Identification of Autistic Risk Candidate Genes and Toxic Chemicals via Multilabel Learning

Zhi-An Huang, Jia Zhang, Zexuan Zhu, *Senior Member, IEEE*, Edmond Q. Wu, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*— As a group of complex neurodevelopmental disorders, autism spectrum disorder (ASD) has been reported to have a high overall prevalence, showing an unprecedented spurt since 2000. Due to the unclear pathomechanism of ASD, it is challenging to diagnose individuals with ASD merely based on clinical observations. Without additional support of biochemical markers, the difficulty of diagnosis could impact therapeutic decisions and, therefore, lead to delayed treatments. Recently, accumulating evidence have shown that both genetic abnormalities and chemical toxicants play important roles in the onset of ASD. In this work, a new multilabel classification (MLC) model is proposed to identify the autistic risk genes and toxic chemicals on a large-scale data set. We first construct the feature matrices and partially labeled networks for autistic risk genes and toxic chemicals from multiple heterogeneous biological databases. Based on both global and local measure metrics, the simulation experiments demonstrate that the proposed model achieves superior classification performance in comparison with the other state-of-the-art MLC methods. Through manual validation with existing studies, 60% and 50% out of the top-20 predicted risk genes are confirmed to have associations with ASD and autistic disorder, respectively. To the best of our knowledge, this is the first computational tool to identify ASD-related risk genes and toxic chemicals, which could lead to better therapeutic decisions of ASD.**

Manuscript received February 22, 2020; revised June 13, 2020; accepted August 8, 2020. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61876162, Grant 61871272, Grant 61671293, and Grant U1933125, in part by the Shenzhen Scientific Research and Development Funding Program under Grant JCYJ20180307123637294 and Grant JCYJ20190808173617147, in part by the Jiangxi Province Key Research and Development Program under Grant 20192BBE50065, in part by the Research Grants Council of the Hong Kong SAR under Grant CityU11202418 and Grant CityU11209219, and in part by the City University of Hong Kong Research Fund under Grant 9610397. *(Corresponding authors: Zexuan Zhu; Kay Chen Tan.)*

Zhi-An Huang and Kay Chen Tan are with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518060, China (e-mail: zahuang2-c@my.cityu.edu.hk; kaytan@cityu.edu.hk).

Jia Zhang is with the Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China (e-mail: j.zhang@stu.xmu.edu.cn).

Zexuan Zhu is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, also with the Shenzhen Pengcheng Laboratory, Shenzhen 518055, China, and also with the SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen University, Shenzhen 518060, China (e-mail: zhuzx@szu.edu.cn).

Edmond Q. Wu is with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Science and Technology on Avionics Integration Laboratory, China National Aeronautical Radio Electronics Research Institute, Shanghai 200000, China (e-mail: edmondqwu@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3016357

***Index Terms*— Autism spectrum disorders (ASDs), autism, autistic biomarkers, chemical toxicants, gene prioritization, gene-environment interaction, multilabel classification (MLC), multilabel learning (MLL).**

I. Introduction

AUTISM spectrum disorder (ASD) is referred to as a range of complex mental disorders (MDs). It is behaviorally characterized by the deficits in communications and socialization along with the presence of repetitive behaviors and/or restrictive interests. Based on the current diagnostic criteria, ASD includes distinct subtypes, such as autistic disorder (AD), Asperger syndrome (AS), and pervasive developmental disorder-not otherwise specified (PDD-NOS). The United States Centers for Disease Control (CDC) reported that 16.8 per 1000 (one in every 59) American children aged eight are diagnosed with ASD, representing an unprecedented spurt of 150% since 2000 [1]. ASD has emerged as an urgent public health concern. Individuals with ASD can benefit more from clinical diagnosis and behavioral interventions at the earlier stage. Unfortunately, the accurate diagnosis of ASD is challenging as it relies on the observation of cognitive phenotypes and behavioral manifestations. No biochemical markers, laboratory tests, or neuroimaging analysis can currently be used as a diagnostic gold standard for ASD [2]. Recently, accumulating evidence [3] has also implicated the involvement of genetic abnormalities and environmental toxicants in the development of ASD.

The key role that genetic factors play in ASD is early recognized and well-accepted. The modern concept of autism considered as a biomedical disorder was primarily coined by the reports that ASD is highly heritable and associated with several genetic syndromes, such as Fragile X and Rett syndrome [4]. Several studies in twins estimate the heritability of 70% of AD, and it could go up to 90% in some cases for the broader ASD phenotype [5]. The current investigations [6] have remarkably advanced our knowledge of the genetic basis and molecular pathophysiology of ASD. For example, macrocephaly (a head circumference >3 standard deviations above the mean) as an autistic characteristic has been repeatedly observed. Researchers found that at least 1% of individuals with ASD and macrocephaly carry the mutations of gene PTEN [7]. Genetic mutations causing ASD have been identified in 20%–25% of the cases using the current approaches. However, each known genetic mutation/variant can merely account for no more than 1%–2% of the probands.

Due to the extraordinary etiological heterogeneity of ASD, there is no single high-risk genetic biomarker identified for the development of ASD so far [4].

Over the past decades, research and clinical studies point to an equal contribution by environmental toxicants [8]. It is well-known that exposures to environmental toxicants (e.g., toluene, arsenic, mercury, lead, and polychlorinated biphenyls) can cause multiple MDs [9]. Nevertheless, the risk for ASD development resulting from environmental toxicants tends to be underestimated. There have been at least 85 000 chemicals manufactured in the United States, but we have little knowledge about their developmental toxicity, including many currently in common use [10]. A recent systematic review [11] has surveyed 190 articles published for examining the environmental toxicants in ASD since 1971, where 89% (170) supports positive associations with the onset of ASD. Chemical toxicants can damage cells by abnormally converging on similar biochemical pathways, which causes adverse effects on disrupting important cellular functions, such as depleting glutathione, increasing oxidative stress, and impairing cellular signaling [12]. Furthermore, individual variability in genetic susceptibility can affect the responses to chemical toxicants and, therefore, exacerbate the vulnerability to ASD. More than 100 "environmental response genes" [13] have been discovered to have complex interactions with certain chemical toxicants. During the critical periods of neurodevelopment, these interactions could elevate the risk of causing ASD in a synergistic or parallel manner [14]. These important observations provide a new window for investigations of the gene-environment interactions, rather than the fixed genetic defects [11].

The ultimate goal of autism studies on genetic and chemical etiology is to discover validated biomarkers, which can facilitate the early diagnosis, prediction of therapeutic response, tracking progression, risk assessment, and drug development. Early intervention can offer the best promise for the well-being of children with ASD by promoting better prognoses, normalizing the brain activity, and reducing secondary behavioral complications [15]. However, the current diagnosis of ASD is based on the clinical manifestations of behavior along with developmental history. Children under the age of two could be subject to some limitations of diagnostic confirmation. Identifying autistic biomarkers potentially enables the early diagnosis and effective intervention of ASD. Nevertheless, it is inefficient to identify ideal biomarkers in the biological "haystack" merely by the traditional experimental validation. The vast majority of experimental biomarkers tend to fail in clinical trials after considerable time, expense, and effort have already been invested.

With the increasing availability of high-throughput sequencing data, developing computational models is imperative and efficient to estimate the confidence of seminal autistic biomarker candidates for further *in vitro* studies. In this regard, here, we integrate multiple heterogeneous biological databases to identify the autistic risk genes and toxic chemicals on a large scale by using a multilabel convolutional neural network (MLCNN). First, through data collection and preprocessing, we construct the developmental brain gene expression profiles as gene expression features and the chemical integrated network as chemical interaction features. The known positive gene-disease associations and chemical-disease associations are integrated as multilabel information. Then, an efficient multilabel feature selection method with manifold regularization called MDFS [16] is employed in this work to select remarkable features with discriminative power across multiple MDs. To tackle the imbalance between labels and within labels, we apply a self-adaptive weighted focal loss function (SW-FL) to the MLCNN classifier. Finally, the network structure and training algorithm of MLCNN are automatically tuned by the Bayesian optimization (BO). The proposed model is validated to achieve reliable performance in the comparison with other state-of-the-art multilabel classification (MLC) methods in the simulation experiments. The main contributions of this article are summarized as follows.

1) We redefine the computer-aided autistic biomarker discovery as an MLC problem for the first time and design an effective MLC framework to identify the autistic risk genes and toxic chemicals.
2) We construct two new benchmark data sets for autistic risk genes and chemicals identification based on a dozen heterogeneous biological databases.
3) We develop a multilabel feature selection method with manifold regularization to inject local and global label correlations for selecting remarkable features.
4) We propose an SW-FL function and an automated hyperparameter tuning of BO to improve the training of the MLCNN model. The comprehensive experiments demonstrate that the SW-FL can significantly lessen the overall loss, and the automated hyperparameter tuning of BO can boost the model performance.

The rest of this article is organized as follows. Section II offers a brief introduction to the related work as well as the CNN model and explains the reasons for the usage of MLC mode in this work. Section III describes the model design, including MDFS, SW-FL, and the MLCNN model equipped with hyperparameter tuning of BO. The simulation experiments and analysis are discussed in Section IV. Section V concludes this article.

II. Background

This section provides the context information discussed throughout this article. First, we briefly review the related works regarding the computer-aided autistic biomarker discovery. Then, we further pinpoint the defects of binary classification adopted in the previous studies and explain why MLC could be more suitable. Finally, the basic principles and hyperparameters of CNN are summarized.

A. Related Work

With the advances of next-generation sequencing and omics-based trials, huge volumes of medical data have accumulated for the systematic investigation of important biological activities. It conduces to the burgeoning of computational approaches for the effective system-level biomarker

screening [17]. In contrast to routine wet-lab experiments, computer-aided autistic biomarker approaches take advantage of diverse data resources and biological knowledge to decipher ASD pathogenesis under a holistic framework. These approaches can be roughly categorized into two classes: statistics-based models and machine learning (ML)-based models.

Statistics-based models are mostly based on the expression difference of molecules between different biological states in considering the mechanism of the significant alteration perturbing system's stability [18]. Statistical tests, such as p-value and the Wilcoxon signed-rank test, are extensively used to guide the candidate biomarkers detection. Shen *et al*. [19] defined a novel proxy measure called autism gene percentage (AGP) representing the percentage of autism-associated genes targeted by a single miRNA to investigate the specificity of these genes. Based on the constructed autism-specific miRNA-mRNA network, 11 candidate autism miRNA biomarkers were screened to have significantly higher AGP values ($p-\text{value} < 0.05$ and the Wilcoxon signed-rank test), and eight out of them (72.7%) were experimentally confirmed to dysfunction in ASD samples via literature validation. Statistics-based models can achieve good performance for generality and applicability due to their statistical evidence as principles. However, such prior knowledge-guided models have to collect, conclude, and distill the general evidence from the accumulation of reports on biomarker discovery.

