DeepIDA: predicting isoform-disease associations by data fusion and deep neural networks

Guoxian Yu, Yeqian Yang, Yangyang Yan, Maozu Guo, Xiangliang Zhang, Jun Wang

Abstract—Alternative splicing produces different isoforms from the same gene locus, it is an important mechanism for regulating gene expression and proteome diversity. Although the prediction of gene(ncRNA)-disease associations has been extensively studied, few (or no) computational solutions have been proposed for the prediction of isoform-disease association (IDA) at a large scale, mainly due to the lack of disease annotations of isoforms. However, increasing evidences confirm the associations between diseases and isoforms, which can more precisely uncover the pathology of complex diseases. Therefore, it is highly desirable to predict IDAs. To bridge this gap, we propose a deep neural network based solution (DeepIDA) to fuse multi-type genomics and transcriptomics data to predict IDAs. Particularly, DeepIDA uses gene-isoform relations to dispatch gene-disease associations to isoforms. In addition, it utilizes two DNN sub-networks with different structures to capture nucleotide and expression features of isoforms, Gene Ontology data and miRNA target data, respectively. After that, these two sub-networks are merged in a dense layer to predict IDAs. The experimental results on public datasets show that DeepIDA can effectively predict IDAs with AUPRC (area under the precision-recall curve) of 0.9141, macro F-measure of 0.9155, G-mean of 0.9278 and balanced accuracy of 0.9303. The results on public datasets show that DeepIDA can effectively predict IDAs with AUPRC (area under the precision-recall curve) of 0.9141, macro F-measure of 0.9155, G-mean of 0.9278 and balanced accuracy of 0.9303., G-mean of 0.9278 and balanced accuracy of 0.9303. Although the nucleotide sequences of individual isoforms spliced from a single gene are homologous (see Fig. (b) for an example), these isoforms may have distinct biological functions and associations with diseases. For example, human heart sarcomeres produce two principal Titin isoforms, a more compliant N2BA isoform and a stiffer N2B isoform. Neagoe et al. measured the N2BA-to-N2B Titin isoform ratio in nonischemic human hearts and nonnecrotic left ventricle of coronary artery disease (CAD) patients. They found that the average N2BA-to-N2B ratio was 47:53 in severely diseased CAD transplanted hearts and 32:68 in nonischemic transplants. To study Alzheimer’s disease, Holtzman et al. detected a direct apoE isoform-specific effect on the mass, structure, and consequences of amyloid deposition in the brain and found that apoE4 had greater effects than apoE3 on $A_\beta$ burden, as well as an even greater effect on fibrillar $A_\beta$ deposition and neuritic plaques.

1 INTRODUCTION

Alternative splicing is a vital mode of genetic regulation in eukaryotes and the variability in splicing patterns leads to protein diversity from the genome [1]. During this process, exons of multi-exon genes are selectively spliced, whose alternative combinations produce different isoforms from a single gene (as illustrated in Fig. 1). For example, the Rho-associated coiled-coil-forming protein kinase (ROCK) gene has two isoforms: ROCK1 and ROCK2, both exhibit similar protection from bleomycin-induced vascular leak, myofibroblast differentiation and fibrosis in mice, but ROCK1-haploinsufficient mice shows greater attenuation of epithelial cell apoptosis [2]. More than 95% of human multi-exon genes undergo alternative splicing [3]. Although the nucleotide sequences of individual isoforms spliced from a single gene are homologous (see Fig. 1(b) for an example), these isoforms may have distinct biological functions and associations with diseases. For example, human heart sarcomeres produce two principal Titin isoforms, a more compliant N2BA isoform and a stiffer N2B isoform. Neagoe et al. measured the N2BA-to-N2B Titin isoform ratio in nonischemic human hearts and nonnecrotic left ventricle of coronary artery disease (CAD) patients. They found that the average N2BA-to-N2B ratio was 47:53 in severely diseased CAD transplanted hearts and 32:68 in nonischemic transplants. To study Alzheimer’s disease, Holtzman et al. detected a direct apoE isoform-specific effect on the mass, structure, and consequences of amyloid deposition in the brain and found that apoE4 had greater effects than apoE3 on $A_\beta$ burden, as well as an even greater effect on fibrillar $A_\beta$ deposition and neuritic plaques.
diseases. The identification of complex diseases related genes is a typical imbalanced classification task. The number of known genes associated with a disease is far smaller than unknown ones, which causes detecting novel gene-disease associations from such imbalanced training samples very hard. To alleviate this problem, clinical data is used to predict the novel disease genes. Luo et al. [13] extracted features from Protein-Protein Interaction (PPI) networks and clinical RNA-Seq data, and then applied the logistic regression classifier to predict new disease genes. By integrating clinical data, a model usually achieves a higher prediction accuracy, but the results are also very dependent on the quality of the clinical data. To address this limitation, basic properties of diseases and genes, such as disease similarities, gene similarities, protein domains, and so on, are fused to predict gene-disease associations. Zeng et al. [14] proposed a probability-based collaborative filtering model by fusing PPI networks, human mRNA co-expression, comparative genomics data, protein complex data, and disease similarity data. Gene Ontology (GO) [15] is also incorporated to improve the prediction. For example, Asif et al. [16] built a functional similarity matrix by integrating gene expression profiles, PPI networks and GO, and adopted multiple basic classifiers to evaluate the importance of features. They found that GO provides a consistent and deeper level of biological information for identifying gene-disease associations than PPI networks or gene expression profiles.

