Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes

Chen Peng, Yang Zheng, and De-Shuang Huang

Abstract—Breast cancer is one of the most common cancers all over the world, which bring about more than 450,000 deaths each year. Although this malignancy has been extensively studied by a large number of researchers, its prognosis is still poor. Since therapeutic advance can be obtained based on gene signatures, there is an urgent need to discover genes related to breast cancer that may help uncover the mechanisms in cancer progression. We propose a deep learning method for the discovery of breast cancer-related genes by using Capsule Network based Modeling of Multi-omics Data (CapsNetMMD). In CapsNetMMD, we make use of known breast cancer-related genes to transform the issue of gene identification into the issue of supervised classification. The features of genes are generated through comprehensive integration of multi-omics data, e.g., mRNA expression, z scores for mRNA expression, DNA methylation, and two forms of DNA copy-number alterations (CNAs). By modeling features based on the capsule network, we identify breast cancer-related genes with a significantly better performance than other existing machine learning methods. The predicted genes with prognostic values play potential important roles in breast cancer and may serve as candidates for biologists and medical scientists in the future studies of biomarkers.

Index Terms—Multi-omics data, capsule network, prediction of cancer-related genes, machine learning, breast cancer

1 Introduction

There are more than 1,300,000 persons around the world suffering from breast cancer each year, which result in greater than 450,000 deaths [1]. As one of the most common cancers, breast cancer is extensively studied by a large number of researchers [2], [6]. Although the diagnosis and treatment of breast cancer are greatly advanced such as the identification of subtype-associated predictors including estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) [1], [7], the prognosis of this malignancy is still poor. Since therapeutic advance can be obtained based on gene signatures [1], there is an urgent need to discover genes related to breast cancer and utilize them to help uncover the mechanisms in the progression of this cancer, thus making an improvement in its therapy.

With the help of known breast cancer-related genes, the discovery of novel genes can be transformed into an issue of supervised classification by regarding various biological characteristics of genes as their features. Common classification methods include classical classifiers such as Neural Network (NN), Support Vector Machine (SVM), Adaboost and K-Nearest Neighbors (KNN), which are widely used in bioinformatics for a long time. For example, Lancashire et al. identify gene transcript signatures predictive for lymph node status and estrogen receptor in breast cancer by means of a stepwise method using artificial neural networks [8]. Guyon et al. utilize SVM with recursive feature elimination to build a new method for genetic diagnosis and drug discovery [9]. Guan et al. present an Adaboost-based prediction tool mirExplore, which can detect miRNAs from both next generation sequencing and genome data [10]. Okun et al. explore the data complexity in cancer classification by taking advantage of ensembles of KNN [11]. Although these methods usually serve as convenient tools in solving bioinformatics problems, they may not be the best choice for the increasingly larger biological data and their performance also need to be improved.

As the development of computer science, many novel classification methods emerge to face the challenges of machine learning brought by enormous data, among which eXtreme Gradient Boosting (XGBoost) [12] and a new GBDT algorithm with GOSS and EFB, i.e., LightGBM [13], are famous due to the state-of-the-art results they achieve in tackling these challenges. Specifically, XGBoost is a scalable end-to-end tree boosting system with a novel sparsity-aware algorithm and a weighted quantile sketch [12]. LightGBM is a gradient boosting decision tree implemented with the techniques of gradient-based one-side sampling and exclusive feature bundling [13]. Recently, there are already some researches in the field of bioinformatics making use of these two methods. As regard to XGBoost, Zhong et al. predict essential proteins with the help of an XGBoost-based framework [14] and

- *The authors are with the Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, No. 4800 Caoan Road, Shanghai 201804, China.*
E-mail: {pcll, zy_yang6354, dshuang}@tongji.edu.cn.

Manuscript received 14 Oct. 2018; revised 22 Feb. 2019; accepted 29 Mar. 2019. Date of publication 9 Apr. 2019; date of current version 7 Oct. 2020.
(Corresponding author: Chen Peng.)
Digital Object Identifier no. 10.1109/TCBB.2019.2909905

Zheng et al. establish miRNA classification model through the combination of SVM and XGBoost [15]. Meanwhile, Wang et al. apply LightGBM to the classification of miRNA in breast cancer patients, which achieves a good performance [16]. In spite of the achievement in disease researches brought by new machine learning methods, these studies do not take full advantage of high-throughput experimental data, which may provide comprehensive and valuable biological information. For example, The Cancer Genome Atlas (TCGA) [1], [17], [18] and International Cancer Genome Consortium (ICGC) [19], [21], which are promoted by the rapid development of DNA sequencing technology, provide massive experimental data of different kinds of cancer in multiple omics such as genomics, epigenomics and transcriptomics [17], [19]. Analysis of these useful information based on effective machine learning methods will be beneficial to the diagnosis, treatment and prevention of cancer.

In this study, we propose a deep learning method for the discovery of breast cancer-related genes by using Capsule Network based Modeling of Multi-omics Data (CapsNetMMD). Capsule network is a novel network structure, which is first put forward in the field of image recognition [22]. It has not yet been applied in bioinformatics, let alone the identification of cancer-related genes. In CapsNetMMD, multi-omics data, e.g., mRNA expression, z scores for mRNA expression, DNA methylation and two forms of DNA copy-number alterations (CNAs) are fully integrated to generate feature matrixes of genes. Then known breast cancer-related genes are incorporated to transform the issue of gene identification into the issue of supervised classification. The evaluation results on several measurements show that CapsNetMMD can achieve the best performance when compared with other machine learning methods, which indicate that the settings of instantiation parameters and dynamic routing mechanism in capsule network are suitable for the discovery of cancer-related genes. The prognostic values of the genes predicted by CapsNetMMD are also explored in the subsequent survival analysis, which may not only corroborate the effectiveness and superiority of CapsNetMMD, but also imply the potential important roles of the genes in the study of breast cancer.