Unlike statistics-based models, ML-based models can guide the computer simulating human learning behaviors to automatically apply arbitrary complex mappings from inputs to outputs. Self-evaluation can be invoked from experience without being explicitly programed. For example, Oh *et al*. [20] conducted a hierarchical cluster analysis to identify the suitable gene expression signatures for classifying subjects with ASD based on the blood gene expression profiles. Cogill and Wang [21] developed a support vector machine (SVM)-based binary classification model to prioritize ASD candidate genes and lncRNAs (long noncoding RNAs) using the brain developmental gene expression data. Following [21], Gök [22] used feature extraction with the Haar wavelet transform, data preprocessing with discretization methods, and binary classification with the Bayes network learning algorithm for prioritization of ASD candidate genes and lncRNAs as well. In spite of the high accuracy achieved by these two models, the assumed negative samples that they inferred may bring a high rate of false-positive results and, therefore, mislead the performance evaluation.

B. Binary Classification Versus Multilabel Classification

To the best of our knowledge, binary classification is the most widely used ML-based models for computer-aided autistic biomarker discovery. Intuitively, using binary classification allows us to obtain a well-trained classifier that can discriminate between ASD and non-ASD risk biomarkers, as long as sufficient positive and negative instances are available. However, both positive and negative autistic risk instances are hard to collect in practice, especially the negative ones. In the previous works, the common solution was to "generate" negative instances based on certain considerations or assumptions that tend to result in a high rate of false positive (FP). For example, Kou *et al*. [23] selected 200 genes to construct negative gene sets based on various criteria. The works [21] and [22] harvested non-ASD genes based on the assumption that a target gene unrelated to ASD or intellectual disability (ID) but associated with other human diseases could be considered as a non-ASD gene. Injecting those assumed negative data in the training process can have inevitably adverse effects on the classification performance by amplifying the systematic bias. It is worse that the binary classification only captures the dependence relation between the ASD-associated and non-ASD-associated instances, ignoring the important relevance between ASD and other related MDs.

In view of the abovementioned issues, MLC surfaces as a suitable alternative. Since the absent labels cannot be necessarily negative, the autistic biomarker discovery problem can be addressed by a positive and unlabeled (PU) classification mode [24]. The interdependence among ASD and its subtypes (e.g., AD and AS) can be included to collaboratively learn the inherent and implied information between multi labels. If an MLC model can achieve reliable results in discriminating ASD and a series of ASD-related MDs, it is more reasonable to demonstrate its capability of identifying seminal autistic risk biomarkers [25], [26]. Therefore, we propose a CNN-based MLC framework to prioritize the risk genes and chemical candidates for ASD and its subtypes.

C. Convolutional Neural Network Model

Inspired by the visual cortex of animals, CNN as one of the most effective deep neural networks has been successfully applied in many domains [27], [28]. The convolutional layer, as the central component of CNN, employs a filter (kernel) to perform convolutional operations on input with multiplication or dot product. One filter, normally viewed as a 2-D matrix smaller than an input, can create a feature map with neuron units to summarize the presence of detected features from the input via horizontal slides and vertical slides in turns until the whole feature matrix is scanned. In this way, filters can do the elementwise multiplication between the filter-sized patch of an input (also known as the concept of "receptive fields" in CNN) and itself for aggregating the output into a single value. Neurons that lie in the same feature map are used across all receptive fields to share their weights so-called parameter sharing and, thereby, can reduce the memory footprint. CNN is capable of automatically learning the stacking filters in parallel along with separate activation functions to explore numerous types of feature representations systematically. A max- or average- pooling layer is commonly used to reduce the spatial size of the representation by shrinking the neuron clusters in feature maps into a single neuron of the next layer. Finally, fully connected (FC) layers are usually combined into a tail of the CNN model to give the final probabilities for each label by flattening the output of previous layers.

It takes more sophisticated skills to tune the hyperparameters and train the model of CNN than that of the traditional

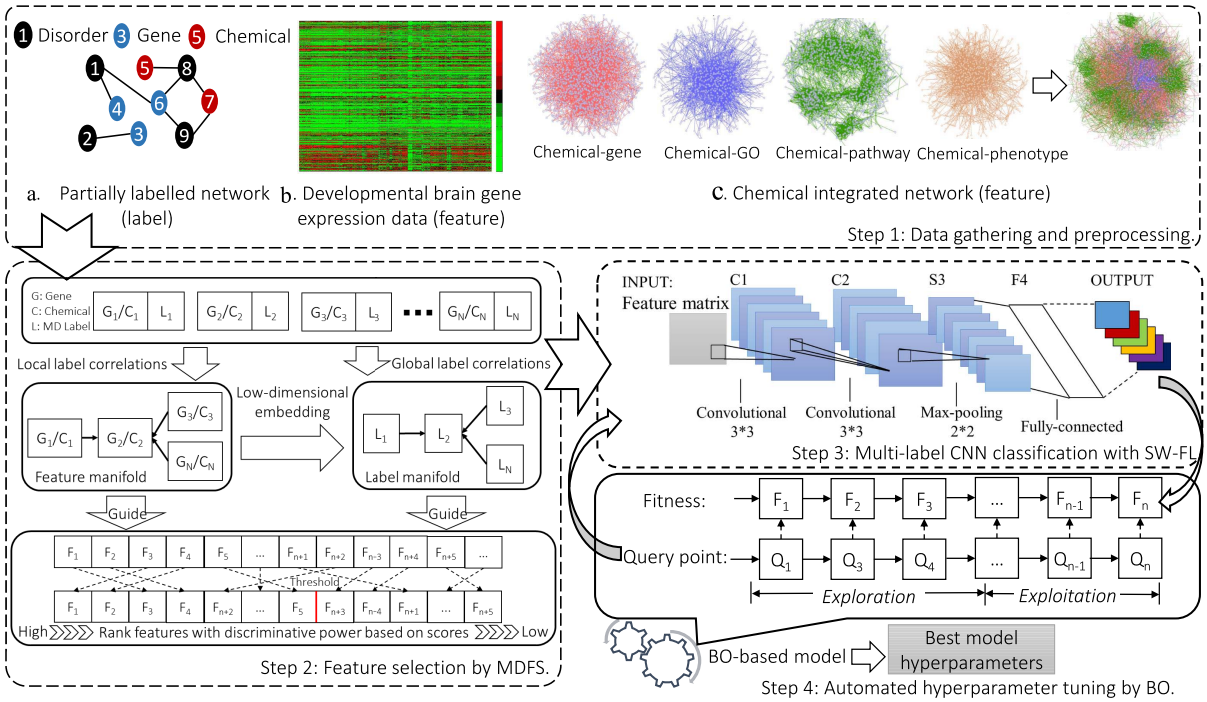


Fig. 1. Illustration of the proposed model. (a) Partially labeled network (label). (b) Developmental brain gene expression data (feature). (c) Chemical integrated network (feature).

ML algorithms. Selecting appropriate hyperparameter sets (or combinations) can have a significant effect on the effectiveness of a CNN model. As manual tuning is rather tedious and impractical in CNN, automated tuning appears to be more effective and feasible. In recent decades, plenty of methods ranging from simplistic procedures, such as random or grid search [29], to more sophisticated surrogate models, such as random forests [30] or Gaussian processes (GPs) [31], have been proposed to automatically optimize the hyperparameter settings. However, as the number of hyperparameter increases, the complexity of the optimization procedure grows exponentially as well [32]. It is challenging to efficiently handle such objectives whose optimization requires massive evaluations. Especially, for the hyperparameter tuning in neural networks, as one kind of typical expensive problems, this could be an issue to rapidly find some competitive models. BO is an efficient framework for well resolving such noisy, expensive black-box problems by globally querying a distribution over functions. Through the construction of a relatively cheap surrogate model, it offers a principled way to accurately model the prediction uncertainty by naturally balanced combinations of exploration and exploitation. The most extensively used model for BO would be the GPs-based model [32] due to its flexibility and simplicity in terms of conditioning and inference. Therefore, an automated hyperparameter tuning technique using GPs-based BO is incorporated into the proposed MLCNN model to find the optimal hyperparameters with no manual effort necessary beyond the initial setup.

III. Model Design

This section describes the details of the proposed methods and procedures used in this work. First, the workflow of the proposed model is illustrated in Fig. 1 where we integrate label information from the partially labeled network [see Fig. 1(a)] and construct instance features from the developmental brain gene expression data [see Fig. 1(b)] and the chemical integrated network [see Fig. 1(c)] in the model. MDFS is utilized to select an optimal subset of significant features discriminating across multiple MDs using multilabel learning (MLL). Then, we extend the focal loss function to the multilabel variant for better measuring the errors in training. Finally, the MLCNN model with BO-based automated hyperparameter tuning is used for the MLC of autistic risk gene and chemical candidates.

A. Problem Formulation

In this section, we briefly describe the setting of the MLC problem. Formally, let a multilabel data set of n instances be denoted as $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, X being the feature matrix, where x_i is the ith instance represented by a d-dimensional feature vector, and Y being the MD label matrix, where y_i having a finite set of q possible disease labels represents the ground-truth label vector of x_i. The element $x_{ij} \in X_i$ (where $X \in \mathbb{R}^{n\times d}, 1 \le j \le d$) denotes the feature value of the ith instance in terms of the jth feature property. The value $y_{ij} \in Y_i$ (where $Y \in \mathbb{R}^{n\times q}, 1 \le j \le q$) represents the status of the ith instance in terms of the jth MD label with $y_{ij} = 1$, indicating that the ith instance is "labeled" with the jth MD, and $y_{ij} = 0$ otherwise.

