# *Daehr*: a Linear Discriminant Analysis Framework for Electronic Health Record Data

## with its Application to Early Detection of Mental Health Disorders

Haoyi Xiong, Jinghe Zhang, Yu Huang, Kevin Leach, and Laura E. Barnes

**Abstract**—Electronic Health Records (EHR) containing a massive number of patients' diagnosis records have been used to predict future or potential diseases according to their past diagnoses. While a number of data mining tools have been adopted for EHR-based early disease detection, Linear Discriminant Analysis (LDA) is one of the most commonly used statistical methods. However, it is difficult to train an accurate LDA model that detects specific diseases when there are too few known patients with the targeted diseases and the EHR data are coded manually with noise, because the covariance matrices used in LDA are usually singular and estimated with large variance. To address these issues, this paper presents *Daehr*, an extended LDA framework using Electronic Health Records. Beyond the existing LDA analyzers, *Daehr* is proposed to 1) eliminate the data noise caused by the manual encoding of EHR data, and 2) lower the variance of the LDA model even when only a few patients' EHR data are given for training. To achieve the two goals, we designed an iterative algorithm to improve the covariance matrix estimation with embedded data noise/variance reduction for LDA. We evaluated *Daehr* extensively using a large-scale real-world EHR dataset, the College Health Surveillance Network (CHSN). Specifically, our experiments compare the performance of LDA to three baselines (i.e., LDA and its derivatives) in terms of identifying high risk college students for mental health disorders from 23 US universities. Experimental results show that *Daehr* significantly outperformed three baselines by achieving 3%–10% higher prediction accuracy, and a 3% –14% higher F1-score.

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

## 1 Introduction

With the rapid development of medical big data, forecasting future or potential ~~disease~~ diseases based on patients' past medical records ~~becomes a promising way to detect and further prevent high risk disease in advance. Instead of paying attentions~~ has emerged as a promising approach towards preventing high-risk diseases. Rather than individualizing patients (e.g., via screening or counseling)~~to all its patient intensively~~, a medical informatics system can predict each patient's potential diseases using his ~~/~~or her past diagnoses as well as ~~the diagnoses records collected from massive~~ diagnoses collected from many other patients. In this way, the medical system can identify ~~high risk patients from the all~~ high-risk patients from a large corpus of patients with low cost~~, then serve patients in a targeted manner, further start prevention~~ . These high-risk patients can then receive targeted care to employ disease prevention techniques in advance. ~~Therefore~~Naturally, the accuracy of ~~disease early detection is a crucial factor to improve~~ such early disease detection is crucial to improving the efficiency of ~~high risk~~ high-risk patient identification and disease prevention.

In this paper~~we present *Daehr*~~, we present *Daehr*—an ~~extending~~ extended linear discriminant analysis (LDA) [1], [2] framework for ~~disease early~~ early disease detection using Electronic Health Records (EHR), which can improve the prediction accuracy of the standard LDA model ~~through reducing the~~ by reducing noise in EHR data and regularizing the estimated covariance matrices. ~~In the rest of this section, we~~ We first discuss the motivations and background of this research, then we formulate a new research problem based on our observations and assumptions. We elaborate the technical challenges of the proposed research~~and finally~~ . Finally, we summarize our technical contributions.

### 1.1 Motivations and Backgrounds

~~In order to~~ To predict patients' potential ~~disease~~ diseases according to their past medical records, a variety of predictive models utilizing heterogeneous medical data have been studied [3]–[5]~~, such as chest imaging for chest cancer early detection~~ . For example, chest imaging has been used for early detection of chest cancer [?] , questionnaire-based assessment (e.g., PHQ-9 [6]) data for ~~mental disorder prediction~~predicting mental disorders, and screening data for ~~heart disease prediction~~predicting heart disease [7]. Among ~~all~~ these medical data, Electronic Health Records (EHR) consisting of the diagnosis records ~~of patients' each visit~~ from patients' visits are used as a general purpose data source that enables ~~massive disease early~~ early disease detection based on the previous diagnoses at a massive scale. Furthermore, this data ~~has a higher accessibility~~ is more accessible to clinicians and researchers, and holds comprehensive information of patients~~medical history~~ ' medical history, especially within the primary care setting. Thus, EHR data ~~also~~ provides a promising opportunity for ~~the disease early~~ early disease detection due to its ~~general-purposeness, accessibility~~generality, accessibility, and standardized use and features.

*Authors are all with Department of Systems and Information Engineering, University of Virginia, VA. Email:{hx6d, xxx, xxx, xxx}@virginia.edu*

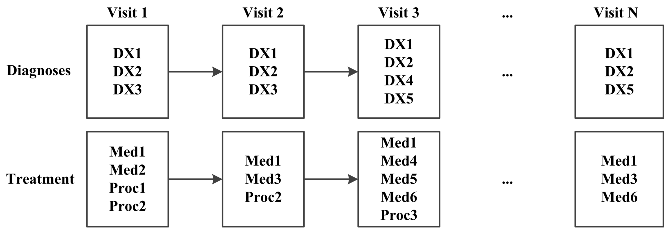

Fig. 1: An Example of a Patient's EHR Data

As shown in ~~Fig~~Figure 1, a patient's EHR data includes all his/her past diagnosis and treatment records, where the diagnosis record includes a sequence of visits~~and each visits~~, and each visit consists of multiple diagnoses. ~~Please note~~ Note that all diagnoses are recorded using ICD-9 codes [8], where each evidence of diagnosis corresponds to a specific ICD-9 code. With diagnosis records in the EHR data, several methods [9]–[12] have been studied to predict the disease of patients. Given a disease as the prediction target (e.g., anxiety/depression) as well as the EHR data of a large population with ~~/out~~ or without the target disease, most ~~of~~ existing methods first represent each given patient's EHR data using a set of features, and then train a predictive model using features and labels (if each patient is diagnosed with the targeted disease) in a supervised manner. Further, given each new patient's EHR data, these models predict if the given patient will develop the targeted disease in near future using the trained predictive model.

**EHR Data Representation for Early Detection.** In terms of representing EHR data, existing approaches include using diagnosis-frequencies [9], [13], [14], pairwise diagnosis transitions [15], [16], graph representations of diagnosis sequences [17], and so on. Among these approaches, the diagnosis-frequency is ~~considered as one of common ways~~ one common way to represent EHR data. 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. ~~Then~~Next, 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., $\langle 1, 0, \ldots, 3\rangle$), where 0 means the ~~$2^{nd}$ diagnoses~~ second diagnosis does not exist in all past visits). In this way, each patient having different number of visits and each visit consisting of multiple diagnoses is represented as a fixed-length data vector, which can be handled by common machine learning algorithms.

~~Please note~~ Note that the diagnosis-frequency representation of EHR data is usually with ultra-high dimensions; for example, there ~~exists~~ are more than 15,000 ICD-9 codes in the EHR scheme, thus the diagnosis-frequency vector using raw ICD-9 codes ~~is usually with~~ contains thousands of dimensions. ~~In order to reduce the dimensinality~~To reduce the dimensionality, clinical professionals may suggest ~~to use clustered code set~~using clustered code sets, where each ICD-9 code can map to one of the 295 clustered codes. ~~In this way~~Thus, each raw diagnosis-frequency vector can be compressed to a vector of around 200 dimensions using clustered codes.

**Supervised Learning for Early Detection.** Given an EHR database and a target disease for early detection, existing ~~method usually needs to~~ methods first select patients both with and without the disease, then use an appropriate representation of their EHR data ~~with appropriate representation~~ to form a training set. ~~In order to~~ To train an accurate predictive model with the training set, ~~a lot of~~ many machine learning methods such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Network, Gaussian Process and Linear Discriminant Analysis (LDA) have been adopted [9], [13]–[18]. Among these machine learning methods, LDA is frequently used as one of the common performance benchmarks in a series of studies [15], [18]–[21] ~~, considering its capacity of dimension reduction~~because it effectively reduces dimensionality. For example, when using diagnosis-frequency vector as the representation of EHR data, a LDA model learns a linear combination of diagnoses (from the all diagnoses) that can optimally separate patients into the two groups (i.e., with/without the disease). Then~~LDA predicts if~~, LDA predicts whether new patients will develop ~~to~~ the targeted disease ~~through~~ by separating their vectors into the two groups using the linear combination.