RNA-seq technology provides another perspective for disease pathology research. More and more isoform sequences and expression profiles have been collected, and biologists start to study disease pathology at the more granular isoform-level. Gandal et al. [17] analyzed three major psychiatric disorders, including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD), based on alternatively spliced transcriptome and isoform/gene-level co-expression networks. Lundberg et al. [18] compared coronary artery disease (CAD) patients with healthy controls by flow cytometry and found that CAD patients expressed significantly more FOX3 isoforms, which are characteristics for activated CD4+ T cells. Dorey et al. [19] applied in vivo and in vitro models to study the role and mechanism of ApoE2 isoform in Alzheimer’s disease.

Most studies of IDAs are still made by wet-lab experiments, although increasing evidences confirm the necessity to computationally predict IDAs. The main bottleneck is the scanty IDAs and there is no public repository collectively storing IDAs. In fact, the computational prediction of isoform functions also suffers the similar bottleneck. To bypass this bottleneck, some efforts [20, 21] have been made toward computational prediction of isoform functions by taking advantage of gene-isofrom relations to dispatch functional annotations at the gene level to isoform level. Wang et al. [22] proposed a data integrative solution called DisoFun to differentiate isoform functions with collaborative matrix factorization. DisoFun assumes the functional annotations of genes are aggregated from those of key isoforms. It collaboratively factorizes the isoform expression data matrix and gene-term data matrix (storing GO annotations of genes) into low-rank matrices to simultaneously explore the latent key isoforms, and accomplishes function prediction by aggregating predictions to their originating genes. Shaw et al. [21] proposed a data integrative solution called DeepIsoFun, that combines multiple instance learning with domain adaptation. DeepIsoFun treats each gene as a bag and isoforms of the gene as instances of that bag. To enable domain adaptation, DeepIsoFun regards a gene as an instance in the gene domain, and also as a bag in the isoform domain. Since GO annotations of genes are known, the adaption technique can be used to transfer functional annotations from the gene domain into the isoform domain. Chen et al. [24] adopted a similar strategy to initialize isoform-level training labels and proposed another deep learning based model called DIFFUSE, which first uses a deep neural network to predict the functions of isoforms from their genomic sequences and then refines the prediction using a conditional random field based on isoform co-expression relationship. From the perspective of the similarity between Gene Ontology (GO) and Disease Ontology (DO) [23], and the application of multi-instance learning, isoform function prediction shares some similarity with the isoform-disease association prediction, but it still has some distinctions. The former mainly focuses on annotating GO terms to isoforms using GO annotations of genes, while the latter more focuses on exploring the relationship between isoforms and diseases via gene-disease associations, which gives a more granular view of isoforms and their relations with diseases. IDA is more close with pathology analysis, since the occurrence of complex diseases is often caused by abnormal alternative splicings. IDAs can give a more clear directional clue for uncovering the pathology, while isoform function prediction just gives broad functional knowledge of alternative spliced isoforms. Last but not least, gene-disease (isoform-disease) associations are much fewer than GO annotations of genes (isoforms). Although GO and DO both adopt a direct acyclic graph to represent the domain knowledge of gene functions and diseases, their inner structure is different and not so overlapped.

Inspired by these efforts, we develop a data fusion based deep learning solution called DeepIDA. Our data fusion solution is motivated by the observation that different types of biological data carry complementary knowledge of IDAs and thus contribute to a more reliable prediction. Particularly, the data we collected can be grouped into two types: (i) data of isoforms including the nucleotide sequence and RNA-seq data, (ii) data of genes including miRNA targets and Gene
Ontology. DeepIDA deploys two DNN sub-networks with different structures, which synchronously extract features from the two types of data. These extracted features are then merged in a fully connected layer, whose output is used as the input of a decision layer with a single neuron to predict the probability of IDAs. To kickoff the learning task, we leverage the obtained gene-isofrom relations and the gene-level disease associations to initialize the isofrom-level disease associations. In addition, DeepIDA adopts the focal loss to alleviate the intrinsic class imbalance problem in predicting IDAs. Given the lack of negative GO annotations of genes, some works take gene function prediction as a positive-unlabeled data learning problem [27] [28], but most other works model this prediction task as a binary or multilabel classification problem [29] [31]. Following the canonical setting in the previous works [20] [22] [32], in this paper we model isofrom-disease association prediction as a binary classification problem.

To evaluate the performance of DeepIDA, we compared DeepIDA against two multi-instance learning based isofrom function prediction methods (Disofun [23] and iMILP [20]) and a representative gene-disease association prediction approach (Prince [33]), and five typical machine learning methods, which are Weighted K-Nearest Neighbor (WKNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) [27]. Since the prediction of IDAs is a typical class imbalance problem, we adopted the macro F-Measure, G-Mean, Balanced Accuracy and AUPRC (area under the precision-recall curve), which are more sensitive to class imbalance, to quantify the performance of these compared methods. Our experimental results show that DeepIDA achieves an average macro F-Measure of 0.9155, G-Mean of 0.9278, Balanced Accuracy of 0.9303 and AUPRC of 0.9141 across 732 diseases, which outperforms the second best performer SVM by 3\% \%, 3.7\%, 3.6\%, 3.3\% respectively. In addition, we also evaluated the performance of DeepIDA and these compared methods on sixteen isofrom-disease associations. DeepIDA has an Accuracy of 0.8125 and Recall of 0.9091, which outperforms the runner-up (iMILP) by 6\% in Accuracy and 9\% in Recall.

The rest of this paper is organized as follows. The datasets we collected and the corresponding preprocess strategy are detailed in Section 2. Our proposed method and its specific implementation are introduced in Section 3. Section 4 reports and analyzes the experimental results of DeepIDA and compared methods. Conclusions and some possible future work are briefly outlined in Section 5.