Given an instance x_i in X, a multilabel classifier can estimate the probability score for each label in Y: $f(x_i) = \{f_1(x_i), f_2(x_i), \ldots, f_q(x_i)\}$ and $f(x_i) \in [0, 1]$. Based on the fundamental concept of empirical risk minimization (ERM) in ML, the proposed multilabel model aims to minimize the

TABLE I
SUMMARY OF THE SYMBOLS USED

Parameter	Symbol	Parameter	Symbol
Number of instances	n	Feature dimension	d
Number of disease labels	q	Feature matrix	X
MD label matrix	Y	Optimization terms of MDFS	V,C,Ω
Low-dimensional embedding	F	Mapping matrix	W
Tradeoff parameters for C, Ω	λ,γ,α,β	Bias term for V	b
Similarity matrix among MD labels	S_0	Diagonal matrix of S_0	D_0
Graph Laplacian ($D_0 - S_0$)	L_0	Estimated probability	p_t
Focusing parameter of FL	ω	Frequency weight factor of FL	μ
Weight augmentation of SW-FL	η	Smoothing factor of SW-FL	ρ
Objective function of BO	f	Acquisition function of BO	h

expectations $\mathbb{E}$ of the loss function $\mathcal{L}$ over the space of (X, Y) following the calculation as follows:

$$\mathcal{L}(f) = \sum_{i}^{n}\sum_{j}^{q}\mathbb{E}_{x_i y_j}[\mathcal{L}(f_j(x_i), y_j)] \tag{1}$$

$$f^* = \text{argmin}\mathcal{L}(f). \tag{2}$$

To ease the understanding of the model description, the used symbols are tabulated in Table I.

B. Feature Selection With Manifold Regularization

The high-dimensionality curse of biological data is the well-known barrier for biomarker discovery. Feature selection is commonly used to find a near-optimal subset of discriminative features by eliminating irrelevant and redundant features [33]. In addition, the MD labels are not independent but inherently correlated. Based on the assumption [34] that functionally related instances tend to be associated with the pathologically similar MDs and vice versa, the appropriate use of the implicit relevance between instances and disease labels is critical to the effectiveness of feature selection for MLC. Inspired by our previous work [16], we develop a multi-label feature selection method MDFS to select discriminating features across multiple MDs. The framework of MDFS is shown in Fig. 1. First, the original feature space of instances is mapped into a low-dimensional embedding based on a geometric learning framework of manifold regularization [35]. With the help of the low-dimensional embedding, we then exploit the local and global label correlations to guide the feature selection process. Aimed at revealing the features that are discriminative for all MD labels, we design an effective optimization framework to take advantage of both local and global label correlations. Let F be the low-dimensional embedding derived from the original feature matrix X, and the optimization framework can be defined as

$$\min_{W,F} V(X, F, W) + \lambda C(F, Y) + \gamma\,\Omega(W) \tag{3}$$

where $W \in \mathbb{R}^{d\times q}$ is the mapping matrix; V, C, and Ω are the optimization terms; and λ and γ are the tradeoff parameters. Ω regulates the complexity of the model with $l_{2,1}$-norm [36] $\Omega(W) = \|W\|_{2,1}$, leading to sparse the optimal solution. The terms V and C are described as follows.

The term V represents the mapping function taking the local label correlations into consideration. The low-dimensional embedding F has the property that helps to recover most of the structures in the feature space X [37], which means that F has similar local structures with X. In other words, the similarity among instances can be decoded and reconstructed by F. Since the mapping F does not help to project test data, we utilize the mapping matrix W to relate F with X as a medium. For the effective exploitation of local label correlations, we can utilize X, F, and W to formulate the first term V as

$$V(X, F, W) = \mathbf{Tr}(F^T L F) + \|XW + 1_n b^T - F\|_F^2 \tag{4}$$

where $1_n \in \mathbb{R}^n$ is a column vector with each element being 1, and $b \in \mathbb{R}^q$ is a bias term. L is the graph Laplacian operator formulated as $L = D - S$, where D is the diagonal matrix of S, i.e., $D_{ii} = \sum_{j=1}^{n} S_{ij}$, $S \in \mathbb{R}^{n\times n}$ denotes the similarity matrix among instances, and each element in S is calculated by a heat kernel [38].

The term C incorporates the low-dimensional embedding F and the ground-truth disease label matrix Y to enforce the MLL with global label correlations. Based on the assumption that if two MD labels have a stronger positive correlation, the prediction on them should be more similar, we further leverage the low-dimensional embedding F to form the label information-based manifold regularizer, thus facilitating the exploitation of global label correlations. Then, a feasible second-order strategy is introduced to infuse the global label correlations. The term C is defined as follows:

$$C(F, Y) = \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{q}[S_0]_{ij}\|[F^T]_i - [F^T]_j\|^2 \tag{5}$$

where the similarity matrix among MD labels $S_0 \in \mathbb{R}^{q\times q}$ is also calculated by a heat kernel, and $[F^T]_i \in \mathbb{R}^n$ represents the predicted result on the ith disease label. With some algebraic steps, (5) can be rewritten as $\mathbf{Tr}(FL_0F^T)$, where $L_0 = D_0 - S_0$, and D_0 is a diagonal matrix of S_0, i.e., $[D_0]ii = \sum_{j=1}^{q}[S_0]_{ij}$. A higher value of $[S_0]_{ij}$ indicates more similar prediction results of the two MD labels $[F^T]_i$ and $[F^T]_j$.

Based on the foregoing, we can clearly elaborate on the rules of the objective function in (3). Given the known ground-truth labels, we design a reliable method to guide the optimization process in a supervised manner. For the sake of the consistency between F and the real label distribution, the objective function of MDFS can be iteratively optimized and rewritten as follows:

$$\min_{F,W,b} \mathbf{Tr}(F^T L F) + \|XW + 1_n b^T - F\|_F^2 + \alpha\|F - Y\|_F^2 + \beta\mathbf{Tr}(FL_0F^T) + \gamma\,\|W\|_{2,1} \tag{6}$$

where F, W, and b are the variables to be learned. α, β, and γ are used to trade off the three terms, respectively. For solving the optimization problem, a common way is to perform an alternating minimization strategy to derive the optimization solution, which is detailedly described in [16]. As such, the optimization solution can be iteratively executed to update F, W, and b until convergence. The value of each feature in $\|W_i\|_2$ $(1 \le i \le d)$ represents the importance of the features. Therefore, we can find the subset of remarkable features based on the top rank of W.

C. Self-Adaptive Weighted Focal Loss Function

The training process of a CNN model requires a decent loss function to evaluate the resulting error of a run. In MLC, a commonly used loss function is binary cross-entropy (CE) loss (also known as log loss). Given the estimated probability p for the positive label of a class, it can be simply calculated by $CE(p_t) = -\log(p_t)$, where $p_t = p$ if the ground-truth class is positive, and $p_t = 1 - p$ otherwise. In fact, computer-aided biomarker identification approaches usually encounter severe class imbalance issues. In this work, such imbalance can be viewed from two perspectives: imbalance within labels and imbalance between labels. In the former case, a large proportion of instances have not been experimentally confirmed to get involved in the development of an MD. The easily classified unlabeled instances comprise the majority of loss leading to degenerate models via dominant gradients. In the latter case, one MD's positive instances may appear ten times more frequently than the other MDs' and then dominate the total loss causing instability in training. As such, severe class imbalance can overwhelm the CE loss without useful learning signals because of the inefficient training.

Lin *et al*. [39] extended CE and proposed a novel loss function called focal loss FL. The original purpose of FL is to address class imbalance for dense object detection, which is a typical binary classification problem. FL enhances the training in terms of accuracy and running time by putting more effort into correcting the hard, misclassified samples in two steps. First, a modulating factor $(1 - p_t)^{\omega}$ is added to the CE loss as follows:

$$FL(p_t) = -(1 - p_t)^{\omega} \log(p_t) \tag{7}$$

where the decay factor $\omega \in [0, 5]$ is a tunable focusing parameter. Intuitively, $(1 - p_t)^{\omega}$ smoothly downweights the loss of the well-classified samples and, in turn, calls for more attention to those misclassified samples. Then, to address the imbalance within labels, the obvious idea is to invert class frequency for both positive classes and negative classes. A weighting factor $\mu \in [0, 1]$ is introduced to define a μ-balanced variant of FL as

$$FL(p_t) = \begin{cases} -\mu(1 - p_t)^{\omega} \log(p_t) & \textbf{if } y = 1 \\ -(1 - \mu)(1 - p_t)^{\omega} \log(p_t) & \textbf{if } y = 0. \end{cases} \tag{8}$$

Extending FL to the multilabel case is straightforward and could perform well, which inspires us to develop a self-adaptive weighted variant called SW-FL with two improvements: weight augmentation and self-adaptive control. With weight augmentation, we aim to complete two tasks: 1) to tackle the imbalance between labels, the less frequent class should obtain higher weight loss and 2) as an autistic biomarker discovery model, ASD and its subtypes should dominate the total loss with the help of learning other MDs as auxiliary tasks to boost the network performance. Therefore, the learning weights of ASD and its subtypes deserve top priorities over other MDs. However, how to smoothly adjust the learning weights is key to balance the loss distribution between labels. Here, we design a weight augmentation vector $\eta \in \mathbb{R}^q$ to address this issue based on class frequency. Suppose that T_{tr}^i is the count of the ith disease's positive labels in training sets, and the expected η_i can be smoothly calculated for the ith disease as

$$\eta_i = \log(\rho T_{tr} / T_{tr}^i) \text{ s.t. } 1 \leq \eta_{\mathrm{i}} \leq 10 \tag{9}$$

where $T_{tr} = \sum_{i=1}^{q} T_{tr}^i$ and $\rho \in [0, 1]$ is a smoothing factor. One notable property of η is that it draws more attention to correct those misclassified samples having insufficient positive labels. The lower bound of η_i is set to 1, which rounds up to the reasonable range. The highest priority weights (i.e., 10) are assigned to the η of ASD and its subtypes accordingly. In this way, the calculated η represents loss weights for all involved MDs.