~~Please be advised that, just like~~ Like many other statistical learning models, the accuracy of a LDA model can be improved ~~,~~when more samples are given for training. ~~It~~ This is because the decision risk of a LDA model is inherited from the variance of its training samples, while *increasing the sample size ~~lower~~ lowers the sample variance* [22], [23]. In contrast, when ~~the training samples are few~~ there are few training samples, the model ~~even~~ cannot produce any valid prediction results ~~. Because~~ because LDA needs to use the *inverse of the covariance matrices* to ~~predict, while in such case~~ make predictions. In such cases, the covariance matrices estimated in LDA are singular ~~or namely *non-inversiable* (i.e., the inverse of the covariance matrix doesn't exist~~(non-invertible) [24], [25].

~~With above backgrounds in mind, we~~ We are motivated to enhance the supervised learning methods ~~on top of EHR data ,~~ building upon EHR data so as to improve the prediction accuracy for ~~disease early~~ early disease detection. Specifically, we ~~attempts at studying~~ study the LDA model using the diagnosis-frequency features ~~, considering~~ because of the relevance of such settings in clinical practices.

## 1.2 Research Assumptions and Objectives

Our research is based on following two observations and two assumptions about EHR data and early detection settings:

**Observation 1. EHR Encoding Variation** ~~-~~ — In terms of ~~EHR data encoding~~encoding EHR data, the diagnosis records are usually inputted manually by clinicians without a unified encoding scheme. Our previous work ~~[?]~~ [?] finds that, for ~~one patient, the~~ a single patient, there may be a higher number of diagnosis records for one disease ~~might be more frequent than the times that the disease has~~ than the number of times that that disease as been diagnosed. For example, ~~*three clinicians–Ann, Bob and Carl are working in the same clinics. Given a patient has been diagnosed with upper respiratory*~~

*~~infection (ICD-9 code: 465.9), Ann may only leave the record of code 465.9 in the first visit when the disease is diagnosed. However Bob may leave the record in the first visit as well as all his/her returning visits to receive screening/treatment for upper respiratory infection; while Carl may leave in the first visit and some of the returning visits that he feels necessary.~~* consider three clinicians: Ann, Bob, and Carl, all working in the same clinic. A single patient has been diagnosed with upper respiratory infection (ICD-9 code: 465.9). Ann may leave only the record of code 465.9 for the first visit in which the disease is diagnosed. However, Bob may leave the record in the first visit as well as all of the patient's returning visits to receive screening or treatment for upper respiratory infection. Carl may leave a record in the first visit and in some of the returning visits at his discretion.

**~~Assumption I. Non-negative Noise in Diagnosis-Frequency Vector Data -~~** ~~—~~ ~~With the first observation in mind, it is reasonable to~~ ***Assumption I. Non-negative Noise in Diagnosis-Frequency Vector Data*** — Based upon the first observation, we assume that each diagnosis is recorded at least one time in ~~EHR and~~ the EHR, and that the number of records might ~~be differing from encoding clinicians~~ differ due to clinician encoding styles (i.e., *frequency of record* $\geq$ *frequency of diagnosis* for each specific disease). ~~In this case, we~~ We further assume the encoding variation of EHR data may cause certain unknown *non-negative data noise* ~~on top of~~ in the diagnosis-frequency vectors.

**Observation 2. Limited Positive Training Samples** ~~-~~ — We find that the total number of patients with a specific disease ~~or namely~~ (*positive samples*~~might be so~~ ) might be too few to train a predictive model for early detection of the disease. For example, *~~a historically black college wants to identify the risk students in terms of mental health disorder using all students' EHR installed in the college clinics. The clinics first sorts all students to two groups (i.e., with/without mental disorder diagnosed), then from each group it selects a subset of students as training samples. However, due to the low utilization rate of psyclinics by African American, the available training samples with at least one type of mental disorders (e.g., depression, anxiety, mood and personality disorders) are so few (e.g., 100 – 500 students) in their school.~~* consider a historically black college that wants to identify the at-risk students in terms of mental health disorders using all students' EHR data in the college clinics. The clinics first separate all students in to two groups (i.e., with/without mental disorder diagnosed). Then, it selects a subset of students from each group as training samples. However, psychiatric clinics are typically underutilized by African Americans [?] , and thus the available training samples that include at least one type of mental disorder are too few (e.g., 100–500 students) in the school.

***Assumption II. Decision Risk of LDA Model for Early Detection of Diseases*** ~~-~~ — Considering the dimension $p$ of diagnosis-frequency vectors (e.g., $p \geq 200$ using clustered code set), we assume that the size of positive samples for LDA training is relatively small (i.e., $0 < n \lll 2^p$), where $n$ refers to the number of positive training samples. ~~Please note that when~~ When $0 < n < p$, the trained LDA model ~~can not~~ cannot produce any valid ~~prediction results~~predictions, since the estimated covariance matrix is singular/~~non-inversiable~~non-invertible; when $p \leq n \lll 2^p$, the trained LDA model might be able to produce ~~the predictionresult~~ a valid prediction, but with large decision risk inherited from the variance of ~~small~~ a small number of training samples.

With above two assumptions in mind, our work attempts ~~at reducing the affect~~ to reduce the effect of noise while lowering the decision risk of the LDA model for early detection of diseases. Specifically~~we use *mental health disorders* as the "~~, we use mental health disorders as the "target disease" in evaluation and experiment design, with respect to *Assumption II*.

## 1.3 Technical Issues and Contributions

In order to improve LDA with respect to the two assumptions, we ~~need to address~~ address the following three technical issues:

1) ***Eliminating the data noise in diagnosis-frequency vectors caused by encoding variation*** ~~-~~ — Given the frequency-diagnosis vectors for training, LDA first estimates sample diagnosis-to-diagnosis covariance matrices using an unbiased estimator ~~like~~ such as *Intrinsic Estimator* or *Maximized Likelihood Estimator (MLE)*, then builds the predictive models using estimated covariance matrices. However, our later analysis shows that the non-negative data noise in the vectors might make the estimated covariance matrices more dense than the noise-free (ideal) one. In this way, ~~there~~ we might need a method to *sparsify* the covariance matrices in order to reduce the ~~affect~~ effect of data noise ~~to~~ on LDA.
2) ***Lowering the decision risk of LDA while guaranteeing non-singularity and positive definiteness of the estimated covariance matrices*** ~~-~~ — ~~In order to~~ To lower the decision risk ~~of~~ associated with LDA, one possible solution is to use the $\ell^1$-penalized estimation of the covariance matrices [26], [27]. However, any modifications (including $\ell^1$-penalty and sparse approximation) to a ~~coavriance matrix might loss~~ covariance matrix might result in loss of its positive definiteness—we cannot use such a modified matrix in the statistics model. ~~In this way, there needs~~ We need an algorithm to obtain the $\ell^1$-penalized estimation of the ~~sparsifed~~ sparsified covariance matrix while ensuring the estimation is non-singular and positive ~~semi-definite~~semidefinite.
3) ***Incorporating the newly-estimated covariance matrices for EHR-based LDA*** ~~-~~ — Given the non-singular/positive-definite $\ell^1$-penalized sparse estimations of the covariance matrices, we might ~~need to~~ use them to replace the covariance matrices originally used in LDA. Thus~~there needs~~, we need a generic framework to extend the original LDA through incorporating the aforementioned covariance matrix estimation algorithms.