2 Methods and Materials

2.1 Multi-omics Data for Breast Cancer

The multi-omics data utilized in this study are derived from TCGA project, which produces tremendous data of cancer genomics. Specifically, we download the data of breast cancer including mRNA expression, DNA methylation and CNA from an open platform: The cBioPortal for Cancer Genomics (http://cbioportal.org) [23], [24], which makes large-scale raw data provided by TCGA more directly and easily available to researchers. Among these data, DNA methylation consists of methylation beta-values that indicate the methylation levels of CpG loci by calculating the intensity ratio between unmethylated and methylated alleles [24]. MRNA expressions obtained from RNA sequencing data are processed via RNA-Seq by Expectation Maximization (RSEM) [25]. CNA data has two forms, i.e., relative linear copy-number values and discrete copy-number calls. In the latter form, $-2, -1, 0, 1$ and 2 respectively represent homozygous deletion, hemizygous deletion, neutral, gain and high level amplification [24]. Besides, z scores for mRNA expression are also used to increase the variety of data. These scores are precomputed for each sample through comparing a gene's expression value to the typical expression for this gene, which is represented by the distribution in a reference population such as expression data in normal tissues [24]. In order to perform survival study, we further download clinical data from cBioPortal, which provides various clinical details about patient samples, e.g., survival and age at diagnosis [24]. As regard to mRNA expression, expression z scores, methylation, CNA and linear CNA, we extract the common 770 patient samples and 10,462 genes in these five types of data for follow-up studies. The information of overall survival for common patient samples is also extracted.

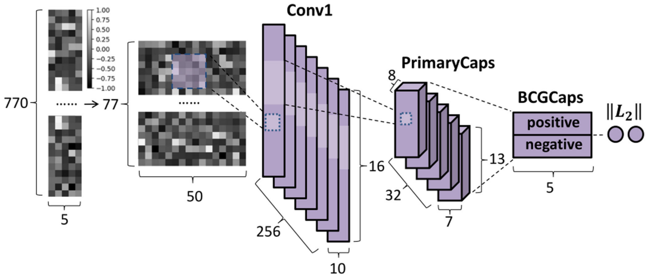

Fig. 1. The structure of capsule network based model for discovery of breast cancer-related genes.

2.2 Known Genes Related to Breast Cancer

To extract known genes related to breast cancer, we resort to a public platform named DisGeNET (http://www.disgenet.org), which collects one of the largest amount of genes and variants that related to human diseases based on scientific literature, GWAS catalogues, expert curated repositories and animal models [26], [27]. The latest version of DisGeNET (v5.0) is used in this study including 561,119 gene-disease associations between 20,370 diseases and 17,074 genes [27]. By searching keywords "Breast Carcinoma", we obtain 4,572 gene-breast cancer associations and the genes in these associations are further regarded as known genes related to breast cancer.

2.3 Capsule Network Based Modeling of Multi-omics Data for Predicting Breast Cancer-Related Genes

In this study, we propose a deep learning method for discovery of breast cancer-related genes by using capsule network based modeling of multi-omics data. The structure of the capsule network based model shown in Fig. 1 mainly consists of two convolutional layers and one fully connected layer [22]. First, for a given gene i, the data of its expression, expression z score, methylation, discrete and linear CNA are respectively normalized ($[0, 1]$ for expression and methylation, $[-1, 1]$ for the others) into five vectors by the following equation:

$$y_{i,norm} = \begin{cases} y_{i,\min} + (x_i - x_{i,\min}) \times \frac{y_{i,\max} - y_{i,\min}}{x_{i,\max} - x_{i,\min}} & \text{if } y_{i,\max} \neq y_{i,\min} \\ y_{i,\min} & \text{if } y_{i,\max} = y_{i,\min}. \end{cases} \tag{1}$$

Then these vectors are integrated to generate a feature matrix with size 770 × 5. For better modeling in the following process, we further reshape the feature matrix into 77 rows and 50 columns. Afterwards, the feature matrix will be referred to as input for the first convolutional layer, which consists of 256 convolution kernels with size 5 × 5 and a stride of 5. The activation function of this layer is Rectified linear units (ReLU) [28]:

$$f(x) = \max(0, x) = \begin{cases} 0, & x < 0 \\ x & x \geq 0 \end{cases}. \tag{2}$$

In this procedure, the initial feature matrix of gene *i* is converted to higher-level and more abstract local features.

The second layer (PrimaryCaps) is a capsule layer that similar to the convolutional capsule layer in [22]. It is a special convolutional layer that consists of 256, 4 × 4 convolution kernels with a stride of 1. The difference between PrimaryCaps and ordinary convolutional layer is that the feature maps after convolution are further transformed to 32 channels of 8D capsules (each capsule is a vector contains 8 convolutional units) in PrimaryCaps. In a word, PrimaryCaps has totally [32 × 7 × 13] capsules and each capsule share their weights with each other in the [7], [13] grid. Besides, PrimaryCaps utilizes a non-linear squashing function [22] to ensure short vectors and long vectors respectively get shrunk to almost zero and slightly below 1. Thus the output length of a capsule represents the probability that the entity exists [22]. The squashing function is computed as follows [22]:

$$O_j = \frac{\left\|I_j\right\|^2}{1 + \left\|I_j\right\|^2} \frac{I_j}{\left\|I_j\right\|}, \tag{3}$$

where I_j is the input of capsule *j* and O_j is the output vector of capsule *j*. Besides the first layer, the input of a capsule in all other layers is calculated by the weighted sum of all "prediction vectors" $\hat{m}_{j|i}$ [22]:

$$I_j = \sum_i r_{j,i} \hat{m}_{j|i}, \tag{4}$$

in which $\hat{m}_{j|i}$ equals to the product of capsule m_i and $W_{i,j}$ as follows:

$$\hat{m}_{j|i} = W_{i,j} m_i, \tag{5}$$

where $W_{i,j}$ is the weight matrix with size 8 × 5 in this study. Coupling coefficients $r_{i,j}$ in (4) are determined by softmax transformation of $p_{i,j}$ with dynamic routing algorithm [22] and the function is shown below:

$$r_{i,j} = \frac{\exp(p_{i,j})}{\sum_k \exp(p_{i,k})}, \tag{6}$$

where $p_{i,j}$ represent the prior probabilities that capsule *i* and *j* are coupled.

The final layer (BCGCaps) has two 5D capsules C_j, which indicate the status of input gene: positive and negative, i.e., whether the input gene is related to breast cancer. The computation process from the second layer PrimaryCaps to the final layer BCGCaps is based on dynamic routing algorithm, which only exist between these two layers. As shown in Fig. 2, PrimaryCaps consists of 2,912 8D capsules (m_i) and the weighted sum of m_i can be obtained with $W_{i,j}$, $\hat{m}_{j|i}$ and $r_{i,j}$ that calculated by equation (5) and (6). Then C_j are computed through non-linear transformation using squashing activation function. Finally, based on the length of C_j, we get the probability of each gene that it is related to breast cancer.

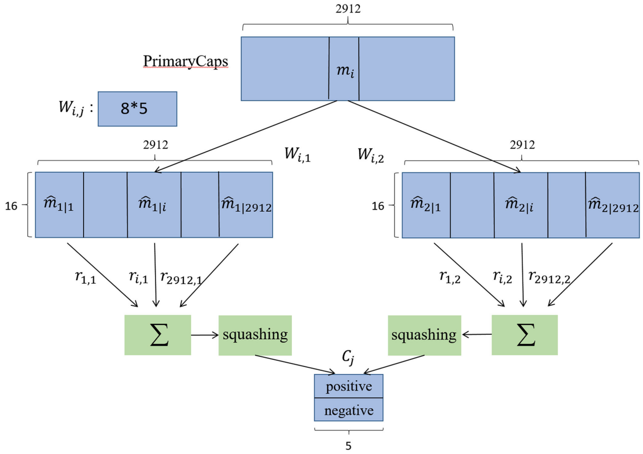

Fig. 2. Computation process from PrimaryCaps to BCGCaps.

Overall, the difference between capsule network and commonly used multi-layer convolution neural network can be described as follows. First, the basic unit of capsule network is scalar while that of commonly used multi-layer convolution neural network is vector. Second, capsule network performs dynamic routing algorithm [28] in the modeling process and in commonly used multi-layer convolution neural network only supervised classification is used. Finally, the activation function of capsule network and commonly used multi-layer convolution neural network are respectively squashing function and tanh function.

2.4 Implementation of Model

In the training procedure, we randomly extract 10 percent data from the training set each time and repeat this process 10 times to get a model until all training data has been picked. Then after 10 × *n* (*n* is set to 10 in this study) experiments, the best model with the highest AUC can be obtained. Besides, according to the suggestion of [22], the dynamic routing mechanism is implemented with three routing iterations and margin loss function, which is shown in equation (7):

$$L_k = Y_k \max(0, m^+ - \|O_k\|)^2 + \lambda(1 - Y_k) \max(0, \|O_k\| - m^-)^2, \tag{7}$$

where Y_k is the true label and $Y_k = 1$ if the input belongs to class *k*. O_k is the output of capsule *k*. The default values of m^+ and m^- are 0.9 and 0.1, respectively [22]. The parameter λ **λ** for down-weighting of the loss is set to 0.5 as recorded in [22]. The total loss of the model is the sum of the losses of *k* classes [22]. In this section, the deep learning model is implemented in TensorFlow 1.2.0, and training and test process of the model are performed on Ubuntu 16.04 LTS work station with processor Intel Xeon(R) CPU e5-2680 v2 @ 2.80 GHz × 40. The main code and data of CapsNetMMD are available from https://github.com/ustcpc/CapsNetMMD.

2.5 Performance Evaluation

To evaluate the performance of CapsNetMMD, we use five-fold cross validation in the test process, i.e., the data are

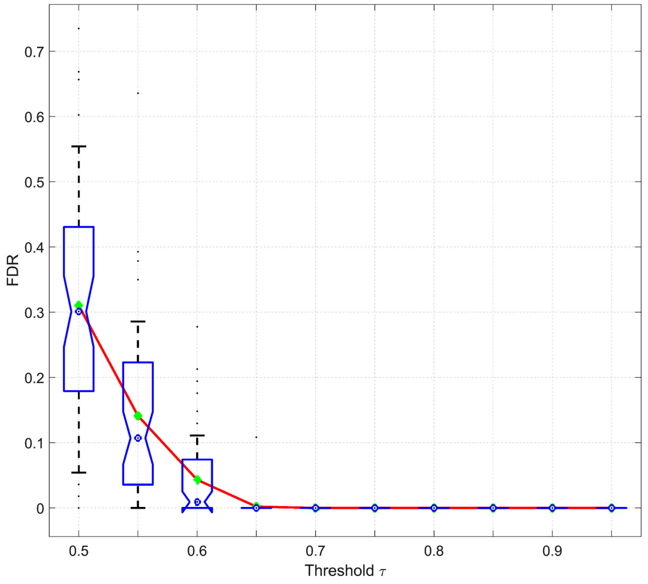

Fig. 3. Boxplot of the FDR of CapsNetMMD. Threshold τ varies from 0.5 to 0.95. The red line with green markers represents variations of the mean values of FDR.

averagely divided into five parts and every part is taken as test data in turn. The probabilities predicted by CapsNetMMD are regarded as the scores of genes and higher scores represent more relevance to breast cancer. First, the false discovery rate (FDR) of our method is estimated following the strategy in [28] and [29]:

$$\widehat{FDR}_{\tau} = \frac{N_{\tau}}{G_{\tau}}, \tag{8}$$

where G_{τ} is the number of genes whose scores are larger than threshold τ in our results. N_{τ} is the gene number in the intersection of G_{τ} and the genes identified at τ with random dataset. Specifically, at the f-th permutation, the feature matrix of gene is permuted to generate a random dataset. Then we run CapsNetMMD on this dataset to identify genes with threshold τ and calculate gene number $\tilde{N}^{(f)}{}_{\tau}$ in the intersection of G_{τ} and G_{τ} the results from random dataset. After F permutations, $\widehat{FDR}_{\tau}$ is finally given by equation (9):

$$\widehat{FDR}_{\tau} = \frac{N_{\tau}}{G_{\tau}} = \frac{\frac{1}{F}\sum_{f=1}^{F} \tilde{N}_{\tau}^{(f)}}{G_{\tau}}. \tag{9}$$

Threshold τ is set to $[0.5, 0.95]$ in this study.

In addition, after ranking the scores of genes in descending order, the top k ranked genes with the highest scores are regarded as breast cancer-related genes in our study. Known genes related to breast cancer are defined as golden standard positive (GSP) and the other genes are defined as golden standard negative (GSN). The interaction of top k ranked genes with GSP and GSN are respectively referred to as true positive (TP) and false positive (FP). The complement of TP with respect to GSP and the complement of FP with respect to GSN are respectively considered as false negative (FN) and true negative (TN). Then sensitivity (Sn) and specificity (Sp) can be computed as:

$$Sp = \frac{TN}{TN + FP} \qquad Sn = \frac{TP}{TP + FN}. \tag{10}$$

Based on these two parameters, we plot Receiver Operating Characteristic curves (ROC curves) and calculate the area under the curve (AUC). x axis and y axis in ROC curves respectively represent 1-Sp and Sn. Moreover, Rank Cutoff curves [30] that measures the proportions of GSP among the top k percent ranked genes are plotted with k in $[0, 20]$. Finally, we draw Precision Recall curves [31] with rank threshold varying from 200 to 2000 according to equation (11):

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}. \tag{11}$$

3 Results

The performance of CapsNetMMD is evaluated and compared with other machine learning methods based on several measurements in this parts.

3.1 FDR of CapsNetMMD

The FDR of CapsNetMMD is calculated with threshold τ varying from 0.5 to 0.95 and the results are shown in Fig. 3. Initially, the mean FDR of CapsNetMMD is 0.31 when τ is 0.5 (*Supplementary* Table I). Then the value of FDR decreases as the threshold increases. When τ is set to 0.6, the mean FDR drops to 0.043, which is smaller than 0.05. Afterwards, the mean values of FDR are stably 0 with threshold τ larger than 0.65. These results potentially indicate the effectiveness of CapsNetMMD and may serve as a guideline for the usage of it.

TABLE 1
The Fractions of Known Breast Cancer-Related Genes with Different Rank Cutoffs and Their Corresponding p-Values in the Results of LightGBM, XGBoost, NN, SVM, Adaboost, and KNN

	Rank cutoff	Top 1%	Top 5%	Top 10%	Top 15%	Top 20%
CapsNetMMD	Fraction	17.5%	58.5%	78.2%	85.7%	89.4%
	p-value	8.1×10^{-125}	4.2×10^{-261}	5.3×10^{-234}	1.1×10^{-182}	3.6×10^{-141}
LightGBM	Fraction	16.9%	53.8%	67.3%	75.9%	82.4%
	p-value	3.0×10^{-114}	2.4×10^{-225}	2.2×10^{-178}	7.4×10^{-145}	1.1×10^{-119}
XGBoost	Fraction	17.0%	52.8%	72.2%	78.1%	81.6%
	p-value	1.0×10^{-116}	5.5×10^{-218}	1.5×10^{-202}	4.4×10^{-153}	3.9×10^{-117}
NN	Fraction	16.5%	51.1%	67.5%	74.4%	79.9%
	p-value	1.4×10^{-109}	5.8×10^{-206}	3.4×10^{-179}	3.0×10^{-139}	4.2×10^{-112}
SVM	Fraction	14.8%	47.1%	64.4%	73.7%	76.6%
	p-value	4.0×10^{-89}	2.9×10^{-178}	1.0×10^{-164}	8.8×10^{-137}	2.8×10^{-102}
Adaboost	Fraction	14.2%	41.7%	55.0%	62.2%	69.7%
	p-value	7.0×10^{-82}	8.2×10^{-144}	2.2×10^{-122}	3.8×10^{-97}	4.7×10^{-83}
KNN	Fraction	7.9%	21.8%	34.6%	40.6%	49.2%
	p-value	7.5×10^{-30}	2.5×10^{-42}	4.7×10^{-47}	1.9×10^{-36}	4.3×10^{-35}

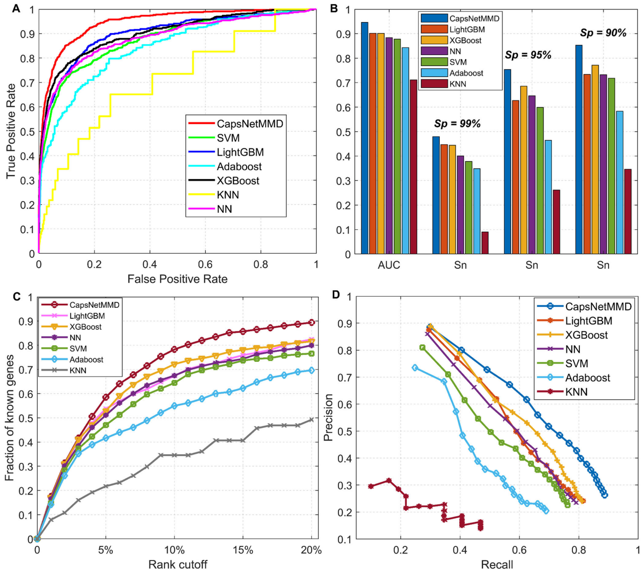

Fig. 4. Performance comparison of CapsNetMMD with LightGBM, XGBoost, NN, SVM, Adaboost, and KNN. A. ROC curves. B. AUC values and Sn values at three stringent levels of Sp. C. Rank Cutoff curves. D. Precision Recall curves.

3.2 Performance Comparison of CapsNetMMD with Other Methods

To obtain a reliable result, we perform CapsNetMMD ten times with the same parameters and use their average scores as the final scores for genes. The performance of CapsNetMMD are compared with other machine learning methods based on several measurements. Here we implement not only well-established methods including common Neural Network (NN), Support Vector Machine (SVM), Adaboost and K Nearest Neighbors (KNN), but also recent famous methods, i.e., XGBoost [12] and LightGBM [13], which achieve good performance on many machine learning challenges. The optimal parameters of these machine learning methods that can achieve the best performance are chosen through cross validation. As shown in Fig. 4A, the ROC curve of CapsNetMMD is obviously above those of other methods. Specifically, the AUC value of CapsNetMMD is 94.6 percent, which is 4.4 and 4.5 percent higher than that of LightGBM and XGBoost, respectively (Fig. 4B). Meanwhile, the AUC values of NN, SVM, Adaboost and KNN are all lower than 90 percent. Besides, at all stringent levels of *Sp*, the *Sn* values of CapsNetMMD are always the largest among these machine learning methods (Fig. 4B). Especially when *Sp* is 95 percent, the *Sn* value of CapsNetMMD (75.4 percent) is 6.8 percent larger than that of XGBoost and more than 10 percent larger than those of the other methods. When *Sp* drops to 90 percent, the *Sn* value of CapsNetMMD reaches 85.3 percent and is still at least 8 percent higher than the others. These phenomenon represent the superiority of CapsNetMMD in detecting true positives and true negatives.

Moreover, the Rank Cutoff curve of CapsNetMMD shown in Fig. 4C is significantly above the curves of LightGBM, XGBoost, NN, SVM, Adaboost and KNN. To prove that the results are not obtained by chance, we also utilize fisher's exact test to calculate their corresponding *p-values*. The detailed results with top 1, 5, 10, 15 and 20 percent ranked genes are listed in Table 1. In top 1 percent ranked genes, CapsNetMMD achieves the best result although the differences between the fractions of CapsNetMMD and other three methods (LightGBM, XGBoost and NN) are not obvious, which are no larger than 1 percent. When rank cutoff increases to top 5 percent, CapsNetMMD begins to show its advantage in identifying breast cancer-related genes. Specifically, CapsNetMMD can predict more than half (58.5 percent) known breast cancer-related genes by ranking them into top 5 percent while the fractions of SVM, Adaboost and KNN are all less than 50 percent. When rank cutoff is set to 10 percent, the fraction of CapsNetMMD grows to 78.2 percent, which is 10.9, 6.0, 10.7, 13.8, 23.2 and 43.6 percent higher than that of LightGBM, XGBoost, NN, SVM, Adaboost and KNN, respectively. The growth of fractions slows down as the range of rank cutoff enlarges. In top 20 percent ranked genes, CapsNetMMD can even predict nearly 90 percent of known breast cancer-related genes, which is far ahead of other methods. Furthermore, the *p-values* of CapsNetMMD at all fractions are always statistically significant (< 0.05) and consistently smaller than those of other methods.

In addition, the Precision Recall curve of CapsNetMMD (Fig. 4D) is above those of other six methods in the whole range. Initially, within the top 200 ranked genes, the precision and recall of CapsNetMMD are respectively 89 and 30 percent, which are similar to those of LightGBM and higher than those of XGBoost, NN, SVM, Adaboost and KNN. Then the variation tendencies of precision and recall are opposite, i.e., precision decreases as recall increases. When gene number enlarges to top 500, the recall of CapsNetMMD is boosted to 56.7 percent, which is respectively 4.4, 4.9, 6.6, 10.7 and 15.9 percent higher than that of LightGBM, XGBoost, NN, SVM, Adaboost and KNN. Meanwhile, the precision of CapsNetMMD reaches 67.2 percent that is at least 5 percent higher than any of other methods. Within the top 2,000 ranked genes, the recall of CapsNetMMD (88.7 percent) can even get close to 90 percent while the precision is still not low (26.3 percent). All results of performance evaluation based on above measurements indicate that CapsNetMMD has a significantly better performance in discovery of breast cancer-related genes than other machine learning methods.

Besides the well-established methods, we also search relevant references and compare the performance of CapsNetMMD with other two studies [32], [33] in discovery of novel breast cancer-related genes. In [32], a computational method is built based on the shortest path algorithm and in [33] a consensus signature from a set of seemingly different gene signatures is constructed by mapping them on a protein interaction network. The performance of these methods are shown by ROC curves and *Sn* values at stringent levels of *Sp* in *Supplementary Figure 1*. From *Supplementary Figure 1A*, we can see that the ROC curve of CapsNetMMD is significantly above those of other two methods. Specifically, the AUC value of CapsNetMMD is respectively 16 and 7 percent higher than that of shortest path based algorithm [32] and network based method [33] (*Supplementary Figure 1B*). When *Sp* is 99 percent, the *Sn* value of CapsNetMMD is 47.9 percent, which is almost nine times larger than that of shortest path algorithm [32] and more than three times larger than that of network based method. Similarly, the gaps between the *Sn* value of CapsNetMMD and those of other two methods are also large when *Sp* is 95 or 90 percent.

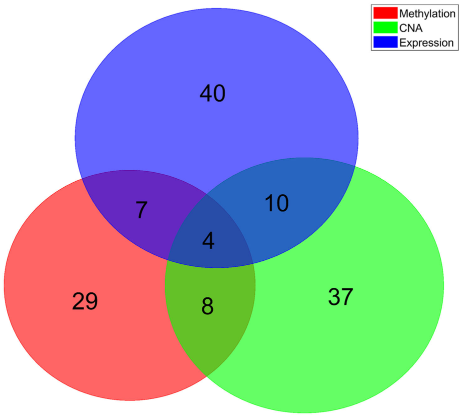

Fig. 5. Venn diagram of the distribution of potential prognostic candidate genes in three categories.

3.3 Discovery of Novel Breast Cancer-Related Genes

To study the results predicted by CapsNetMMD, we exact the top 100 ranked genes that do not contain known breast cancer-related genes and perform survival analysis on them using their features. Specifically, the relationships between genes and clinical outcome of patient samples is evaluated by Kaplan-Meier (KM) plot based on mRNA expression, CNA and DNA methylation of this gene. If the survival plots of a gene using a kind of features are significantly different ($p < 0.05$), we regard this gene as prognostic candidate gene [34]. Since there are three kinds of features, a gene may serve as prognostic candidate gene in any of the three categories: expression, CNA and methylation. The details of the evaluation results are shown in *Supplementary Table II*, from which we see that more than 2/3 (73) genes that ranked in top 100 are discovered as potential prognostic candidate genes. Among these genes, 4 genes are found to be prognostic candidate genes in all three categories, i.e., the information of CNA, expression and methylation of these genes can all be used to help classify patient samples into long survival and short survival. Besides, there are 25 genes being identified as prognostic candidate genes in two of the three categories and 44 genes in one of the three categories. The distribution of potential prognostic candidate genes in three categories is displayed as Venn diagram (Fig. 5), in which the number of genes in the category of expression is largest. The results of survival analysis show the prognostic values of the genes predicted by CapsNetMMD, which may corroborate the effectiveness and superiority of CapsNetMMD.

Furthermore, we analyze the results of survival analysis on top 10 ranked genes in detail. The KM plots with *p-values* smaller than 0.05 in CNA, DNA methylation and mRNA expression are shown in Fig. 6, *Supplementary Figure 21, 3*, respectively. Among the top 10 ranked genes, 8 genes are identified as potential prognostic candidate genes in breast cancer: 4 genes in CNA ($p - value = 2.2 \times 10^{-2}$ for RPL14, $p\text{-}value = 2.1 \times 10^{-3}$ for CHD4, $p - value = 1.6 \times 10^{-2}$ for LAPTM4A and $p - value = 2.3 \times 10^{-2}$ for DYNC1H1), 4 genes in methylation ($p - value = 8.0 \times 10^{-3}$ for AEBP1, $p\text{-}value = 1.6 \times 10^{-3}$ for DSTN, $p\text{-}value = 4.8 \times 10^{-2}$ for ORAOV1 and $p\text{-}value = 4.2 \times 10^{-2}$ for DYNC1H1) and 5 genes in expression ($p - value = 1.1 \times 10^{-2}$ for RPL14, $p\text{-}value = 2.1 \times 10^{-2}$ for AEBP1, $p\text{-}value = 2.4 \times 10^{-3}$ for LAPTM4A, $p\text{-}value = 4.3 \times 10^{-2}$ for DYNC1H1 and $p\text{-}value = 3.8 \times 10^{-2}$ for RNF213). In order to analyze these genes conveniently, their names and scores are also listed in *Supplementary Table III*.

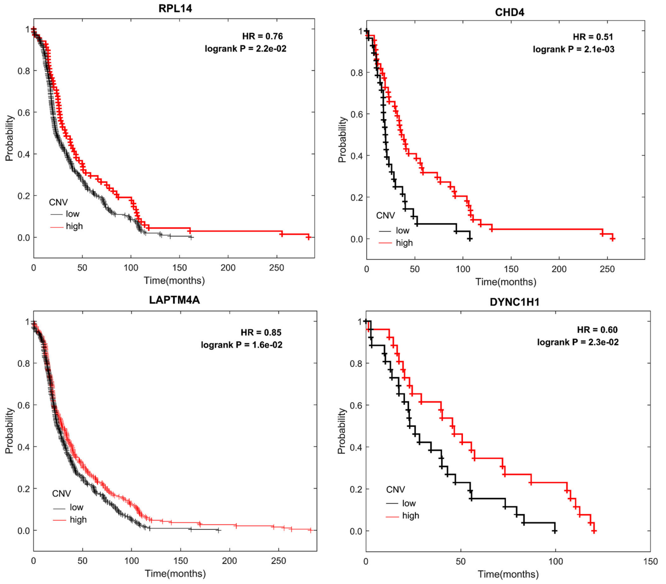

Fig. 6. The KM plots of RPL14, CHD4, LAPTM4A, and DYNC1H1 using CNA.

Based on the information provided by *Supplementary Table III*, we can see that gene RPL14 ranks first with the highest normalized score. Besides, from Fig. 6 and *Supplementary Figure 3*, RPL14 is also identified as potential prognostic candidate gene in both CNA ($p\text{-}value = 2.2 \times 10^{-2}$) and expression ($p\text{-}value = 1.1 \times 10^{-2}$). By consulting literatures, significant loss of heterozygosity of RPL14 are detected in several tumors such as non-small cell lung cancer according to [35] and similar results are also obtained in the study of [36], which suggests that the analysis to RPL14 may be used as a potential molecular marker of esophageal squamous cell carcinomas [36]. These discoveries are to some extent in accordance with the above results of our study, which indicate that gene RPL14 may play important roles in cancer progression. In addition, the second-ranked gene AEBP1 simultaneously appears in categories of methylation ($p\text{-}value = 8.0 \times 10^{-3}$) and expression ($p\text{-}value = 2.1 \times 10^{-2}$) while the fourth-ranked gene LAPTM4A in categories of CNA ($p\text{-}value = 1.6 \times 10^{-2}$) and expression ($p\text{-}value = 2.4 \times 10^{-3}$).

Moreover, according to Fig. 6, *Supplementary Figure 2, 3*, the eighth-ranked gene DYNC1H1 is found to be potential prognostic candidate gene in all three categories ($p\text{-}value = 2.3\times 10^{-2}$ for CNA, $p\text{-}value = 4.2 \times 10^{-2}$ for methylation and $p\text{-}value = 4.3 \times 10^{-2}$ for methylation). Recently, there are researches finding that the missense mutations of DYNC1H1 are related to deleterious cancer including pancreatic cancer, colorectal cancer, etc [37]. Meanwhile, earlier researches indicate that DYNC1H1 mutation can cause spinal muscular atrophy [38] and Charcot-Marie-Tooth disease [39]. Although the relationship between DYNC1H1 and breast cancer is not discussed in these references, the functions of DYNC1H1 in other diseases especially other cancers may imply its potential values in the researches of breast cancer.

4 Conclusion and Discussion

We present a deep learning method named CapsNetMMD to identify breast cancer-related genes by modeling multi-omics data based on capsule network. Commonly, with the help of known disease-related genes, prediction of novel genes can be transformed into an issue of supervised classification [40]. Therefore, appropriate features and suitable classifier are crucial for the results of the research. In this study, we integrate multi-omics data of breast cancer from TCGA database to generate the feature matrix of genes, which include mRNA expression, z scores for mRNA expression, DNA methylation and two forms of CNAs. These five kinds of data comprehensively provide useful information of genes in different omics and the features are further reshaped to obtain a valid input for the classifier. Besides, the settings of instantiation parameters and dynamic routing mechanism in the modeling process make the classifier learn effectively from the training dataset and precisely predict novel disease-related genes. In all, the significantly better performance of CapsNetMMD compared with other existing machine learning method is attributed to not only the integration of multi-omics data but also the modeling based on capsule network. The predicted genes with prognostic values in breast cancer may serve as candidates for biologists and medical scientists in the future studies of biomarkers of breast cancer.

In spite of the superiority of CapsNetMMD in discovery of breast cancer-related genes, there are still some limitations in its generalization to other diseases. For example, since supervised classification is performed based on known breast cancer-related genes, CapsNetMMD cannot be applied to the diseases that have no known genes. In that case, the identification of disease-related genes will be transformed into the issue of unsupervised classification [41], [42], [43], [44], which is usually implemented with clustering methods and the results will be hard to evaluate. Alternatively, the network-based algorithms [31], [45], [46], [47] can be used by introducing similarities between diseases. Specifically, the relationships between candidate genes and the disease that to be studied are assessed through exploring the associations between candidate genes and known genes related to similar diseases [48], [49], [50]. In addition, due to the fact that a critical factor leading to the success of CapsNetMMD is the integration of multi-omics data of breast cancer, this method is unsuitable for diseases whose multi-omics data are incomplete even unavailable. Nonetheless, the reduced cost and rapid development of high-throughput technologies (e.g., next generation sequencing, microarrays, etc.) greatly promote the researches of multi-omics data, thus it is predictable that more complete data will be available in the future. Moreover, many other information about genes including their interactions with non-coding RNAs such as microRNAs (miRNAs) or long non-coding RNAs (lncRNAs), which have been reported to play importation roles in human cancers [51], [52], are not taken into account in this study. We will incorporate these data into CapsNetMMD in future works to generate a more comprehensive model in uncovering the mechanism under human diseases.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China, Nos. 61702371, 61520106006, 61861146002, 61732012, 61532008, U1611265, 61672382, 61772370, 61772357, and 61873270, China Postdoctoral Science Foundation Grant, No. 2017M611619, by the Scientific & Technological Base and Talent Special Program of the Guangxi Zhuang Autonomous Region, GuiKe AD18126015, and by the "BAGUI Scholar" Program of Guangxi Province of China.

References

[1] Network C. G. A., "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, 2012, Art. no. 61.

[2] C. H. Barcenas, et al., "Outcomes in patients with early-stage breast cancer who underwent a 21-gene expression assay," *Cancer*, vol. 123, no. 13, pp. 2422–2431, 2017.

[3] M. V. Iorio, et al., "MicroRNA gene expression deregulation in human breast cancer," *Cancer Res.*, vol. 65, no. 16, pp. 7065–7070, 2005.

[4] A. Müller, et al., "Involvement of chemokine receptors in breast cancer metastasis," *Nature*, vol. 410, no. 6824, 2001, Art. no. 50.

[5] S. Nik-Zainal, et al., "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," *Nature*, vol. 534, no. 7605, 2016, Art. no. 47.

[6] L. J. Van't Veer, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, 2002, Art. no. 530.

[7] P. L. Nguyen, et al., "Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy," *J. Clin. Oncology*, vol. 26, no. 14, pp. 2373–2378, 2008.

[8] L. J. Lancashire, R. C. Rees, and G. R. Ball, "Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach," *Artif. Intell. Med.*, vol. 43, no. 2, pp. 99–111, 2008.

[9] I. Guyon, et al., "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002.

[10] D.-G. Guan, et al., "mirExplorer: Detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features," *RNA Biol.*, vol. 8, no. 5, pp. 922–934, 2011.

[11] O. Okun and H. Priisalu, "Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors," *Artif. Intell. Med.*, vol. 45, no. 2-3, pp. 151–162, 2009.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016. pp. 785–794.

[13] G. Ke, et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[14] J. Zhong, et al., "XGBFEMF: An XGBoost-based framework for essential protein prediction," *IEEE Trans. NanoBiosci.*, vol. 17, no. 3, pp. 243–250, Jul. 2018.

[15] S. Zheng and M. Zeng, "MicroRNA Prediction Based on Machine Learning Algorithm," *Comput. Sci.*, vol. 42, no. 2, pp. 2–17, 2015.

[16] D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An effective miRNA classification method in breast cancer patients," in *Proc. Int. Conf. Comput. Biol. Bioinf.*, 2017. pp. 7–11.

[17] Network, C. G. A. R., "Comprehensive molecular characterization of urothelial bladder carcinoma," *Nature*, vol. 507, no. 7492, 2014, Art. no. 315.

[18] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome Atlas (TCGA): An immeasurable source of knowledge," *Contemporary Oncology*, vol. 19, no. 1A, 2015, Art. no. A68.

[19] Consortium, I. C. G., "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, 2010, Art. no. 993.

[20] Y. Joly, et al., "Data sharing in the post-genomic world: the experience of the international cancer genome consortium (ICGC) data access compliance office (DACO)," *PLoS Comput. Biol.*, vol. 8, no. 7, 2012, Art. no. e1002549.

[21] J. Zhang, et al., "International cancer genome consortium data portal—A one-stop shop for cancer genomics data," *Database*, 2011, 2011, Art. no. bar026.

[22] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[23] E. Cerami, et al., "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer Discov.*, vol. 2, pp. 401–404, 2012.

[24] J. Gao, et al., "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signal.*, vol. 6, no. 269, pp. pl1–pl1, 2013.

[25] Z. Ding, S. Zu and J. Gu, "Evaluating the molecule-based prediction of clinical drug responses in cancer," *Bioinf.*, vol. 32, no. 19, pp. 2891–2895, 2016.

[26] J. Piñero, et al., "DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res.*, vol. 45, pp. D833–D839, 2016.

[27] J. Piñero, et al., "DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes," *Database*, 2015, 2015, Art. no. bav028.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010. pp. 807–814.

[29] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. 9, no. 1, pp. 18–29, 2007.

[30] B. Linghu, et al., "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome Biol.*, vol. 10, no. 9, 2009, Art. no. R91.

[31] C. Peng, A. Li and M. Wang, "Discovery of bladder cancer-related genes using integrative heterogeneous network modeling of multi-omics data," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 15639.

[32] F. Holger, "Network based consensus gene signatures for biomarker discovery in breast cancer," *Plos One*, vol. 6, no. 10, Oct. 2011, Art. no. e25364.

[33] L. Chen, X. Z. Hao, et al., "Application of the shortest path algorithm for the discovery of breast cancer-related genes," *Current Bioinf.*, vol. 11, no. 1, pp. 51–58, Feb. 2016.

[34] B. Györffy, et al., "An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients," *Breast Cancer Res. Treatment*, vol. 123, no. 3, pp. 725–731, 2010.

[35] S. P. Shriver, et al., "Trinucleotide repeat length variation in the human ribosomal protein L14 gene (RPL14): Localization to 3p21. 3 and loss of heterozygosity in lung and oral cancers," *Mutation Res./Mutation Res. Genomics*, vol. 406, no. 1, pp. 9–23, 1998.

[36] X.-P. Huang, et al., "Alteration of RPL14 in squamous cell carcinomas and preneoplastic lesions of the esophagus," *Gene*, vol. 366, no. 1, pp. 161–168, 2006.

[37] C. Sucularli and M. Arslantas, "Computational prediction and analysis of deleterious cancer associated missense mutations in DYNC1H1," *Mol. Cellular Probes*, vol. 34, pp. 21–29, 2017.

[38] M. Harms, et al., "Mutations in the tail domain of DYNC1H1 cause dominant spinal muscular atrophy," *Neurology*, vol. 78, pp. 1714–1720, 2012.

[39] M. N. Weedon, et al., "Exome sequencing identifies a DYNC1H1 mutation in a large pedigree with dominant axonal Charcot-Marie-Tooth disease," *Amer. J. Hum. Genetics*, vol. 89, no. 2, pp. 308–312, 2011.

[40] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinf.*, vol. 22, no. 15, pp. 1855–1862, 2006.

[41] D.-S. Huang, et al., "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein Peptide Sci.*, vol. 15, no. 6, pp. 553–560, 2014.

[42] C.-H. Zheng, et al., "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 599–607, Jul. 2009.

[43] C.-H. Zheng, et al., "Molecular pattern discovery based on penalized matrix decomposition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 6, pp. 1592–1603, Nov./Dec. 2011.

[44] L. Zhu, et al., "ChIP-PIT: Enhancing the analysis of ChIP-Seq data using convex-relaxed pair-wise interaction tensor decomposition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 1, pp. 55–63, Jan./Feb. 2016.

[45] S.-P. Deng, L. Zhu, and D.-S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," in *BMC Genomics*, vol. 16, 2015, Art. no. S4.

[46] C. Peng and A. Li, "A heterogeneous network based method for identifying GBM-related genes by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 713–720, May/Jun. 2017.

[47] O. Vanunu, et al., "Associating genes and protein complexes with disease via network propagation," *PLoS Comput. Biol.*, vol. 6, no. 1, 2010, Art. no. e1000641.

[48] S.-P. Deng, L. Zhu, and D.-S. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 1, pp. 27–35, Jan./Feb. 2016.

[49] L. Zhu, S.-P. Deng, and D.-S. Huang, "A two-stage geometric method for pruning unreliable links in protein-protein networks," *IEEE Trans. Nanobiosci.*, vol. 14, no. 5, pp. 528–534, Jul. 2015.

[50] L. Zhu, et al., "t-LSE: A novel robust geometric approach for modeling protein-protein interaction networks," *PLoS One*, vol. 8, no. 4, 2013, Art. no. e58368.

[51] W. Peng, P. Koirala, and Y. Mo, "LncRNA-mediated regulation of cell signaling in cancer," *Oncogene*, vol. 36, no. 41, 2017, Art. no. 5661.

[52] N. W. Wong, et al., "OncomiR: An online resource for exploring pan-cancer microRNA dysregulation," *Bioinf.*, vol. 34, no. 4, pp. 713–715, 2017.

Chen Peng received the BSc degree in microelectronics from the School of Computer Science and Technology, North University of China (NUC), in 2011, and the PhD degree in biomedical engineering from the School of Information Science and Technology, USTC, in 2016. She is now employed by Tongji University as a postdoctoral fellow at the Institute of Machine Learning and Systems Biology. Her current research interests include computational cancer genomics, bioinformatics, statistical machine learning, and artificial intelligence.

Yang Zheng received the BSc degree in computer science and technology from Zhejiang Normal University (ZJNU), ZheJiang, China, in 2017. Now, he is working toward the master's degree in the College of Electronics and Information Engineering, Tongji University, China. His research interests include bioinformatics and machine learning.

De-Shuang Huang received the BSc, MSc, and PhD degrees all in electronic engineering from the Institute of Electronic Engineering, Hefei, China, National Defense University of Science and Technology, Changsha, China, and Xidian University, Xian, China, in 1986, 1989 and 1993, respectively. During the 1993-1997 period, he was a postdoctoral student, respectively, in the Beijing Institute of Technology and in the National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In Sept. 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences, as the Recipient of "Hundred Talents Program of CAS". In Sept. 2011, he entered into Tongji University as a chaired professor. From Sept. 2000 to Mar. 2001, he worked as a research associate at Hong Kong Polytechnic University. From Aug. to Sept. 2003, he visited the George Washington University, a Washington DC, USA, as a visiting professor. From July to Dec. 2004, he worked as the University fellow at Hong Kong Baptist University. From Mar. 2005 to Mar. 2006, he worked as a research fellow at the Chinese University of Hong Kong. From Mar. to July 2006, he worked as a visiting professor at the Queen's University of Belfast, UK. In 2007, 2008, and 2009, he worked as a visiting professor at Inha University, Korea, respectively. At present, he is the director of the Institute of Machines Learning and Systems Biology, Tongji University. He is currently an IAPR fellow. He has published more than 180 journal papers. His current research interest includes bioinformatics, pattern recognition, and machine learning. He is a senior member of the IEEE.