Note that the weighting factor μ in (8) is originally set as a decimal for binary classification. In MLC, as the significantly unequal distribution covers dozens of labels, there is no one-size-fits-all value of μ to be fixed for alleviating the imbalance within labels. Here, self-adaptive control is introduced to automatically invert frequencies for positive and negative classes. μ is redefined as a vector of size q to separately calculate the frequency for each MD label. Given the training MD label matrix denoted as $Y_{tr} \in \mathbb{R}^{n_{tr} \times q}$, we use λ_t to fuse η and μ as follows:

$$\lambda_t = \begin{cases} Y_{tr}\eta(1_q - \mu) & \textbf{if } y = 1 \\ (1_{n_{tr} \times q} - Y_{tr})\eta\mu & \textbf{if } y = 0. \end{cases} \tag{10}$$

Finally, λ_t is added to (7) to reach the final SW-FL as follows:

$$SW\text{-}FL(p_t) = -\lambda_t(1 - p_t)^{\omega} \log(p_t). \tag{11}$$

D. MLCNN Model With Automated Hyperparameter Tuning

Based on the proposed SW-FL loss function, the proposed MLCNN model can precisely measure the training error over the iterative optimization process. As shown in Fig. 2, we present a multilabel CNN model with automated hyperparameter tuning to perform MLC for high-risk autistic genes and chemicals. The remarkable features of the instances are ranked by MDFS and then fed to the convolutional operators on CNN. The proposed MLCNN model is a relatively small and more compact variant of the well-known VGGNet network [40]. The representative VGGNet-like architectures that we use are characterized by: 1) only using 3×3 convolutional layers throughout the whole net, namely, a stack of two 3×3 convolutional layers without spatial pooling in between; 2) reducing the volume size via a 2×2 max-pooling layer; and 3) placing the FC layer at the end of the network prior to the output activation function. The primary difference lies in the output activation function where we replace softmax with sigmoid, which empirically works better in MLC. We use the ReLU activation function [41] followed by batch normalization for all hidden layers. To avoid overfitting, dropout is adopted in our network architecture by randomly disconnecting nodes between layers.

The performance of the MLCNN model is sensitive to the setting of their hyperparameters. The automated hyperparameter tuning technique enables to yield settings competitive with those found by human experts. BO, a model-based

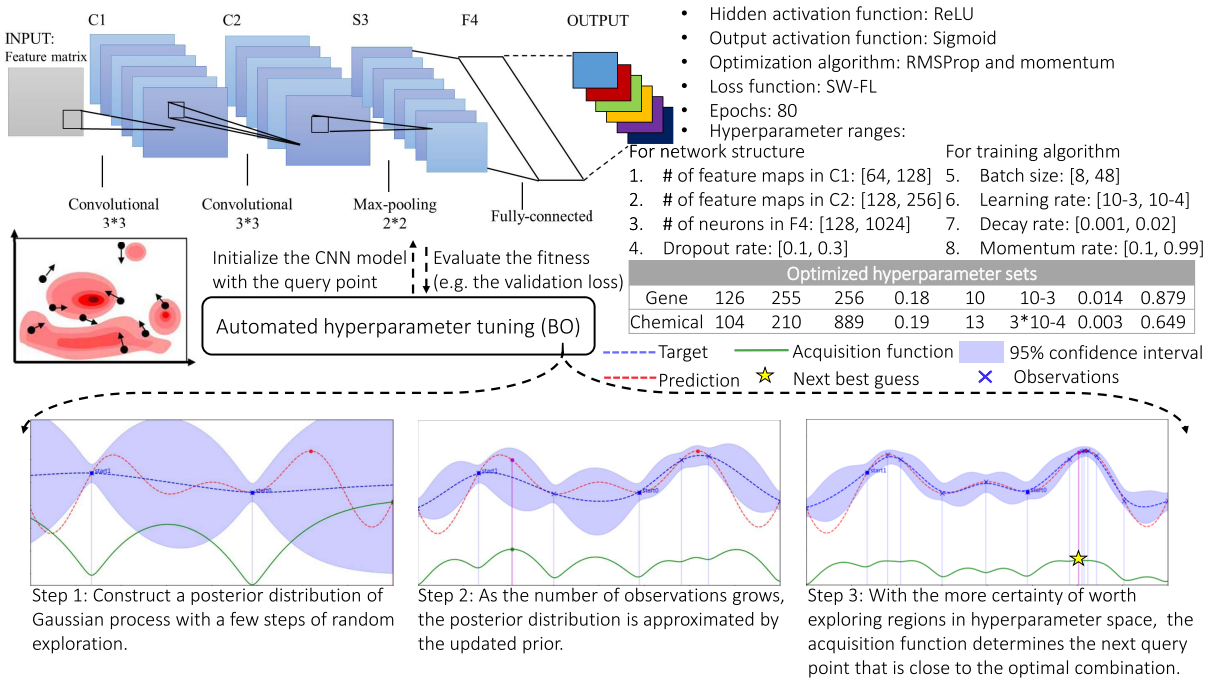


Fig. 2. MLCNN model with automated hyperparameter tuning of BO.

algorithm for solving black-box optimization problems, has been demonstrated to achieve better performance with fewer iterations than random search [42]. In this work, BO with GPs is employed to automatically optimize the hyperparameter settings for network structure and training algorithm, including four continuous values and four discrete values (see Fig. 2). The aim is to globally optimize the unknown objective function f in a minimum number of steps by building a surrogate function (probability model) based on the previous observations. We cannot take derivatives or find an analytical expression for f, and the evaluation of f is restricted to proposing sampling points in potential hyperparameter space (done by the acquisition function denoted as h), thus getting possibly noisy responses. Mathematically, f is sampled at $x_t = \arg\max_x h(x|\mathcal{O}_{1:t-1})$, where $\mathcal{O}_{1:t-1} = \{(x_i, y_i)|1 \leq i \leq t-1\}$ are the previous $t-1$ samples drawn from f. As a surrogate model, the GPs-based BO is used to define a prior over f for proposing a next sampling point. The procedure of BO is repeated until it reaches the maximum number of iterations. For example, x_t is the next sampling point predicted by h over the current GP posterior. Due to the prediction uncertainty, we obtain a possibly noisy sample result $y_t = f(x_t) + \epsilon_t$ from f. The latest sample (x_t, y_t) is added to previous samples $\mathcal{O}_{1:t} = \{\mathcal{O}_{1:t-1}, (x_t, y_t)\}$ to update the GP posterior for better approximating f, i.e., reducing ϵ.

Evaluating f is expensive because it requires to complete the whole training of a new MLCNN model. Thus, we trade off the exploration and exploitation to find the global optimum in limited steps. Exploration allows the random search to diversify the hyperparameter space while exploitation concentrates on promising areas where an improvement over the current best observation is estimated. Given the predefined ranges that the hyperparameter values can be searched over, BO can be used to iteratively tune the hyperparameters of MLCNN models. The resulted hyperparameter sets optimized by BO are shown in Fig. 2 for the readers' reference. The pseudocode of the whole proposed model is shown in Algorithm 1.

Algorithm 1 Pseudocode of the Proposed Model

Input: Feature matrix X, label matrix Y, model boundary ω.
Output: Predicted probabilities of test set Y_{te}^{pred}.

$W \leftarrow MDFS(X, Y)$
$X' \leftarrow Feature\_selection(X, W)$
$[X_{tr,v,te}, Y_{tr,v,te}] \leftarrow Split(X', Y)$ //with size of 6:2:2
Initialization: BO(ω)$\leftarrow \varnothing$, $Lowest\_test\_loss \leftarrow +\infty$
for $i = 1, \ldots, 30$ **do**//1-20:exploration, 21-30:exploitation
  $\Theta \leftarrow$ BO.predict(ω)
  Initialize MLCNN model: MLCNN(Θ)$\leftarrow \varnothing$
  Create model: (Conv$\Rightarrow$ReLU)$\times$2 $\rightarrow$Max_POOL$\rightarrow$
  (FC$\Rightarrow$ReLU)$\rightarrow$(FC$\Rightarrow$Sigmoid)
  MLCNN.complie(loss=*SW_FL*,Optimizer=*RMSProp*)
  Val_loss$\leftarrow$MLCNN.train([X_{tr}, Y_{tr}],[X_v, Y_v])
  BO.update(Θ,Val_loss)
  If Lowest_val_loss $>$ Val_loss **then**:
    Lowest_val_loss$\leftarrow$Val_loss
    File_model$\leftarrow$MLCNN.save_weights()
  end
end
MLCNN.load_weights(File_model)
$Y_{te}^{pred} \leftarrow$MLCNN.predict(X_{te})

IV. Experimental Setting

In order to fully assess the performance of the proposed model, a series of simulation experiments are conducted based on both global and local measure metrics. First, we introduce

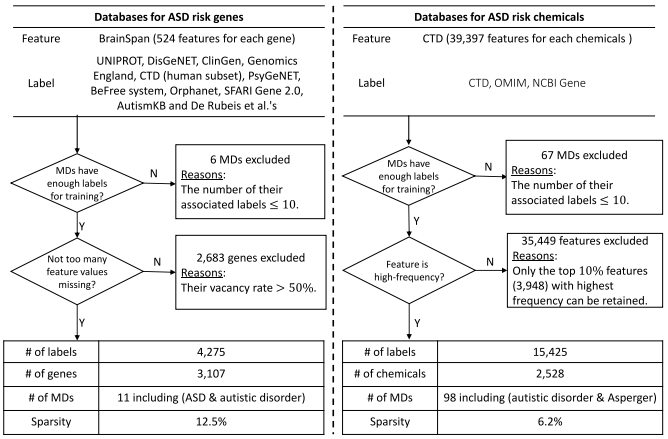


Fig. 3. Data gathering and preprocessing.

the integrated data sets and various evaluation metrics. Especially, the criteria applied for data gathering and preprocessing are also described. Then, the state-of-the-art multilabel classifiers are used to compare with the proposed model, so as to verify its performance. We perform the effect analysis to evaluate the effectiveness of the proposed methods: MDFS, SW-FL, and the automated hyperparameter tuning of BO. Finally, a case study is conducted to estimate the practical effects of the predicted autistic risk genes by manual validation of published studies.

A. Data Sets

As shown in Fig. 3, we integrate and consolidate the information for instance features and labels that are fed into the proposed model afterward. A series of MDs closely related to ASD are considered for training based on the hierarchical disease relationships of MeSH descriptors [43]. In this work, since we aim to identify two kinds of ASD high-risk instances (i.e., genes and chemicals) using different databases, the ways to preprocess raw data are also different.

To build gene feature matrices, developmental brain gene expression data are downloaded from the BrainSpan database [44], which provides the most comprehensive transcriptome of the human developing brain. Each archived gene instance embraces 524 features by representing the expression levels of RNA-sequencing reads in the units of Reads Per Kilobase of transcript per Million mapped reads (RPKM). They are involved in different developmental stages ranging from eight weeks postconception to 40 years of age over 26 brain structures. To construct an effective classifier, we filter out those genes with high sparsity ($>50\%$) in features, resulting in 3107 gene instances. A $\log_2(\text{RPKM}+1)$ transformation is conducted to normalize the data to a smaller range. Acting as label information, the positive gene-MD associations are collected from multiple versatile databases. Six MD classes having insufficient positive labels with gene instances (≤ 10) are excluded. As a result, we take 11 MD classes (including ASD and AD) and 3107 gene instances into consideration in this study.