2 DATA PROCESSING

As mentioned above, DeepIDA integrates two types of biological data, data of isofroms and data of genes. The relevant data preprocessing strategies are summarized in Figure 2 (a) and more details are given in this section.

2.1 Data of isofroms

Isoform sequences contain multiple functional sites, disease- associated single nucleotide polymorphism (SNP) sites, single-molecule peptides and motifs [24], which can help us to reveal the association between isofroms and diseases. Given that, we first collected sequence data from Ensembl Database [35] and NCBI RefSeq Database [36]. Then we adopted the Conjoint Triad (CT) approach [37] to numerically encode sequences. CT is originally designed to encode protein sequence information, which considers the properties of one amino acid and its vicinal amino acids. CT regards any three continuous amino acids as a unit. Thus, the triads can be differentiated according to the classes of amino acids in a unit. The amino acids of a protein are divided into 7 classes, so a 7^3-dimensional feature vector can be generated for a protein. The CT encoding mechanism has been widely used for encoding amino acids sequences [38] [39]. Here, we use CT to encode nucleotide sequence of isofroms. However, there are only four bases in the nucleotide sequence, and thus CT can only generate 4\^3-dimensional sequence features, which is too short to encode the nucleotide sequence data and to support follow-up analysis. Therefore, we implemented 4-mer CT [38], which can generate the 4\^4-dimensional vector to encode more comprehensive features. First, we number the four bases by the following rules: A=0, C=1, T=2 and G=3. Then, every four bases are regarded as a group, and each group can have 4^4 cases. In the end, the number of occurrences of each group in a sequence is counted and normalized. The exemplar 4-mer CT for a nucleotide sequence is shown in Fig. 3.

Genetic variants can strongly affect RNA splicing, a critical step in gene expression, whose disruption contributes to many diseases [40] [41]. Alike proteins, an isofrom with the same sequence features can be associated with different diseases and with different expression profiles across different diseases. Thus, we collected and processed RNA-Seq data from the ENCODE project [42] (access date: 2019-11-10). Particularly, we downloaded 596 RNA-seq runs (of total 298 samples from different tissues and conditions) of Human. We adopted the canonical Fragments Per Kilobase of exon per Million fragments mapped fragments (FPKM) values to quantify the expression of isofroms. The datasets are heterogeneous in terms of library preparation procedures and sequencing platform. Following the pre-process done in [20] [23], for each tissue, we control the quality of these RNA-seq datasets and quantify the expression value of isofroms as follows:

1. We firstly align the short-reads of each RNA-seq dataset of the Human genome (build GRCh38.90) from Ensembl using HISAT2 (v.2-2.1.0) [43], and a GTF annotation file of the same build with an option of no-novel-junction.

2. Then, we use StringTie (v.1.3.3b) [44] to calculate the relative abundance of the transcript as FPKM. We separately compute the FPKM values of a total of 57,964 genes with 219,288 isofroms for each sample.

3. The FPKM values of very short isofroms are exceptionally higher. Therefore, we discard the isofroms with less than 100 nucleotides.

4. To further control the quality of isofroms, we use known protein coding gene names to map those genes obtained in step (3). In addition, in order to fuse transcriptomics and genomics data, we intersect these genes with the genes in the Gene Ontology data and with the collected miRNA target data. Finally, we obtain 12,663 genes with 31,408 isofroms.
potential effects on almost every genetic pathway. Studies have shown that miRNA mutations or mis-expressions are associated with various human cancers, and miRNAs can function as tumour suppressors and oncomines \[49\].

To overcome the inconsistency in the naming conventions of biomolecules across these databases, we aligned the genes across different data sources through the name mapping files to filter genes absent in any of these databases. After filtering, a total of 6,428 GO terms (4,724 in Biological Process, 1,000 in Molecular Function, and 704 in Cellular Component Ontology), and 495 miRNAs were retained. Next, the above multilevel biomolecular networks composed with genes, miRNAs, GO terms and their inter-relations are encoded by one-hot coding.

Due to the lack of IDAs, the conventional supervised methods cannot be directly applied. Most isoform function prediction solutions leverage the gene-isoform relations to initialize the functional annotations of isoforms by transferring GO annotations at the gene level \[21\] \[24\]. This strategy can also be applied to initialize IDAs. So we downloaded the gene-isoform relations from ENCODE project and gene-disease associations from Disgenet citepinero2016disgenet. Suppose \( Z \in \mathbb{R}^{m \times d} \) records the latent isoform-disease associations for \( m \) isoforms and \( d \) types of diseases. Motivated by the principle that the diseases associated with a gene are responsible by individual isoforms spliced from this gene, we can reversely dispatch the disease annotations of genes to individual isoforms and initialize the disease annotations of isoforms as follows:

\[
Z = A^T D \quad (1)
\]

where \( A \in \mathbb{R}^{n \times m} \) is the one-hot coding gene-isoform relational data matrix and \( D \in \mathbb{R}^{n \times d} \) stores gene-disease associations. Based on the obtained 12,663 genes from multiple RNA-seq datasets, we collected 1,018 diseases associated with these 12,663 genes from Disgenet \[50\]. To remove sparse disease tags, we further selected 732 tags, each of which is annotated to at least 100 genes for experiments.

