CRLEDD: Regularized Causalities Learning for Early Detection of Diseases Using Electronic Health Record (EHR) Data
Jiang Bian, Sijia Yang, Haoyi Xiong, Licheng Wang, Yanjie Fu, Zeyi Sun, and Zhishan Guo

Abstract—The availability of Electronic Health Records (EHR) in health care settings has provided tremendous opportunities for early disease detection. While many supervised learning models have been adopted for EHR-based disease early detection, the ill-posed inverse problem in the parameter learning has imposed a significant challenge on improving the accuracy of these algorithms. In this paper, we propose CRLEDD – Causality-Regularized Learning for Early Detection of Disease, an algorithm to improve the performance of Linear Discriminant Analysis (LDA) on top of diagnosis-frequency vector data representation. While most existing regularization methods exploit sparsity regularization to improve detection performance, CRLEDD provides a unique perspective by ensuring positive semi-definiteness of the sparsified precision matrix used in LDA which is different from the regular regularization method (e.g., L2 regularization). To achieve this goal, CRLEDD employs Graphical Lasso to estimate the precision matrix in the ill-posed settings for enhanced accuracy of LDA classifiers. We perform extensive evaluation of CRLEDD using a large-scale real-world EHR dataset to predict mental health disorders (e.g., depression and anxiety) of college students from 10 universities in the U.S. We compare CRLEDD with other regularized LDA and downstream classifiers. The result shows that CRLEDD outperforms all baselines in terms of accuracy and F1 scores.

Index Terms—Classification algorithms, detection algorithms, linear discriminant analysis.

I. INTRODUCTION

THE early disease detection is one of the most prevalent tasks in statistical learning and machine learning, and it plays an important role in modern medical diagnosis and pre-treatment systems. From the aspect of feature extraction, image is the mainstream data type for discovering the latent correlation among the factor of diseases and thereby helps us recognize or classify them. For example, [1], [2] propose to use SAR [3] image data to process the object recognition and the target segmentation, where the statistical-based texture features such as KWE [4] and KCE [5] are well-studied [6] as the basis to support the classification. From the aspect of learning model, [7] propose a hierarchical learning architecture which integrates the well-known CNN [8] and MLP [9] to recognize the target image object. However, most of these preliminary work are based on the image data, where sometimes it is difficult to collect such highly related image data in disease detection task due to the privacy and technical issue (e.g., for some disease, we do not even know the source of the lesion). Fortunately, for general diagnosis, we still have the common electronic health records associated with each patient, which has been wide-used in the medical systems.

Electronic Health Records (EHR) [10] play a critical role in modern health information management and service innovations. A patient’s EHR contains his/her medical visit history, medication, diagnoses, treatment plans, allergies and so on. One significant feature is the interchangeability of EHR, as a standard protocol for medical/health data generation, storage and communication. The health information is built and managed by authorized institutions in a unified digital format (e.g., ICD-9/10, CPT-9/10 used in EHR standards) such that researchers and scientists can share and analyze the EHR data to enable innovative health services, such as providing computer-assisted diagnosis and offering medication advice. Among these services, early detection of diseases, using their past longitudinal health information of the EHR system, has recently attracted significant attention from the research community. There has been a series of works [10]–[15], which attempt to predict future disease of patients, through data mining techniques using EHR data. Prior literature usually first selected important features, such as diagnosis-frequencies [10], pairwise diagnosis transitions [13], and graphs of diagnosis sequences [15], to represent the EHR data of the patients. Then, a wide range of supervised learning algorithms were adopted to build predictive models for early disease detection, on top of well-represented EHR data.

Specifically, supervised learning tools such as Linear Classification, Logistic Regression, Linear Discriminant Analysis
(LDA), Decision Tree (DT), Random Forest (RF), and Bayesian Network [10], [13] have been adopted to train various predictive models, where a critical step is to learn model parameters from training dataset. However, from the viewpoint of “inverse problem” [1], [16], [17], learning parameters from training data is frequently ill-posed [18]. It is difficult to recover the patterns of causalities between variables (e.g., evidence of diagnosis in EHR data), when the number of training samples is limited but the dimension of EHR data (e.g., types of evidence used in prediction) is large. Such causalities consist of discriminative information and thus are the keys to build predictive models. For example, to train a linear classifier for discriminant projection, we need to first learn an optimal Slope Vector. Literature [19] has shown that when the size of training data is less than the dimension of the data (aka EHR data), the estimated slope vector would be “ill-posed” with weak capacity of discrimination, when using traditional Ordinary Least Squares (OLS) or Maximum Likelihood Estimation (MLE) estimator [20], [21]. In this case, the performance of such linear classifiers with ill-posed estimation of parameters will be degraded significantly [22]. Thus, we consider the key challenge of training predictive models for EHR-based early detection of diseases as a type of ill-posed inverse problem.

To understand the ill-posed inverse problem in machine learning, Vapnik and Chervonenkis proposed Structural Risk Minimization (SRM) theory [23]. The SRM theory decomposes the error of predictive model into two parts: training error and generalization error. According to the SRM theory [24], the training of traditional models mainly focuses on minimizing the training error over the training set, without appropriately controlling the generalization error. To understand the generalizability of the model, they further proposed VC dimension [25] (Vapnik-Chervonenkis dimension) as a measure of potential generalization error, leveraging the complexity of the model. More recently, they proposed the regularization method to balance training error and generalization error, with respect to the VC dimension of the trained model, to tackle the ill-posed inverse problem in parameter learning. Usually, these regularization methods intend to approximate the \( \text{sparse}(st) \) parameter estimation, while lowering the training error [26].

For example, to regularize linear classification, Support Vector Machine (SVM) [27] has been proposed to leverage the sparse estimation of the slope vector for discriminative linear projection, where a Lasso [28] estimator is used to balance the training error and \( \ell_1 \)-norm of the slope vector [29] (which is closely related to the VC dimension of linear classification model). Another example, to improve the performance of Logistic Regression [30], \( \ell_1 \)-norm regularization has been applied to balance the trade-off between training error and generalization error. Further, even for more complicated classification tools such as neural network [31], the regularization is frequently used to avoid over-fitting (control the generalization error) of the model.