Systematic integration is used as a molecular basis representation for chemical features construction. Four functionally heterogeneous networks are concerned: the interactions of chemical-gene, chemical-pathway, chemical-GO term, and chemical-phenotype. Based on the information curated from the CTD database [45], there are 39 397 binary variables in the chemical integrated network, taking two possible states labeled as 0 and 1. However, a large number of redundant properties make the prediction difficult; therefore, we trim down 90% features with low frequency to retain the final 3948 remarkable features. Likewise, 67 MD classes having insufficient positive labels with chemical instances (≤ 10) are excluded as well. Consequently, the remaining 2528 chemical instances and 98 MD classes (including AD and AS) are used in this work.

B. Evaluation Metrics

The performance of the proposed model is evaluated in terms of both global and local measures in this work. The global measures concern the classification effectiveness for all involved MD classes, while the local measures merely consider that of ASD and its subtypes using more concrete metrics as a single-label validation manner. Given a test set consisting of t instances denoted as $\mathcal{U} = \{(x_i, y_i)|1 \leq i \leq t\}$, let $\mathcal{Z}_i = \mathcal{H}(x_i)$ be the set of labels predicted by a multilabel classifier $\mathcal{H}$ for a test instance x_i. Six key global measure metrics [46] in MLC can be calculated as follows.

1) *Micro-F1:* This metric computes the microaveraged precision and microaveraged recall and then combines the both as the F-measure averaging over the prediction matrix

$$\text{MiF}(\mathcal{Z}, \mathcal{U}) = \frac{2\sum_{i=1}^{t} |\mathcal{Z}(x_i) \cap y_i|_1}{\sum_{i=1}^{t} |y_i|_1 + \sum_{i=1}^{t} |\mathcal{Z}(x_i)|_1} \tag{12}$$

where $|.|_1$ denotes the l_1-norm.

2) *Average Precision:* This metric summarizes the average fraction of the related labels ranked higher than a particular label $l_k \in y_i$

$$\text{AP}(\mathcal{Z}, \mathcal{U}) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{|y_i|}\sum_{l_j, l_k \in y_i}\frac{|\{l_j|\text{rank}(x_i, l_j) \leq \text{rank}(x_i, l_k)\}|}{\text{rank}(x_i, l_k)} \tag{13}$$

where $rank(x_i, l_k) = \sum_{j=1}^{q} \delta(\mathcal{Z}_j(x_i) \geq \mathcal{Z}_k(x_i))$ returns the rank of l_k sorted in the descending order. Assume that $\delta()$ is a binary indicator and $|.|$ denotes a cardinality of the set.

3) *Hamming Loss:* This metric reflects the fraction of labels that are wrongly estimated

$$HL(\mathcal{Z}, \mathcal{U}) = \frac{1}{t}\sum_{i=1}^{t}\frac{|\mathcal{Z}(x_i) \triangle y_i|_1}{q} \tag{14}$$

where $\triangle$ performs the XOR operation between two sets.

4) *One Error:* This metric measures the probability that the top-ranked label is not in the set of possible labels

$$OE(\mathcal{Z}, \mathcal{U}) = \frac{1}{t}\sum_{i=1}^{t}\delta(\arg\max \mathcal{Z}(x_i) \notin y_i). \tag{15}$$

5) *Ranking Loss:* This metric returns the fraction of reversely ordered label pairs

$$\mathrm{RL}(\mathcal{Z},\mathcal{U}) = \frac{1}{t}\sum_{i=1}^{t}\frac{|\{(l_j,l_k)|\mathcal{Z}_j(x_i)\le\mathcal{Z}_k(x_i),(l_j,l_k)\in y_i\times\bar{y}_i\}|}{|y_i||\bar{y}_i|}. \tag{16}$$

6) *Coverage:* This metric evaluates the average depth in the ranking, so as to cover all the ground-truth labels associated with an instance

$$\mathrm{COV}(\mathcal{Z},\mathcal{U}) = \frac{1}{t}\sum_{i=1}^{t}\max_{l_k\in y_i}\mathrm{rank}(x_i,l_k)-1. \tag{17}$$

In fact, MLC can be viewed as a set of binary classification problems (i.e., one for each class) based on the confusion matrix, which is derived from the following four categories: true positives (TPs), FPs, true negatives (TNs), and false negatives (FNs). To better evaluate the classification performance on ASD and its subtypes (i.e., AD and AS), five common local measure metrics [47] are also considered as follows.

1) *Accuracy:* This metric calculates the fraction of all correct predictions identified as either TP or TN

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{FN}}. \tag{18}$$

2) *Specificity:* This metric calculates the number of correct negative predictions out of the total actual negatives

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \tag{19}$$

3) *Recall:* This metric calculates the number of correct positive predictions out of the total actual positives

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{20}$$

4) *Precision:* This metric calculates the number of correct positive predictions out of the total predictions identified as positive

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{21}$$

5) *F1-Score:* This metric calculates a harmonic mean of the precision and recall

$$F1 = \frac{2\times\mathrm{REC}\times\mathrm{PRE}}{\mathrm{REC}+\mathrm{PRE}}. \tag{22}$$

C. Performance Comparison

To evaluate the proposed model, we implement fivefold cross validations (CVs) by dividing the data sets into five folds where each fold is used to test the model, while the rest are used for training the model. This process is repeated five rounds until all folds are used for testing in turns. All the experimental results shown in the following are averaged from ten times of fivefold CVs to reduce random sampling bias. To the best of our knowledge, there are no MLC-based computational approaches developed for the identification of ASD risk genes or chemicals so far. For the performance evaluation,

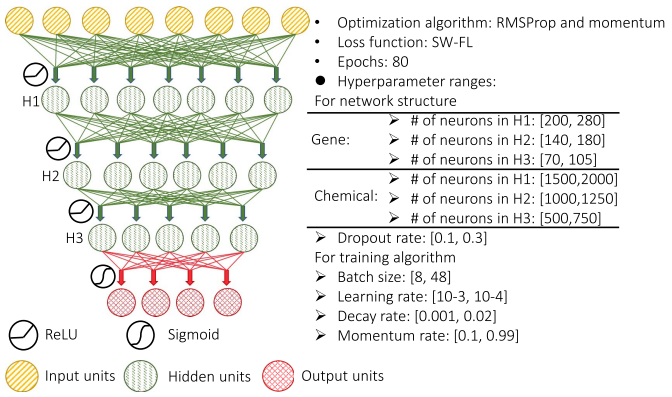


Fig. 4. Architecture and setting of the MLANN model.

we compare the proposed model against the three state-of-the-art MLC algorithms: MLARAM [48], MLKNN [49], and MLTSVM [50]. MLARAM is the MLC variant of adaptive resonance theory (ART)-based neural networks. MLKNN and MLTSVM are MLC extensions to the traditional ML algorithms: K-nearest neighbor (KNN) and SVM, respectively. To be fair, their parameters are optimized from all feasible parameter combinations via grid search with CVs. Since their hyperparameter settings are relatively simple with fewer tunable arguments, the searching space is much smaller than that of MLCNN. Furthermore, they all provide default values and suggested values as prior knowledge for easy application. Therefore, based on prior knowledge, it is efficient to find the decent hyperparameter sets for every combination of various hyperparameters via grid search. The multilabel artificial neural network (MLANN) is also used for the performance comparison as a baseline of deep learning (DL) classifiers (its architecture and setting are shown in Fig. 4). Likewise, the setting of its hyperparameters is also optimized by BO with GPs, where we control the same amount of steps to perform exploration (20 iterations) and exploitation (ten iterations) as the proposed MLCNN model does. The experiment results are obtained based on the best scenarios by varying the remarkable feature combinations from 250/1000 to 400/2500 features (with an interval of 50/500), which are in the top rank derived from MDFS for downstream classifications of risk genes/chemicals.

The performance comparison is tabulated in Section A in Table II for risk genes and chemicals in terms of global measure metrics. It is shown that the MLCNN and MLANN models tend to obtain superior classification performance compared with the other state-of-the-art MLC algorithms. Especially, on the chemical data set, it is more challenging to make a reliable prediction due to its higher sparsity and more MD classes involved. A great number of easily classified unlabeled samples overwhelm the training and lead to degenerate the performance of MLARAM, MLKNN, and MLTSVM. The problem becomes even more pronounced in MLKNN and MLTSVM since both of them are nonneural network MLC algorithms. They could fail to capture different levels of abstraction by "shallow" or linear models. MLARAM considers each disease label as a unique class that could be less