These initial disease annotations may mislead the model to learn the common features and thus not helpful to differentiate isoform-disease associations. In practice, some isoform function prediction methods (i.e., iMILP \[20\] and DeepIsoFun \[21\]) also initialize isoform annotations using their genes’ annotations in a similar way, and then kick off multi-instance label propagation or deep learning to predict...
the isoform-level annotations. DeepIsoFun and iMIILP force the model to predict at least one isoform as positive if the gene is labeled as positive. In contrast, DeepIDA updates the initial disease annotations of isoforms by leveraging isoform-level feature information. Our motivation is that isoforms spliced from the same genes have different expression profiles across RNA-seq datasets, and different genes are annotated with diverse GO terms, the initial isoform annotations of two similar isoforms spliced from different genes can be also quite different. As such, these initial annotations can be corrected by model training. As illustrated in Fig. 4, three gene bags include diverse isoforms, the isoforms in each bag are separately initialized with the same positive/negative annotations. To minimize the training loss, some initial positive/negative annotations of isoforms will be updated/predicted as negative/positive ones during model training/prediction. In this way, DeepIDA can differentiate the disease annotations of individual isoforms. In addition, deep learning models can also absorb some noisy labels for training [51]. Given that, we initialize the isoform disease annotations using annotations of hosting genes.

The similar strategy can be applied to dispatch the above preprocessed gene-level GO data and miRNA-target data to isoform-level also as follows:

\[
G = A^T H, \quad M = A^T I
\]  

where \(H \in \mathbb{R}^{n \times g}\) and \(I \in \mathbb{R}^{n \times h}\) are gene-level GO data and miRNA-target data for \(n\) genes, \(g\) and \(h\) are the cardinality of GO terms and miRNA targets. \(G \in \mathbb{R}^{m \times g}\) and \(M \in \mathbb{R}^{m \times h}\) are dispatched GO data and miRNA-target data at the isoform-level, respectively. The following Section elaborates on how DeepIDA makes the prediction of IDAs.

### 3 The Proposed DeepIDA

#### 3.1 Overview

To fuse the preprocessed transcriptomics and genomics data, DeepIDA adopts two deep neural networks (DNNs) with different structures, which are deployed in parallel to synchronously extract features from two types of data and eventually merged by a fully connected layer with 256 neurons to integrate the data. The output of the above fully connected layer is used as the input of a decision layer with a single neuron and Sigmoid activation function, which finally predicts IDAs.

### 3.2 Imbalanced Loss Function

Isoform-disease association prediction is an extreme class imbalance problem. In our collected data, the ratio between nonpathogenic genes and pathogenic ones are 300 to 1, which brings a great challenge to the correct prediction of IDAs. Under such imbalanced data, the length of the minority class gradient component is much smaller than that of the majority class. In other words, the majority class is essentially dominating the network gradient that is responsible for updating the model’s weights. This reduces the error of the majority group very quickly during early iterations, but often increases the error of the minority group and causes the network to get stuck in a slow convergence mode. In deep learning, there are two typical ways to solve the class imbalance problem, one is the data-level technology, and the other is balancing the weight of the positive and negative classes. The data-level technique refers to obtaining balanced data by sampling, such as undersampling and oversampling. But the distribution of sampled data and of real data are actually different. The second technique changes the weight of each training sample to balance the loss of each class. The most common way is setting the weight of a class as the inverse class frequency. But choosing the right class weights is complicated and adopting simple inverse class frequency weight is not always useful. Given that, we adopt the focal loss [26] instead of the standard binary cross entropy to address this challenge.

The focal loss function is a dynamic scaling cross-entropy loss function, it down-weights the loss assigned to easy-to-classify samples, and guides the model more focused on the hard-to-classify samples [26]. It is an improvement to the standard cross-entropy loss function, which is defined as follows:

\[
L = -z_{ij}\log \hat{z}_{ij} - (1 - z_{ij})\log (1 - \hat{z}_{ij})
\]  

\[
L = \begin{cases} 
- \log \hat{z}_{ij}, & z_{ij} = 1, \\
-\log (1 - \hat{z}_{ij}), & z_{ij} = 0.
\end{cases}
\]

where \(z_{ij} \in \{0, 1\}, 1 \leq i \leq m, 1 \leq j \leq d\) stores the association between the \(i\)-th isoform and the \(j\)-th disease, \(\hat{z}_{ij} \in [0, 1]\) is the predicted output after the Sigmoid activation function. It can be inferred from Eq. 4 that the loss decreases as \(\hat{z}_{ij}\) becomes smaller when \(z_{ij} = 0\), and the loss decreases as \(\hat{z}_{ij}\) becomes larger when \(z_{ij} = 1\). Since the standard cross-entropy loss function gives the same weight to the positive and negative examples, a large number of easy-to-classify negative samples gives a considerable loss and dominates the direction of gradient descent. The effect of a few positive samples becomes trivial. This eventually causes the loss function to converge slowly during the iteration and stuck to a local optimum.

To address the above issue, focal loss function first adds a balance factor \(\alpha (0 < \alpha < 1)\) to balance the weights of positive and negative classes as follows:

\[
L_{fl} = \begin{cases} 
-\alpha \log \hat{z}_{ij}, & z_{ij} = 1, \\
-(1 - \alpha)\log (1 - \hat{z}_{ij}), & z_{ij} = 0.
\end{cases}
\]

Here \(\alpha\) only balances the importance of positive/negative classes, it does not differentiate between easy/hard samples.
Thus, the focal loss function introduces a modulating factor $\gamma$ ($\gamma > 0$) to reduce the loss caused by easy-to-classify samples and guide the model to more focus on difficult and mis-classified samples.

$$L_{fl} = \begin{cases} -(1 - \tilde{z}_{ij})^\gamma \log \tilde{z}_{ij}, & z_{ij} = 1, \\ -\tilde{z}_{ij}^\gamma \log(1 - \tilde{z}_{ij}), & z_{ij} = 0. \end{cases}$$

When a positive sample is mis-classified and $\tilde{z}_{ij}$ is small, the value of $(1 - \tilde{z}_{ij})^\gamma$ is near 1 and the loss is almost unaffected. As $\tilde{z}_{ij} \to 1$, the value of $(1 - \tilde{z}_{ij})^\gamma$ goes to 0 and the loss for easy-to-classify samples is down-weighted. The situation is similar if a negative sample is mis-classified. It can be clearly seen that the introduction of $\gamma$ can exponentially reduce the loss caused by a large number of well-classified samples. When $\gamma = 0$, the focal loss function is equivalent to cross-entropy loss function, and the effect of the modulating factor is also increased as $\gamma$ increase. The complete focal loss function is shown below.

$$L_{fl} = \begin{cases} -\alpha(1 - \tilde{z}_{ij})^\gamma \log \tilde{z}_{ij}, & z_{ij} = 1, \\ -\alpha \gamma \tilde{z}_{ij}^\gamma \log(1 - \tilde{z}_{ij}), & z_{ij} = 0. \end{cases}$$

In this way, we can solve the class imbalance problem through the coordination of $\alpha$ and $\gamma$. Parameter sensitivity analysis of $\alpha$ and $\gamma$ is reported in the Supplementary file.

### 3.3 Network structure and activation function

DeepIDA consists of two DNN sub-networks. One is used to extract features from nucleotide sequences $S \in \mathbb{R}^{m \times e}$ and isoform expression data $E \in \mathbb{R}^{m \times f}$ with respect to isoforms, and the other is used to capture features from GO $G \in \mathbb{R}^{n \times g}$ and miRNA targets $M \in \mathbb{R}^{n \times h}$ with respect to genes. $e$ and $f$ are the dimensions of nucleotide sequences and isoform expression data, and $g$ and $h$ are the number of GO terms and miRNAs, respectively. The specific structure of DeepIDA is shown in Fig. 2(b).

Since the initial dimensions of the two types of data are different (one is 6,923 and the other is 554), the numbers of neurons in each layer in two sub-networks are also different. In fact, after the dimension of the input layer and that of the output layer are determined, how to choose the number of hidden layers and the number of neurons in each layer is still a difficult problem. If the number of hidden layer neurons is too small, the network cannot have the necessary learning and information processing capabilities. Conversely, if too many, the number of parameters will explode and the network structure will become too complicated, which slows down the learning rate of the model. Therefore, DeepIDA adopts a mode of gradually decreasing the number of neurons exponentially in each layer. For GO data and miRNA target data, a dense layer with 1,024 neurons is used to capture the underlying features, the batch normalization (BN) [52] and dropout [53] are used to accelerate convergence and handle the over-fitting problem. Then dense layers with fewer neurons are used to extract more abstract features. For sequence and expression data of isoforms, a dense layer with 512 neurons is used to capture the underlying features, since there are only 554-dimensional features. When both sub-networks project features to 256 dimensions, the two sub-networks are merged at a concatenation layer.

The activation functions of all the above layers are ReLU function. ReLU is a piecewise linear function through which all positive values remain unchanged and all negative values become 0. This unilateral inhibition makes the features extracted by ReLU more representative, which enables the model having a stronger generalization ability. The ReLU function can be formalized as:

$$ReLU(x) = \begin{cases} x, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

The output of each layer in the two sub-networks can be summarized as:

$$X^1_p = ReLU(w_p X^1_{p-1} + b_p)$$
$$X^2_q = ReLU(w_q X^2_{q-1} + b_q)$$

where $p$ and $q$ are the layer of two sub-networks, $w$ and $b$ are the weight and bias. $X^1$ and $X^2$ are initialized as $[S; E]$ and $[G; M]$, respectively.

Suppose $X^1_c$ and $X^2_c$ record the output of the two sub-networks, then we can obtain the concatenated features $C$ as follows:

$$X_c = [X^1_c, X^2_c]$$

Consequently, we can obtain the output of the above concatenation layer as:

$$Y = ReLU(w_c X_c + b_c)$$

$Y \in \mathbb{R}^{256}$ is used as the input of the decision layer with a single neuron and Sigmoid activation function as follows:

$$Z = Sigmoid(w_d Y + b_d)$$

where $w_d$ and $b_d$ are the weight and bias of the decision layer, they convert the prediction of isoform-disease associations into a probability estimation $Z \in [0, 1]$.

It is critical to choose the right optimizer for a deep learning model. Traditional gradient descent optimizers, such as SGD (Stochastic Gradient Descent) optimizer, BGD (Batch Gradient Descent) optimizer and MBGD (Mini-Batch Gradient Descent) optimizer, are susceptible to the impact of the learning rate and oscillate near the local extreme point, resulting in a slow convergence. Here, DeepIDA adopts the Adam optimizer [54], which computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. Adam optimizer can adaptively adjust model parameters from two angles: gradient mean and gradient square, instead of directly determining parameters from the current gradient, which is adopted by traditional gradient descent optimizers. Adam optimizer can automatically adjust the learning rate, the choice of the initial learning rate is not so stringent as other optimizers.

### 4 Experimental Results and Analysis

In this section, we study the performance of DeepIDA and compare it against related methods. In addition, we investigate the effectiveness of DeepIDA and that of compared methods on case IDAs, along with the ablation study.
4.1 Experimental Setup

We adopt five-fold cross validation to evaluate the performance of DeepIDA and of compared methods on the preprocessed 31,408 isoforms spliced from 12,663 genes, which are associated with 732 diseases (as we introduced in Section 2). The initial disease annotations of isoforms were inherited from their hosting genes, as formulated in Eq. 1. These annotations are partially reliable, given the incompleteness and intrinsic uncertainty of gene-disease associations. There are 456,273 associations in the positive training set and 22,534,383 associations in the negative training set. It is easy to find that the class imbalance exists in our following experiments. In order to evaluate the prediction results under class imbalance, we abandoned the typical Accuracy and AUROC evaluation metrics, which may mislead the analyst with high scores that incorrectly indicate a good performance. Given a binary dataset with a positive class distribution of 1%, a naive learner that always outputs the negative class label for all inputs has an Accuracy of 99%. So we adopt the evaluation metrics, Precision, Recall, F-Measure, G-Mean, Balanced Accuracy and AUPRC (area under the precision-recall curve), which are more sensitive to class imbalance. Given the lack of ground-truth labels on the isoforms level, we follow the typical surrogate evaluation in the previous studies [20-22] that aggregates the isoform-disease associations to their hosting genes, whose associations with diseases are known, to evaluate the prediction from the isoform level.

We adopt the Keras framework, a high-level neural networks API written in Python to quickly and efficiently deploy our model. We use a combination of experience guidance and grid optimization to determine deep network model parameters on the training set. Partial neural network parameters are given in Figure 2b. The grid search range and finally adopted input values of parameters are given in Table S1 of the Supplementary file. Based on the parameter sensitivity analysis of $\gamma$ and $\alpha$ (used for the focal loss function in Eq. 7) in the Supplementary file, DeepIDA adopts $\gamma = 2$ and $\alpha = 0.8$ for experiments.

To the authors’ knowledge, there is no public computational solution that targets for predicting IDAs. To make a comprehensive evaluation, we compare the performance of DeepIDA against two representative multi-instance learning based isoform function prediction methods (iMILP [20] and DisoFun [8]), a classical gene-disease association prediction approach (Prince) [33], five typical classification methods. (i) iMILP models the gene-isoform relationships as multiple instance data and develops a novel multiple instance-based label propagation to predict functions of isoforms. In this paper, we adopt it for predicting isoform-disease associations using gene-disease associations and isoform expression data. (ii) DisoFun collaboratively factorizes the isoform expression data matrix and gene functional annotation data matrix, along with the gene interaction data, to infer the key isoforms and their functional annotations, and then transfers inferred annotations to other isoforms to achieve isoform function prediction. Here, we adopt it for predicting isoform-disease associations using the gene-disease associations, data of isoforms and genes.

(iii) Prince is a global network-based method for prioritizing genes and protein complexes for a disease of interest. (iv) Weighted K-Nearest Neighbor (WKNN) predicts the diseases associated with a target isoform based on diseases associated with its k nearest isoforms by majority voting. The weighted KNN gives larger weights to closer isoforms and smaller weights to farther ones based on the Euclidean distance between them. (v) Support Vector Machine (SVM) optimizes a decision hyperplane that best separates the positive and negative classes with maximum margins, it is widely adopted in computational biology. (vi) Logistic Regression (LR) is another widely used classifier for binary classification, it adopts a Sigmoid function to map the prediction of linear regression into $[0, 1]$. (vii) Random Forest (RF) is an ensemble learning method that builds a forest of decision trees. Each tree in RF is a separate weak model that is trained with a randomly selected subset and each node of the tree is split by random selection of features, which allows RF to create many less correlated weak models. RF then combines these trees to generate a robust and stable model. (viii) eXtreme Gradient Boosting (XGBoost) [34] is a boosting algorithm with additive model, it adopts the greedy mechanism to generate new base tree model to remedy the bias of the previous models, and thus progressively reduce the bias of the overall model.

We use grid search to find the optimal parameters of the above methods against the performance on the training set. The grid search ranges and selected values of input parameters are given in Table S2 of the Supplementary file.

4.2 Performance comparison with other methods

Based on the above experimental setups, we apply five-fold cross validation to study the performance of DeepIDA and compare it against eight compared methods. The overall results are shown in Table 1.

DeepIDA gives better results than the compared methods across most evaluation metrics. This prominent improvement shows that DeepIDA can more reliably predict the associations between isoforms and diseases by fusing nucleotide sequence, RNA-seq data, miRNA-target and GO data. In fact, most of typical classifiers also adopt the fused gene and isoform data, and present good prediction results, which again confirm that the features we chose are closely related with the prediction of IDAs. Both iMILP and DisoFun model the gene-isoform associations during the prediction of isoform-disease associations, they still lose to SVM. That is because iMILP only uses isoform expression data and DisoFun depends on the identification of key isoforms. Similarly, Prince mainly uses the gene-level data and thus is outperformed by iMILP and DisoFun by a large margin, and not to mention DeepIDA.

When typical classifiers encounter imbalanced data, the minority class will be more biased toward the majority one, due to the influence of sample distribution. In disease isoform prediction, the cost of misdianosing the disease isoform is very high. Therefore, in SVM and LR, the value of the penalty coefficient $C$ is relatively large to prevent the model from mis-classifying such isoforms. We can find that WKNN, SVM and LR have good prediction results, but the two ensemble


<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>XGBoost</th>
<th>DeepIDA</th>
<th>iMILP</th>
<th>Prince</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.9861 ± 0.0007</td>
<td>0.9861 ± 0.0007</td>
<td>0.9861 ± 0.0007</td>
<td>0.9861 ± 0.0007</td>
<td>0.9861 ± 0.0007</td>
</tr>
<tr>
<td>Recall</td>
<td>0.7909 ± 0.0073</td>
<td>0.8210 ± 0.0140</td>
<td>0.8210 ± 0.0001</td>
<td>0.5522 ± 0.0130</td>
<td>0.5864 ± 0.0001</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.8955 ± 0.0050</td>
<td>0.8955 ± 0.0080</td>
<td>0.8955 ± 0.0001</td>
<td>0.7112 ± 0.0180</td>
<td>0.7220 ± 0.0001</td>
</tr>
<tr>
<td>G-mean</td>
<td>0.8584 ± 0.0041</td>
<td>0.9012 ± 0.0001</td>
<td>0.8943 ± 0.0001</td>
<td>0.7260 ± 0.0001</td>
<td>0.7260 ± 0.0001</td>
</tr>
<tr>
<td>Balanced Acc</td>
<td>0.8653 ± 0.0035</td>
<td>0.8943 ± 0.0001</td>
<td>0.8943 ± 0.0001</td>
<td>0.7260 ± 0.0001</td>
<td>0.7260 ± 0.0001</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.8764 ± 0.0070</td>
<td>0.8812 ± 0.0001</td>
<td>0.8423 ± 0.0001</td>
<td>0.8828 ± 0.0001</td>
<td>0.8470 ± 0.0001</td>
</tr>
</tbody>
</table>

(1) DeepIDA's prediction divergence

(a) DeepIDA's prediction divergence

(b) DeepIDA vs. iMILP on prediction divergence

We analyzed the predicted association divergence from the disease aspect. Particularly, for each disease we computed the standard deviation of predicted isoform-disease association probabilities of a multi-isoform gene, and further converted the deviation smaller than 0.01 into 0. We took this standard deviation as the predicted association divergence and each multi-isoform gene has a divergence value. Next, we use boxplot to visualize the divergence values of 7,897 multi-isoform genes with respect to three groups of diseases in Fig. 5(a). The first group includes three most frequent diseases (acinar cell carcinoma (DOID:3025), breast carcinoma (DOID:3459) and major depressive disorder (DOID:1470)), which are associated with the largest numbers of genes; the second group includes three moderate frequent diseases (pterygium (DOID:0002116), multiple endocrine neoplasia type 2A (DOID:0050430) and attention deficit hyperactivity disorder (DOID:1094)) that are associated with a median number of genes; and the third group includes three least frequent diseases (multiple endocrine neoplasia type 1 (DOID:10017), Huntington’s disease (DOID:12858) and ovarian disease (DOID:1100)) that are associated with the smallest number of genes. There are 2,435, 781, 179 non-zero values of these free group respectively. We find the more frequent the disease, the larger the prediction divergence among the isoforms is. That is because more genes/isoforms are associated with the more frequent diseases.

For reference and comparison, we also computed the predicted association divergence of iMILP in a similar way. Fig. 5(b) presents the boxplots of DeepIDA and iMILP with respect to the aforementioned 9 diseases. Both DeepIDA and iMILP use the gene-level annotations to initialize the isoform-level annotations. Unlike iMILP, DeepIDA does not force the model to predict at least one isoform as positive if the gene is labeled as positive, it still obtains larger prediction divergences than iMILP by fusing isoform-level features.

In our used dataset, there are 7,897 multi-isoform genes.

learning methods (RF and XGBoost) are less effective. This is mainly because both RF and XGBoost adopt random row sampling (sample sampling) and column sampling (feature sampling) for base classifiers. After random row sampling, most base classifiers still face the problem of class imbalance and there are fewer samples in the training subset. In fact, the problem of class imbalance in the training subset is usually more severe than that in the original data set, which ultimately causes the model bias toward the majority class. The class imbalance usually leads to the skew of feature distribution, which also affects the random column sampling and leads to a degenerated performance. It is interesting to find that XGBoost has the highest Precision but lowest Recall. That is due to the known dilemma between Precision and Recall, of data sources and the combinations of them, the input value other compared methods under different threshold values.

Supplementary file. DeepIDA again consistently outperforms methods with the threshold as 0.5, 0.7 or 0.8 in Table S3 of the reported the results of DeepIDA and those of other compared methods with the threshold as 0.5, 0.7 or 0.8 in Table S3 of the Supplementary file. DeepIDA again consistently outperforms methods with the threshold as 0.5, 0.7 or 0.8 in Table S3 of the Supplementary file.

To quantify the performance using these metrics, we need to convert the predicted isoform-disease association probability (Z) into binary. For the results in Table 1, we simply adopted a threshold as 0.6 for all compared methods. We also reported the results of DeepIDA and those of other compared methods with the threshold as 0.5, 0.7 or 0.8 in Table S3 of the Supplementary file. DeepIDA again consistently outperforms other compared methods under different threshold values. In addition, we also studied the impact of different types of data sources and the combinations of them, the input value of k for k-mer conjoint triad. Due to page limit, we provided the results and analysis in the Supplementary file. Overall, the results show that these data sources are complementary for each other.

In our used dataset, there are 7,897 multi-isoform genes.
and minimizing the focal loss. The average divergence of DeepIDA is 0.0450, while that of iMILP is 0.0408. Given that, we conclude that DeepIDA can differentiate the disease annotations of individual isoforms spliced from the same gene.

### 4.3 Case Study

Given the scarcity of well organized or stored isoform-disease associations, it is hard to count the unique associations of the same gene and of the spliced isoforms. To further investigate the reliability of DeepIDA in predicting novel isoform-disease associations, we did a manual literature search and collected eight genes with known isoform-disease associations with respect to five types of diseases, and listed the details in Table 2. Following the experimental setup in the previous subsection, we applied DeepIDA and those compared methods to predict the associations between sixteen isoforms spliced from these genes, and reported the results in Table 2.

We observe that out of 16 IDs in the fourth column of Table 2, DeepIDA correctly identifies 13 of them, which is significantly more accurate than the other methods. SVM, RF, iMILP all correctly predict 12 IDs, while WKNN, LR, and XGBoost all make 11 IDs, respectively. These reported cases suggest that the predictions of DeepIDA are more consistent with the evidences from biomedical literature. In addition, for the prediction of IDs, we are more concerned about how many pathogenic samples are correctly detected, so we also reported the Recall for each method. DeepIDA achieves a Recall of 90.91%, outperforms the second best methods (iMILP and RF) by 9.09%. These case studies again confirm the effectiveness of DeepIDA on identifying IDs. We observe that DeepIDA can more credibly predict the isoform-level disease associations than iMILP and Disofun. This study further proves the effectiveness of DeepIDA to differentiate isoforms associations with diseases.

### 4.4 Ablation Study

To investigate the impact of the focal loss function and network structure on the prediction results, we also introduced two variants of DeepIDA: DeepIDA-without focal loss and DeepIDA-without sub-networks. DeepIDA-without focal loss adopts the same network structure as DeepIDA but uses the cross-entropy loss function. DeepIDA-without sub-networks utilizes the focal loss function but concatenates all features together with one network to directly extract features, instead of using two sub-networks with different structures to separately extract the features from multi-type data.

DeepIDA has a largely superior performance to its variants. This is mainly because in the extremely unbalanced disease data, if DeepIDA uses the standard cross-entropy loss function, the positive and negative samples will be given with same weight and the loss of a few positive samples will be overwhelmed by a large number of negative samples, which eventually causes the model failing to detect the pathogenic isoforms. In addition, if the model does not take the encoding methods and preprocessing operations of different data into consideration, it will not be able to effectively capture key features. This observation supports our motivation to adopt focal loss function to address class imbalance problem and two sub-networks to extract important features from different types of biological data.

### 5 DISCUSSION AND CONCLUSION

Although the evidences that isoforms trigger diverse complex diseases are fast growing, the large-scale computational prediction of isoform-disease associations has not been well studied yet, mainly due to the lack of known isoform-disease associations. To bridge this gap, we introduce DeepIDA that leverages deep learning and data fusion. DeepIDA extracts and fuses the complementary information of different types of biological data by two subnets, and handles the intrinsic class-imbalance issue of predicting IDs by focusing more on the hard-to-classify samples. Our extensive empirical study confirms the advantage of DeepIDA to other related representative methods, proves the necessity of fusing different types of data, and the rationality of adopting two subnets and accounting for the class-imbalance.

Our DeepIDA depends on the initial disease annotations of isoforms induced from gene-level. We will study how to collaboratively leverage the incomplete disease annotations of genes and the inter-relations between genes and isoforms to boost the prediction of IDs.

### REFERENCES


<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Isiform</th>
<th>Literature evidence</th>
<th>WKNN</th>
<th>SVM</th>
<th>LR</th>
<th>RF</th>
<th>XGBoost</th>
<th>Disofun</th>
<th>tMLP</th>
<th>DeepIDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s disease</td>
<td>APOE</td>
<td>APOE2</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>APOE3</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>APOE4</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Idiopathic pulmonary fibrosis</td>
<td>ROCK</td>
<td>ROCK1</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROCK2</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Restless legs syndrome</td>
<td>MEIS1</td>
<td>MEIS1-001</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Cardiomyopathy</td>
<td>VEGF-A</td>
<td>VEGFA-001</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>VEGF-B</td>
<td>VEGFB-001</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>VEGF-C</td>
<td>VEGFC-001</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Colitis</td>
<td>HNF4A</td>
<td>HNF4A-002</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HNF4A-006</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HNF4A-004</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HNF4G-001</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td>68.75%</td>
<td>75.00%</td>
<td>68.75%</td>
<td>75.00%</td>
<td>68.75%</td>
<td>75.00%</td>
<td>81.25%</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td></td>
<td></td>
<td>72.73%</td>
<td>81.82%</td>
<td>72.73%</td>
<td>81.82%</td>
<td>63.64%</td>
<td>72.73%</td>
<td>81.82%</td>
<td>90.91%</td>
</tr>
</tbody>
</table>


[34] T. Chen and C. Guestrin, “Xgboost: A scalable tree boost-


Guoxian Yu is a Professor at the School of Software, Shandong University, Jinan, China. He received the Ph.D. in Computer Science from South China University of Technology, Guangzhou, China in 2013. His current research interests include data mining and bioinformatics. He serves as reviewers for KDD, ICDM, IJCAI, AAAI, TKDE, TNNLS, TCBB, Bioinformatics and other prestigious conferences and journals.

Yeqian Yang is an MEng. student at the School of Software, Shandong University, Jinan, China. She received B.Sc. degree in Digital Media from Qingdao University, Qingdao, China in 2020. Her research interests include data mining and bioinformatics.

Yangyang Yan is an Engineer in UnionPay, Shanghai, China. He received M.Eng. in Computer Technology from Southwest University, Chongqing, China in 2020, and B.Sc degree in Software Engineering from Nantong University, Nantong, China in 2017. His research interests include machine learning and bioinformatics.

Maozu Guo is a professor at the College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China. He received the Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology. His research interests include bioinformatics, machine learning, and data mining.

Xiangliang Zhang is an Associate Professor and directs the Machine Intelligence and Knowledge Engineering (MINE) Laboratory in King Abdullah University of Science and Technology (KAUST). She earned her PhD degree in Computer Science with great honors from INRIA-University Paris-Sud 11, France, in 2010. Her main research interests and experiences are in diverse areas of machine learning and data mining.

Jun Wang is a Professor with the Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University. She received B.Sc. degree in Computer Science, MPhil degree in Computer Science and Ph.D. in Artificial Intelligence from Harbin Institute of Technology, Harbin, China in 2004, 2006 and 2010, respectively. Her current research interests include data mining and their applications in bioinformatics.