In this paper, we focus on another commonly-used linear classification model LDA in early detection of diseases [10], [13]. In LDA, two parameters [32] need to be estimated, i.e., the mean vectors of the training samples, and the precision matrix which represents the causalities between variables (label or diagnosis of diseases). The precision matrix can be estimated by the inverse of sample covariance matrix in LDA. However, when the dimensionality of the training samples is larger than the sample size, the sample covariance matrix becomes singular/noninvertible. Even if the dimension of the training samples is less than the training sample size, this matrix is invertible but still ill-posed [18], [33]. As mentioned before, the regularization techniques are assumed to be able to improve the estimation through sparsifying the parameters. Traditionally, the \( \ell_1 \)-norm regularization (e.g., Lasso) is an option to handle the ill-posed problem such as \( \ell_1 \)-norm SVM. Unfortunately, it cannot guarantee the Symmetric Positive Definiteness (SPD) [34] of the covariance matrix. Another common regularized method – Shrinkage [35] can ensure the SPD of the covariance matrix, whereas it may not be optimal. In order to better estimate the precision matrix, the estimation result is required to be optimal and the SPD property needs to be satisfied as well.

To achieve the above goals, we propose Causality-Regularized Learning [1] for Early Detection of Disease based on Linear Discriminant Analysis. This algorithm aims to improve the performance of LDA on top of diagnosis-frequency vector data representation. Specifically, to achieve both the positive semi-definiteness and the optimal estimation, \( \text{CRLedd} \) employs Graphical Lasso to estimate the precision matrix, and then boost the accuracy of LDA classifiers. The paper is organized as follows: We first introduce some preliminaries and problem formulation in Section II. The framework of \( \text{CRLedd} \) in early detection of disease is described in Section III. Section IV presents the details of the key algorithm used in \( \text{CRLedd} \). Then, we evaluate \( \text{CRLedd} \) using large-scale empirical EHR dataset for predicting mental health disorders in Section V. Based on the comparison results presented in Section V, we further analyze and discuss both advantage and disadvantage of \( \text{CRLedd} \) in Section VI. Finally, in Section VII, we conclude that \( \text{CRLedd} \) clearly outperforms the baseline algorithms in terms of overall accuracy and F1-score in all settings, with lower precision matrix estimation error.

II. PRELIMINARY WORK AND PROBLEM FORMULATION

In this section, we first summarize previous studies and background related to this paper from two aspects, i.e., data mining approaches to early detection of diseases, and linear models for classification and the regularization for linear models. Then we formulate our research problem on top of the existing works.

A. Data Mining Approaches to EHR-Based Early Detection of Disease

Given the raw EHR data, existing data mining efforts to EHR-based early detection first learn a set of features from EHR data to represent each patient. Specifically, the EHR data of each patient was represented as a vector consisting of the frequency of each diagnosis code that has been discovered in previous visits [10]–[12]. EHR data can also be represented using N-gram-like [37] graphs, through counting the pairwise transitions between each

\( ^1 \)This work is an extension of our previous paper [36].
pair of diagnosis codes in every visit [13], [14]. Most recently, Liu et al. proposed to represent the EHR of a patient using the temporal graphs, in order to preserve the temporal order of diagnoses partially [15]. To reduce the dimensionality of EHR data, clustered ICD-9 codes [38] have been frequently used in practice, where each ICD-9 diagnosis code can map to one of 295 groups, compressing each raw diagnosis-frequency vector (≥15,000 dimensions) to roughly 295 dimensions. Liu et al. discussed the method of dimensionality reduction for temporal EHR graphs through edge selection [15].

Given EHR data represented as vectors and graphs, researchers have proposed to predict the target disease through supervised learning, using downstream classifiers [13] or similarity search [10]–[12]. Given EHR data with rich structures, sub-sequential pattern matching and sub-graph pattern matching are also leveraged to identify the disease risk of patients [14], [15].

B. Linear Models for Binary Classification and Regularization

Given the training set \( \{(x_i, y_i), i = 1, 2, \ldots, m\} \), \( x_i \in \mathbb{R}^d \). Let \( X = [x_1, x_2, \ldots, x_m]^T \), \( Y = [y_1, y_2, \ldots, y_m] \) and \( \beta = [\beta_1, \beta_2, \ldots, \beta_d]^T \). Normally, the basic linear model is as follow,

\[
    Y = \beta^T X
\]

where \( \beta \) is the slope vector which can represent discriminative property between variables. Based on the above model, the linear classification model can be described here as \( Y = \text{Sign} (\beta^T X) \).

In order to build the linear classification model, we need to estimate the \( \beta \) from the training samples.

**Direct \( \beta \)-based Regularization** — Typically, the optimal \( \beta^* \) is estimated by an optimization problem formulated as below,

\[
    \beta^* = \arg \min_{\beta} \left( \sum_{i=1}^{m} \left( \text{Sign}(\beta^T x_i) - y_i \right)^2 + \lambda |\beta| \right)
\]

where \( m \) is the number of training samples. Specifically, when \( \lambda = 0 \), Eq. (2) is a standard form without any regularization. To further improve the performance of the model, a weighted \( \ell_1 \)-norm (also Lasso) regularization is adopted (\( \lambda \neq 0 \)) to constrain (sparsifying) the \( \beta \) in the estimation [39]. For example, the linear SVM is a type of \( \beta \)-based linear classification model. Correspondingly, the \( \ell_1 \)-norm SVM can sparsify the parameter estimation by shrinking the small coefficients \( \beta \) of the hyperplane to exactly zero. Especially in High Dimension Low Sample Size (HDLSS) settings, the \( \ell_1 \)-norm SVM performs well compared to the normal linear SVM [40].

**Covariance-based Regularization** — In this paper, we focus on another well-studied linear model – Linear Discriminant Analysis (LDA). Based on the basic linear model,

\[
    \beta = (X^T X)^{-1} X^T Y
\]

where \( X^T X \) can be considered the covariance matrix \( \Sigma \). Instead of using \( \beta \) vector in linear model, the LDA leverages precision matrix (inverse covariance matrix) \( \Theta = \Sigma^{-1} \) as the parameter to represent the causality between the variables. However, the estimation of \( \Theta \) is invalid in HDLSS settings due to the loss of positive semi-definite property of \( \Sigma \). Thus, ensuring the SPD of the covariance matrix \( \Sigma \) is critical for estimating \( \Theta \). Although a common regularized technique of shrinkage can guarantee the SPD of \( \Sigma \) [41], the results of the estimation are usually not optimal.

C. Problem Formulation

To estimate the precision matrix \( \Theta \), the maximum likelihood estimation problem is shown below,

\[
    \Theta^* = \arg \min_{\Theta \in \mathcal{P}_{pp}} \left( -\sum_{i=1}^{m} \log p(x_i | \mu, \Theta) + \lambda |\Theta|_1 \right),
\]

where \( \mu \) is the mean vector of the \( m \) samples. Specifically, the first term \( -\sum_{i=1}^{m} \log p(x_i | \mu, \Theta) \) is an expression of negative Logarithm maximum likelihood which can be expanded as

\[
    -\log(p(x_1 | \mu, \Theta) \times p(x_2 | \mu, \Theta) \times \cdots \times p(x_m | \mu, \Theta)).
\]

However, the optimization problem (4) is intractable. To address this problem, we reduce the intractable formulation to:

\[
    \text{tr}(\Sigma \Theta) - \log \det(\Theta),
\]

where \( \text{tr}(\cdot) \) is the trace of square matrix and \( \Sigma \) is the sample covariance matrix. In the estimation of \( \Theta \), the \( \ell_1 \)-norm regularization is adopted to provide a sparse precision matrix which can be used as an optimal approximation of the inverse sample covariance matrix \( \Sigma^{-1} \). This method is also called Graphical Lasso [42], it can ensure the SPD of sample covariance matrix and while simultaneously provide an optimal solution for the precision matrix. When using EHR data with HDLSS settings as training samples, we employ Graphical Lasso to improve the performance of original LDA model by providing a better estimation of precision matrix.

III. FRAMEWORK

In this section, we introduce the CRLEDD framework. CRLEDD consists of three phases as shown in Fig. 1. First, we use diagnosis-frequency vectors to represent the EHR data. Then, we estimate the covariance matrices used in LDA with respect to our problem formulation and estimate sparse covariance matrix via Graphical Lasso. After that, we adopt LDA with newly estimated parameters to predict whether the new patient will develop the targeted disease.

**Phase I: EHR Data Representation** — There are many existing approaches to represent EHR data including the use of diagnosis-frequencies [10], [11], pairwise diagnosis transition [13], and graph representations of diagnosis sequences [15]. Among these approaches, the diagnosis-frequency is a common way to represent EHR data.

Given each patient’s EHR data, the proposed method first retrieves the diagnosis codes [43] recorded during each visit. Next, the frequency of each diagnosis in all past visits is counted, followed by further transforming the frequency of each diagnosis into a vector of frequencies. For example, \( \{1, 0, 0, \ldots, 3\} \), where 0 means that the second disease has not been diagnosed in any of the past visits. In this paper, we denote the dimension of diagnosis-frequency vectors as \( p \). Note that the dimension...
Fig. 1. Overview of the three-phase framework: CRLEDD – Regularized Causalities Learning for Early Detection of Diseases using Electronic Health Record (EHR) Data. Depending on the functionality, the framework are divided into three phases which are Data Representation, Correlation Analysis, Supervised Learning and Prediction.

$p \geq 15,000$ when using ICD-9 codes, $p \geq 250$ even when using clustered ICD-9 codes [38], while the number of samples for training $m$ is significantly less than $p$.

**Phase II: Correlation Analysis** — Given the patients’ EHR data as a training set, this phase estimates the sparse precision matrices for each type of the disease for two classes of patients (diagnosed with target disease or not) with following two steps:

1) **Sample Covariance Matrix Estimation With Extracted Diagnosis-frequency Vector** — CRLEDD combines diagnosis-frequency vector for each patient with his/her label (indicating whether the patient has been diagnosed with the targeted disease). Then we estimate the sample covariance matrices using maximized likelihood estimator.

2) **Sparse Precision Matrix Estimation Using Graphical Lasso** — Given sample covariance matrices $\bar{\Sigma}$, CRLEDD estimates the sparse precision matrix using Graphical Lasso estimator.

Note that the covariance matrices for the two classes of patients are estimated in this phase through a unified process.

**Phase III: Supervised Learning and Prediction** — Given the estimated matrices $\bar{\Sigma}$ as well as the training samples, this phase first trains the optimal model for LDA prediction. Then, it uses the LDA model for new patient prediction.

Given all parameters $\bar{\Sigma}$, $\bar{\mu}_{+1}$ (the mean vector of the sample consisting of the patients with target disease), and $\bar{\mu}_{-1}$ (the mean vector of sample consisting of the patients without target disease), the LDA model classifies a new patient’s data $x$ as the result of:

$$\arg\max_{l \in \{+1, -1\}} \left( x^T \bar{\Sigma}^{-1} \bar{\mu}_l - \frac{1}{2} \bar{\mu}_l^T \bar{\Sigma}^{-1} \bar{\mu}_l + \log \alpha_l \right),$$  (7)

where $l$ is the label needs to be identified to predict if a certain patient is diagnosed with the target disease or not. $l$ can be either positive one or negative one. Positive one means the patient will be predicted to have the target disease, while negative one means the patient will not be predicted to have the target disease. $\alpha_{+1}$ and $\alpha_{-1}$ refer to the empirical frequencies of positive samples (i.e., patients with the target disease) and negative samples (i.e., patients without the target disease) in the whole population.

**IV. KEY ALGORITHM OF CRLEDD**

In this section, we introduce the design of implementation of key algorithm used in CRLEDD. Section A describes the Causality-Regularized LDA classifier, and Section B presents the Graphical Lasso algorithm.

**A. Causalities-Regularized LDA for Diagnosis Frequency Vector Classification**

Given $m$ samples (i.e., EHR frequency vectors) which will be used to train the estimator along with corresponding labels, i.e., $(x_0, l_0) \ldots (x_{m-1}, l_{m-1})$, the early disease detection procedure is designed to determine if a new patient with its data vector $x$ would develop into the target disease by projecting the vector $x$ to $+1$ (positive) or $-1$ (negative).

To enable the classification with LDA, CRLEDD first estimates the sample covariance matrix using the pooled maximized likelihood estimator:

$$\bar{\Sigma} = \frac{1}{m} \sum_{l \in \{+1, -1\}} \sum_{x_i \in X_l} (x_i - \bar{\mu}_l)(x_i - \bar{\mu}_l)^T,$$  (8)
where \( X_l \) refers to the set of patients with the label \( l \) (i.e., \( l \in \{-1, +1\} \) referring to the patients without/with the target diseases respectively), specifically, \( X_l = \{ x_i | (x_i, l_i) \} \) and \( l_i = l \).

Given the sample covariance matrix \( \Sigma \), this method estimates a sparse precision matrix \( \Theta \) using the Lasso-alike regularization estimator:

\[
\hat{\Theta} = \arg\min_{\Theta \in \mathbb{P}_N} \left( \text{tr}(\Sigma \Theta) - \log \det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right). \tag{9}
\]

It is considered a \( \ell_1 \)-penalized negative log-likelihood minimization estimator, which can be implemented as Graphical Lasso under SPD constraint.

### B. Implementation of the Log-Divergence Minimization Algorithm via Graphical Lasso

Suppose we have \( m \) samples with dimension \( p \) and sample covariance matrix \( \Sigma \). In order to solve the optimization problem in Eq. (9) to obtain the \( \hat{\Theta} \), the Graphical Lasso algorithm [42] is used to estimate \( \Theta^{-1} \) and recover \( \hat{\Theta} \) after convergence. The details of this algorithm are listed as follows.

Let \( \mathbf{W} = \Theta^{-1} \) and \( \mathbf{S} = \Sigma \), then partitioning \( \mathbf{W} \) and \( \mathbf{S} \)

\[
\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & \mathbf{w}_{22} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}^T & \mathbf{S}_{22} \end{pmatrix} \tag{10}
\]

The solution for \( w_{12} \) satisfies

\[
w_{12} = \arg\min_y \left\{ \mathbf{y}^T \mathbf{W}_{11}\mathbf{y} : \| y - s_{12} \|_{\infty} \leq \lambda \right\} \tag{11}
\]

This is a box-constrained quadratic program that was once solved by Banerjee et al. [44] using an interior point procedure. It has been illustrated that the iterates in this procedure remain positive definite and invertible, even if \( P > N \) when the procedure is initialized with a positive definite matrix. Thus, here the SPD of \( \mathbf{W} \) can be ensured.

Using convex duality, Banerjee et al. [44] showed that solving Eq. (11) is equivalent to solving the dual problem

\[
\min_{\beta} \left\{ \frac{1}{2} \| \mathbf{W}_{11}^2 \beta - b \|_2^2 + \lambda \| \beta \|_1 \right\}, \tag{12}
\]

where \( b = \mathbf{W}_{11}^T s_{12} \). If \( \beta \) solves Eq. (12), then \( w_{12} = \mathbf{W}_{11} \beta \) solves Eq. (11). Expression of Eq. (12) resembles a Lasso form, and is the basis for the approach of Graphical Lasso.

To verify the equivalence of the solutions between Eq. (9) and Eq. (12) directly, the relation \( \mathbf{W}\Theta = \mathbf{I} \) can be expanded as below:

\[
\begin{pmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & \mathbf{w}_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12}^T & \Theta_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & 1 \end{pmatrix}. \tag{13}
\]

Now the sub-gradient equation [45] for the maximization of the log-likelihood of Eq. (9) is

\[
\mathbf{W} - \mathbf{S} - \lambda \text{Sign}(\Theta) = 0, \tag{14}
\]

using the fact that the derivative of \( \log \det(\Theta) \) equals \( \Theta^{-1} = \mathbf{W} \).

**Algorithm 1:** The \( \ell_1 \)-Penalized Log-Divergence Minimization via Graphical Lasso.

1. **Initialize** \( \mathbf{W} = \mathbf{S} + \lambda \mathbf{I} \). The diagonal of \( \mathbf{W} \) remains unchanged in what follows.
2. **Repeat** for \( j = 1, 2, \ldots, p \), until convergence:
   - **Partition** the matrix \( \mathbf{W} \) into two parts. Part 1: all but the \( j \)th row and column. Part 2: the \( j \)th row and column.
   - **Solve** the estimating equation
     \[
     \mathbf{W}_{11} \beta_{12} - s_{12} + \lambda \text{Sign}(\beta_{12}) = 0
     \]
     using the cyclical coordinate-descent algorithm for the modified Lasso.
   - **Update** \( w_{12} = \mathbf{W}_{11} \beta_{12} \).
3. In the final cycle (for each \( j \)) solve for \( \hat{\beta}_{12} = -\beta_{12} \hat{\Theta}_{22} \), with
   \[
   1/\hat{\Theta}_{22} = w_{22} - w_{12}^T \beta_{12}.
   \]

The upper right block of the gradient equation from Eq. (14) is

\[
w_{12} - s_{12} - \lambda \text{Sign}(s_{12}) = 0. \tag{15}
\]

On the other hand, the sub-gradient equation from Eq. (12) works out to be

\[
\mathbf{W}_{11} \beta_{12} - s_{12} + \lambda \text{Sign}(\beta_{12}) = 0, \tag{16}
\]

where \( w_{12} = -\mathbf{W}_{11} \beta_{12}/\hat{\Theta}_{22} = \mathbf{W}_{11} \beta \). The equivalence of the first two terms is obvious. For the sign terms, since \( \mathbf{W}_{11} \beta_{12} + w_{12} \beta_{22} = 0 \) from Eq. (14), we have that \( \hat{\beta}_{12} = -\hat{\Theta}_{22} \mathbf{W}_{11}^{-1} w_{12} \). Since \( \hat{\Theta}_{22} > 0 \), it follows that \( \text{Sign}(\hat{\beta}_{12}) = -\text{Sign}(\mathbf{W}_{11}^{-1} w_{12}) = -\text{Sign}(\beta_{12}) \). This proves the equivalence. Thus, we can solve the Lasso problem Eq. (12) instead of solving the original Eq. (9).

In terms of inner products, the lasso estimates for the \( p \)th variable on the others take \( \mathbf{S}_{11} \) and \( s_{12} \) as the input data, where \( p \) is the dimension of the samples. To solve Eq. (12), we instead use \( \mathbf{W}_{11} \) and \( s_{12} \), where \( \mathbf{W}_{11} \) is our current estimate of the upper block of \( \mathbf{W} \). We then update \( w_{12} \) and cycle through all of the variables until convergence. The main steps of this estimation process are shown in Algorithm 1.

Note that the Lasso [28] problem in step (b) above can be efficiently solved by cyclical coordinate-descent algorithm [46]. Here are the details. Let \( V = \mathbf{W}_{11} \), then the update has the form

\[
\hat{\beta}_i \leftarrow S((s_{12})_j - \sum_{k \neq j} V_{kj} \hat{\beta}_k, \lambda)/V_{jj} \tag{17}
\]

for \( j = 1, 2, \ldots, p \), where \( S \) is the soft-thresholding operator:

\[
S(x,t) = \text{sign}(x)(|x| - t)_+. \tag{18}
\]

It cycles through the predictors until convergence. Although step 2 has estimated \( \hat{\Theta}^{-1} = \mathbf{W} \), it can recover \( \hat{\Theta} = \mathbf{W}^{-1} \) relatively cheaply. Note that from the partitioning.
in Eq. (14), we have
\begin{align}
    W_{11} \theta_{12} + w_{12} \theta_{22} &= 0 \\
    w_{12}^T \theta_{12} + w_{22} \theta_{22} &= 1,
\end{align}
from which we derive the standard partitioned inverse expressions
\begin{align}
    \theta_{12} &= -W_{11}^{-1} w_{12} \theta_{22} \\
    \theta_{22} &= 1/(w_{22} - w_{12}^T W_{11}^{-1} w_{12}).
\end{align}

According to Eq. (20), \( \theta_{22} \) and \( \theta_{12} \) can be easily computed in step 3. The Graphical Lasso algorithm stores all the coefficients \( \beta \) for each of the \( p \) problems in a \( p \times p \) matrix, and compute \( \hat{\theta} \) after convergence. As was discussed in [44], the estimator \( \hat{\theta} \) should be Symmetric Positive-Definite (SPD) and Sparse. Furthermore, the recent work [47] leverages the similar method to estimate covariance matrix and proves its superiority under HDLSS settings.

V. EXPERIMENTAL RESULTS

In this section, we first introduce the data preprocessing based on the raw EHR data. After that, the existing algorithms that will be used as the baseline settings when comparing with CRLEDD are given. Then, the experimental results are demonstrated and discussed.

A. Data Preparation

To evaluate CRLEDD, we select the de-identified EHR data of 10 participating schools from the entire dataset including 31 student health centers across the U.S. with totally over 1 million patients and 6 million visits records provided by the College Health Surveillance Network (CHSN) [48]. 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 U.S. The demography of enrolled students (sex, race/ethnicity, age, undergraduate/graduate status) in the selected dataset closely matches the demography of the students in the universities throughout the U.S.

We select the most common mental health disorders, anxiety and mood disorders from primary care data, as the target disease for early detection. Thousands of ICD-9 codes are clustered into 283 categories according to the AHRQ Clinical Classification Software and expert opinions [38]. We use his/her diagnosis-frequency vector based on the clustered code set to represent each patient, where four clustered codes (i.e., 651, 657, 658, 662) represent anxiety and mood disorders.

Note that in our research, we do not predict these four types of mental disorders separately, as these four disorders are often co-occurring in clinical practices [49]. Further, patients with less than two visits were excluded from the analysis.

Notably, the visit data and corresponding diagnosis information within one-month of the first diagnosis of anxiety/depression in the target group is excluded for the aim of early detection at least one to three-month prior to diagnosis. The diagnosis-frequency vectors are used as predictors in our experiment and only include the diagnosis frequency of non-mental health diagnoses with all mental health related information removed. In this case, our experiment is equivalent to predicting whether a patient is likely to have or develop a mental health disorder based on their diagnosis history.

B. Baseline Algorithms and Comparison Settings

To understand the performance impact of CRLEDD beyond classic LDA, we first propose two kinds of baseline approaches to compare against CRLEDD, then two types of discriminative learning models are prepared for the comparison:

Regularized LDA Classifiers (three algorithms) – First, we use the typical LDA classifier, which employs the sample covariance estimation. Then, we consider the Shrinkage LDA [50] using shrinkage covariance estimator with the sparsity parameter \( \beta \). Finally, we propose to use DIAG–a special Shrinkage with \( \beta = 0.0 \).

Downstream Classifiers (four algorithms) – We start with Support Vector Machine (SVM, with regularization parameter \( C = 1.0 \)) [10], and then use Logistic Regression (Log. Reg.) [51]. Finally, we adopt two Adaboost classifiers ensemble 10 and 50 logistic regression classifiers (AdaBoost(10) and AdaBoost(50)).

With the seven baseline algorithms, we perform experiments with training samples and testing samples. We randomly select 50, 100, 150, 200, and 250 patients with mental health disorders as the positive training samples, and randomly select the same number of patients without a mental health diagnosis as negative training samples to maintain the balance. In terms of testing samples, we randomly select 500, 1000, 1500, and 2000 patients from each of the two patient classes (positive/negative) to build the testing set.

Then, we reveal the initial settings of some key parameters in proposed CRLEDD algorithm. The L1 regularization parameter \( \lambda \) is set to be 1, 10, 100 for comparison. The tolerance to declare convergence for graphical lasso is set to be \( 10^{-4} \), and the number of maximum iteration for its optimization is set to be 100. For each setting, we execute the seven algorithms and repeat 30 times. Then, we compare the accuracy and F1-Score of different algorithms.

Also we perform an experiment to compare \( \ell_1 \)-norm error of estimator between LDA and CRLEDD with different sample sizes. Specifically, we randomly select 100 and 200 patients from each of the two patient classes (positive/negative) to build the testing samples.

C. Experiment Results

In this experiment, two types of comparison results are demonstrated:

1) Accuracy and F1-Score Comparison: Figs. 2 and 3 present the performance in terms of accuracy and F1-score of our method and baselines with various sizes of testing samples given different training sample sizes (more results are attached in the appendix). As can be seen from the experiment results, CRLEDD clearly outperforms the baseline algorithms in terms of overall
Fig. 2. Accuracy Performance Comparison between CRLEDD and Baselines with Small Training Datasets (Testing Sample Size = 500 × 2, 1000 × 2, 1500 × 2, 2000 × 2 from left top to right bottom, 90 days in advance).

2) Sensitivity and Specificity Comparison: Table I additionally presents the performance with regards to sensitivity and specificity. The sensitivity is the percentage of patients who are correctly diagnosed as having the corresponding disease. As can be seen in the table, when training sample is 100 and testing sample is 1000, the sensitivity of CRLEDD is 0.842 in average, obviously higher than the sensitivity of SVM that have the highest value 0.633 among other baseline algorithms, which can explain that CRLEDD has greater ability to correctly detect patients than the other baseline algorithms. The specificity which measures the proportion of people who are correctly identified as not having the disease, provided by the CRLEDD is lower than the other baseline algorithms. According to the table, CRLEDD achieves the highest value of the specificity as 0.510 when λ = 1, which is still lower than the LDA that have the lowest value 0.571 among other baseline algorithm. Similarly, this also occurs when the training sample is 500 and the testing sample is 4000. While, the CRLEDD is not better than the baseline algorithms in regards to specificity, it performs better with regards to correctly identifying those individuals with the disease. Further, we expect a high number of false positives because 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.

Trade-off: Moreover, we can observe that the CRLEDD sacrifices some specificity to achieve high sensitivity to some degree (33% gain in sensitivity VS 17% loss in Specificity when comparing with LDA). However, we see the utility of CRLEDD as an opportunity to perform psychological screening (e.g.; PHQ-9 [52]) in a primary care setting which could further identify the student’s risk of a mental health disorder. Because of this, we focus more on correctly diagnosing those patients with the target disease.

3) Estimator Error Comparison: We assume CRLEDD improves LDA because that the sparse precision matrix used in CRLEDD is more “precise” than the sample precision matrix used in simple LDA models when the training sample size is limited. Thus, we compare the ℓ₁-norm error of these two estimators accuracy, and F1-score, in all settings. Specifically, CRLEDD achieves 3.1%–20.9% higher accuracy and 11.7%–31.9% higher F1-score, compared to the typical LDA; CRLEDD achieves 7.5%–15.7% higher accuracy and 13.8%–41.9% higher F1-score, compared to the DIAG; CRLEDD achieves 6.7%–19.2% higher accuracy and 12.3%–71.6% higher F1-score, compared to the Shrinkage. Compared to those robust classifiers such as SVM, Logistic Regression, and AdaBoost, CRLEDD still clearly outperforms these baseline algorithms. Thus we can conclude that CRLEDD overall outperforms the baseline algorithms in all experimental settings.
and the results show that CRLEDD can always outperform with less error in different sample sizes. Fig. 4 presents the average error between precision matrices in $\ell_1$-norm. The results show that, compared to LDA ($\Sigma^{-1}$), the precision matrix estimated in CRLEDD ($\hat{\Theta}$) using small samples is closer to the precision matrix estimated using large samples. Note that we repeat the comparison in each setting for 30 times to estimate the average errors.

4) Causality Graph Visualization: To validate the key algorithm of CRLEDD, we draw a causality graph based on the precision matrix in Eq. (9). Specifically, we randomly select a training set with 4000 balanced samples and threshold [53] the Graphical Lasso ($\lambda = 0.1$) at level 

$$
\Phi^{-1}\left(1 - \frac{\alpha}{p(p-1)}\right)\hat{\sigma}_{ij}/\sqrt{n}
$$

where $\alpha = 0.05$ and $\hat{\sigma}_{ij}^2 = \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}_{ij}^2$. We leverage this threshold to pick up the strong causalities node pairs at the 95% significance level. As shown in Fig. 5, each node in the graph represents a category of disorder and the thickness of the edge shows the intensity of the causality. Further, we present the undirected disorder pairs by ranking their causality in the Table II. According to our results, we speculate that the disorders can be grouped into those that are related to anxiety and mood disorders such as other upper respiratory infections, other connective tissue diseases, and administrative/social admission. Other diagnoses are the ones that are unrelated to anxiety/depression such as immunizations and screening for infectious disease, and contraceptive and procreative management. We hypothesize that in the highest risk level that their are pairs of which both or only one of diagnoses are related to anxiety/depression in the higher.
TABLE I
SENSITIVITY AND SPECIFICITY COMPARISON

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Set: 50×2, Testing Set: 500×2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaBoost (×10)</td>
<td>0.627 ± 0.045</td>
<td>0.562 ± 0.091</td>
<td>0.498 ± 0.135</td>
<td>0.756 ± 0.073</td>
</tr>
<tr>
<td>AdaBoost (×50)</td>
<td>0.627 ± 0.036</td>
<td>0.547 ± 0.091</td>
<td>0.471 ± 0.137</td>
<td>0.783 ± 0.072</td>
</tr>
<tr>
<td>CRLEDD (λ = 1.0)</td>
<td>0.651 ± 0.026</td>
<td>0.694 ± 0.026</td>
<td>0.793 ± 0.057</td>
<td>0.510 ± 0.060</td>
</tr>
<tr>
<td>CRLEDD (λ = 10.0)</td>
<td>0.656 ± 0.017</td>
<td>0.713 ± 0.008</td>
<td>0.855 ± 0.027</td>
<td>0.456 ± 0.055</td>
</tr>
<tr>
<td>CRLEDD (λ = 100.0)</td>
<td>0.640 ± 0.029</td>
<td>0.710 ± 0.010</td>
<td>0.878 ± 0.034</td>
<td>0.402 ± 0.088</td>
</tr>
<tr>
<td>LDA</td>
<td>0.560 ± 0.020</td>
<td>0.554 ± 0.032</td>
<td>0.548 ± 0.054</td>
<td>0.571 ± 0.046</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.621 ± 0.037</td>
<td>0.523 ± 0.100</td>
<td>0.440 ± 0.148</td>
<td>0.801 ± 0.081</td>
</tr>
<tr>
<td>SVM</td>
<td>0.616 ± 0.017</td>
<td>0.621 ± 0.029</td>
<td>0.633 ± 0.065</td>
<td>0.599 ± 0.064</td>
</tr>
<tr>
<td>DIAG</td>
<td>0.573 ± 0.023</td>
<td>0.528 ± 0.050</td>
<td>0.484 ± 0.076</td>
<td>0.662 ± 0.060</td>
</tr>
<tr>
<td>Shrinkage (β = 0.25)</td>
<td>0.569 ± 0.028</td>
<td>0.495 ± 0.169</td>
<td>0.469 ± 0.169</td>
<td>0.670 ± 0.122</td>
</tr>
<tr>
<td>Shrinkage (β = 0.5)</td>
<td>0.566 ± 0.025</td>
<td>0.488 ± 0.166</td>
<td>0.459 ± 0.164</td>
<td>0.672 ± 0.118</td>
</tr>
<tr>
<td>Shrinkage (β = 0.75)</td>
<td>0.570 ± 0.016</td>
<td>0.540 ± 0.039</td>
<td>0.509 ± 0.063</td>
<td>0.630 ± 0.045</td>
</tr>
<tr>
<td></td>
<td>Training Set: 250×2, Testing Set: 2000×2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaBoost (×10)</td>
<td>0.633 ± 0.027</td>
<td>0.536 ± 0.089</td>
<td>0.447 ± 0.140</td>
<td>0.818 ± 0.086</td>
</tr>
<tr>
<td>AdaBoost (×50)</td>
<td>0.631 ± 0.026</td>
<td>0.535 ± 0.087</td>
<td>0.445 ± 0.137</td>
<td>0.818 ± 0.085</td>
</tr>
<tr>
<td>CRLEDD (λ = 1.0)</td>
<td>0.686 ± 0.006</td>
<td>0.721 ± 0.009</td>
<td>0.813 ± 0.029</td>
<td>0.558 ± 0.026</td>
</tr>
<tr>
<td>CRLEDD (λ = 10.0)</td>
<td>0.675 ± 0.007</td>
<td>0.720 ± 0.006</td>
<td>0.838 ± 0.021</td>
<td>0.512 ± 0.028</td>
</tr>
<tr>
<td>CRLEDD (λ = 100.0)</td>
<td>0.671 ± 0.009</td>
<td>0.719 ± 0.004</td>
<td>0.844 ± 0.028</td>
<td>0.497 ± 0.043</td>
</tr>
<tr>
<td>LDA</td>
<td>0.648 ± 0.009</td>
<td>0.648 ± 0.018</td>
<td>0.651 ± 0.037</td>
<td>0.644 ± 0.025</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.628 ± 0.028</td>
<td>0.520 ± 0.095</td>
<td>0.427 ± 0.146</td>
<td>0.828 ± 0.090</td>
</tr>
<tr>
<td>SVM</td>
<td>0.666 ± 0.009</td>
<td>0.672 ± 0.014</td>
<td>0.687 ± 0.030</td>
<td>0.644 ± 0.023</td>
</tr>
<tr>
<td>DIAG</td>
<td>0.635 ± 0.015</td>
<td>0.621 ± 0.030</td>
<td>0.601 ± 0.053</td>
<td>0.668 ± 0.032</td>
</tr>
<tr>
<td>Shrinkage (β = 0.25)</td>
<td>0.638 ± 0.012</td>
<td>0.631 ± 0.027</td>
<td>0.621 ± 0.051</td>
<td>0.656 ± 0.032</td>
</tr>
<tr>
<td>Shrinkage (β = 0.5)</td>
<td>0.642 ± 0.011</td>
<td>0.635 ± 0.026</td>
<td>0.626 ± 0.050</td>
<td>0.657 ± 0.032</td>
</tr>
<tr>
<td>Shrinkage (β = 0.75)</td>
<td>0.641 ± 0.010</td>
<td>0.635 ± 0.024</td>
<td>0.628 ± 0.046</td>
<td>0.655 ± 0.030</td>
</tr>
</tbody>
</table>

D. Conclusion on Experiment Results

In the experiments, we evaluate CRLEDD using the empirical EHR datasets, and compare the algorithm with other classifiers under the same balanced dataset settings. The overall evaluation result shows that our algorithm significantly outperforms the existing linear discriminant analysis classifiers and other downstream classifiers, with both higher accuracy and F1-score. The case studies based on the estimated precision matrix show that the Graphical Lasso estimator used in CRLEDD can reduce the ℓ1-norm estimation error and improve the accuracy of classification, on top of the classical LDA classifiers. Further, we visualize the graph of casualties discovered from the data, which makes sense in the medical contexts [54], [55]. It is reasonable to conclude that, through lowering the error of precision matrix estimation, CRLEDD efficiently recovers the casualties between diagnoses related to the social anxiety/depression population from the data, then it improves the classification accuracy/F1-score by incorporating the well-recovered casualties. Note that our algorithm, along with all other baseline algorithms, is evaluated under balanced settings.

Efficiency Comparison: Also, we compare the time consumption of CRLEDD algorithm with most competitive algorithm SVM (500 patients for training and 2000 patients for testing). On average, CRLEDD takes 334.75 seconds for training and testing.
TABLE II
CAUSALITY RANKING OF DISORDERS PAIRS (UNDIRECTED)

<table>
<thead>
<tr>
<th>Level</th>
<th>Code One</th>
<th>Code Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest</td>
<td>(156)Medical examination/evaluation</td>
<td>(81)Contraceptive and procreative management</td>
</tr>
<tr>
<td>Higher</td>
<td>(112)Other connective tissue disease</td>
<td>(104)Other non-traumatic joint disorders</td>
</tr>
<tr>
<td>Higher</td>
<td>(153)Allergic reaction</td>
<td>(32)Asthma</td>
</tr>
<tr>
<td>Higher</td>
<td>(204)Nutritional deficiencies</td>
<td>(37)Other upper respiratory disease</td>
</tr>
<tr>
<td>Higher</td>
<td>(30)Other upper respiratory infections</td>
<td>(36)Other lower respiratory disease</td>
</tr>
<tr>
<td>Middle</td>
<td>(153)Allergic reaction</td>
<td>(37)Other upper respiratory disease</td>
</tr>
<tr>
<td>Middle</td>
<td>(132)Sprains and Strains</td>
<td>(104)Other non-traumatic joint disorders</td>
</tr>
<tr>
<td>Middle</td>
<td>(132)Sprains and Strains</td>
<td>(159)Residual codes; unclassified</td>
</tr>
<tr>
<td>Middle</td>
<td>(76)Menstrual disorders</td>
<td>(81)Contraceptive and procreative management</td>
</tr>
<tr>
<td>Middle</td>
<td>(157)Other aftercare</td>
<td>(155)Administrative/Social admission</td>
</tr>
<tr>
<td>Lower</td>
<td>(156)Medical examination/examination</td>
<td>(2)Immunization and Screening for infectious</td>
</tr>
<tr>
<td>Lower</td>
<td>(155)Administrative/Social admission</td>
<td>(2)Immunization and Screening for infectious</td>
</tr>
<tr>
<td>Lowest</td>
<td>(105)Spondylosis; Intervertebral disc disorders; other back prob</td>
<td>(104)Other non-traumatic joint disorders</td>
</tr>
</tbody>
</table>

which is slightly more than the SVM algorithm (295.21 seconds) but achieve 15% better accuracy. (The experiment platform is Windows OS with 2.8 GHz CPU).

VI. DISCUSSION

In this section, we conclude the superiorities and limitations of CRLEDD. Further, we come up with the plans to improve CRLEDD in the future work.

1) Clinical Relevance and the Motivation: It is well known that the mental health diseases and disorders are hard to be successfully diagnosed in clinical practice at its earlier stage [56]. It can take months, and sometimes years, for clinicians to accurately diagnose a mental illness partially due to the fact that the sample size of the existing EHR data in terms of mental issues is limited compared to the dimensionality of each individual sample, which leads to huge errors in prediction using some existing LDA based classifiers. Motivated by the status quo, CRLEDD “Causality-Regularized Learning for Early Detection of Disease” is proposed in this paper to help clinical practitioners to implement effective initial screening to support and improve an early detection of mental diseases. It can help clinicians save diagnosis time and improve the efficiency. Specifically, the empirical EHR data can be used to train the proposed estimator and test its performance with the labeled diagnosis. Then, the clinicians can use the predictions as an auxiliary diagnosis so that further actions can be determined and adopted along with other clinical information to treat the patients at early stage effectively.

2) Comparing to Existing Algorithm: The aforementioned regularized method is a possible way to address the ill-posed inverse problem. For example, the Shrinkage LDA is a regularized LDA estimator which leverages sparsity of the shrinkage covariance matrix. In this paper, we proposed CRLEDD which employs graphical lasso to estimate the precision matrix to achieve both the positive semi-definiteness and the optimal estimation. In order to demonstrate the superiority of CRLEDD we compare it with the regularized LDA (Shrinkage LDA and DIAG) and other frequently-used algorithms (SVM, Decision Tree, Logistic Regression and AdaBoost) in the experiment. The results show the Accuracy and F1-score of CRLEDD is higher than the above baseline algorithms. Although CRLEDD sacrifice its Specificity comparing to the baseline algorithms, which can be regarded as a trade-offs between Sensitivity and Specificity, CRLEDD still overall outperforms other baseline algorithms especially when applying to “early detection of disease” background, where CRLEDD can try the best to find the patients with the target disease (high Sensitivity) and further filter out the misclassified patient through a traditional diagnosis.

Another possible way is to treat the problem as a sparse data recovering. Since the sparse data are more commonly encountered in industrial applications rather than complete data, the researchers have investigated and proposed some efficient algorithms [57]–[61] to solve the problem. We plan to initial some comparisons between the sparse data recovering techniques (e.g., matrix factorization [62]) and co-variance regularized operations in the future work to deep understand the correlation and the merits which can be leveraged in different scenarios.

3) Data and Matrices: In the experiment section, we train the estimator using the balanced sample data, where in the future, we can directly work on unbalanced datasets [63] which are more common in real-world datasets. Moreover, in our work, we use the ICD-9 raw data by filtering out short time observations, where CRLEDD has achieved a good performance in terms of accuracy and F1 scores. In order to improve CRLEDD from the aspect of data preparation, we consider the following measures in our future work: (1) for the data pre-processing, we can use the normalization techniques to avoid small snapshot of the patient visit (e.g., length of the observation) or side-effect of social factors (e.g., health care access); (2) for the data coding itself, we can consider combining different quality levels of source data (e.g., coarse level provided by CPT) to address more complicated diagnosis records.

4) Computing Method and Optimization Algorithm: Our proposed algorithm CRLEDD first collect all the relevant data from each medical institution, then train and test the target classifier
based on these aggregated data in a single institution. Due to the practical situations, such as the privacy concern when gathering all the data in one center/hub, the possibly large workload in single institution and the needs for real-time medical data updating, we might plan to improve the current CRLEDD with distributed computing patterns. Further, we can also apply the on-line optimization algorithm (e.g., on-line stochastic gradient descent) to obtain the optimal coefficient parameters of the estimator, which the real-time data updating can be achieved.

5) Compare to the previous version: This work is the extended version of previous work [36]. We mainly target on the following three new parts to supplement the previous work: (1) we present the detailed techniques of Graphical Lasso in Section IV-B to show how it benefit to the traditional LDA, where Graphical Lasso ensures positive semi-definiteness of the sparsified precision matrix; (2) To validate the key algorithm of CREDD, we present the causality graph visualization in the Experimental Results section. Moreover, we also conduct more experiments with the results based on a large range of parameters. Note that the complete experimental results are listed in the Appendix A; (3) We extend to discuss the superiorities and the limitation of the proposed CREDD in this section.

VII. CONCLUSION

In this paper, CRLEDD is designed to lower the expected error rate of LDA model for high-dimensional EHR data, through regularizing the precision matrix with Graphical Lasso. The experimental results using real-world EHR dataset show that CRLEDD has better performance compared with all baseline algorithms in terms of overall accuracy, and F1-score in all settings. Also, we compare the $\ell_1$-norm error of LDA and CRLEDD, the results show that CRLEDD always outperforms other methods with less error in different sample sizes. Furthermore, we visualize the causality between the disorders based on the precision matrix to validate the algorithm in CRLEDD.

ACKNOWLEDGMENT

The authors would like to thank Smart Start-up Funding from Missouri University of Science and Technology, and the anonymous reviewers for their valuable comments.

REFERENCES


Jian Bian received the B.Eng. degree in logistics systems engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014, and the M.Sc. degree in industrial systems engineering from the University of Florida at Gainesville, FL, USA, in 2016. He is currently working toward the Ph.D. degree with the Department of Computer and Electrical Engineering, University of Central Florida, Orlando, FL, USA, under Co-supervision of Dr. Zhihan Guo and Dr. Haoyi Xiong. From 2016 to 2018, he spent the first two years of his Ph.D. study with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA. His research interests include human-subject data learning, ubiquitous computing, intelligent cyber-physical systems.

Sijia Yang received the M.Sc. degree in information, communications and technology business management from Telecom École de Management, Paris, France, in 2015 and the B.Eng. degree in food science and engineering from Zhejiang Gongshang University, Zhejiang, China, in 2011. She is currently working toward the Ph.D. degree at the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA (formerly known as University of Missouri at Rolla). From 2015 to 2016, she was a Post-Doctoral Research Associate with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. He is currently with the Big Data Laboratory, Baidu Research, Beijing, China. His current research interests include cloud computing, artificial intelligence, and big data processing.

Haoyi Xiong received the Ph.D. degree in computer science from Telecom SudParis, University of Paris VI, Paris, France, in 2015. From 2016 to 2018, he was an Assistant Professor with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA (formerly known as University of Missouri at Rolla). From 2015 to 2016, he was a Post-Doctoral Research Associate with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. He is currently with the Big Data Laboratory, Baidu Research, Beijing, China. His current research interests include cloud computing, artificial intelligence, and big data processing.
Licheng Wang received the Ph.D. degree from Shanghai Jiao Tong University, in 2007. He is currently an Associate Professor at the Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include modern cryptography, network security, and trust management.

Yanjie Fu received the B.E. degree from the University of Science and Technology of China, in 2008, the M.E. degree from the Chinese Academy of Sciences, in 2011, and the Ph.D. degree from Rutgers University, in 2016. He is currently an Assistant Professor with the Missouri University of Science and Technology. His general interests are data mining and big data analytics. He has published prolifically in refereed journals and conference proceedings, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the ACM Transactions on Knowledge Discovery from Data, the IEEE TRANSACTIONS ON MOBILE COMPUTING, and ACM SIGKDD.

Zhishan Guo received the B.Eng. degree in computer science and technology from Tsinghua University, Beijing, China, in 2009, the M.Phil. degree in mechanical and automation engineering from The Chinese University of Hong Kong, Hong Kong, in 2011, and the Ph.D. degree in computer science from the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, in 2016. He is currently an Assistant Professor with the Department of Computer and Electrical Engineering, University of Central Florida, Orlando, FL, USA. From 2016 to 2018, he was an Assistant Professor with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA (formerly known as University of Missouri at Rolla). His current research interests include real-time and cyber-physical systems, neural networks, and computational intelligence.