TABLE II
PERFORMANCE COMPARISON FOR RISK GENES AND CHEMICALS USING GLOBAL MEASURE METRICS

Model	MiF(%) ↑	AP(%) ↑	HL(%) ↓	OE(%) ↓	RL(%) ↓	COV(%) ↓	MiF(%) ↑	AP(%) ↑	HL(%) ↓	OE(%) ↓	RL(%) ↓	COV(%) ↓
Section A	For best performance mode on risk genes						For best performance mode on risk chemicals					
MLCNN	**0.893**	**0.888**	**0.026**	**0.059**	**0.029**	**1.773**	**0.850**	0.819	**0.017**	0.293	0.043	**16.678**
MLANN	0.705	0.614	0.069	0.247	0.065	2.171	0.827	**0.838**	0.019	**0.262**	**0.040**	16.775
MLARAM	0.610	0.596	0.085	0.277	0.089	2.369	0.524	0.485	0.047	0.415	0.079	23.130
MLKNN	0.391	0.550	0.249	0.330	0.099	2.541	0.337	0.259	0.130	0.542	0.085	23.741
MLTSVM	0.611	0.095	0.085	0.903	0.342	4.896	0.196	0.104	0.062	0.952	0.685	79.983
Section B	For effect analysis on risk genes						For effect analysis on risk chemicals					
MLCNN^a	0.848(-5.7)	0.867(-2.4)	0.036(-38.5)	0.113(-91.5)	0.038(-31.0)	1.866(-5.25)	0.838(-1.4)	0.788(-3.8)	0.020(-17.6)	0.342(-16.7)	0.055(-27.9)	19.04(-14.2)
MLANN^a	0.673(-4.5)	0.680(+10.7)	0.072(-4.3)	0.243(+1.6)	0.066(-1.5)	2.176(-2.3)	0.802(-3.0)	0.825(-1.6)	0.028(-47.4)	0.314(-19.8)	0.045(-12.5)	15.89(+5.3)
MLARAM^a	0.607(-0.5)	0.594(-0.3)	0.085(0)	0.278(-0.4)	0.093(-4.5)	2.415(-1.9)	0.499(-4.8)	0.458(-5.6)	0.049(-4.3)	0.414(-0.2)	0.077(+2.5)	22.06(+4.6)
MLKNN^a	0.391(0)	0.548(-0.4)	0.248(-0.4)	0.331(-0.3)	0.100(-1.0)	2.549(-0.3)	0.323(-4.2)	0.252(-2.7)	0.132(-1.5)	0.564(-4.1)	0.089(-4.7)	24.36(-2.6)
MLTSVM^a	0.613(+0.3)	0.148(+55.8)	0.085(0)	0.910(-0.8)	0.363(-6.1)	5.104(-4.2)	0.171(-12.8)	0.039(-62.5)	0.062(0)	0.911(+4.3)	0.731(-6.7)	83.02(-3.8)
MLCNN^b	0.815(-8.7)	0.794(-10.6)	0.044(-69.2)	0.181(-200.7)	0.056(-93.1)	2.104(-18.7)	0.814(-4.2)	0.745(-9.0)	0.032(-88.2)	0.412(-40.6)	0.062(-44.2)	20.19(-21.1)
MLANN^b	0.658(-6.7)	0.583(-5.0)	0.089(-28.9)	0.273(-10.5)	0.071(-9.2)	2.290(-5.5)	0.792(-4.2)	0.806(-3.8)	0.025(-31.6)	0.361(-37.8)	0.048(-20.0)	18.20(-8.5)
MLCNN$^{10}_{20}$	0.901(+0.9)	0.869(-2.1)	0.026(0)	0.064(-8.5)	0.031(-6.9)	1.853(-4.5)	0.841(-1.1)	0.798(-2.6)	0.018(-5.9)	0.344(-7.0)	0.046(-7.0)	17.478(-4.8)
MLCNN$^{15}_{15}$	0.897(+0.4)	0.876(-1.4)	0.027(-3.8)	0.056(+5.1)	0.033(-13.8)	1.869(-5.4)	0.840(-1.2)	0.806(-1.6)	0.019(-11.8)	0.314(-7.2)	0.048(-11.6)	18.243(-9.4)
MLANN$^{10}_{20}$	0.703(-0.3)	0.609(-0.8)	0.069(0)	0.248(-0.4)	0.066(-1.5)	2.173(-0.1)	0.837(+1.2)	0.848(+1.2)	0.018(+5.3)	0.257(+1.9)	0.041(-2.5)	16.649(+0.8)
MLANN$^{15}_{15}$	0.706(+0.1)	0.612(-0.3)	0.070(-1.4)	0.248(-0.4)	0.067(-3.1)	2.181(-0.5)	0.829(+0.2)	0.833(-0.6)	0.019(0)	0.242(+7.6)	0.042(-5.0)	17.280(-3.0)

Note: ↑ means the larger the better; ↓ means the smaller the better. (-%/+%) represents the worse/better difference compared to the best performance.
a means no feature selection by MDFS; b means no automated tuning by BO. Model^x_y means automated tuning by BO with x/y steps for exploration/exploitation.

effective in ignoring pairwise neuropathological relationships between disease labels. Generally speaking, the proposed MLCNN model maintains the best overall performance on both data sets. The results highlight its excellent capability of automatically extracting the remarkable features from such heterogeneous biological data. For the classification of risk chemicals, the MLANN model obtains comparable performance to the MLCNN model, representing an equal 3-3 share in global measure metrics. It could be attributed to the larger amount of neuron units employed in MLANN as the number of selected features increases. The number of trainable parameters in MLANN reaches almost two times larger than that in MLCNN. Furthermore, MLANN could have stronger expertise in coping with such discrete features of risk chemicals.

We make an extra comparison analysis in terms of local measure metrics to focus on ASD and its subtypes (AD and AS) on both data sets. As shown in Fig. 5, almost all compared classifiers achieve excellent results in terms of accuracy and specificity. In such imbalanced data sets, accuracy and specificity can be misleading since the easily classified unlabeled samples account for the majority. By contrast, recall, precision, and F1-score are more appropriate to really estimate the performance of the algorithms. Especially, for uneven class distribution, F1-score is usually more useful than accuracy by seeking a balance between recall and precision. As expected, the MLCNN and MLANN models outperform MLARAM, MLKNN, and MLTSVM in terms of recall, precision, and F1-score. The MLCNN model attains the highest F1-scores of 0.863 and 0.802 for risk genes in ASD and AD, respectively. The MLANN model suffers from an issue of low recall and, therefore, achieves the inferior F1-scores of 0.626 and 0.402 for risk genes in ASD and AD, respectively. MLKNN manages to obtain some tradeoffs between recall and precision. With regard to the MLC of risk chemicals, MLANN and MLARAM show a substantial improvement in the three metrics. Particularly, the MLANN model takes advantage of the full interconnection between layers for modeling complex relationships between features and labels. The yielded F1-scores reach the highest values of 0.722 and 0.935 for risk chemicals in AS and AD, respectively. It is worth highlighting that both MLANN and MLCNN models can still maintain a reliable classification for AS, which is rather short of positive samples.

To further evaluate the performance of all compared classifiers from both global and local perspectives, the precision–recall curve (PRC) is used for the visual interpretation and comparison on such imbalanced data sets. PRC represents the tradeoff between precision and recall by changing different thresholds. A high area under PRC (AUPRC) represents an accurate classification with both high precision and high recall. Fig. 6 demonstrates the comparison performance for risk genes and chemical prediction in three scenarios, namely, the global estimation over all disease classes as well as the local estimation over ASD and its subtypes. From the results of risk genes shown in Fig. 6(a)–(c), the MLCNN model achieves the highest AUPRC of 0.810 over all disease labels, as well as the highest AUPRCs of 0.910 and 0.830 for ASD and AD, respectively. The MLANN, MLARAM, and MLKNN models all tend to produce a decent global performance for most MDs but fail to maintain in ASD and AD. From the results of the risk chemicals in Fig. 6(c) and (d), the MLANN model reaches the highest AUPRC of 0.836 over all disease labels, exceeding that of the MLCNN model (0.725). Due to the positive samples insufficient in AS, MLARAM, MLKNN, and MLTSVM models seem to be overwhelmed by the dominant unlabeled samples. For risk chemicals in AD, MLCNN,

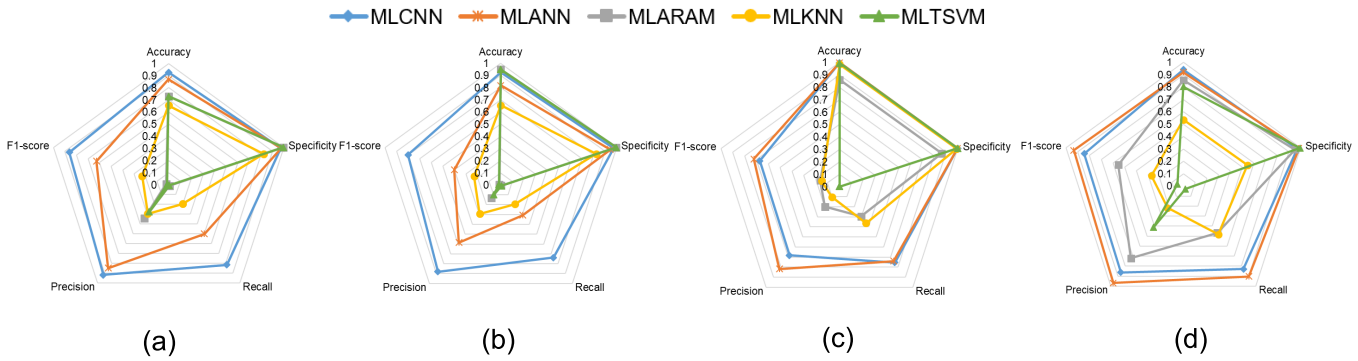


Fig. 5. Performance comparison for risk genes and chemicals on ASD and its subtypes in terms of local measure metrics. The bigger values in each criterion, i.e., farther away from the centroid, indicate better performance. (#) represents the number of positive samples. (a) For risk genes in ASD (841). (b) For risk genes in AD (601). (c) For risk chemicals in AS (11). (d) For risk chemicals in AD (497).

