visits of 213,112 patients from which we extracted the ICD-9 codes in each of their visits. Target and control groups are created for the experiments, which contain 21,097 patients with anxiety/depression and 327,198 patients without any mental health disorder in their medical history. Patients with less than two visits from the control group were excluded from the analysis. Likewise, for target groups, there must be at least two visits in the three months before a patient's first diagnosis of anxiety/depression. Notably, the diagnosis information from within three months of the first diagnosis of anxiety/depression in the target group is excluded for the aim of early detection. Consequently, there are 7,322 and 205,790 patients in the finalized target and control groups, respectively.

B. Experimental Design

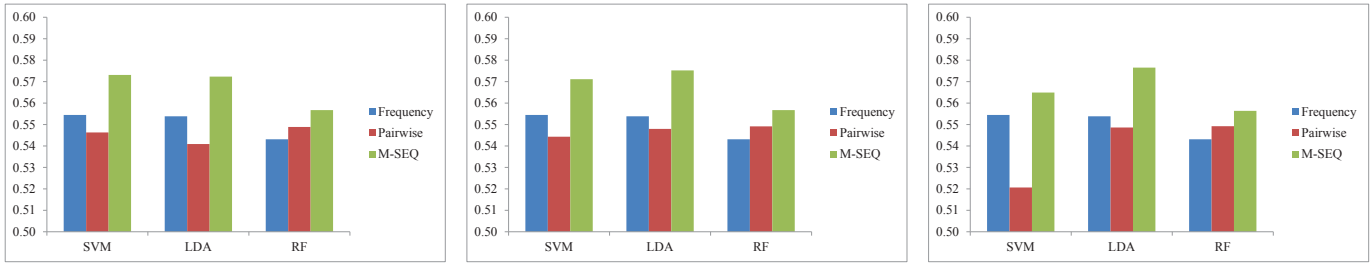
The size of control group is much larger than that of the target group, which causes a class imbalance problem. We adopted the EasyEnsemble approach to prevent the majority class from dominating the learning process and adversely affecting the performance of the classifiers. In the EasyEnsemble method, we randomly sampled from the majority class multiple times and trained the classifiers on each of the samples. The final output is averaged over all the learners [29].

For supervised learning, we build three different classifiers: linear SVM, LDA, and RF with 20 trees. In each experiment, we used 5-fold cross validation for training and testing the performance of the models. Furthermore, we consider multiple values of the decay factor k: 0.3, 0.5, 1.0, 1.5, and 2.0.

C. Experimental Results

For all experiments based on the three vector representations (i.e., frequency, pairwise and *M-SEQ*), we conducted 5-fold cross validation with different values of k using all three classifiers: SVM, LDA and RF. The accuracy of the three classifiers are presented in Table III. Specifically, Figure 3 shows the accuracy of SVM, LDA, and RF based on the three vector representations: Frequency, Pairwise, and *M-SEQ* when k is 0.3, 1.0, and 2.0.

Generally, the results shown in Figure 3 and Table III demonstrate that our method improves the accuracy for all values of k. Furthermore, the p values of all the paired t-tests on the accuracy of all the predictive models are significant at the 0.05 level indicating that the performance improvement by using *M-SEQ* is statistically significant. With each k tested in the experiments, the *M-SEQ* models achieve a higher accuracy than both the frequency and pairwise models. When $k = 0.3$, the SVM classifier based on *M-SEQ* improved the performance by 1.86% compared to using the *frequency vector* and achieves



(a) $k = 0.3$ (b) $k = 1.0$ (c) $k = 2.0$

Fig. 3: Prediction accuracy of classifiers based on *frequency*, *pairwise*, and *M-SEQ* representations

TABLE III: Prediction accuracy of classifiers based on *frequency*, *pairwise*, and *M-SEQ* representations

k	Classifier	Accuracy		
		Frequency	Pairwise	M-SEQ
0.3	SVM	55.45%	54.63%	**57.31%**
	LDA	55.38%	54.09%	**57.23%**
	RF	54.31%	54.89%	**55.62%**
0.5	SVM	55.45%	54.63%	**57.21%**
	LDA	55.38%	54.71%	**57.35%**
	RF	54.31%	54.88%	**55.65%**
1.0	SVM	55.45%	54.43%	**57.11%**
	LDA	55.38%	54.80%	**57.52%**
	RF	54.31%	54.92%	**55.67%**
1.5	SVM	55.45%	52.68%	**56.86%**
	LDA	55.38%	54.85%	**57.63%**
	RF	54.31%	54.91%	**55.64%**
2.0	SVM	55.45%	52.07%	**56.49%**
	LDA	55.38%	54.86%	**57.65%**
	RF	54.31%	54.92%	**55.64%**

a 2.68% higher accuracy than using the *pairwise transitions*. Similarly, the LDA model based on *M-SEQ* presents an improvement between 1.85% and 3.14% in accuracy than using the *frequency* and *pairwise* representations, respectively. The RF classifier based on the proposed method improves the accuracy by 1.36% and 0.78% compared to using the other two representations.

Furthermore, the performance of classifiers on the three representations are shown in Figure 4 in terms of accuracy, sensitivity, and specificity, when $k = 0.3$. Even though the classifiers using the *pairwise transition* vectors have a much higher sensitivity, they achieve extremely low specificity. In other words, the former models are biased towards the positive class, while the *M-SEQ* shows more balanced performance on both positive and negative classes. Besides accuracy, *M-SEQ* achieves an F1 measure of 62.53% using SVM, 63.12% using LDA, and 57.25% using RF when $k = 1.0$.

According to Table III, we found that there is very little change in the accuracy with different values of k in the proposed method. Thus, it demonstrates that the performance is fairly stable for different values of the decay factor. Overall, the experimental results indicate that the proposed methods can improve the accuracy of early detection of anxiety/depression for college students using their primary care visit data.

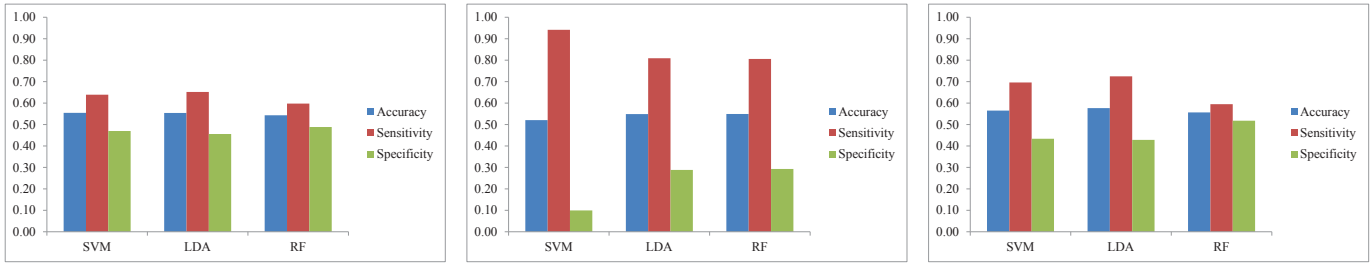
TABLE IV: The selected diagnoses

Index	Diagnoses
1	Malaise and fatigue
2	Contraceptive and procreative management
3	Poisoning by other medications and drugs
4	Headache; including migraine
5	Nausea and vomiting
6	Abdominal pain
7	Other upper respiratory infections
8	Menstrual disorders
9	Spondylosis; intervertebral disc disorders; other back problems
10	Other gastrointestinal disorders
11	Conditions associated with dizziness or vertigo
12	Other female genital disorders
13	Genitourinary symptoms and ill-defined conditions
14	Cardiac dysrhythmias
15	Noninfectious gastroenteritis
16	Administrative/social admission
17	Other skin disorders
18	Esophageal disorders
19	Other nutritional; endocrine; and metabolic disorders
20	Immunizations and screening for infectious disease
21	Deficiency and other anemia
22	Urinary tract infections
23	Mycoses
24	Other nervous system disorders
25	Asthma
26	Other connective tissue disease
27	Other endocrine disorders
28	Other lower respiratory disease
29	Inflammatory diseases of female pelvic organs
30	Nonspecific chest pain

D. Case Study 1: selected diagnoses with maximal information gain

We further investigate the diagnoses selected from all 283 groups, listed in Table IV.

By observing the selected features, we speculate that the above diagnoses are selected as predictive features of anxiety/depression due to different reasons. Some of the diagnoses in Table IV are depression-relevant somatic symptoms, which include abdominal pain, headaches, nausea, chest pain, back pain, dizziness, malaise and fatigue, and menstrual disorder.



(a) Frequency (b) Pairwise (k=2.0) (c) *M-SEQ* (k=2.0)

Fig. 4: Accuracy, sensitivity, and specificity of classifiers based on *frequency*, *pairwise*, and *M-SEQ* representations (k=2.0)

According to [30]–[38], somatic symptoms of general aches, pains, and fatigue, present frequently among patients with anxiety or depression, especially in primary care. Accordingly, these physical diagnoses are recognized as being highly correlated with anxiety/depression.

Diagnoses of contraceptive and procreative management, menstrual disorders, and female genital disorder occur purely or mostly among females rather than males. Females are also much more likely to suffer from mental health disorders—the incidence rates of anxiety and depression in females are 60% more than in males [39]. The CHSN data also demonstrates that anxiety and depression are more prevalent in the female population than in males. Overall, the selected diagnoses provide helpful information for clinicians to use in the early detection of anxiety/depression.

E. Case Study 2: selected pairwise transition features

We further investigate the 9 pairwise transitions selected from the 900 original transition pairs:

1) Contraceptive and procreative management
   → Contraceptive and procreative management
2) Immunizations and screening for infectious disease
   → Other upper respiratory infections
3) Immunizations and screening for infectious disease
   → Immunizations and screening for infectious disease
4) Immunizations and screening for infectious disease
   → Administrative/social admission
5) Immunizations and screening for infectious disease
   → Poisoning by other medications and drugs
6) Other upper respiratory infections
   → Poisoning by other medications and drugs
7) Poisoning by other medications and drugs
   → Other upper respiratory infections
8) Poisoning by other medications and drugs
   → Immunizations and screening for infectious disease
9) Administrative/social admission
   → Immunizations and screening for infectious disease

According to our analysis on the 30 diagnoses with maximal information gain, we speculate that the 9 selected pairwise transitions are constructed by two categories of diagnoses:

- Unrelated diagnoses to anxiety/depression: immunizations and screening for infectious disease, contraceptive and procreative management, etc.; and
- Potentially related diagnoses of anxiety/depression: other upper respiratory infections, administrative/social admission, etc.

Thus, there are two important types of transitions among the selected pairwise features:

- Transition from unrelated diagnoses to potentially related diagnoses of anxiety/depression, such as from immunizations and screening for infectious disease to other upper respiratory infections and from immunizations and screening for infectious disease to administrative/social admission.
- Transition from unrelated diagnoses to unrelated diagnoses of anxiety/depression, such as from immunizations and screening for infectious disease to immunizations and screening for infectious disease.

We speculate that the transitions from unrelated to related diagnoses imply high risk, while transitions from unrelated to unrelated diagnoses imply low risk of anxiety/depression. Administrative/social admission includes health service encounters in psychosocial circumstances, convalescence and palliative care, persons seeking consultation, etc. Prior epidemiological studies suggest that upper respiratory infections affect mood and cognition, and psychological stress which is a significant risk factor for upper respiratory infections [40], [41]. Additionally, we found that the transitions from and to medication and drug poisoning play an important role in the early detection of anxiety/depression. Further clinical investigation is needed to fully understand these transitions. In general, these findings on important pairwise transitions are informative for the early detection of anxiety/depression.

V. Discussion and Conclusion

A. Discussion

In this research, we studied the problem of early detection of anxiety/depression using patients' medical history. Data in medical records are complicated due to their heterogeneity and different operational contexts. It is common to use frequencies of clinical events to represent an aspect of a patient of interest. However, using merely frequencies omits the temporal orders of events, which might include vital information. There are

various approaches to processing sequential data, such as using a sliding window approach. In this study, we created pairwise transitions between diagnoses as additional features of the frequency vector. In this way, we include the information contained in temporal diagnoses orders in addition to commonly used frequencies. The experimental results demonstrate an improvement in the accuracy comparing to using frequencies alone.

The data used to learn the patterns for prediction on anxiety/depression suffer from the class imbalance problem, where there are many more patients in the control group. We adopted the EasyEnsemble approach to prevent the majority class from dominating the learning process and affecting the performance of the classifiers [29]. Apart from this work, there are many other methods to address the class imbalance problem. For example, SMOTE (Synthetic Minority Over-sampling TEchnique) is an over-sampling approach that introduces synthetic examples by joining the nearest neighbors of the minority class near the boundary [42]. We will examine other sampling and heuristic methods to balance the training data in future work.

In this research, we used more general diagnosis categories of ICD-9 codes from the AHRQ classification scheme. To improve clinical meaning and to address the data quality issues are the major motivations of collapsing individual ICD-9 codes into general categories. Data quality varies among different student health centers which affects the accuracy of the ICD-9 coding. Thus, using more general groups rather than specific codes alleviates the inaccuracy in ICD-9 codes to some extend and provides more clinical meaning. In future work, we will explore other approaches to incorporate ICD-9 codes into the predictive models with less information loss.

Currently, we only consider the transitions between each pair of diagnoses, while other common features such as demographic information and notes are not included. This is due to the limited data types contained in the CHSN database and miscoding of some demographics, such as ethnicity. We will consider the associate procedures in a visit as well as patient demographics available in the database such as geographic region, age, and student standing. Additionally, future work will incorporate transitions between more than two diagnoses as well as co-occurring diagnoses. Preliminary feature selection is performed to remove unrelated and redundant diagnoses codes before predictive modeling; however, metric learning approaches could be used for automatic feature extraction and classification together which may increase model performance. Some of these methods have been applied to solve classification problems on health data of which we will explore in future work [11].

Mental health disorders are often unrecognized in primary care settings such as the student health centers. This oversight leads to adverse outcomes and higher costs when patients with anxiety/depression cannot receive proper treatment on time. Thus, this work on the early detection of anxiety/depression could potentially aid student health centers in identifying high-risk patients in advance and referring them to behavioral health services. In addition to assisting clinicians with detecting patients at high risk, the outcome of this research could also be implemented as a practical tool for college students to understand their own risk of developing anxiety/depression according to their medical history.

B. Conclusion

In this study, we proposed a framework, *M-SEQ*, for the early detection of anxiety/depressive disorders based on the primary care visit data of patients in 10 student health centers. We expanded existing work in this domain by considering the temporal diagnoses patterns in models. The developed models could be used to aid clinicians in identifying patients at a high risk of anxiety/depression in non-mental health settings. Practical tools supported by algorithms in this research could also be developed to help students understand their risk of anxiety or depression according to their medical history.

ACKNOWLEDGMENT

The authors wish to thank the leadership of the participating schools for their commitment of time and resources to the College Health Surveillance Project and to their vision for furthering a better understanding of the epidemiology and health care needs of America's college students.

REFERENCES

[1] "Any mental illness (AMI) among adults. NIH national institute of mental health," 2015. [Online]. Available: http://www.nimh.nih.gov/

[2] "Prevalence, severity, and unmet need for treatment of mental disorders in the world health organization world mental health surveys," *JAMA*, vol. 291, no. 21, p. 2581, Jun. 2004.

[3] American College Health Association, "American College Health Association National College Health Assessment," *Spring 2014 Reference Group Executive Summary*, 2014. [Online]. Available: http://www.ijme.net/archive/2/communication-training-and-perceived-patient-similarity/

[4] R. Ettema, D. Grobbee, and M. Schuurmans, "In-hospital risk prediction for post-stroke depression: Development and validation of the depres," *Early Detection of Post-Stroke Depression*, p. 117, 2012.

[5] K. S. Kendler, R. C. Kessler, M. C. Neale, A. C. Heath, and L. J. Eaves, "The prediction of major depression in women: toward an integrated etiologic model," *American Journal of Psychiatry*, vol. 150, pp. 1139–1139, 1993.

[6] I. Yaroslavsky, J. Rottenberg, and M. Kovacs, "The utility of combining rsa indices in depression prediction." *Journal of abnormal psychology*, vol. 122, no. 2, p. 314, 2013.

[7] J. Parks, D. Svendsen, P. Singer, M. E. Foti, and B. Mauer, "Morbidity and mortality in people with serious mental illness," 2006.

[8] S. W. Smith and R. Koppel, "Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ," *Journal of the American Medical Informatics Association*, vol. 21, no. 1, pp. 117–131, 2014.

[9] D. M. Ndetei and R. Jenkins, "The implementation of mental health information systems in developing countries: Challenges and opportunities," *Epidemiologia e Psichiatria Sociale*, vol. 18, no. 01, pp. 12–16, 2009.

[10] K. Kroenke and R. L. Spitzer, "The PHQ-9: a new depression diagnostic and severity measure," *Psychiatr Ann*, vol. 32, no. 9, pp. 1–7, 2002.

[11] J. H. F. W. Kenney Ng, Jimeng Sun, "Personalized predictive modeling and risk factor identification using patient similarity," *AMIA Summit on Clinical Research Informatics (CRI)*, 2015.

[12] 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.

[13] J. Pittman, E. Huang, H. Dressman, C.-F. Horng, S. H. Cheng, M.-H. Tsou, C.-M. Chen, A. Bild, E. S. Iversen, A. T. Huang *et al.*, "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 22, pp. 8431–8436, 2004.

[14] 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.

[15] A. P. Association, A. P. Association *et al.*, "Diagnostic and statistical manual of mental disorders (dsm)," *Washington, DC: American psychiatric association*, pp. 143–7, 1994.

[16] P. I. Chow and B. W. Roberts, "Examining the relationship between changes in personality and changes in depression," *Journal of Research in Personality*, vol. 51, pp. 38–46, 2014.

[17] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. d. Mendona, "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Research Notes*, vol. 4, no. 1, p. 299, Aug. 2011.

[18] 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," *Journal of the American Medical Informatics Association: JAMIA*, vol. 21, no. 6, pp. 1069–1075, Dec. 2014.

[19] J. Lindstrom and J. Tuomilehto, "The diabetes risk score: A practical tool to predict type 2 diabetes risk," *Diabetes Care*, vol. 26, no. 3, pp. 725–731, 2003.

[20] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. A. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review," *BMJ*, vol. 344, 2012.

[21] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki, "Predictive modeling of hospital readmissions using metaheuristics and data mining," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7110–7120, Nov. 2015.

[22] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, May 2011.

[23] S. M. Domchek, A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, and B. L. Weber, "Application of breast cancer risk prediction models in clinical practice," *Journal of Clinical Oncology*, vol. 21, no. 4, pp. 593–601, 2003.

[24] "Clinical classifications software (CCS) for ICD-9-CM," 2015. [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

[25] J. Han, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2006.

[26] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, pp. 1289–1305, 2003.

[27] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[28] J. C. Turner and A. Keller, "College Health Surveillance Network: Epidemiology and Health Care Utilization of College Students at U.S. 4-Year Universities," *Journal of American college health: J of ACH*, p. 0, Jun. 2015.

[29] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[30] L. J. Kirmayer, J. M. Robbins, M. Dworkind, and M. J. Yaffe, "Somatization and the recognition of depression and anxiety in primary care," *The American Journal of Psychiatry*, vol. 150, no. 5, pp. 734–741, May 1993.

[31] B. Lwe, R. L. Spitzer, J. B. W. Williams, M. Mussell, D. Schellberg, and K. Kroenke, "Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment," *General Hospital Psychiatry*, vol. 30, no. 3, pp. 191–199, May 2008.

[32] P. Henningsen, T. Zimmermann, and H. Sattel, "Medically Unexplained Physical Symptoms, Anxiety, and Depression: A Meta-Analytic Review," *Psychosomatic Medicine*, vol. 65, no. 4, pp. 528–533, Jul. 2003.

[33] A. L. Vaccarino, T. L. Sills, K. R. Evans, and A. H. Kalali, "Prevalence and association of somatic symptoms in patients with Major Depressive Disorder," *Journal of Affective Disorders*, vol. 110, no. 3, pp. 270–276, Oct. 2008.

[34] K. Demyttenaere, A. Bonnewyn, R. Bruffaerts, T. Brugha, R. De Graaf, and J. Alonso, "Comorbid painful physical symptoms and depression: Prevalence, work loss, and help seeking," *Journal of Affective Disorders*, vol. 92, no. 23, pp. 185–193, Jun. 2006.

[35] G. E. Simon, M. VonKorff, M. Piccinelli, C. Fullerton, and J. Ormel, "An International Study of the Relation between Somatic Symptoms and Depression," *New England Journal of Medicine*, vol. 341, no. 18, pp. 1329–1335, Oct. 1999.

[36] A. Tylee and P. Gandhi, "The Importance of Somatic Symptoms in Depression in Primary Care," *Primary Care Companion to The Journal of Clinical Psychiatry*, vol. 7, no. 4, pp. 167–176, 2005.

[37] D. S. Baldwin and G. I. Papakostas, "Symptoms of fatigue and sleepiness in major depressive disorder," *The Journal of Clinical Psychiatry*, vol. 67 Suppl 6, pp. 9–15, 2006.

[38] M. M. Ohayon and A. F. Schatzberg, "Using chronic pain to predict depressive morbidity in the general population," *Archives of General Psychiatry*, vol. 60, no. 1, pp. 39–47, Jan. 2003.

[39] W. Narrow, "One-year prevalence of depressive disorders among adults 18 and over in the us: Nimh eca prospective data. population estimates based on us census estimated residential population age 18 and over on july 1, 1998," *Unpublished table*, 1998.

[40] R. S. Bucks, Y. Gidron, P. Harris, J. Teeling, K. A. Wesnes, and V. H. Perry, "Selective effects of upper respiratory tract infection on cognition, mood and emotion processing: A prospective study," *Brain, Behavior, and Immunity*, vol. 22, no. 3, pp. 399–407, Mar. 2008.

[41] S. Cohen, "Psychological Stress and Susceptibility to Upper Respiratory Infections," *Am J Respir Crit Care Med*, vol. 152, no. 4_pt_2, pp. S53–S58, Oct. 1995.

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.