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 theses 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

Manuscript received October 10, 2019; revised March 24, 2020 and May 15, 2020; accepted June 24, 2020. Date of publication August 10, 2020; date of current version July 22, 2021. This work was supported by the NSF: CRII (+REU): CSR: NeuroMC Parallel Online Scheduling of Mixed-Criticality Real-Time Systems via Neural Networks. *(Corresponding author: Jiang Bian.)*

Jiang Bian and Zhishan Guo are with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: bjbj11111@knights.ucf.edu; zsguo@ucf.edu).

Sijia Yang and Licheng Wang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: ysjhhh@gmail.com; wanglc2012@126.com).

Haoyi Xiong is with the Big Data Lab, Baidu Inc., Beijing 100049, China (e-mail: haoyi.xiong.fr@ieee.org).

Yanjie Fu is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: yanjie.fu@ucf.edu).

Zeyi Sun is with the Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: sunze@mst.edu).

Digital Object Identifier 10.1109/TETCI.2020.3010017

(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 Chervonekis 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 *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 ℓ_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], ℓ_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 ℓ_1-norm regularization (e.g., *Lasso*) is an option to handle the ill-posed problem such as ℓ_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 property 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 Causalities-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, *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 *CRLEDD* in early detection of disease is described in Section III. Section IV presents the details of the key algorithm used in *CRLEDD*. Then, we evaluate *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 *CRLEDD* in Section VI. Finally, in Section VII, we conclude that *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 ($\geq$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 $\{(x_i, y_i), i = 1, 2, \ldots, m\}$, $x_i \in R^d$. Let $X = [x_1^T, x_2^T, \ldots, x_m^T]^T$, $Y = [y_1, y_2, \ldots, y_m]$ and $\beta = [\beta_1, \beta_2, \ldots, \beta_d]^T$. Normally, the basic linear model is as follow,

$$Y = \beta^T X \tag{1}$$

where β 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 $Y = Sign(\beta^T X)$. In order to build the linear classification model, we need to estimate the β from the training samples.

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

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{m} \left(Sign(\beta^T x_i) - y_i \right)^2 + \lambda |\beta| \right) \tag{2}$$

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 ℓ_1-norm (also Lasso) regularization is adopted ($\lambda \neq 0$) to constrain (sparsifying) the β in the estimation [39]. For example, the linear SVM is a type of β-based linear classification model. Correspondingly, the ℓ_1-norm SVM can sparsify the parameter estimation by shrinking the small coefficients β of the hyperplane to exactly zero. Especially in High Dimension Low Sample Size (HDLSS) settings, the ℓ_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,

$$\beta = (X^T X)^{-1} X^T Y \tag{3}$$

where $X^T X$ can be considered the covariance matrix Σ. Instead of using β 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 Θ is invalid in HDLSS settings due to the loss of positive semi-definite property of Σ. Thus, ensuring the SPD of the covariance matrix Σ is critical for estimating Θ. Although a common regularized technique of shrinkage can guarantee the SPD of Σ [41], the results of the estimation are usually not optimal.

C. Problem Formulation

To estimate the precision matrix Θ, the maximum likelihood estimation problem is shown below,

$$\Theta^* = \underset{\Theta \in I^+_{p \times p}}{\operatorname{argmin}} \left(- \sum_{i=1}^{m} \log p(x_i | \mu, \Theta) + \lambda |\Theta|_1 \right), \tag{4}$$

where μ 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

$$-log(p(x_1|\mu, \Theta) \times p(x_2|\mu, \Theta) \times \cdots \times p(x_m|\mu, \Theta)). \tag{5}$$

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

$$\operatorname{tr}(\bar{\Sigma}\Theta) - \log \det(\Theta), \tag{6}$$

where $\operatorname{tr}()$ is the trace of square matrix and $\bar{\Sigma}$ is the sample covariance matrix. In the estimation of Θ, the ℓ_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 Σ^{-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, $\langle 1, 0, \ldots, 3 \rangle$, 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