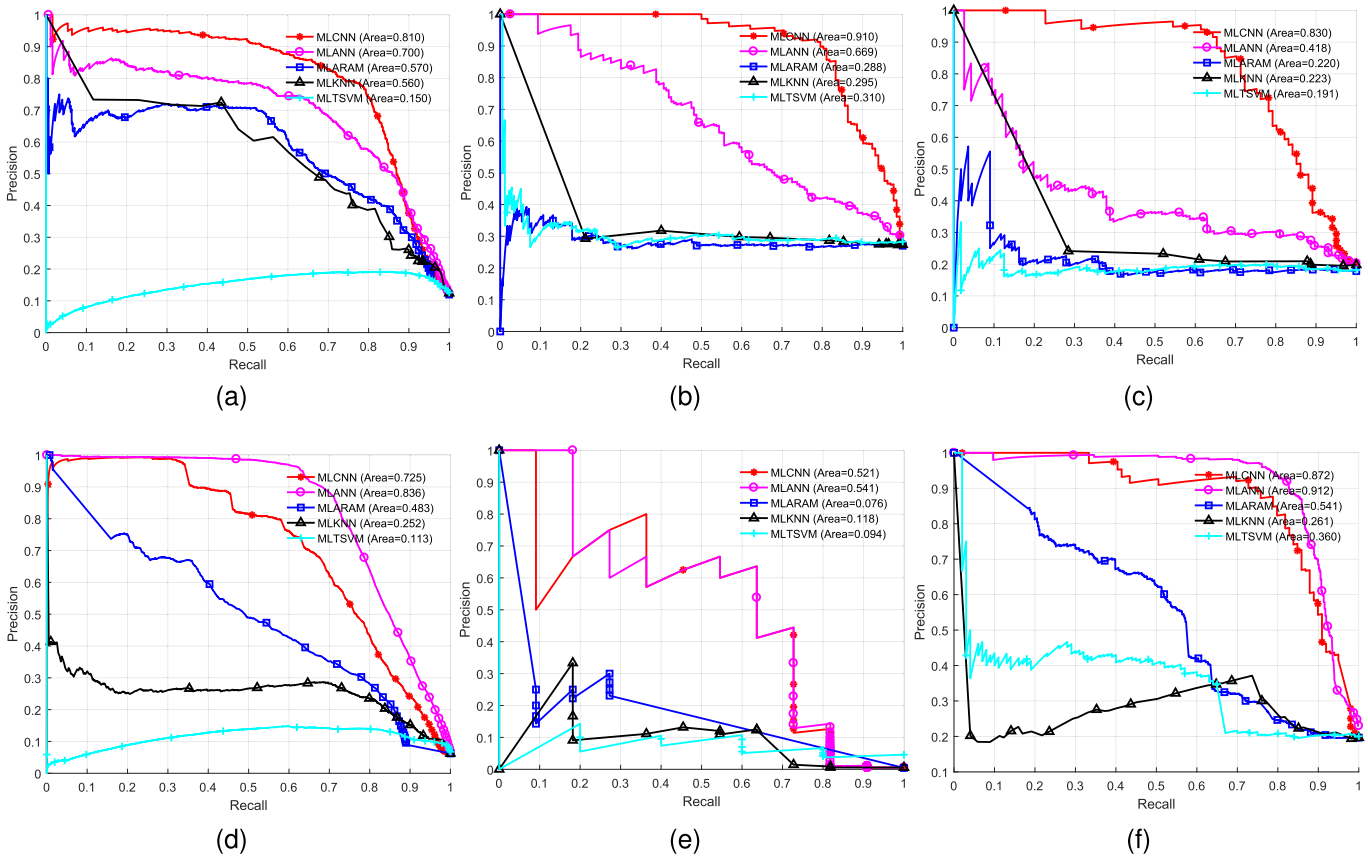


Fig. 6. PRC is applied to the comparison performance for risk genes and chemicals in three scenarios: the global estimation over all disease labels as well as the local estimation over ASD and its subtypes. (a) For risk genes over all diseases. (b) For risk genes in ASD. (c) For risk genes in AD. (d) For risk chemicals over all diseases. (e) For risk chemicals in AS. (f) For risk chemicals in AD.

MLANN, and MLARAM models outperform the other two nonneural network models (MLKNN and MLTSVM) to achieve the reliable AUPRCs of 0.872, 0.912, and 0.541, respectively.

The capability of the proposed model to handle such large-scale data can be attributed to the MDFS method and the MLCNN model equipped with SW-FL. MDFS performs a multilabel feature selection with manifold regularization to inject local and global label correlations, so as to select remarkable features discriminating across MD classes. Furthermore, SW-FL not only effectively addresses the imbalance within labels and between labels but also smoothly adjusts the learning weights to balance the loss distribution between MD labels. In this way, the appropriate measures of learning loss can guide the MLCNN model to put more focus on the hard, misclassified instances on large-scale data.

D. Effect Analysis of MDFS, BO, and SW-FL

In this section, we further conduct the performance effect analysis of MDFS, BO, and SW-FL. We first evaluate the

performance of all compared models using the whole feature sets based on global measure metrics in Section B in Table II (symbolized by "*a*"). Without feature selection by MDFS, all compared models show different levels of performance degradation in most global measure metrics. It is shown that the effectiveness of feature selection by MDFS tends to be more remarkable in the MLCNN, MLANN, and MLTSVM models. In other words, MDFS makes a greater contribution to their performance improvement. Next, the MLCNN and MLANN models with a fixed hyperparameter set (without automated tuning by BO) are also evaluated based on global measure metrics in Section B in Table II (symbolized by "*b*"). As we expect, automated hyperparameter tuning of BO has an overall positive effect on their performance by efficiently optimizing their network structure and training algorithm. We note that the automated hyperparameter tuning of BO brings more improvement to the performance of the MLCNN model compared with the MLANN model. That could be attributed to the more improvement potential in MLCNN as holding much more sophisticated settings in network structures.

To evaluate the feasibility and practicability of BO within finite time, the total number of iterations for optimization is fixed at 30. To determine the specific amount of steps for exploration and exploitation, we specify three possible combinations that can achieve a reasonable tradeoff as 10/15/20 steps for exploration followed by 20/15/10 steps for exploitation, respectively. As we can see in Section B in Table II, the combination of 20 steps of exploration and ten steps of exploitation achieves the highest probabilities to obtain decent optimization performance. Hence, this combination is chosen in our work.

To understand the proposed SW-FL better, we attempt to train the MLCNN model with standard binary CE loss or SW-FL, respectively. As illustrated in Fig. 7, their training performance is compared on both data sets of risk genes and chemicals. The convergence of SW-FL is guaranteed on both training and validation sets. We observe that the loss yielded by SW-FL falls more quickly than the binary CE and then reaches a stable stage. Furthermore, SW-FL enables more effective learning to dramatically reduce more overall loss. It could be attributed to the fact that SW-FL can effectively discount the effect of easy unlabeled samples, putting more focus on the hard positive labeled samples.

E. Case Study: Manual Validation for Autistic Risk Genes via Published Studies

It is necessary to assess the predictive capability of the proposed model in practice. We attempt to manually validate the top-20 predicted risk genes for ASD and AD via published studies, respectively. It needs to note that all labeled autistic risk genes are not overlapped in the prediction list. As we observe in Table III, 60% and 50% out of the top-20 predicted risk genes have been demonstrated to have associations with ASD and AD, respectively. For example, SEMA5A (fourth in the prediction list of ASD) has been found to have lower expression levels in cell lines and brains from individuals with ASD [51]. ANK1 has been reported to be highly positively correlated with the expression with mercury levels in the AD group but not in the typical control group [52]. These predicted risk genes are expected to serve as potential biomarkers for measuring and tracking the aberrances in autistic neurological and biological functions.

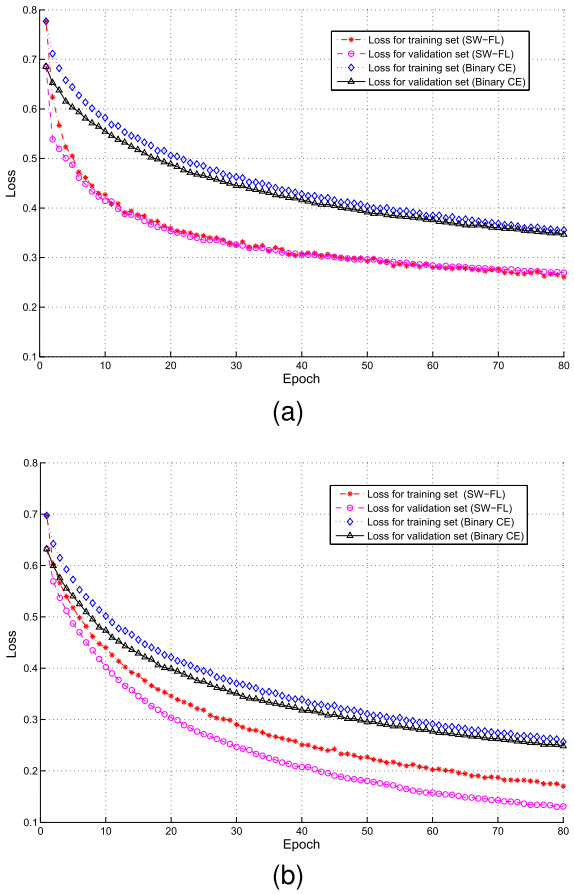


Fig. 7. SW-FL is compared with the binary CE on the both data sets of risk genes and chemicals. (a) For risk genes. (b) For risk chemicals.

TABLE III
MANUAL VALIDATION FOR PREDICTED RISK GENES VIA STUDIES

For ASD risk genes (12/20)			For AD risk genes (10/20)		
Gene	Evidence	Weight	Gene	Evidence	Weight
NPPA	**unconfirmed**	0.9752	*BCL6*	15121991[a]	0.7542
FBXL7	26777411[a]	0.9695	*ANK1*	19937285[a]	0.7050
MGP	**unconfirmed**	0.9663	*CPNE8*	**unconfirmed**	0.5875
SEMA5A	21805639[a]	0.9655	*PTHLH*	19937285[a]	0.5790
FAT1	24188901[a]	0.9637	*RORA*	21359227[a]	0.5776
CDH8	25294932[a]	0.9623	*CDH5*	26881141[a]	0.5763
NXPH1	22016809[a]	0.9616	*ZDHHC2*	**unconfirmed**	0.5685
TGFB2	**unconfirmed**	0.9575	*SLC13A5*	**unconfirmed**	0.5633
KDM5D	**unconfirmed**	0.9570	*SLIT3*	30483073[a]	0.5421
LAMB2	**unconfirmed**	0.9503	*RDH12*	**unconfirmed**	0.5086
UPF3B	20479756[a]	0.9385	*KIF17*	**unconfirmed**	0.4966
PLN	18252227[a]	0.9332	*CHPT1*	**unconfirmed**	0.4928
FBLN5	**unconfirmed**	0.9325	*TESC*	**unconfirmed**	0.4903
NR3C2	10.1101/338855[b]	0.9314	*ARX*	17044103[a]	0.4857
CTTNBP2	27909399[a]	0.9273	*SLC12A7*	**unconfirmed**	0.4776
PTCHD1	20844286[a]	0.9270	*TNNI3*	**unconfirmed**	0.4746
FOXP1	23815876[a]	0.9266	*FAM53B*	**unconfirmed**	0.4622
GLRA2	26370147[a]	0.9243	*BTN2A1*	16845580[a]	0.4418
NPR3	**unconfirmed**	0.9223	*SACS*	25623060[a]	0.4123
FAM193A	**unconfirmed**	0.9219	*HCCS*	21701786[a]	0.4031

Note: a means support by PMID; b means support by DOI.

V. Conclusion

We have presented an efficient MLC model to identify the potential autistic risk genes and chemicals on large-scale data sets. The feature matrices and partially labeled networks were first constructed from the heterogeneous biological databases for risk genes and chemicals. Then, a multilabel manifold regularized feature selection method MDFS was introduced to select the remarkable features discriminating across multiple MDs. A new loss function SW-FL was proposed to address the imbalance within labels and between labels by putting more focus on the hard classified samples. Finally, based on the proposed SW-FL, an MLCNN model equipped with automated hyperparameter tuning of BO was used for the identification of autistic risk genes and chemicals. A series of simulation experiments were conducted to evaluate the performance of the proposed model in terms of both global and local measure metrics. The experiment results demonstrated that our model can achieve superior performance compared with the state-of-the-art MLC methods. To the best of our knowledge, this is the first work to computationally identify the autistic risk genes and chemicals. It is anticipated that this work can facilitate the discovery of autistic biomarkers and aid future research efforts toward the investigation of gene-environment interactions to better understand the pathomechanism of ASD.