With the aforementioned research challenges in mind, we ~~made~~ make following technical contributions in this study:

- In this work, we studied the problem ~~to improve~~ of improving the existing Linear Discriminant Analysis (LDA) for ~~disease early~~ early disease detection based on our two assumptions. To the best of our knowledge, this paper is the first work for LDA-based ~~disease early detection on top of EHR data ,~~ early disease detection built upon EHR data by addressing the issues of encoding variation and low training sample size~~issues~~.
- In order to ~~tackle the technical challengesaforementioned, we proposed *Daehr*~~address these technical challenges, we propose *Daehr*—an ~~extending~~ extended LDA framework. It takes a novel approach to eliminate the ~~affect~~ effect of data noise and lower the decision risk of LDA ~~model~~ models through estimating sparse and non-singular diagnosis-to-diagnosis covariance matrices from diagnosis-frequency vectors. Theoretical analysis shows that, with low computational complexity, the proposed algorithm can approximate the $\ell^1$-penalized near-sparsest estimation of the diagnosis-to-diagnosis covariance matrices with non-singularity and positive semi-definiteness guaranteed, even when a very limited number of diagnosis-frequency vectors are given for LDA training.
- We evaluated ~~*Daehr*~~*Daehr* using a real-world dataset, CHSN, which contains more than 300,000 students' EHR records collected from 23 US universities ~~in~~ over the past three years. We designed a set of experiments based on CHSN for large-scale ~~mental health disorder early detection . The evaluation results show *Daehr*significantly outperformed~~ early detection of mental health disorders. The experimental results show *Daehr* significantly outperforms three baselines (i.e., LDA and its derivatives) by achieving 3%–10% higher prediction accuracy, and a 3%–14% higher F1-score.

The paper is structured as follows: Section 2 discusses the previous studies that have been done in the data mining approaches to early detection of disease and LDA extensions. Section 3 introduces the problem formulation of our study and introduces the ~~*Daehr*~~*Daehr* framework to solved the problem. Section 4 describes two core algorithms used in ~~*Daehr*~~*Daehr*. Section 5 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 6.

## 2 Related Work

In this section, we summarize ~~the~~ previous studies related to this paper from ~~following~~ two aspects: *data mining approaches to early detection of diseases* and *extensions to LDA ~~learner~~learning*.

### 2.1 Big Data Approaches to ~~Disease~~ Early Disease Detection

Various analytical methods have been used to study the causes, prevention, progression, and interventions of diseases~~, among which~~ . Among these methods, machine learning has ~~become very promising~~ emerged as a promising technique in the prediction of diseases [28], [29]. In this section, we will discuss ~~the related works in terms of the~~ previous work in two areas: *predictive modeling* and *data representation* approaches.

~~**Predictive Models for Early Detection of Disease**~~

#### 2.1.1 Predictive Models for Early Detection of Disease

Predictive models have become popular in the early detection of diseases, such as breast cancer, type II diabetes, ~~cardiovascular disease, etc.~~and cardiovascular disease [30]–[33]~~The outcome~~ . The outcomes of the predictive models are beneficial to both care providers and patients. Accurate prediction of diseases can assist clinicians in identifying high-risk patients in an early stage, ultimately leading to more timely ~~diagnosis and focusing the resources to deliver effective treatment~~ diagnoses and more focused delivery of effective treatments to those patients. ~~In essence, the~~ The early detection of diseases can be viewed as a classification problem so that well-established classifiers can be used to perform the task. Among the studies on the early detection of mental disorders, a LASSO logistic regression model has been applied to predict the depression severity to help personalize treatment for high-risk patients [29]. In this work, the feature vector used for prediction includes gender, ICD-9 codes, disease and drug ingredient terms, and average number of visits. However, the predictive model is more accurate in recognizing low risk-patients and achieves a 90% specificity, while the sensitivity ~~are~~ is 25% using the information 12 months before the diagnosis and 50% at the time of diagnosis~~, respectively~~ [29].

~~**EHR Data Representation for Predictive Models.**~~

#### 2.1.2 EHR Data Representation for Predictive Models

Electronic health data is highly accessible in health care institutions and has become a promising data source for public health research. However, EHR data is heterogeneous and cannot be readily expressed in a unified vector space. Thus, an appropriate representation of ~~those~~ this data is critical for further advancements in analytics and modeling. Many data representation approaches ~~has~~ have been developed to preserve useful information from the raw data. Usually, frequency is used as the ~~representations~~ representation for categorical features of an instance, ~~where~~ while presence or absence is used for binary variables [29], [34]. However, this representation omits the temporal ~~orders~~ ordering of clinical events~~and an attempt~~ . Attempts has been made to incorporate ~~the~~ temporal information by introducing pairwise transitions of diagnoses in addition to the widely used frequency features [15]. ~~In addition~~Furthermore, some novel frameworks ~~are proposed to~~ learn the temporal knowledge in patients' sequences [17], [35], [36]. In [36], ~~it~~ Wang et al. uses a spatial-temporal matrix to represent ~~the~~ a sequence of events in which the two dimensions ~~represents~~ represent the event type and time information~~, respectively~~. In [17], ~~the~~ Liu et al. considers events in a patient's EHR is represented by a temporal graph and basis graphs are learned as the features to represent patients. Furthermore, frequent sequence mining has been utilized to uncover the most important event sequences [37]–[39]. In [37], ~~it~~ Gotz et al. combines the episode definition and temporal pattern mining techniques to support the visual exploration of the ~~most impactful~~ clinical event patterns ~~to outcome~~with the most impact. To address high dimensional data, FeaFiner~~is proposed for~~ [40] uses simultaneous feature grouping and selection. ~~Thus, it~~ It extracts relevant and non-overlapping feature concepts in a low dimensional space, where the prediction accuracy is improved when applied to predicting Alzheimer's Disease-related scores [40].

### 2.2 Extensions to LDA Model

Regarding the application of LDA to EHR-based early ~~detection of diseases~~disease detection, here we mainly introduce several LDA extensions ~~under~~ in High Dimension Low Sample Size (HDLSS) settings. As ~~aforementioned~~discussed above, when LDA works in HDLSS, there might exist two major technical issues: 1) LDA ~~needs the inverse of~~ requires inverting covariance matrices for classification~~while~~ , but these covariance matrices estimated from small numbers of samples are usually singular (~~non-inversiable~~non-invertible), and 2) large decision risk is inherited from the variance of small samples ~~,~~ through classical LDA training. In order to handle the singular (~~non-inversiable~~non-invertible) covariance matrix issues, ~~[41] proposed to use~~ Ye et al. [41] uses the Pseudo-inverse of the singular covariance matrix, while Direct LDA [25], [42] ~~proposed to use~~ uses the *simultaneous diagonalization* of covariance matrices, which are non-singular, to replace the original covariance matrices. On the other hand, several works [23], [43], [44] have been proposed to ~~order to~~ lower the decision risk through regularizing the estimated covariance matrices. ~~.~~

~~*Daehr*is different from above related work in the following aspects~~*Daehr* is distinct in three ways. First, compared to other data mining approaches to early detection of disease ~~such as~~(e.g., [30]–[33]~~, *Daehr*~~), *Daehr* is the first work that intends to improve the performance of LDA model by addressing data noise and small positive training sample size issues. Second, our contribution is complementary with ~~those work~~ these works in EHR data representation [17], [35], [36]~~and we are open to further improve *Daehr*~~, and we can further improve *Daehr* by incorporating advanced EHR data representation methods. Third, when compared to ~~the~~ existing LDA extensions, ~~*Daehr*~~*Daehr* re-estimates the covariance matrices to (1) eliminate the ~~affect~~ effect of data noise to LDA model, (2) lower the decision risk inherited from small positive training samples, and (3) guarantee ~~the~~ non-singularity of covariance matrices, while [23], [25], [41]–[44] ~~focuses~~ all focus on regularizing the covariance matrices to enable LDA in a general HDLSS setting. Thus, the estimation/optimization problems considered in ~~any single one~~ each of the previous studies are mathematically different from ours with different objectives and assumptions.

## 3 *Daehr* System Model

In this section, we first formulate the research problem of our study~~; then we propose *Daehr*~~, then we describe the *Daehr* framework to solve the formulated problem.

### 3.1 Problem Formulations

According to our research assumptions, ~~in this section,~~ we make two definitions and introduce several preliminary studies ~~used in our studies~~that we use. Further, we formulate our research problem based on ~~all above~~ these definitions and preliminaries.

***Definition I.*** *Diagnosis-frequency Vector and Non-negative Noise Vector* ~~-~~— Given EHR data of $m$ patients (both with and without the targeted disease), we can extract $m$ diagnosis-frequency vectors $X_0, X_1 \ldots X_{m-1}$. Each vector (e.g., $X_i =< 1, 0, \ldots, 3 >$) consists of two parts: (1) $\hat{X}_1$, the vector of true diagnosis frequencies (not diagnosis record frequencies)~~and~~ , and (2) $E_i$, the non-negative noise vector:

$$X_i = \hat{X}_1 + E_i \tag{1}$$

***Preliminary I.*** *Generalized Two-class LDA and Covariance Matrices* — ~~According to the common implementation of a~~ In typical implementations of an LDA classifier [45], given $m$ training samples as well as the labels ~~i.e.,~~ $(X_0, l_0) \ldots (X_{m-1}, l_{m-1})$, where $l_i \in \{-1, +1\}$ ~~refers to~~ indicates whether the patient $i$ has been diagnosed with the target disease (i.e., positive sample or negative sample), a two-class LDA model first sorts each sample into two groups according to the label, and estimates covariance matrix/mean vector of the two classes, ~~i.e.,~~ $(\Sigma_+, \mu_+)$ and $(\Sigma_-, \mu_-)$ ~~,~~ using the positive samples and negative samples, respectively. Then, generalized two-class LDA ~~determine if~~ determines whether a new patient $(X')$ would develop to the targeted disease, using

$$\begin{aligned}&(X' - \mu_-)^T \Sigma_-^{-1} (X' - \mu_-) + ln|\Sigma_-| - \\ &(X' - \mu_+)^T \Sigma_+^{-1} (X' - \mu_+) - ln|\Sigma_+| < T,\end{aligned} \tag{2}$$

where $T$ is an optimal threshold based on the training samples. However, as illustrated in ~~our observation~~ Observation 2, when positive sample size is relatively small (e.g., for ~~the~~ a rare disease in the database), $Rank(\Sigma_+) < p$, $\Sigma_+$ is singular and $\Sigma_+^{-1}$ ~~doesn't~~ does not exist. In this case, Equation 2 might not work.

~~Please note that in the rest of paper we name the~~ Note that hereafter, we refer to both $\Sigma_+$ and $\Sigma_-$ as ~~a~~ *covariance* ~~*matrix*~~*matrices* ~~simply, since they are considered equally~~ because they are both considered equal in our problem formulation and solution design~~; in contrast, the covariance matrix may refer to the either $\Sigma_+$ or $\Sigma_-$~~.

***Definition II.*** *Sample Diagnosis-to-Diagnosis Covariance Matrix Estimation and Disturbance of Non-negative Noise* — With the above settings in mind, we further define $\Sigma$ as the sample diagnosis-to-diagnosis covariance matrix based on noisy data, $\hat{\Sigma}$, as the sample covariance matrix based on "noisy-free" vectors, and $\Delta = \Sigma - \hat{\Sigma}$ as the disturbance of non-negative noise to covariance estimation.

$$\begin{aligned}\Sigma &= \frac{1}{n}\sum_{i=0}^{n-1} X_i X_i^T = \frac{1}{n}\sum_{i=0}^{n-1} (\hat{X}_i + E_i)(\hat{X}_i + E_i)^T \\ &= \hat{\Sigma} + \Delta\end{aligned} \tag{3}$$

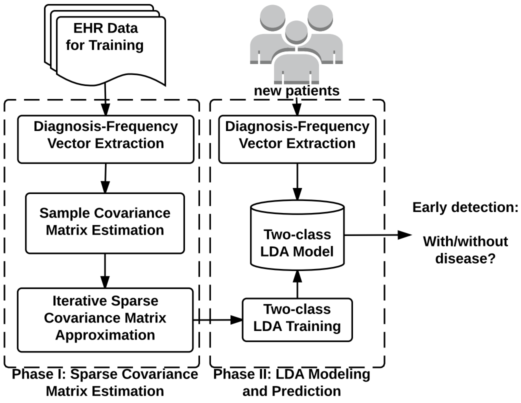
EHR Data for Training
new patients
Diagnosis-Frequency Vector Extraction
Sample Covariance Matrix Estimation
Iterative Sparse Covariance Matrix Approximation
Diagnosis-Frequency Vector Extraction
Two-class LDA Model
Two-class LDA Training
Early detection:
With/without disease?
Phase I: Sparse Covariance Matrix Estimation
Phase II: LDA Modeling and Prediction

Fig. 2: *Daehr* Framework

As the sample covariance matrix estimation shown in 3, the disturbance should be:

$$\Delta = \frac{1}{n}\sum_{i=0}^{n-1}(2\hat{X}_iE_i^T + E_iE_i^T).$$

According to our definition, $\hat{X}_i$ and $E_i$ are both non-negative matrices~~, it is not hard to~~. From this, we find that $\Delta = \Sigma - \hat{\Sigma} \geq \mathbf{0}$ is a non-negative matrix and $||\Sigma|| \geq ||\hat{\Sigma}||$. Thus, we can ~~roughly~~ conclude that $\hat{\Sigma}$ might be a sparse estimation of $\Sigma$.

***Preliminary II.*** *Minimax decision risk estimation of the covariance matrix in HDLSS settings* — Previous work [26], [27] showed that it is possible to achieve *minimax risk* covariance matrix estimation from a few samples, using the *minimal $\ell^1$-normal estimation* of the original sample covariance matrix. In this case, in terms of lowering variance of LDA, we can assume that the optimal [26] covariance matrix $\tilde{\Sigma}$ should be a $\ell^1$-penalized sparse estimation of $\hat{\Sigma}$.

**Problem Formulation.** According to ~~above~~ the definitions and preliminaries above, this paper considers ~~a~~ the problem of finding the positive-definite sparse estimation of $\hat{\Sigma}$—the ~~noisy-free~~ noise-free diagnosis-to-diagnosis covariance matrices, to improve the performance of LDA for early detection of disease. ~~Hereby, we~~ We define our research problem that ~~, Given~~ in the following way: given $n$ diagnosis-frequency vectors $X_0, X_1 \ldots X_{n-1}$, our problem is to estimate $\tilde{\Sigma}$:

$$\text{min. } |\tilde{\Sigma}|_1 \text{ s.t. } ||\tilde{\Sigma} - \hat{\Sigma}||_F^2 \leq \epsilon \text{ and } \tilde{\Sigma} \in I^+ \tag{4}$$

where $I^+$ refers to the overall set of positive ~~semi-definite matrices. Please note~~ semidefinite matrices. Note that $\hat{\Sigma}$ is not foreknown due to the unknown data noise.

Intuitively, it is possible to solve the formulated problem through sparsifying and regularizing the sample diagnosis-to-diagnosis covariance matrix $\Sigma$ ~~subject to the positive semi-definite and non-singularity constraint~~that is positive semidefinite and non-singular.

### 3.2 *Daehr* Framework

In this section, we introduce the framework design of ~~*Daehr*. The framework of *Daehr*~~*Daehr*. *Daehr* consists of two phases~~,~~ ~~which first uses~~. First, we use the EHR data for training to estimate the covariance matrices used in LDA ~~w.r.t~~ with respect to our problem formulation~~, then adopts LDA with newly-estimated~~. Next, we adopt LDA with newly estimated parameters to predict ~~if~~ whether the new patient will develop the targeted disease.

*Phase I: Sparse Covariance Matrix Estimation* — Given the patients' EHR data as a training set, this phase estimates the sparse covariance matrices for two classes of patients with following two steps:

1) **Diagnosis-frequency Vector Extraction and Sample Covariance Matrix Estimation —** ~~*Daehr*first convert~~ *Daehr* first converts each patient's EHR data to a diagnosis-frequency vector and ~~combine~~ combines it with his/her label (indicating ~~if~~ whether the patient has been diagnosed with ~~/without~~ the targeted disease)~~i.e. ,~~. Specifically, we acquire $(X_0, l_0)\ldots(X_{m-1}, l_{m-1})$, where $l_i \in \{-1, +1\}$ ~~as~~ is the label of the $i^{th}$ patient. With the vectors ~~of~~ corresponding to each of the two classes, ~~*Daehr*~~*Daehr* then estimates the sample covariance matrices for the two classes $\Sigma_+$ and $\Sigma_-$ using ~~Eq.~~Equation 3.
2) **Iterative Sparse Covariance Matrix Approximation —** Given sample covariance matrices $\Sigma_+$ and $\Sigma_-$, ~~*Daehr*~~*Daehr* estimates the positive-definite $\ell^1$-penalized estimation of both $\Sigma_+$ and $\Sigma_-$ using a unified iterative approximation process, where ~~*Daehr*~~*Daehr* treats $\Sigma_+$ and $\Sigma_-$ equally. As shown in ~~Alg.~~Algorithm 1, given an input sample covariance matrix $\Sigma_0 = \Sigma_+$ or $\Sigma_-$, the process iteratively approximates to the positive definite $\ell^1$-penalized estimation of $\Sigma_0$ ~~, through alternating two algorithms —~~ through alternating between two algorithms—*$\ell^1$-penalized Sparse Matrix Estimation* and *Nearest Positive ~~Semi-Definite~~ Semidefinite Matrix Approximation* in each iteration. In ~~Alg~~Algorithm 1, $\Delta' = \frac{||\Sigma_{t+1} - \Sigma_t||_\infty}{||\Sigma_t||_\infty}$ and *tol* is a threshold characterizing the tolerance of convergence. Specifically, in each (~~e.g~~i.e., the $t^{th}$, $t \geq 0$) iteration~~iteration~~, the process obtains an improved result $\Sigma_{t+1}$ using the previous result $\Sigma_t$. With the result improved ~~iteration by~~ each iteration, the algorithm stops only when the predefined convergence ~~achieved~~ is achieved ($\Delta'' < tol$~~or maximal iterations reached~~) or after iterating $maxit'$ times (i.e., $t > maxit'$).

~~Please note~~ Note that the covariance matrices for the two classes of patients are estimated in this phase through ~~an unified process, we~~ a unified process. We denote the new covariance matrices as $\Sigma_+^*$ and $\Sigma_-^*$ for the positive and negative classes, respectively.

*Phase II: LDA Modelling and Prediction* — Given the two estimated matrices $\Sigma_+$ and $\Sigma_-$ as well as the training samples, this phase first trains the optimal model for LDA prediction~~and then~~. Then, it uses the LDA model for new ~~patients~~ patient prediction. This phase consists of following two steps:

1) **LDA Model Training —** Given the two estimated covariance matrices $\Sigma_+^*$ and $\Sigma_-^*$ as well as training samples $(X_0, l_0)\ldots(X_{m-1}, l_{m-1})$, ~~*Daehr*searches~~ *Daehr*

**Algorithm 1:** Iterative Approximation Process for Sparse Covariance Matrix Estimation

```
Data: $\Sigma_0$ — the sample covariance matrix i.e., $\Sigma_+$ or $\Sigma_-$
Result: $\Sigma_{t+1}$ — the positive definite $\ell^1$-penalized
        estimation of $\Sigma_0$
begin
    while $\Delta' \geq tol,\ or\ 0 \leq t \leq maxit'$ do
        $\Sigma_{t+\frac{1}{2}} \leftarrow \ell^1$-penalized sparse estimation of $\Sigma_t$
        $\Sigma_{t+1} \leftarrow$ the nearest positive semidefinite
        approximation to $\Sigma_{t+\frac{1}{2}}$
    end
    return $\Sigma_{t+1}$
end
```

searches for the optimal threshold $T^*$ that can maximally classify the two classes of samples ~~with Eq.~~using Equation 2. In this case, ~~*Daehr*models~~ *Daehr* uses a LDA model as $(\Sigma^*_+, \mu_+, \Sigma^*_-, \mu_-, T^*)$.

2) **LDA-based new Patient Prediction —** Given a new patient's EHR data, ~~*Daehr*first convert~~ *Daehr* first converts her data to a diagnosis-frequency vector (e.g., $X'$~~. Then together~~). Combined with the LDA model described as $(\Sigma^*_+, \mu_+, \Sigma^*_-, \mu_-, T^*)$, ~~*Daehr*predict if~~ *Daehr* predicts whether the patient will develop the targeted disease using the criterion in ~~Eq.~~Equation 2.

After the above two phases terminate, ~~*Daehr*has~~ *Daehr* will have (1) learned a LDA model with advanced covariance ~~matrices estimation, then~~ matrix estimation, and (2) adopted the LDA model to enable the early detection of targeted disease. Though the ~~whole framework has beens sketched~~architecture of the framework is discussed here, the design of ~~some algorithms have not yet been introduced. The design of~~ the aforementioned $\ell^1$*-penalized Sparse Matrix Estimation* and *Nearest Positive Semi-Definite Matrix Approximation* algorithms are discussed in following sections.

# 4 *Daehr* Core Algorithms

In this section, we first introduce the two core algorithm used in *Daehr*, then analyzes the performance of the proposed algorithms.

## 4.1 $\ell^1$-penalized Sparse Matrix Estimation

Given the covariance matrix estimated in the previous iteration $\Sigma_t$, this algorithm estimates $\Sigma_{t+\frac{1}{2}}$ – the $\ell^1$-penalized sparse estimation of $\Sigma_t$, using the Proximal Gradient Descent algorithm [46] with following objective function:

$$min.\ \frac{1}{2}||\Sigma_{t+\frac{1}{2}} - \Sigma_t||_F^2 + \tau|\Sigma_{t+\frac{1}{2}}|_1, \tag{5}$$

where $\tau$ is a Lagrange multiplier [47]. When $\tau \geq 0$, the Eq. 5 is a *convex function with sparse input* which can be optimally converged using proximal gradient descent [46]. Please note that $\Sigma_{t+\frac{1}{2}}$ is neither symmetric nor positive semi-definite.

## 4.2 Nearest Positive Semi-Definite Matrix Approximation

Given the sparse matrix $\Sigma_{t+\frac{1}{2}}$, we intend to approximate its nearest positive-definite matrix $\Sigma_t$ (the output of the $t^{th}$ iteration) as Equation 6.

$$min.\ ||\Sigma_{t+1} - \Sigma_{t+\frac{1}{2}}||_F^2\ s.t.\ \Sigma_{t+1} \in I^+ \tag{6}$$

In order to achieve the goal, we use the Alternating Projection Algorithm [48] shown in Alg 2. Specifically, the projection $P_S(A) = \frac{1}{2}(V\lambda_+V^T + (V\lambda_+V^T)^T)$ and $\lambda_+ = \langle min\{\lambda_0, 0\}, min\{\lambda_1, 0\}\ldots\rangle$, where $V, \lambda_i$ is the eigenvalue decomposition of $A$; the projection $P_U(A) = A'$, where $A'_{i,j} = 1$ when $i = j$, and $A'_{i,j} = A_{i,j}$ when $i \neq j$; the stopping criterion $\Delta'' = max\{\frac{||H_{k+1}-H_k||_\infty}{||H_k||_\infty}, \frac{||H^*_{k+1}-H^*_k||_\infty}{||H^*_k||_\infty}, \frac{||H^*_{k+1}-H^*_k||_\infty}{||H_k||_\infty}\}$.

The algorithm stops when the predefined convergence achieved $\Delta'' < tol$, or maximal iterations reached $k = maxit''$. Please note that when the algorithm stops at any $k > 0$, the output $\Sigma_{t+1}$ must be a positive semi-definite matrix; while when $k \to +\infty$, the output $\Sigma_{t+1}$ could converge to the optimal solution [49] of the optimization problem addressed in Eq. 6.

**Algorithm 2:** Alternating Projection Algorithm for Nearest Positive Definite Matrix Approximation

```
Data: $\Sigma_{t+\frac{1}{2}}$ – the $\ell^1$-penalized sparse estimation of $\Sigma_t$,
      $tol$ – the tolerance of convergence
Result: $\Sigma_{t+1}$ – the nearest positive definite
        approximation to $\Sigma_{t+\frac{1}{2}}$
begin
    initialization:
    $H_0 = \frac{1}{2}(\Sigma_{t+\frac{1}{2}} + \Sigma^T_{t+\frac{1}{2}})$, $k = 1$, $I_{mod_0} = 0$, $\Delta = 1$;
    while $\Delta'' \geq tol,\ or\ 0 \leq k \leq maxit''$ do
        $R_{k+1} = H_k - I_{mod_k}$,
        $H^*_{k+1} = P_S(R_{k+1})$;
        $I_{mod_{k+1}} = H^*_{k+1} - R_{k+1}$;
        $H_{k+1} = P_U(H^*_{k+1})$;
    end
    $\Sigma_{t+1} = H_{k+1}$
    return $\Sigma_{t+1}$
end
```

## 4.3 Algorithm Analysis

# 5 Evaluation

In this section, we ~~first introduce the experiment~~ introduce the experimental design of our evaluation~~, then we introduce the experiment~~. Then, we present the experimental results, including the performance comparison between ~~*Daehr*~~the *Daehr* framework and original LDA baselines~~, as well as the performance comparison between *Daehr*~~. Additionally, we present performance comparisons between *Daehr* and other predictive models. Finally, we compare the time ~~consumption of *Daehr*to~~ consumed by *Daehr* with other models.

### 5.1 ~~Experiment~~ Experimental Design

We first present the datasets used for ~~*Daehr*~~our evaluation, then introduce the targeted diseases for the early detection~~, further we~~. We also specify the settings of early detection.

**Dataset for Evaluation** ~~-~~— In this study, to evaluate ~~*Daehr*~~*Daehr*, we plan use the de-identified EHR data from the College Health Surveillance Network (CHSN), which contains over 1 million patients and 6 million visits from 31 student health centers across the ~~US~~United States [50]. In the experiments, we use the EHR data from 10 participating schools. The available information includes ICD-9 diagnostic codes, CPT procedural codes, and limited demographic information. There are over 200,000 enrolled students in those 10 schools representing all geographic regions of the US. The demography of enrolled students (sex, race/ethnicity, age, undergraduate/graduate status) closely matched the demography for the population of US universities.

**Targeted Disease for Early Detection** ~~-~~— Among all diseases recorded in CHSN, we choose mental health disorders, including *anxiety disorders, mood disorders, depression disorders, and other related disorders*, as the targeted disease for early detection. Specifically, we plan to evaluate ~~*Daehr*~~*Daehr* using the early detection of mental health disorders in *college students*, considering following issues:

1) ~~*Emergency*~~ *Emergence of early detection of mental health disorders* ~~-~~— Mental health disorders have become a severe problem in the United States and many other countries that 18.6% adults ~~are with~~ have at least one mental ~~disorders~~disorder. According to the Spring 2014 American College Health Association's National College Health Assessment report, approximately half of the college students have had the feeling of hopeless and overwhelming anxiety [51].
2) *Difficulty ~~to recognize~~ recognizing mental health disorders in early ~~stage~~ -stages* — Mental health disorders are ~~unrecognized frequently~~ frequently unrecognized in primary care~~that untimely~~. Untimely treatment results in emotional, physical, economic, and social burdens to patients and others.
3) ~~*Limitation*~~ *Limitations of common approaches ~~to~~ for early detection of mental health disorders* ~~-~~— Questionnaires are commonly used to detect mental health disorders. Usually, specific questionnaires, interviews, or standard ~~measurement~~ measurements are designed by researchers to collect patients' behavioral information targeting ~~on~~ a particular psychiatric disorder. In particular, psychological screening, PHQ-9, is used to evaluate a patient's risk of mental health disorders [6]. However, these approaches are not generally applicable in primary care thus cannot detect mental disorders at an early stage.

~~With all above in mind, we~~ We are motivated to use EHR data for the early detection of mental health disorders, considering the accessibility and information contained in EHR data.

**Early Detection Settings** ~~-~~— From the CHSN datasets, we select 21,097 patients with anxiety/depression in the target group and 327,198 patients without any mental health disorder in the control group. We represent each patient using his/her diagnosis-frequency vector based on the clustered codeset, where four clustered codes (i.e., xxx, xxx, xxx, xxx) are considered to represent the diagnoses of mental health disorders. Specifically, if a patient has any of these four codes in his/her EHR, we ~~consider~~ say that he/she has been diagnosed with mental health disorders as ground truth. ~~Please note~~ Note that in our research, we ~~don't~~ do not intend to predict these four types of mental disorders separately, as these four disorders are usually correlated and heavily overlapped in clinical practices.

### 5.2 Comparison to LDA Baselines

In order to understand the performance improvement of ~~*Daehr*~~*Daehr* beyond classic LDA, we first propose three LDA baseline approaches that we compare ~~to *Daehr*~~against *Daehr*:

- **LDA** ~~-~~— This algorithm is based on the common implementation of generalized linear discriminant analysis using sample covariance matrix estimation and ~~Eq.~~Equation 2. ~~To handle the singular covariance matrices, this algorithm uses~~ This algorithm uses the pseudo-inverse [41] to replace the matrix inverse in ~~Eq. 2 ,~~ Equation 2 when the sample covariance matrix is singular.
- **Shrinkage** ~~-~~— This algorithm is based on the aforementioned **LDA** implementation (using pseudo-inverse). However, rather than using the sample covariance matrix, this algorithm adopts the sparse estimation of the covariance matrix $\Sigma^* = \beta * \Sigma + (1 - \beta) * diag(\Sigma)$, where $\Sigma$ refers to the given sample covariance matrix, $diag(\Sigma)$ refers to a $p \times p$ matrix preserving the diagonal elements of $\Sigma$ only, and $\beta \geq 0$ is a tuning parameter. ~~**Shrinkage**~~ The Shrinkage algorithm can be considered as a heuristic approach to the optimization problem addressed in ~~Eq.~~Equation 4.
- **DIAG** ~~-~~— This algorithm is based on the ~~**Shrinkage**~~ Shrinkage approach with $\beta = 0.0$, which means the sparse estimation of the covariance matrix $\Sigma^* = diag(\Sigma)$ used in LDA only includes the diagonal information of the sample covariance matrix.

~~Please note~~ Note that the implementation of ~~*Daehr*~~*Daehr* as well as above baselines are derived from the Java implementation of LDA released by Psychometrica[1].

With the four algorithms, we perform experiments with following settings:

- **Training Samples** ~~-~~— we randomly select 50, 100, 150, 200, 250, 300, 350, and 400 patients from the target group as the positive training samples, then randomly select the same number of patients from the control group as negative training samples; here, the training set of the two classes of patients is balanced~~;~~.
- **Testing Samples** ~~-~~— we randomly select 200 and 1000 unselected patients (not included in the training set) from the target group as well as the same number of unselected patients from the control group as the testing set; here, the testing set is also balanced.

[1] Java-Implementation of the Linear Discriminant Analysis, Institute for Psychological Diagnosis, http://www.psychometrica.de/lda.html

TABLE 1: Performance Comparison between *Daehr* and LDA Baselines (Testing Sample Size =200 × 2)

| | | Training Set ×2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 50 | | 150 | | 250 | | 350 | |
| Algorithm | Parameters | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score |
| LDA | N/A | 0.547 | 0.539 | 0.617 | 0.612 | 0.639 | 0.644 | 0.661 | 0.670 |
| DIAG | N/A | 0.592 | 0.591 | 0.635 | 0.635 | 0.639 | 0.639 | 0.653 | 0.660 |
| Shrinkage($\beta$) | 0.25 | 0.593 | 0.592 | 0.636 | 0.638 | 0.640 | 0.643 | 0.656 | 0.665 |
| | 0.50 | 0.594 | 0.592 | 0.630 | 0.630 | 0.641 | 0.645 | 0.660 | 0.669 |
| | 0.75 | 0.592 | 0.590 | 0.626 | 0.624 | 0.639 | 0.643 | 0.662 | 0.672 |
| *Daehr*($\tau$) | $0.005 * 0.5^0$ | 0.644 | 0.692 | **0.667** | 0.714 | **0.662** | **0.716** | **0.670** | **0.722** |
| | $0.005 * 0.5^1$ | 0.645 | 0.694 | 0.666 | 0.713 | **0.662** | **0.716** | **0.670** | **0.722** |
| | $0.005 * 0.5^2$ | **0.646** | **0.697** | 0.663 | 0.714 | **0.662** | **0.716** | **0.670** | **0.722** |
| | $0.005 * 0.5^3$ | **0.646** | 0.694 | 0.661 | 0.712 | **0.662** | **0.716** | **0.670** | **0.722** |
| | $0.005 * 0.5^4$ | **0.646** | 0.696 | 0.662 | **0.715** | **0.662** | **0.716** | **0.670** | **0.722** |

TABLE 2: Performance Comparison between *Daehr* and LDA Baselines (Testing Sample Size =1000 × 2)

| | | Training Set ×2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 50 | | 150 | | 250 | | 350 | |
| Algorithm | Parameters | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score |
| LDA | N/A | 0.552 | 0.545 | 0.619 | 0.620 | 0.644 | 0.648 | 0.656 | 0.663 |
| DIAG | N/A | 0.595 | 0.588 | 0.624 | 0.625 | 0.641 | 0.642 | 0.653 | 0.662 |
| Shrinkage($\beta$) | 0.25 | 0.596 | 0.592 | 0.629 | 0.631 | 0.644 | 0.648 | 0.657 | 0.667 |
| | 0.50 | 0.594 | 0.589 | 0.630 | 0.633 | 0.646 | 0.649 | 0.660 | 0.670 |
| | 0.75 | 0.590 | 0.584 | 0.629 | 0.632 | 0.647 | 0.650 | 0.660 | 0.668 |
| *Daehr*($\tau$) | $0.005 * 0.5^0$ | 0.653 | 0.711 | **0.655** | **0.716** | 0.666 | 0.718 | **0.667** | **0.720** |
| | $0.005 * 0.5^1$ | 0.653 | 0.711 | **0.655** | **0.716** | 0.666 | 0.718 | **0.667** | **0.720** |
| | $0.005 * 0.5^2$ | **0.653** | **0.712** | **0.655** | **0.716** | 0.666 | **0.720** | **0.667** | **0.720** |
| | $0.005 * 0.5^3$ | 0.652 | 0.710 | **0.655** | **0.716** | 0.666 | 0.719 | **0.667** | **0.720** |
| | $0.005 * 0.5^4$ | 0.652 | 0.710 | **0.655** | **0.716** | **0.667** | **0.720** | **0.667** | **0.720** |

For each setting, we ~~evaluate~~ execute the four algorithms and repeat 30 times. ~~Particularly~~In particular, we are interested in measuring following metrics:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN}, \\ \text{F1-score} &= \frac{2*TP}{2*TP+FP+FN} \end{aligned} \tag{7}$$

where $TP$, $TN$, $FP$, and $FN$ refer to the true-positive, true-negative, false-positive, and false-negative classification samples in early detection of mental health disorders respectively. Specifically, the ~~metric Accuracy characterizes~~ Accuracy metric characterizes the proportion of patients who are accurately classified in the early detection of mental disorders~~; while~~. The F1-Score measures both correctness and completeness of the early detection.

Table 1 and Table 2 present ~~a part of~~ part of the comparison results. The results show that under all settings~~*Daehr*outperform~~, *Daehr* outperforms the three baseline algorithms in terms of overall accuracy and F1-score. Compared to LDA, ~~*Daehr*~~*Daehr* achieves 1.4%–18.3% higher accuracy and 7.6%–29.3% higher F1-score. Compared to Shrinkage and DIAG, ~~*Daehr*~~*Daehr* achieves 1.5%–9.7% higher accuracy and 7.9%–21.1% higher F1-score.

TABLE 3: Performance Comparison between *Daehr* and other Predictive Models

| | Training Set ×2 | | | |
|---|---|---|---|---|
| | 50 | | 250 | |
| Algorithm | Accuracy | F1-Score | Accuracy | F1-Score |
| LDA | 0.551 | 0.549 | 0.639 | 0.641 |
| Logit. Reg. | 0.614 | 0.521 | 0.615 | 0.501 |
| SVM | 0.614 | 0.608 | 0.660 | 0.669 |
| AdaBoost-10 | 0.643 | 0.599 | 0.629 | 0.538 |
| AdaBoost-50 | 0.633 | 0.568 | 0.633 | 0.550 |
| *Daehr* | **0.658** | 0.695 | **0.684** | 0.719 |

Further, it is ~~obvious~~ clear that decreasing the ~~training samples, larger the improvement of~~ quantity of training samples results in a larger improvement in accuracy and F1-score~~obtained~~. In this case, we can conclude that ~~*Daehr*~~*Daehr* significantly improves the accuracy and F1-score from the classic LDA, especially when the training sample size is small~~;~~ ~~while *Daehr*~~. *Daehr* outperforms all other baselines derived from LDA ~~,~~ in terms of accuracy and F1-score.

## 5.3 Comparison to other predictive models

In order to understand the performance of ~~*Daehr*~~*Daehr*, we compare it to other predictive models frequently used for early detection of diseases. Specifically, we consider to use following algorithms for the comparison:

- *Support Vector Machine (SVM)* ~~–~~— Inspired by [?], we use a linear binary SVM classifier with fine-tuned parameters.
- *Logistic Regression (Logit. Reg.)* ~~–~~— Inspired by [?], we use a Logistic Regression classifier.
- *AdaBoost-10* and *AdaBoost-50* ~~– In order to compare to ensemble~~ — To compare an ensemble of learning methods, we use AdaBoost to ensemble multiple Logistic Regression classifiers, where AdaBoost-10 refers to the AdaBoost classifier based on 10 Logistic Regression instances and AdaBoost-50 refers to the one with 50 Logistic Regression instances.

~~Together~~ Combined with LDA and ~~*Daehr*~~*Daehr* ($\tau = 0.005 * 0.5^2$), we evaluate these six algorithms using the experiment settings introduced in Section 5.2. The comparison results are shown in Table 3.[2]

~~Comparing with~~ Compared to LDA, SVM, Logistic Regression and AdaBoost can achieve 11.4%–16.7% higher accuracy and 3.5%–10.8% higher F1-score (the only exception is the F1-score of Logistic Regression, which is 5% lower than LDA) with a relatively small training set (Training Set = 50)~~;~~. On a large training set (~~Traning~~ Training set = 250), SVM still ~~has~~ attains better performance than LDA~~while LDA has almost equal performance on accuracy and better F1-score comparing~~. The performance of LDA is nearly equal to Logistic Regression and AdaBoost ~~, while, also compared~~ in terms of accuracy, while achieving a better F1-score. Compared to SVM, Logistic Regression, and AdaBoost, ~~*Daehr*~~*Daehr* can achieve 2.3%–19.4% higher accuracy and 7.5%–43.5% higher F1-score. In this case, we can conclude that the classic LDA model cannot perform as ~~good~~ well as many other predictive models such as SVM and AdaBoost~~, however, *Daehr*~~. However, *Daehr* significantly outperforms all ~~other five~~ five baseline algorithms in all settings. ~~The conclusion indicates that *Daehr*~~These results indicate that *Daehr* not only improves LDA, but ~~*Daehr*itself also is~~ that *Daehr* is also a leading predictive model for early detection of mental health disorders.

## 5.4 Two Case Studies

In order to further understand the performance of ~~*Daehr*, we here use~~ *Daehr*, we present two case studies to ~~first~~ show the time consumption of ~~*Daehr*~~*Daehr*, then analyze the reason ~~why *Daehr*could~~ how *Daehr* can outperform LDA baselines.

**Computational Time Analysis** ~~–~~— We measure computational time consumption of the six algorithms in the experiments introduced ~~by~~ in Section 5. We carried out the experiments using a laptop with an Intel Core i7-2630QM ~~Quart-Core~~ Quad-Core CPU and 8GB memory. All algorithms used in our experiments were implemented with the Java SE platform on a Java HotSpot(TM) 64-Bit Server VM. Table 4 shows the computational time comparison between ~~*Daehr*~~*Daehr* and the rest of methods, where the *“Training”* row refers to the average time consumption of the six algorithms to train a model~~, while the~~. The average time consumption to classify each patient of the testing set is shown in the *“Testing”* row. Among these six algorithms, ~~*Daehr*consumes~~ *Daehr* takes the longest time to ~~train, however~~ train—however, the average time consumption to train a model with 250 × 2 = 500 samples is less than 12 seconds~~which is fairly~~, which is acceptable. On the other hand, the average time consumption to classify a patient using ~~*Daehr*~~*Daehr* is similar to LDA, as these two algorithms are equivalent in terms of prediction. ~~Besides~~In any case, the time consumption of all these six algorithms to classify patients is quite tolerable (i.e., thousands patients per second). ~~In this case, we could~~ We conclude that all ~~these algorithms including *Daehr*~~of the algorithms described here, including *Daehr*, are computationally efficient, in terms of model training and early detection of diseases.

TABLE 4: Computation Time Comparison (in Milliseconds, Training Samples: 250 × 2), “AB ”refers to AdaBoost

| | LDA | *Daehr* | SVM | Logit. Reg. | AB-10 | AB-50 |
|---|---|---|---|---|---|---|
| Training | 249.1 | 11076.3 | 830.97 | 44.97 | 484.2 | 2631.0 |
| Testing | 0.098 | 0.098 | 0.001 | 0.002 | 0.016 | 0.077 |

TABLE 5: Performance Comparison between *Daehr* and other Predictive Models

| | Training Set ×2 | | | |
|---|---|---|---|---|
| | 50 | | 250 | |
| Algorithm | $\lvert\Sigma - \Sigma_l\rvert_1$ | $\lvert\lvert\Sigma - \Sigma_l\rvert\rvert_F^2$ | $\lvert\Sigma - \Sigma_l\rvert_1$ | $\lvert\lvert\Sigma - \Sigma_l\rvert\rvert_F^2$ |
| LDA | 0.551 | 0.549 | 982.56 | 421.58 |
| *Daehr* | **0.658** | 0.695 | **862.5** | **224.24** |

**Covariance Matrix Estimation Analysis** ~~–~~— ~~In our research, we assume *Daehr*improves LDA model,~~ We assume *Daehr* improves LDA the model because the sparse covariance matrix used in ~~*Daehr*~~*Daehr* is more “accurate” than the sample covariance matrix used in LDA when the training sample size is limited. In order to verify our hypothesis, we (1) ~~we first~~ gather the EHR data of all 21,097 patients with mental health disorders from CHSN (4 years EHR of 22 US Universities); (2) ~~then, we~~ randomly select 10,000 patients from them to estimate covariance matrix $\Sigma_l$, (3) ~~we~~ randomly select another 50 or 250 samples to train LDA and *Daehr*; and (4) ~~we~~ further compare $\Sigma_l$ to the covariance matrices estimated in LDA and ~~*Daehr*~~*Daehr* separately through measuring the error of matrices. We repeat ~~above step~~ steps 1 ~~to~~ through 4 for ~~totally~~ a total 30 ~~times,~~ trials so as to obtain the average error between the covariance matrices. Table 5 ~~present~~ presents the average error between covariance matrices in $\ell^1$/Frobenius-norm. The results show that, compared to LDA, the covariance matrix estimated in ~~*Daehr*~~*Daehr* using small samples is **more closed** to the covariance matrix estimated using large samples.

[2] Please note that the results of LDA and ~~*Daehr*~~*Daehr* in Table 3 are slightly different from those in Table 1 and Table 2, since we ~~do~~ conduct the two sets of experiments separately.

In this case, we ~~could conclude that *Daehr*~~conclude that *Daehr* can accurately estimate the covariance matrix for linear discriminant analysis, even when a small number of samples are given for model training.

~~Please note~~ Note that in our experiment, we simulate a training set with a relatively large sample size (i.e., 10,000)~~,~~ ~~however~~. However, for realistic predictive model training, such a large number of samples ~~are~~ is usually not available.

~~Due to spacelimitation, some *Daehr*evaluation~~ To conserve space, some results are not reported here. Readers are encouraged to see the Appendix for additional details, including the evaluation results under more evaluation settings and more ~~experiment~~ experimental insights.

## 6 Discussions & Conclusions

## References

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[2] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2004, vol. 544.

[3] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[4] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*. IEEE, 2008, pp. 108–115.

[5] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction 1," 2011.

[6] 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.

[7] R. B. D'Agostino Sr, S. Grundy, L. M. Sullivan, P. Wilson *et al.*, "Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation," *Jama*, vol. 286, no. 2, pp. 180–187, 2001.

[8] E. R. Dubberke, K. A. Reske, L. C. McDonald, and V. J. Fraser, "Icd-9 codes and surveillance for clostridium difficile–associated disease," *Emerging infectious diseases*, vol. 12, no. 10, p. 1576, 2006.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 16–24, 2012.

[14] F. Wang and J. Sun, "Psf: A unified patient similarity evaluation framework through metric learning with weak supervision," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 3, pp. 1053–1060, May 2015.

[15] Y. H. H. W. K. L. L. E. B. Jinghe Zhang, Haoyi Xiong, "MSEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data," in *Big Data), 2015 International Conference on*. IEEE, 2015.

[16] S. Jensen and U. SPSS, "Mining medical data for predictive and sequential patterns: Pkdd 2001," in *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2001.

[17] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: ACM, 2015, pp. 705–714. [Online]. Available: http://doi.acm.org/10.1145/2783258.2783352

[18] L. Cazzanti and M. R. Gupta, "Local Similarity Discriminant Analysis," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 137–144.

[19] J. Kalina, L. Seidl, K. Zvára, H. Grünfeldová, D. Slovák, and J. Zvárová, "Selecting relevant information for medical decision support with application to cardiology," *European Journal for Biomedical Informatics*, vol. 9, no. 1, pp. 2–6, 2013.

[20] I. Karlsson and H. Bostrom, "Handling sparsity with random forests when predicting adverse drug events from electronic health records," in *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*. IEEE, 2014, pp. 17–22.

[21] F. Wang, P. Zhang, X. Wang, and J. Hu, "Clinical risk prediction by exploring high-order feature correlations," in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 1170.

[22] P.-L. Hsu and H. Robbins, "Complete convergence and the law of large numbers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 33, no. 2, p. 25, 1947.

[23] Z. Qiao, L. Zhou, and J. Z. Huang, "Effective linear discriminant analysis for high dimensional, low sample size data," in *Proceeding of the World Congress on Engineering*, vol. 2. Citeseer, 2008, pp. 2–4.

[24] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of lda," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3. IEEE, 2002, pp. 29–32.

[25] H. Gao and J. W. Davis, "Why direct lda is not equivalent to lda," *Pattern Recognition*, vol. 39, no. 5, pp. 1002–1006, 2006.

[26] T. T. Cai and H. H. Zhou, "Minimax estimation of large covariance matrices under l1 norm," *Statistica Sinica*, vol. 22, no. 4, pp. 1319–1378, 2012.

[27] L. Xue, S. Ma, and H. Zou, "Positive-definite 1-penalized estimation of large covariance matrices," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1480–1491, 2012.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] K. Ng, J. Sun, J. Hu, and F. Wang, "Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity," *AMIA Summits on Translational Science Proceedings*, vol. 2015, pp. 132–136, Mar. 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525240/

[35] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 453–461. [Online]. Available: http://dl.acm.org/citation.cfm?id=2339605

[36] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. Laine, "A Framework for Mining Signatures from Event Sequences and Its Applications in Healthcare Data," 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6200289

[37] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of Biomedical Informatics*, vol. 48, pp. 148–159, Apr. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046414000094

[38] A. Perer and F. Wang, "Frequence: interactive mining and visualization of temporal frequent event sequences," in *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014, pp. 153–162. [Online]. Available: http://dl.acm.org/citation.cfm?id=2557508

[39] A. Perer, F. Wang, and J. Hu, "Mining and exploring care pathways from electronic medical records with visual analytics," *Journal of Biomedical Informatics*, vol. 56, pp. 369–378, Aug. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046415001306

[40] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, and J. Ye, "FeaFiner: biomarker identification from medical data through feature generalization and selection," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1034–1042. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487671

[41] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 8, pp. 982–994, 2004.

[42] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using lda-based algorithms," *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 195–200, 2003.

[43] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, 2011.

[44] J. Shao, Y. Wang, X. Deng, S. Wang *et al.*, "Sparse linear discriminant analysis by thresholding for high dimensional data," *The Annals of statistics*, vol. 39, no. 2, pp. 1241–1265, 2011.

[45] E. R. Ziegel, "Modern applied statistics with s," *Technometrics*, vol. 45, no. 1, p. 111, 2003.

[46] Y. Nesterov, *Introductory lectures on convex optimization*. Springer Science & Business Media, 2004, vol. 87.

[47] H.-C. Wu, "The karush–kuhn–tucker optimality conditions in multiobjective programming problems with interval-valued objective functions," *European Journal of Operational Research*, vol. 196, no. 1, pp. 49–60, 2009.

[48] N. J. Higham, "Computing the nearest correlation matrixa problem from finance," *IMA journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[49] R. L. Dykstra, "An algorithm for restricted least squares regression," *Journal of the American Statistical Association*, vol. 78, no. 384, pp. 837–842, 1983.

[50] 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.

[51] 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/