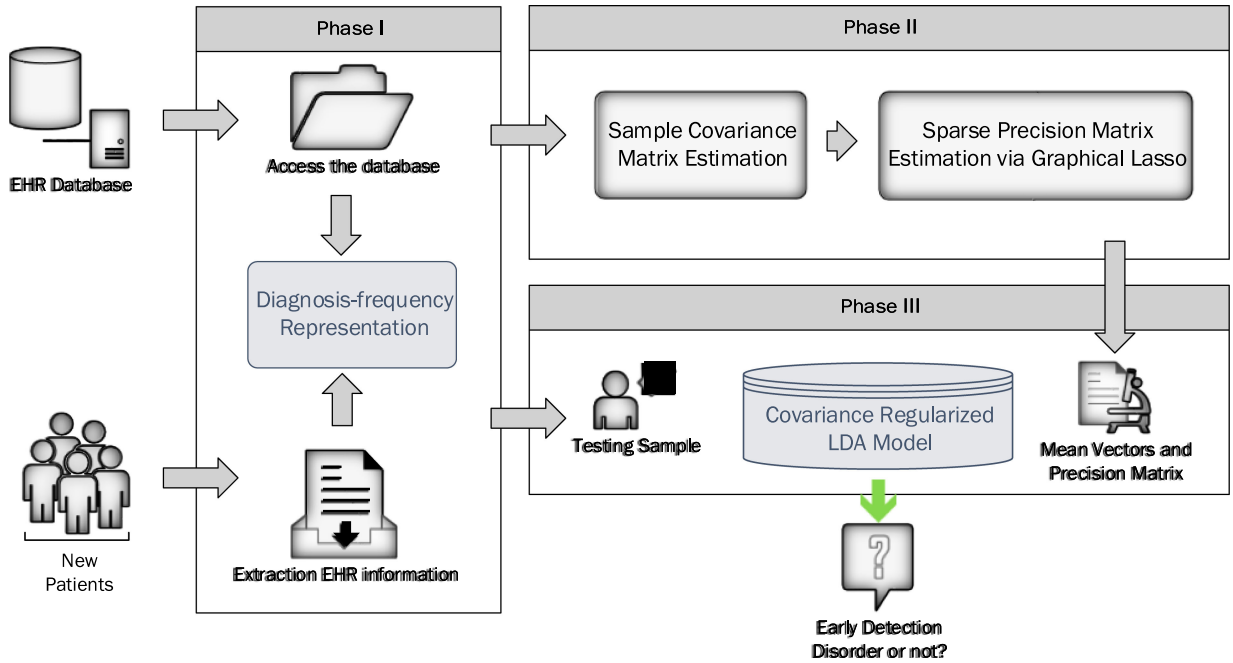


Fig. 1. Overview of the three-phase framework: *CRLEDD* – Regularized Causalities Learning for Early Detection of Diseases using Electronic Health Record (EHR) Data. Depending on the functionality, the framework are divided into three phases which are **Data Representation**, **Correlation Analysis**, **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:

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

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. α_{+1} and α_{-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., $(x_0, l_0) \ldots (x_{m-1}, l_{m-1})$ — $l_i \in \{-1, +1\}$, 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:

$$\bar{\Sigma} = \frac{1}{m} \sum_{l\in\{+1,-1\}} \sum_{x_i \in X_l} (x_i - \bar{\mu}_l)(x_i - \bar{\mu}_l)^T, \tag{8}$$

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 $\bar{\Sigma}$, this method estimates a sparse precision matrix $\hat{\Theta}$ using the Lasso-alike regularization estimator:

$$\hat{\Theta} = \underset{\Theta \in I^+_{p\times p}}{\operatorname{argmin}} \left(\operatorname{tr}(\bar{\Sigma}\Theta) - \log\det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right). \quad (9)$$

It is considered a ℓ_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 $\bar{\Sigma}$. In order to solve the optimization problem in Eq. (9) to obtain the $\hat{\Theta}$, the Graphical Lasso algorithm [42] is used to estimate $\hat{\Theta}^{-1}$ and recover $\hat{\Theta}$ after convergence. The details of this algorithm are listed as follow.

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

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix} \quad (10)$$

The solution for w_{12} satisfies

$$w_{12} = \arg\min_y \left\{ y^T \mathbf{W}_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \lambda \right\} \quad (11)$$

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

$$\min_\beta \left\{ \frac{1}{2} \left\| \mathbf{W}_{11}^{\frac{1}{2}} \beta - b \right\|^2 + \lambda \|\beta\|_1 \right\}, \quad (12)$$

where $b = \mathbf{W}_{11}^{\frac{1}{2}} s_{12}$. If β 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:

$$\begin{pmatrix} \mathbf{W}_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0^T & 1 \end{pmatrix}. \quad (13)$$

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

$$\mathbf{W} - \mathbf{S} - \lambda Sign(\Theta) = 0, \quad (14)$$

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

Algorithm 1: The ℓ_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 $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ **until** convergence:
(a) Partition the matrix $\mathbf{W}$ into two parts.
Part 1: all but the jth row and column.
Part 2: the jth row and column.
(b) Solve the estimating equation
$\mathbf{W}_{11}\beta - s_{12} + \lambda Sign(\beta) = 0$
using the cyclical coordinate-descent algorithm for the modified *Lasso*.
(c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$.
3, In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

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

$$w_{12} - s_{12} - \lambda Sign(\theta_{12}) = 0. \quad (15)$$

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

$$\mathbf{W}_{11}\beta - s_{12} + \lambda Sign(\beta) = 0, \quad (16)$$

where $w_{12} = -\mathbf{W}_{11}\theta_{12}/\theta_{22} = \mathbf{W}_{11}\beta$. The equivalence of the first two terms is obvious. For the sign terms, since $\mathbf{W}_{11}\theta_{12} + w_{12}\theta_{22} = 0$ from Eq. (14), we have that $\theta_{12} = -\theta_{22}\mathbf{W}_{11}^{-1}w_{12}$. Since $\theta_{22} > 0$, it follows that $Sign(\theta_{12}) = -Sign(\mathbf{W}_{11}^{-1}w_{12}) = -Sign(\beta)$. 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 pth 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 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

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

for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$, where S is the soft-threshold operator:

$$S(x, t) = sign(x)(|x| - t)_+. \quad (18)$$

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

$$\begin{aligned} \mathbf{W}_{11}\theta_{12} + w_{12}\theta_{22} &= 0 \\ w_{12}^T\theta_{12} + w_{22}\theta_{22} &= 1, \end{aligned} \tag{19}$$

from which we derive the standard partitioned inverse expressions

$$\begin{aligned} \theta_{12} &= -\mathbf{W}_{11}^{-1}w_{12}\theta_{22} \\ \theta_{22} &= 1/(w_{22} - w_{12}^T\mathbf{W}_{11}^{-1}w_{12}). \end{aligned} \tag{20}$$

According to Eq. (20), $\hat{\theta}_{22}$ and $\hat{\theta}_{12}$ can be easily computed in step 3. The Graphical Lasso algorithm stores all the coefficients β 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 β. 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 ensembling 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.

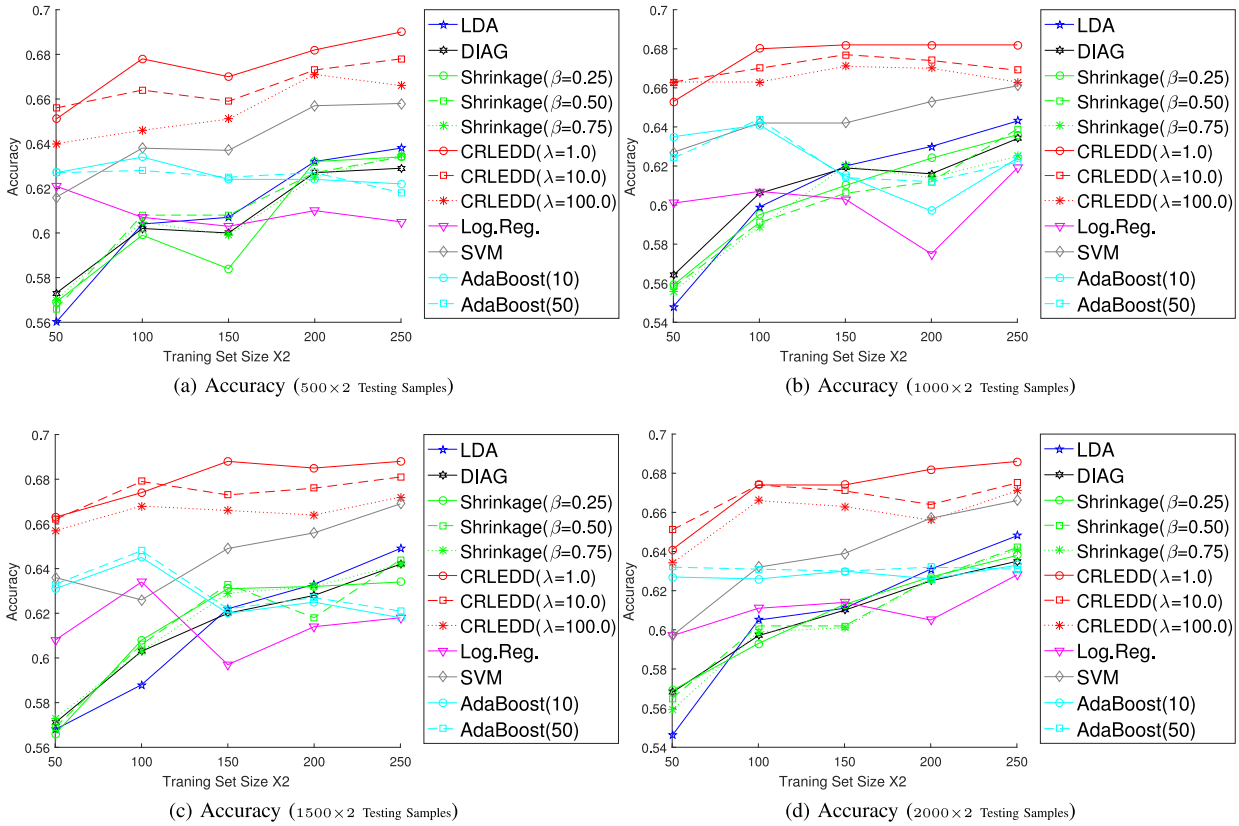
Then, we reveal the initial settings of some key parameters in proposed *CRLEDD* algorithm. The L1 regularization parameter λ is set to be 1, 10, 100 for comparison. The tolerance to declare convergence for graphical lasso is set to be 10^{-4}, and the number of maximum iteration for its optimization 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 ℓ_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



(a) Accuracy (500×2 Testing Samples) (b) Accuracy (1000×2 Testing Samples)

(c) Accuracy (1500×2 Testing Samples) (d) Accuracy (2000×2 Testing Samples)

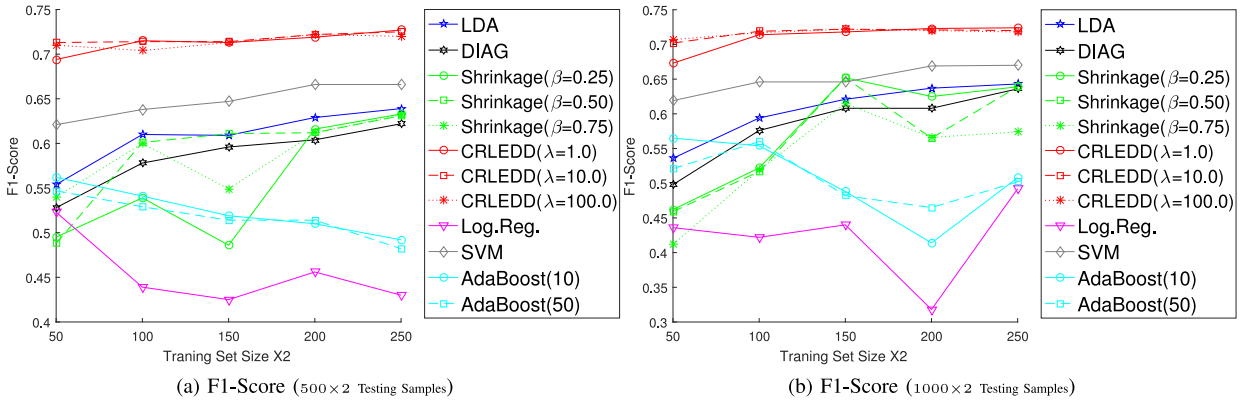
Fig. 2. Accuracy Performance Comparison between *CRLEDD* and Baselines with Small Training Datasets (Testing Sample Size $= 500 \times 2, 1000 \times 2, 1500 \times 2, 2000 \times 2$ from left top to right bottom, 90 days in advance).

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.

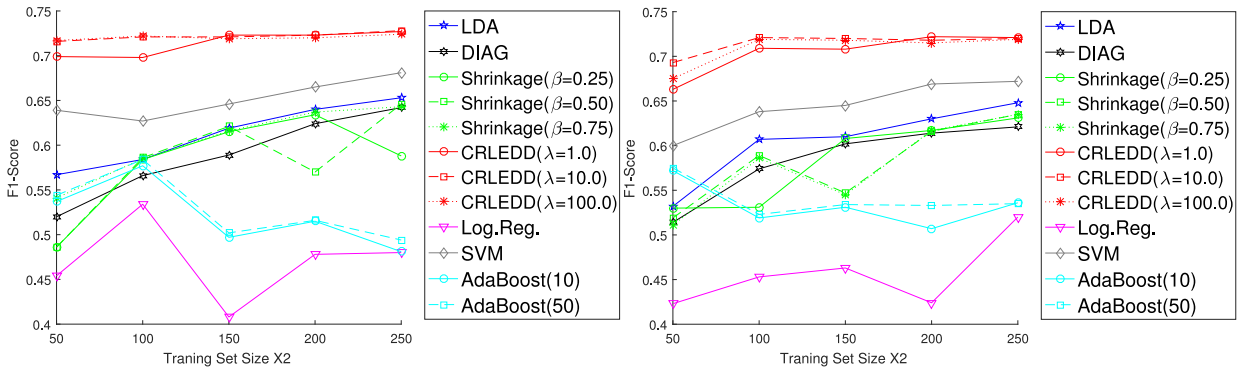
2) Sensitivity and Specificity Comparison: Table I additionally presents the performance with regards to sensitivity and specificity. The sensitivity is the percentage of patients who are correctly diagnosed as having the corresponding disease. As can be seen in the table, when training sample is 100 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 the *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 $\lambda = 1$, 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 ℓ_1-norm error of these two estimators



(a) F1-Score (500×2 Testing Samples) (b) F1-Score (1000×2 Testing Samples)



(c) F1-Score (1500×2 Testing Samples) (d) F1-Score (2000×2 Testing Samples)

Fig. 3. F1-Score Performance Comparison between *CRLEDD* and Baselines with Small Training Datasets (Testing Sample Size $= 500 \times 2, 1000 \times 2, 1500 \times 2, 2000 \times 2$ from left top to right bottom, 90 days in advance).

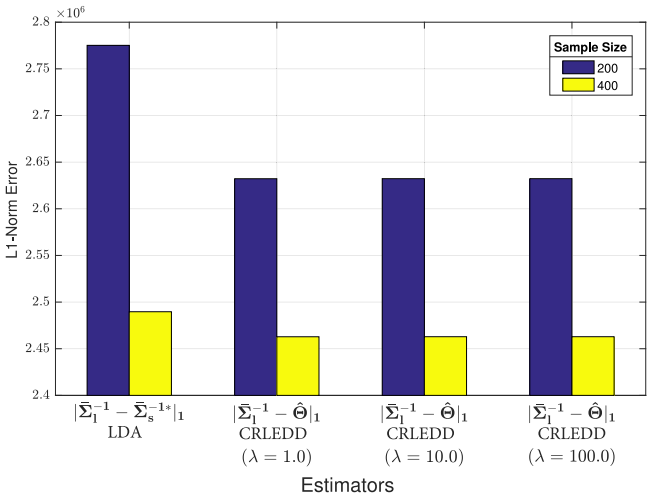


Fig. 4. ℓ_1-norm Error Comparisons of Estimators on Different Sample Sizes.

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 ℓ_1-norm. The results show that, compared to LDA (Σ_S^{-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

$$\Phi^{-1}\left(1 - \frac{\alpha}{p(p-1)}\right)\hat{\sigma}_{ij}/\sqrt{n} \tag{21}$$

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

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

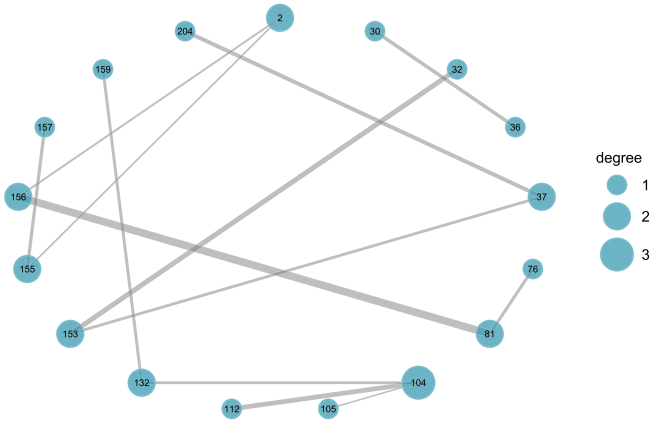


Fig. 5. Causality Graph.

risk group. For example, prior epidemiological studies suggest that upper respiratory infections affect mood and cognition, and psychological stress which is a significant risk factor for upper respiratory infections [54], [55]. Further clinical investigation is needed to fully understand these disorder pairs, but in general, these findings are informative for the early detection of anxiety/depression.

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 ℓ_1-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)

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

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 ℓ_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

[1] Z. Tirandaz, G. Akbarizadeh, and H. Kaabi, "Polsar image segmentation based on feature extraction and data compression using weighted neighborhood filter bank and hidden markov random field-expectation maximization," *Measurement*, vol. 153, 2020, Art. no. 107432.

[2] F. Samadi, G. Akbarizadeh, and H. Kaabi, "Change detection in SAR images using deep belief network: A new training approach based on morphological images," *IET Image Process.*, vol. 13, no. 12, pp. 2255–2264, 2019.

[3] C. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar Images*. SciTech Publishing, 2004.

[4] G. Akbarizadeh, "A new statistical-based kurtosis wavelet energy feature for texture recognition of SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4358–4368, Nov. 2012.

[5] Z. Tirandaz and G. Akbarizadeh, "A two-phase algorithm based on kurtosis curvelet energy and unsupervised spectral regression for segmentation of sar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 3, pp. 1244–1264, Mar. 2016.

[6] G. Akbarizadeh and Z. Tirandaz, "Segmentation parameter estimation algorithm based on curvelet transform coefficients energy for feature extraction and texture description of sar images," in *Proc. IEEE 7th Conf. Inf. Knowl. Technol.*, 2015, pp. 1–4.

[7] F. Sharifzadeh, G. Akbarizadeh, and Y. S. Kavian, "Ship classification in SAR images using a new hybrid CNN–MLP classifier," *J. Indian Soc. Remote Sens.*, vol. 47, no. 4, pp. 551–562, 2019.

[8] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[9] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)-A review of applications in the atmospheric sciences," *Atmospheric Environ.*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[10] F. Wang and J. Sun, "PSF: A unified patient similarity evaluation framework through metric learning with weak supervision," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 3, pp. 1053–1060, May 2015.

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

[12] K. Ng, J. Sun, J. Hu, and F. Wang, "Personalized predictive modeling and risk factor identification using patient similarity," in *Proc. AMIA Summit Transl. Sci.*, 2015, p. 132.

[13] J. Zhang, H. Xiong, Y. Huang, H. Wu, K. Leach, and L. E. Barnes, "MSEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data," in *Proc. IEEE Int. Conf. Big Data (Workshop)*, 2015.

[14] S. Jensen and U. SPSS, "Mining medical data for predictive and sequential patterns: PKDD 2001," in *Proc. 5th Eur. Conf. Princ. Pract. Knowl. Discovery Databases*, 2001.

[15] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal Phenotyping from longitudinal electronic health records: A graph based framework," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 705–714. [Online]. Available: http://doi.acm.org/10.1145/2783258.2783352

[16] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, PA, USA: SIAM, 2005.

[17] D. Panda, S. Singh, S. Mukherjee, and S. Chakraborty, "Comparing and analysis of different optimization techniques on sparse multi-class data," in *Proc. IEEE Int. Conf. Comput. Intell. Knowl. Economy*, 2019, pp. 528–531.

[18] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statistical Sci.*, vol. 1, pp. 502–518, 1986.

[19] Z. Qiao, L. Zhou, and J. Z. Huang, "Effective linear discriminant analysis for high dimensional, low sample size data," in *Proc. World Congr. Eng.*, 2008, vol. 2, pp. 2–4.

[20] F. Scholz, "Maximum likelihood estimation," *Encyclopedia of Statistical Sciences*, 1985.

[21] S. Puntanen and G. P. Styan, "The equality of the ordinary least squares estimator and the best linear unbiased estimator," *Amer. Statist.*, vol. 43, no. 3, pp. 153–161, 1989.

[22] H. W. Engl and C. W. Groetsch, *Inverse and Ill-Posed Problems*. vol. 4. New York, NY, USA: Elsevier, 2014.

[23] V. N. Vapnik and V. Vapnik, *Statistical Learning Theory*. vol. 1. Hoboken, NJ, USA: Wiley, 1998.

[24] V. Vapnik, "Princ. of risk minimization for learning theory," in *Proc. Neural Inf. Process. Syst.*, 1991, pp. 831–838.

[25] V. Vapnik, E. Levin, and Y. Le Cun, "Measuring the VC-dimension of a learning machine," *Neural Computation*, vol. 6, no. 5, pp. 851–876, 1994.

[26] K. van den Doel and U. M. Ascher, "On level set regularization for highly ill-posed distributed parameter estimation problems," *J. Comput. Phys.*, vol. 216, no. 2, pp. 707–723, 2006.

[27] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[29] X. L. Tan, "Optimal estimation of slope vector in high-dimensional linear transformation model," *J. Multivariate Anal.*, vol. 169, 2019, pp. 179–204.

[30] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. vol. 398. Hoboken, NJ, USA: Wiley, 2013.

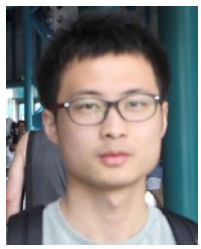
[31] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[32] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*, Berlin, Germany: Springer, 2013, pp. 237–280.

[33] T. T. Cai *et al.*, "Geometric inference for general high-dimensional linear inverse problems," *Ann. Statist.*, vol. 44, no. 4, pp. 1536–1563, 2016.

[34] S. S. Mahil, "On the application of Lagrange's method to the description of dynamic systems," *IEEE Trans. Syst., Man Cybern.*, vol. 12, no. 6, pp. 877–889, Nov. 1982.

[35] R. Peck and J. V. Ness, "The use of shrinkage estimators in linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-4, no. 5, pp. 530–537, Sep. 1982.
[36] J. Bian, L. E. Barnes, and G. Chen, "Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2017, pp. 457–460.
[37] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based N-gram models of natural language," *Comput. Linguis.*, vol. 18, no. 4, pp. 467–479, 1992.
[38] HCUP, "Appendix a - Clinical classification software-diagnoses (january 1980 through september 2014)," 2014. [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt
[39] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Roy. Statist. Soc.: Ser. B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.
[40] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Proc. Neural Inf. Process. Syst.*, 2003, vol. 15, pp. 49–56.
[41] O. Ledoit and M. Wolf, "Honey, I shrunk the sample covariance matrix," *J. Portfolio Manage.*, vol. 30, no. 4, pp. 110–119, 2004.
[42] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
[43] E. R. Dubberke, K. A. Reske, L. C. McDonald, and V. J. Fraser, "ICD-9 codes and surveillance for clostridium difficile–associated disease," *Emerg. Infectious Diseases*, vol. 12, no. 10, pp. 1576–1579, 2006.
[44] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.
[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[46] J. Friedman *et al.*, "Pathwise coordinate optimization," *Ann. Appl. Statist.*, vol. 1, no. 2, pp. 302–332, 2007.
[47] B. Bayar, N. Bouaynaya, and R. Shterenberg, "SMURC: High-dimension small-sample multivariate regression with covariance estimation," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 573–581, Mar. 2017.
[48] J. C. Turner and A. Keller, "College health surveillance network: Epidemiology and health care utilization of college students at U.S. 4-year universities," *J. Amer. College Health: J. ACH*, vol. 63, no. 8, pp. 530–538, Jun. 2015.
[49] K. S. Kendler, J. M. Hettema, F. Butera, C. O. Gardner, and C. A. Prescott, "Life event dimensions of loss, humiliation, entrapment, and danger in the prediction of onsets of major depression and generalized anxiety," *Archives General Psychiatry*, vol. 60, no. 8, pp. 789–796, 2003.
[50] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[51] S. H. Huang, P. LePendu, S. V. Iyer, M. Tai-Seale, D. Carrell, and N. H. Shah, "Toward personalizing treatment for depression: Predicting diagnosis and severity," *J. Amer. Med. Informat. Assoc.*, vol. 21, no. 6, pp. 1069–1075, 2014.
[52] K. Kroenke and R. L. Spitzer, "The PHQ-9: A new depression diagnostic and severity measure," *Psychiatric Ann.*, vol. 32, no. 9, pp. 1–7, 2002.
[53] J. Jankova *et al.*, "Confidence intervals for high-dimensional inverse covariance estimation," *Electron. J. Statist.*, vol. 9, no. 1, pp. 1205–1229, 2015.
[54] R. S. Bucks, Y. Gidron, P. Harris, J. Teeling, K. A. Wesnes, and V. H. Perry, "Selective effects of upper respiratory tract infection on cognition, mood and emotion processing: A prospective study," *Brain, Behav. Immunity*, vol. 22, no. 3, pp. 399–407, Mar. 2008.
[55] S. Cohen, "Psychological stress and susceptibility to upper respiratory infections," *Am. J. Respiratory Crit. Care Med.*, vol. 152, no. 4_pt_2, pp. S53–S58, Oct. 1995.
[56] B. G. Link, J. C. Phelan, M. Bresnahan, A. Stueve, and B. A. Pescosolido, "Public conceptions of mental illness: Labels, causes, dangerousness, and social distance," *Amer. J. Public Health*, vol. 89, no. 9, pp. 1328–1333, 1999.
[57] X. Luo, M. Zhou, S. Li, L. Hu, and M. Shang, "Non-negativity constrained missing data estimation for high-dimensional and sparse matrices from industrial applications," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 1844–1855, May 2020.
[58] T. D. Pham, K. Wardell, A. Eklund, and G. Salerud, "Classification of short time series in early parkinsons disease with deep learning of fuzzy recurrence plots," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 6, pp. 1306–1317, Nov. 2019.
[59] J. Jokinen, T. Raty, and T. Lintonen, "Clustering structure analysis in time-series data with density-based clusterability measure," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 6, pp. 1332–1343, Nov. 2019.
[60] B. Xia, W. Yuan, N. Xie, and C. Li, "A novel statistical manifold algorithm for position estimation," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 6, pp. 1513–1518, Nov. 2019.
[61] X. Luo, Z. Wang, and M. Shang, "An instance-frequency-weighted regularization scheme for non-negative latent factor analysis on high-dimensional and sparse data," *IEEE Trans. Syst., Man, Cybern.: Syst.*, to be published.
[62] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
[63] G. E. Batista, A. C. Carvalho, and M. C. Monard, "Applying one-sided selection to unbalanced datasets," in *Proc. Mex. Int. Conf. Artif. Intell.*, 2000, vol. 2000, pp. 315–325.

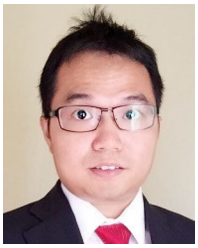


Jiang Bian received the B.Eng. degree in logistics systems engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014, and the M.Sc. degree in industrial systems engineering from the University of Florida at Gainesville, FL, USA, in 2016. He is currently working toward the Ph.D. degree with the Department of Computer and Electrical Engineering, University of Central Florida, Orlando, FL, USA, under Co-supervision of Dr. Zhishan 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 Ecole 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 in cyberspace security from Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include cyber security, data analytics, and machine learning.

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 automated deep learning (AutoDL), ubiquitous computing, artificial intelligence, and cloud computing. He has published more than 60 papers in top computer science conferences and journals, such as UbiComp, ICLR, RTSS, AAAI, IJCAI, ICDM, PerCom, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON COMPUTERS, *ACM Transactions on Intelligent Systems and Technology, ACM Transactions on Knowledge Discovery from Databases*, and etc. He gave keynote speak in a series of academic and industrial activities, such as the industrial session of the *19th IEEE International Conference on Data Mining (ICDM'19)*, and served as Poster Co-chair for the *2019 IEEE International Conference on Big Data (IEEE Big Data'19)*. Dr. Xiong was a recipient of the Best Paper Award from IEEE UIC 2012, the Outstanding Ph.D. Thesis Runner Up Award from CNRS SAMOVAR 2015. He was the co-recipient of Science & Technology Advancement Award from the Chinese Institute of Electronics 2019.

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.

Zeyi Sun received the B.Eng. degree in material science and engineering from Tongji University, Shanghai, China, in 2002, the M.Eng. degree in manufacturing from the University of Michigan Ann Arbor, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in industrial engineering and operations research from the University of Illinois at Chicago, Chicago, IL, USA, in 2015. Since 2015, he has been an Assistant Professor with the Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO, USA. His research interests include sustainable manufacturing from the perspective of system-level modeling and optimization, integration of aggregated electric vehicles in frequency regulation in the smart grid, and supply chain restructuring for the secondgeneration biofuel manufacturing.

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.