References

[1] J. Baio *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014," *MMWR Surveill. Summaries*, vol. 67, no. 6, p. 1, 2018.

[2] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state fMRI using deep belief network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 21, 2020, doi: 10.1109/TNNLS.2020.3007943.

[3] Y. S. Kim and B. L. Leventhal, "Genetic epidemiology and insights into interactive genetic and environmental effects in autism spectrum disorders," *Biol. Psychiatry*, vol. 77, no. 1, pp. 66–74, Jan. 2015.

[4] D. H. Geschwind, "Advances in autism," *Annu. Rev. Med.*, vol. 60, no. 1, pp. 367–380, 2009.

[5] S. Sandin, P. Lichtenstein, R. Kuja-Halkola, C. Hultman, H. Larsson, and A. Reichenberg, "The heritability of autism spectrum disorder," *Jama*, vol. 318, no. 12, pp. 1182–1184, 2017.

[6] M. Alarcón *et al.*, "Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene," *Amer. J. Hum. Genet.*, vol. 82, no. 1, pp. 150–159, Jan. 2008.

[7] E. A. Varga, M. Pastore, T. Prior, G. E. Herman, and K. L. McBride, "The prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly," *Genet. Med.*, vol. 11, no. 2, pp. 111–117, 2009.

[8] J. Hallmayer *et al.*, "Genetic heritability and shared environmental factors among twin pairs with autism," *Arch. Gen. Psychiatry*, vol. 68, no. 11, pp. 1095–1102, 2011.

[9] P. Grandjean and P. Landrigan, "Developmental neurotoxicity of industrial chemicals," *Lancet*, vol. 368, no. 9553, pp. 2167–2178, Dec. 2006.

[10] L. R. Goldman and S. Koduru, "Chemicals in the environment and developmental toxicity to children: A public health and policy perspective," *Environ. Health Perspect.*, vol. 108, no. 3, pp. 443–448, 2000.

[11] D. A. Rossignol and R. E. Frye, "A review of research trends in physiological abnormalities in autism spectrum disorders: Immune dysregulation, inflammation, oxidative stress, mitochondrial dysfunction and environmental toxicant exposures," *Mol. Psychiatry*, vol. 17, no. 4, pp. 389–401, 2012.

[12] Z. Li, T. Dong, C. Pröschel, and M. Noble, "Chemically diverse toxicants converge on Fyn and c-Cbl to disrupt precursor cell function," *PLoS Biol.*, vol. 5, no. 2, p. e35, Feb. 2007.

[13] M. R. Herbert *et al.*, "Autism and environmental genomics," *Neurotoxicology*, vol. 27, no. 5, pp. 671–684, 2006.

[14] D. A. Rossignol, S. J. Genuis, and R. E. Frye, "Environmental toxicants and autism spectrum disorders: A systematic review," *Transl. Psychiatry*, vol. 4, no. 2, p. e360, Feb. 2014.

[15] G. Dawson *et al.*, "Early behavioral intervention is associated with normalized brain activity in young children with autism," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 51, no. 11, pp. 1150–1159, Nov. 2012.

[16] J. Zhang, Z. Luo, C. Li, C. Zhou, and S. Li, "Manifold regularized discriminative feature selection for multi-label learning," *Pattern Recognit.*, vol. 95, pp. 136–150, Nov. 2019.

[17] Y. Lin, F. Qian, L. Shen, F. Chen, J. Chen, and B. Shen, "Computer-aided biomarker discovery for precision medicine: Data resources, models and applications," *Briefings Bioinf.*, vol. 20, no. 3, pp. 952–975, May 2019.

[18] T. Zeng, W. Zhang, X. Yu, X. Liu, M. Li, and L. Chen, "Big-data-based edge biomarkers: Study on dynamical drug sensitivity and resistance in individuals," *Briefings Bioinf.*, vol. 17, no. 4, pp. 576–592, Jul. 2016.

[19] L. Shen, Y. Lin, Z. Sun, X. Yuan, L. Chen, and B. Shen, "Knowledge-guided bioinformatics model for identifying autism spectrum disorder diagnostic MicroRNA biomarkers," *Sci. Rep.*, vol. 6, no. 1, p. 39663, Dec. 2016.

[20] D. H. Oh, I. B. Kim, S. H. Kim, and D. H. Ahn, "Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning," *Clin. Psychopharmacol. Neurosci.*, vol. 15, no. 1, p. 47, 2017.

[21] S. Cogill and L. Wang, "Support vector machine model of developmental brain gene expression data for prioritization of autism risk gene candidates," *Bioinformatics*, vol. 32, no. 23, pp. 3611–3618, 2016.

[22] M. Gök, "A novel machine learning model to predict autism spectrum disorders risk gene," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6711–6717, Oct. 2019.

[23] Y. Kou, C. Betancur, H. Xu, J. D. Buxbaum, and A. Ma'Ayan, "Network- and attribute-based classifiers can prioritize genes and pathways for autism spectrum disorders and intellectual disability," *Amer. J. Med. Genet. C, Seminars Med. Genet.*, vol. 160, no. 2, pp. 130–142, 2012.

[24] A. Kanehira and T. Harada, "Multi-label ranking from positive and unlabeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5138–5146.

[25] J. Zhang, C. Li, Z. Sun, Z. Luo, C. Zhou, and S. Li, "Towards a unified multi-source-based optimization framework for multi-label learning," *Appl. Soft Comput.*, vol. 76, pp. 425–435, Mar. 2019.

[26] J. Zhang *et al.*, "Multi-label learning with label-specific features by resolving label correlations," *Knowl.-Based Syst.*, vol. 159, pp. 148–157, Nov. 2018.

[27] W. Luo, J. Li, J. Yang, W. Xu, and J. Zhang, "Convolutional sparse autoencoders for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3289–3294, Jul. 2018.

[28] J.-Y. Park and J.-H. Kim, "Online incremental classification resonance network and its application to human–robot interaction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1426–1436, May 2020.

[29] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[30] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proc. Int. Conf. Learn. Intell. Optim.* Berlin, Germany: Springer, 2011, pp. 507–523.

[31] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2951–2959.

[32] J. Snoek *et al.*, "Scalable Bayesian optimization using deep neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2171–2180.

[33] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1496–1509, Jun. 2017.

[34] Z.-H. You *et al.*, "PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction," *PLoS Comput. Biol.*, vol. 13, no. 3, Mar. 2017, Art. no. e1005455.

[35] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[36] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ_2, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[37] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[38] R. Huang, W. Jiang, and G. Sun, "Manifold-based constraint Laplacian score for multi-label feature selection," *Pattern Recognit. Lett.*, vol. 112, pp. 346–352, Sep. 2018.
[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[41] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with ReLU activation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 597–607.
[42] M. Feurer, J. T. Springenberg, and F. Hutter, "Initializing Bayesian hyperparameter optimization via meta-learning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–8.
[43] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bull. Med. Library Assoc.*, vol. 88, no. 3, p. 265, 2000.
[44] M. J. Hawrylycz *et al.*, "An anatomically comprehensive atlas of the adult human brain transcriptome," *Nature*, vol. 489, no. 7416, pp. 391–399, 2012.
[45] A. P. Davis *et al.*, "The comparative toxicogenomics database: Update 2019," *Nucleic Acids Res.*, vol. 47, no. 1, pp. D948–D954, Jan. 2019.
[46] E. Gibaja and S. Ventura, "A tutorial on multi-label learning," *ACM Comput. Surv.*, vol. 47, pp. 1–38, Apr. 2015.
[47] D. Brzezinski, J. Stefanowski, R. Susmaga, and I. Szczech, "On the dynamics of classification measures for imbalanced and streaming data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2868–2878, Aug. 2020.
[48] F. Benites and E. Sapozhnikova, "HARAM: A hierarchical ARAM neural network for large-scale text classification," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 847–854.
[49] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[50] W.-J. Chen, Y.-H. Shao, C.-N. Li, and N.-Y. Deng, "MLTSVM: A novel twin support vector machine to multi-label learning," *Pattern Recognit.*, vol. 52, pp. 61–74, Apr. 2016.
[51] R. Holt and A. P. Monaco, "Links between genetics and pathophysiology in the autism spectrum disorders," *EMBO Mol. Med.*, vol. 3, no. 8, pp. 438–450, Aug. 2011.
[52] B. Stamova *et al.*, "Correlations between gene expression and mercury levels in blood of boys with and without autism," *Neurotoxicity Res.*, vol. 19, no. 1, pp. 31–48, Jan. 2011.

Zhi-An Huang received the B.Eng. degree in software engineering from Shenzhen University, Shenzhen, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong.

His research interests include artificial intelligence, machine learning, bioinformatics, and medical imaging analysis.

Jia Zhang received the M.S. degree from the School of Computer Science, Minnan Normal University, Zhangzhou, China, in 2016. He is currently pursuing the Ph.D. degree with the Artificial Intelligence Department, Xiamen University, Xiamen, China.

He is broadly interested in machine learning, data mining, and artificial intelligence. He is currently working on multilabel learning, data fusion, feature selection, and weakly supervised learning.

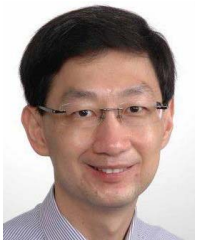
Zexuan Zhu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Fudan University, Shanghai, China, in 2003, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2008.

He is currently a Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include computational intelligence, machine learning, and bioinformatics.

Dr. Zhu is also the Chair of the IEEE CIS Emergent Technologies Task Force on Memetic Computing. He is also an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.

Edmond Q. Wu received the Ph.D. degree in controlling theory and engineering from Southeast University, Nanjing, China, in 2009.

He is currently with the Science and Technology on Avionics Integration Laboratory, China National Aeronautical Radio Electronics Research Institute, Shanghai, China. He is also an Associate Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, Shanghai. His research interests include deep learning, fatigue recognition, and human–machine interaction.



Kay Chen Tan (Fellow, IEEE) received the B.Eng. degree (Hons.) in electronics and electrical engineering and the Ph.D. degree in evolutionary computation and control systems from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively.

He is currently a Full Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. He has authored or coauthored over 200 refereed articles and six books.

Dr. Tan was an elected member of the IEEE Computational Intelligence Society Administrative Committee (CIS AdCom) from 2017 to 2019. He also serves as the editorial board member for over ten journals. He was the Editor-in-Chief of the *IEEE Computational Intelligence Magazine* from 2010 to 2013. He is also the Editor-in-Chief of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION.