Accepted Manuscript

A Deep Learning-based Multi-model Ensemble Method for Cancer Prediction

Yawen Xiao, Jun Wu, Zongli Lin, Xiaodong Zhao

PII: S0169-2607(17)30494-7
DOI: 10.1016/j.cmpb.2017.09.005
Reference: COMM 4491

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 26 April 2017
Revised date: 7 August 2017
Accepted date: 6 September 2017

Please cite this article as: Yawen Xiao, Jun Wu, Zongli Lin, Xiaodong Zhao, A Deep Learning-based Multi-model Ensemble Method for Cancer Prediction, *Computer Methods and Programs in Biomedicine* (2017), doi: 10.1016/j.cmpb.2017.09.005

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- An ensemble of multiple machine learning models outperforms single classifiers.
- We propose a deep learning-based ensemble method for cancer prediction.
- We select differentially expressed genes from gene expression data.
- We present prediction results on lung, stomach and breast cancer data.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

A Deep Learning-based Multi-model Ensemble Method for Cancer Prediction

Yawen Xiao* Jun Wu† Zongli Lin‡ Xiaodong Zhao§

Abstract

Background and Objective: Cancer is a complex worldwide health problem associated with high mortality. With the rapid development of the high-throughput sequencing technology and the application of various machine learning methods that have emerged in recent years, progress in cancer prediction has been increasingly made based on gene expression, providing insight into effective and accurate treatment decision making. Thus, developing machine learning methods, which can successfully distinguish cancer patients from healthy persons, is of great current interest. However, among the classification methods applied to cancer prediction so far, no one method outperforms all the others.

Methods: In this paper, we demonstrate a new strategy, which applies deep learning to an ensemble approach that incorporates multiple different machine learning models. We supply informative gene data selected by differential gene expression analysis to five different classification models. Then, a deep learning method is employed to ensemble the outputs of the five classifiers.

Results: The proposed deep learning-based multi-model ensemble method was tested on three public RNA-seq data sets of three kinds of cancers, Lung Adenocarcinoma, Stomach Adenocarcinoma and Breast Invasive Carcinoma. The test results indicate that it increases the prediction accuracy of cancer for all the tested RNA-seq data sets as compared to using a single classifier or the majority voting algorithm.

Conclusions: By taking full advantage of different classifiers, the proposed deep learning-based multi-model ensemble method is shown to be accurate and effective for cancer prediction.

Keywords: Multi-model ensemble; deep learning; gene expression; feature selection; cancer prediction.

1 Introduction

Cancer has been characterized as a collection of related diseases involving abnormal cell growth with the potential to divide without stopping and spread into surrounding tissues [1]. According to the GLOBOCAN project [2], in 2012 alone, about 14.1 million new cases of cancer occurred globally (not including skin cancer other than melanoma), which caused about 14.6% of the death. Since cancer is a major cause of morbidity and mortality, diagnosis and detection of cancer in its early stage is of great importance for its cure. Over the past decades, a continuous evolution of cancer research has been performed [3]. Among the diverse methods and techniques developed for cancer prediction, the utilization of gene expression level is one of the research hotspots in this field. Data analysis on gene expression level has facilitated cancer diagnosis and treatment to a great extent. Accurate prediction of cancer is one of the most critical and urgent tasks for physicians [4].

With the rapid development of computer-aided techniques in recent years, application of machine learning methods is playing an increasingly important role in the cancer diagnosis, and various prediction algorithms are being explored continuously by researchers. Sayed et al. [5] conducted a comparative study on feature selection and

*Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai 200240, China, Email: foreverxyw@sjtu.edu.cn

†School of Biomedical Engineering Shanghai Jiao Tong University, Shanghai 200240, China, Email: junwu302@gmail.com

‡Corresponding author, Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, P.O. Box 400743, Charlottesville, VA 22904-4743, U.S.A., Email: zl5y@virginia.edu, Tel +1 434 924 6342

§School of Biomedical Engineering Shanghai Jiao Tong University, Shanghai 200240, China, Email: xiaodong122@yahoo.com

ACCEPTED MANUSCRIPT

classification using data collected from the central database of the National Cancer Registry Program of Egypt, and three classifiers were applied, including support vector machines (SVMs), k-nearest neighbour (kNN) and Naive Bayes (NBs). The results showed that SVMs with polynomial kernel functions yielded higher classification accuracy compared with kNN and NBs. Statnikov et al. [6] carried a comprehensive comparison of random forests (RFs) and SVMs for cancer diagnosis. The results were obtained that SVMs outperformed RFs in fifteen data sets, RFs outperformed SVMs in four data sets, and the two algorithms performed the same in three data sets. These results were obtained by using full set of genes. Similar results were derived based on the gene selection method. From a large body of literature in cancer prediction research, none of these machine learning methods is fully accurate and each method may be lacking in different facets in the classification procedure. For instance, it is difficult for SVMs to figure out an appropriate kernel function, and although RFs have solved the over-fitting of decision trees (DTs), RFs may lead the classification result to the category with more samples.

In view of the fact that each machine learning method may outperform others or have defects in different cases, it is thus natural to expect that a method that takes advantages of multiple machine learning methods would lead to superior performance. To this end, several studies have been reported in the literature that aim to integrate models to increase the accuracy of the prediction. For example, Breiman [7] introduced *Bagging*, which combines outputs from decision trees generated by several randomly selected subsets of the training data and votes for the final outcome. Freund and Schapire [8] introduced *Boosting*, which updates the weights of training samples after each iteration of training and combines the classification outputs by weighted votes. Wolpert [9] proposed to use linear regression to combine outputs of the neural networks, which was later known as *Stacking*. Aik et al. [10] applied *Bagging* and *Boosting* on cancerous microarray data for cancer classification. Cho and Won [11] applied the majority voting algorithm to combine four classifiers using three benchmark cancer data sets. The *Stacking* and majority voting take advantages of different machine learning methods. Although the majority voting algorithm is the most common in classification tasks, it is still too simple a combination strategy to discover complex information from different classifiers. *Stacking*, through the use of a learning method in the combination stage, is a much more powerful ensemble technique. Given that the small number of deep learning studies in biomedicine have shown success with this method [12], deep learning has become a strong learning method with many advantages. Unlike the majority voting which only considers the linear relationships among classifiers and requires for manual participation, deep learning has the ability to "learn" the intricate structures, especially nonlinear structures, from the original large data sets automatically. Thus, in order to better describe the unknown relationships among different classifiers, we adopt deep learning in the *Stacking*-based ensemble learning of multiple classifiers.

In this paper, we attempt to use deep neural networks to ensemble five classification models, which are kNN, SVMs, DTs, RFs and gradient boosting decision trees (GBDTs), to construct a multi-model ensemble model to predict cancer in normal and tumor conditions. To avoid over-fitting, we employ the differential gene expression analysis to select important and informative genes. The selected genes are then supplied to the five classification models. After that, a deep neural network is used to ensemble the outputs of the five classification models to obtain the final prediction result. We evaluate the proposed method on three public RNA-seq data sets from lung tissues, stomach tissues and breast tissues, respectively. The final results indicate that the proposed deep learning-based multi-model ensemble method makes more effective use of the information of the limited clinical data and generates more accurate prediction than single classifiers or the majority voting algorithm.

ACCEPTED MANUSCRIPT

2 Methods

The flowchart of the proposed deep learning-based ensemble strategy is shown in Fig. 1. Initially, differential expression analysis is used to select the significantly differentially expressed genes, namely the most informative features, which are then fed to the following classification process. Then, we employ the technique of S-fold cross validation to divide the initial data into S groups of training and testing data sets. After that, multiple classifiers (first-stage models) are learned from the training sets, each of which consists of $S-1$ of the S groups, and then applied to the corresponding test set, which is the remaining group of the S groups, to output the predicted class of the samples. Finally, we use a deep neural network classifier (second-stage ensemble model) to combine the predictions in the first stage with the aim of reducing the generalization error and procuring a more accurate outcome.

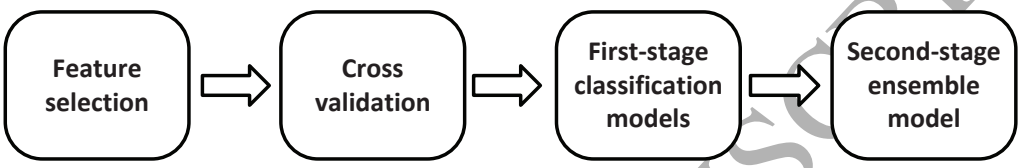


Figure 1: Flowchart of the proposed deep learning-based multi-model ensemble method.

2.1 Feature selection

The use of gene expression data with an increasing number of features (*e.g.*, genes) and information makes it more challenging to develop classification models. In clinical practice, the number of cancer samples available is rather small in comparison with the number of features, resulting in higher risk of over-fitting and degradation of the classification performance. Feature selection is a good way to address these challenges [13]. By reducing the entire feature space to a subset of features, over-fitting of the classification model can be avoided, thus mitigating the challenges arising from a small sample size and a high data dimensionality.

In this paper, we employ the DESeq [14] method to select informative genes for the downstream classification. The DESeq method is usually used to decide whether, for a given gene, an observed difference in read count is significant, that is, whether it is greater than what would be expected just due to natural random variation [14]. In differential expression analysis, by setting the thresholds of the BH-adjusted p-value and the fold change level, the significantly differentially expressed genes are screened and selected.

2.2 Cross validation

For many classification models, the complexity may be governed by multiple parameters. In order to achieve the best prediction performance on new data, we wish to find appropriate values of the complexity parameters that lead to the optimal model for a particular application.

If data are plentiful, then a simple way for model selection is to divide the entire data into three subsets, the training set, the validation set and the test set. A range of models are trained on the training set, compared and selected on the validation set, and finally evaluated on the test set. Among the diverse complex models that have been trained, the one having the best predictive performance is selected, which is an effective model validated by the data in the validation set. In a practical application, however, the supply of data for training and testing is limited, leading to an increase of the generalization error. An approach to reducing the generalization error and preventing over-fitting is to use cross validation [15].

ACCEPTED MANUSCRIPT

The technique of S-fold cross validation [15] used in this paper is illustrated in Fig. 2 for the case of $S = 4$. S-fold cross validation partitions the available data set D into S disjoint groups, $D_1, D_2, \ldots, D_S$, with all subsets maintaining consistency in the data distribution. After that, S-1 groups are used as the training set and the remaining group is used as the test set. The procedure is then repeated for all S possible choices of the $S - 1$ groups, and the performance scores resulting from the S runs are then averaged. In our study, we not only utilize S-fold cross validation to implement model selection for every single classifier separately, but also generate new data sets for the ensemble stage by using S-fold cross validation on the initial data sets in order to avoid over-fitting.

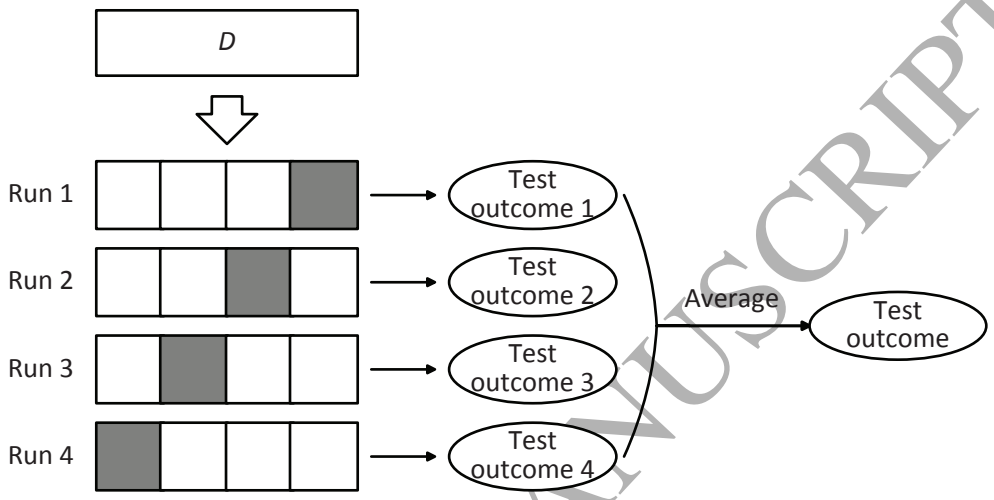


Figure 2: The technique of S-fold cross validation ($S = 4$).

2.3 Classification methods

After preprocessing of the data sets, we assess the prediction performance of five popular classification methods towards the discrimination between normal and tumor samples. Specifically, we apply k-nearest-neighbor (kNN), support vector machines (SVMs), decision trees (DTs), random forests (RFs), and gradient boosting decision trees (GBDTs) as first-stage classification models. All of these five classification methods are of high accuracy in practical applications and are reviewed as follows [4, 13, 16, 17].

kNN is a non-parametric classification method that is used when there is little or no prior knowledge about the distribution of the data. kNN classifiers transform samples to a metric space, where distances between samples are determined. The distance function between a test sample and the training samples is the basis, by calculating which, kNN classifies a test sample based upon the most common class in its k-nearest training samples.

SVMs initially map the input vector into a feature space of higher dimensionality and identify a hyperplane that separates the data points into two categories. The gap between the two categories is as wide as possible. New samples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall on with higher confidence.

DTs have tree-like structures in which the nodes represent the input variables and the leaves correspond to decision outcomes. When traversing the tree for the classification of a new sample, we are able to predict the category of the data with adequate reasoning due to the specific architecture.

RFs have only recently been applied in the field of cancer prediction. RFs are an ensemble learning method that combines tree predictors, each of which depends on the values of a random vector sampled independently and with the same distribution. The final outcome is the most popular class that receives the majority of votes from

ACCEPTED MANUSCRIPT

the trees in the forest, hence yielding an overall better model.

GBDTs are a machine learning technique that combines an ensemble of decision trees into a stronger prediction model. GBDTs build the model in a stage-wise fashion like other boosting methods do, and implement a generalization by allowing the optimization of an arbitrary differentiable loss function.

Three classical methods (*i.e.*, kNN, SVMs and DTs) and two advanced algorithms (*i.e.*, RFs and GBDTs) are introduced. As suggested in the literature, kNN is one of the simplest classification methods especially for distribution-unknown data. But kNN is sensitive to redundant features and requires effective feature selection prior to classification, and the choice of the number k can greatly affect the performance of classifier. SVMs can be considered as the most effective and common algorithm for cancer classification. Nevertheless, it is a challenge for SVMs to figure out an appropriate kernel for specific issues. In particular, there is no general solution for nonlinear cases, thus the prediction accuracy can not be guaranteed. DTs, as the the most fundamental and widely used classification method in various fields, however, tend to be a weak classifier to distinguish normal and cancer samples since it often over-fits the model. The latter two methods, RFs and GBDTs are evolutionary approaches which are ensembles of DTs, overcoming the over-fitting problem to an extent, but may lead to the classification result tending to the category with more samples. Considering that each method has its own shortcomings relative to others, we come up with an ensemble strategy to make use of the advantages of the multiple methods and avoid the shortcomings. Here, we select both the fundamental and evolutionary methods in order to increase the diversity of our ensemble model.

2.4 Multi-model ensemble based on deep learning

In practice, several classification models are available for cancer prediction, but none of these is fully accurate and each method may be making mistakes in different facets. Stacking of multiple different classification methods may lead to performance improvement over individual models. Multi-model ensemble is a technique in which the predictions of a collection of models are given as inputs to a second-stage learning model. The second-stage model is trained to combine the predictions from first stage models optimally to form a final set of predictions.

In this paper, we adopt deep learning as the ensemble model to stack the multiple classifiers. Neural networks are inspired by how the brain works and is widely used in many applications. A neural network is trained to generate an output as a combination among the input variables. Given a set of features and a target, it can learn to be a nonlinear function approximator, where, between the input and output layers, there can be one or more nonlinear layers, called hidden layers. Deep learning involves deep neural networks with many hierarchical hidden layers of nonlinear information processing which endow the capabilities to learn complex patterns from high dimensional raw data with little guidance [12].

Shown in Fig. 3 is a neural network. The leftmost layer is the input layer with neurons being called input neurons. The rightmost layer is called the output layer with an output neuron. The middle layers are hidden layers that are made up of hidden neurons. In order to classify samples correctly, we define an objective function, which computes the error between the predicted scores and the actual scores. Then, by training with the training samples, the machine modifies the values of its internal adjustable parameters that define the input-output function to reduce the error. In practice, the stochastic gradient descent (SGD) algorithm is most commonly used in this machine learning procedure. In a deep neural network, we denote the number of layers as n_l and layer l as L_l, so layer L_1 is the input layer and layer L_{n_l} is the output layer. We also let s_l denote the number of neurons in layer l. The neural network has parameters $W = \{W^1, W^2, \ldots, W^{n_l}\}$ and $b = \{b^1, b^2, \ldots, b^{n_l}\}$, where W^l_{ij}, $j = 1, 2, \ldots, s_{l-1}, i = 1, 2, \ldots, s_l, l = 2, 3, \ldots, n_l$, denotes the weight associated the connection between unit j in

ACCEPTED MANUSCRIPT

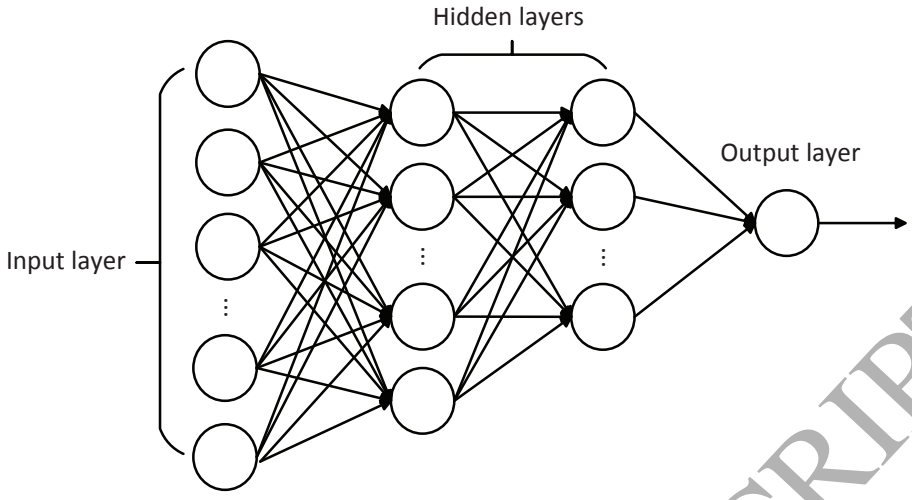


Figure 3: An illustration of the neural network structure.

layer $l-1$ and unit i in layer l, and b_i^l, $i = 1, 2, \ldots, s_l, l = 2, 3, \ldots, n_l$, denotes the bias of unit i in layer l. Suppose that we have a training set $\{(x^1, y^1), (x^2, y^2), \ldots, (x^m, y^m)\}$ of m samples, with which we train the neural network using the SGD. We define the cost function (the objective function mentioned above) as,

$$\begin{aligned} J(W,b) &= \frac{1}{m}\sum_{i=1}^{m} J(W,b;x^i,y^i) + \frac{\lambda}{2}\sum_{l=2}^{n_l}\sum_{j=1}^{s_{l-1}}\sum_{i=1}^{s_l}(W_{ij}^l)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h_{W,b}(x^i) - y^i\|^2\right) + \frac{\lambda}{2}\sum_{l=2}^{n_l}\sum_{j=1}^{s_{l-1}}\sum_{i=1}^{s_l}(W_{ij}^l)^2, \end{aligned} \tag{1}$$

where the first term is a mean square error term and the second term is a regulation term used to constrain the magnitudes of the weights and prevent over-fitting, and λ is the weight decay parameter that regulates the relative importance of the two terms. The nonlinear hypothesis $h_{W,b}(x)$ of the neural network is defined as,

$$h_{W,b}(x) = f(W^{\mathrm{T}}x + b), \tag{2}$$

where $f : \mathbb{R} \to \mathbb{R}$ is called the activation function. In recent years, the most popular nonlinear function used here is the rectified linear unit (ReLU) $f(z) = \max\{0, z\}$ which typically learns much faster in multi-layer deep neural networks than the more conventional hyperbolic tangent and logistic sigmoid function [18]. For one sample, we define the activation (output value) of unit i in layer l as a_i^l and the weighted sum as z_i^l, so that

$$a_i^l = f(z_i^l) = f\left(W_{i1}^{l-1}a_1^{l-1} + W_{i2}^{l-1}a_2^{l-1} + \ldots + W_{is_{l-1}}^{l-1}a_{s_{l-1}}^{l-1} + b_i^{l-1}\right), \tag{3}$$

and with x_i as the unit i in the input layer L_1, *i.e.*,

$$a_i^1 = x_i. \tag{4}$$

Thus the activation of the unit in the output layer is

$$h_{W,b}(x) = a_i^{n_l} = f\left(W_{i1}^{n_l-1}a_1^{n_l-1} + W_{i2}^{n_l-1}a_2^{n_l-1} + \ldots + W_{is_{n_l-1}}^{n_l-1}a_{s_{n_l-1}}^{n_l-1} + b_i^{n_l-1}\right). \tag{5}$$

This step to compute the activation of each unit is called the forward propagation. In SGD, our goal is to minimize $J(W, b)$ by adjusting parameters W and b. We first initialize each W_{ij}^l and b_i^l to a small random value near zero

ACCEPTED MANUSCRIPT

and then update the parameters in each iteration of SGD as,

$$W_{ij}^{l} = W_{ij}^{l} - \alpha \frac{\partial}{\partial W_{ij}^{l}} J(W, b), \tag{6}$$

$$b_{i}^{l} = b_{i}^{l} - \alpha \frac{\partial}{\partial b_{i}^{l}} J(W, b), \tag{7}$$

where α is the learning rate. Then we employ the back propagation algorithm to compute the partial derivatives. In detail, given a training sample (x, y), the back propagation algorithm can be described as follows.

1. Conduct the forward propagation calculations to compute the activation of each unit in layer L_2 up to the output layer L_{n_l}.

2. For each unit i in layer n_l, calculate the residual

$$\delta_{i}^{n_l} = \frac{\partial}{\partial z_{i}^{n_l}} \frac{1}{2} \|y - h_{W,b}(x)\|^2 = -(y_i - a_{i}^{n_l}) f'(z_{i}^{n_l}). \tag{8}$$

3. For each unit i in layer l, $l = n_l - 1, n_l - 2, ..., 2$, calculate the residual

$$\delta_{i}^{l} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{l} \delta_{j}^{l+1} \right) f'(z_{i}^{l}). \tag{9}$$

4. Calculate the desired partial derivatives

$$\frac{\partial}{\partial W_{ij}^{l}} J(W, b; x, y) = a_{j}^{l} \delta_{i}^{l+1}, \tag{10}$$

$$\frac{\partial}{\partial b_{i}^{l}} J(W, b; x, y) = \delta_{i}^{l+1}. \tag{11}$$

The process of deriving from back forward is the intention of back propagation. By repeating the iterative steps of the SGD, we can decrease the cost function $J(W, b)$ so as to train the neural network.

In our study, we have proposed a deep learning-based multi-model ensemble method with a 5-fold stacking. The overall algorithm is shown in Fig. 4. In the first stage, we divide the given data set D into five subsets, D_1, D_2, $\ldots$, D_5, where $D_i = \{x_i, y_i\}$, $i = 1, 2, \ldots, 5$, contains labeled points drawn independent and identically distributed according to the same distribution. In the first round, the union of D_2, D_3, D_4 and D_5 is used as the training set, and $D_1 = \{x_1, y_1\}$ is used as the test set. Given the input x_1, five classification models in this stage propose corresponding hypotheses $h_1(x_1), h_2(x_1), \ldots, h_5(x_1)$, where $h_i(x_i)$ is a binary variable, and the subscript i of $h_i(x_1)$ is referred to as the ith model. After the classifications in the first round, we assemble the predictions of each model into $H_1 = [h_1(x_1), h_2(x_1), \ldots, h_5(x_1)]$, which is merged with the corresponding label y_1 to form a new data set D'_1, for use in the second stage. This procedure is then repeated for five times, according to the S-fold cross validation technique discussed in the section of Materials and methods. After all this, we obtain five new data sets, D'_1, D'_2, $\ldots$, D'_5, where $D'_i = \{H_i, y_i\}$, $i = 1, 2, \ldots, 5$. In the second stage, we apply a deep neural network as the ensemble model. To classify normal and tumor samples, we use a five-layer neural network. The input layer of the network contains five neurons, which represent features of samples in the new data set. In the hidden layers, we experiment with different numbers of nodes in each layer for better classification performance. The output layer of the network contains one neuron whose output was 0 or 1, denoting normal or tumor, respectively. In this stage, we also employ 5-fold cross validation and take the mean value to obtain the outcome.

ACCEPTED MANUSCRIPT

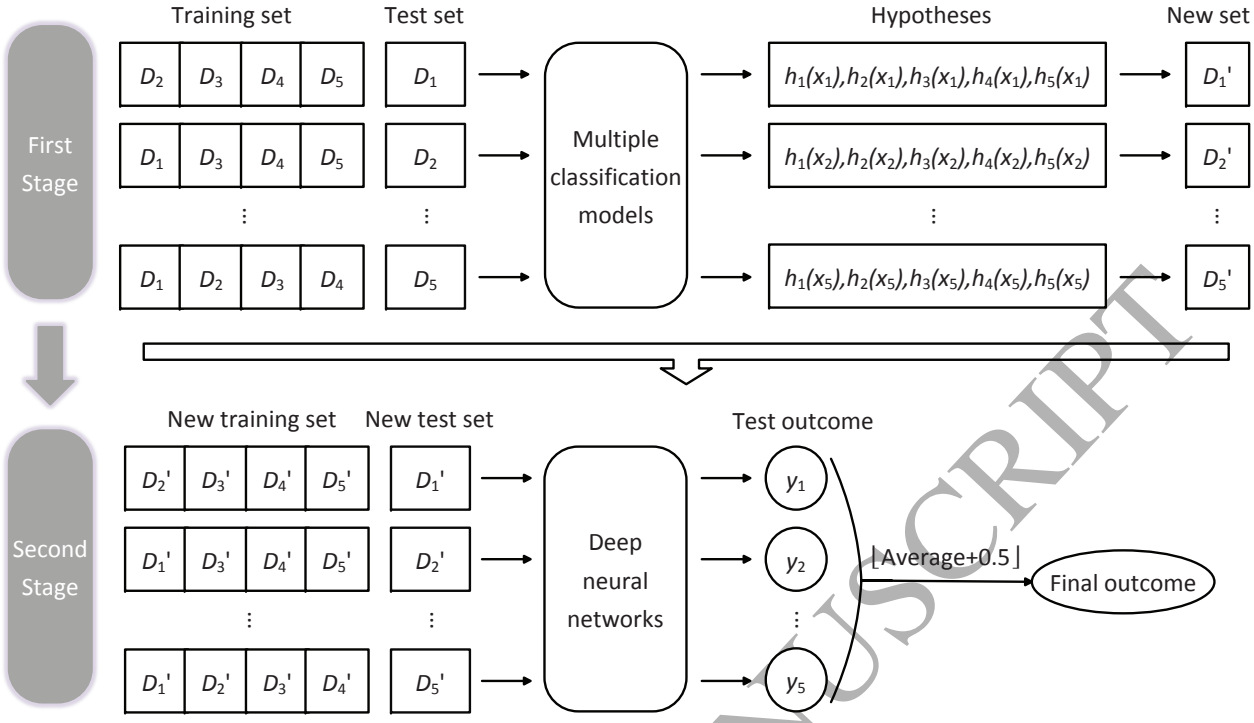


Figure 4: The deep learning-based ensemble method.

Unlike the commonly used weighted averaging and majority voting algorithms in general ensemble strategy, which merely consider the linear relationships among classifiers and need for manual participation, the deep learning-based ensemble model "learn" the relationships automatically. Generally, the relationships between the multiple classifiers and the labels of test samples are unknown, and the reliability of the prediction can not be guaranteed if only a simple linear relation is taken into account. However, the deep learning used in the second stage in our method has the ability to automatically learn intricate relationships, especially nonlinear relationships, and requires very little engineering by hand. Thus, the deep learning-based multi-model ensemble method can make full use of the information provided by data and guarantee the prediction results.

3 Results

3.1 Data collection

We evaluated the proposed method on three RNA-seq data sets of three kinds of cancers, including Lung Adenocarcinoma (LUAD), Stomach Adenocarcinoma (STAD) and Breast Invasive Carcinoma (BRCA). The gene expression data were obtained from the TCGA project web page [19]. These data sets, which include all stages of cancers, were collected from subjects of various clinical conditions and different ages, genders and races. As described in the profile [20], the tumor tissues from patients not treated with prior chemotherapy or radiotherapy were selected. The specific information of the data sets is shown in Table 1. In our procedure, we used both the raw count data and the normalized fragments per kilobase per million (FPKM) data. The raw count data were used to selected the significantly differentially expressed genes and the normalized FPKM data were used to the following classification and ensemble procedure.

ACCEPTED MANUSCRIPT

Table 1: Data sets information

Data set	Genes	Samples		
		Tumor	Normal	Total
LUAD	20532	125	37	162
STAD	29699	238	33	271
BRCA	20532	775	103	878

3.2 Gene selection of normal and tumor samples

We analyzed gene differential expression between normal samples and cancer samples in three data sets. In order to filter the significantly differentially expressed features, the technique of DESeq was employed. Through the analysis and comparison, the genes that satisfied the following conditions were considered most differentially expressed [21]: (1) the BH-adjusted p-value less than 0.01; (2) the fold change threshold of 4; (3) the mean FPKM of each gene in all the samples larger than 2. For the LUAD data, 1385 differentially expressed genes were selected. For the STAD data, 801 genes were selected. For the BRCA data, 934 genes were selected. The selected genes satisfying the estimate threshold settings were significantly differentially expressed in different cancer data sets where the difference was greater than that in randomly selected genes.

We utilized the three mostly used evaluation metrics to compare the prediction performance with the entire set of genes and the differentially expressed genes identified previously: precision, recall and accuracy. Precision is defined as the fraction of correctly identified cancer patients, recall measures the proportion of predicted cancer patients to all the people sampled, and accuracy is the weighted average of precision and recall, denoting the overall correctness. Both the mean values and the standard deviations were calculated for multiple test sets in cross validation. These evaluation metrics were also used in the following assessment of classification models.

We applied a DT classifier to train and test the entire data and selected data separately. The obtained results are shown in Table 2. We observe that better accuracy and a better tradeoff between recall and precision are represented by feature selection. Besides, the classification performance is more stable on the selected data. The computation time is also compared, revealing the importance of selecting features. Therefore, we would use the selected data as the input of the subsequent procedure, in consideration of a more accurate prediction of cancer as well as a greatly reduced running time of classification.

Table 2: The precision, recall, accuracy and CPU time of the entire data and selected data analyzed by DTs

Data set	Data size	Precision(%)	Recall(%)	Accuracy(%)	CPU time(s)
LUAD	Entire set	97.10(±4.29)	97.37(±2.63)	95.60(±2.19)	0.0532
	Selected set	**98.46(±2.29)**	**97.37(±2.63)**	**96.80(±2.28)**	**0.0039**
STAD	Entire set	96.69(±1.22)	**97.22(±1.39)**	94.63(±2.04)	0.8608
	Selected set	**99.42(±0.80)**	96.67(±3.75)	**96.59(±3.80)**	**0.0182**
BRCA	Entire set	96.60(±1.26)	97.34(±0.98)	94.62(±1.64)	0.0591
	Selected set	**97.77(±0.96)**	**97.42(±0.61)**	**95.76(±0.94)**	**0.0040**

ACCEPTED MANUSCRIPT

3.3 Multi-model ensemble based on neural networks

We first applied five classification methods in the first stage individually, which were k-nearest neighbor, support vector machines, decision trees, random forests and gradient boosting decision trees, and then averaged the predictions derived from these methods after using 5-fold cross validation technique. Then, we went a step further to employ the multi-model ensemble method to integrate all the first-stage predictions using a deep neural network. With 5-fold cross validation, the dimensionality of the new data set has decreased and the sample size has increased significantly, which provides the possibility for the application of deep neural networks and also provides more information for the prediction.

Table 3: The predictive accuracy (%) of individual and ensemble methods on three data sets

Classification algorithm	LUAD	STAD	BRCA
kNN	88.00(±5.10)	93.90(±2.99)	95.08(±0.89)
SVM	97.20(±2.28)	81.22(±22.50)	79.55(±19.22)
DT	96.80(±2.28)	96.59(±3.80)	95.76(±0.94)
RF	93.20(±1.79)	96.83(±1.85)	94.17(±1.53)
GBDT	96.80(±2.28)	96.59(±2.64)	95.76(±4.46)
Majority voting	97.20(±1.79)	98.54(±1.34)	98.18(±0.73)
Proposed method	**98.80(±1.79)**	**98.78(±1.44)**	**98.41(±0.41)**

We compare the predictive accuracy of each individual method, the majority voting and our proposed ensemble method on three data sets, which are the LUAD, STAD and BRCA data sets. The receiver operating characteristic (ROC) curve is also used to compare the performance of different methods. In statistic, the ROC curve is a graphical plot that illustrates the performance of a binary classifier as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (*i.e.*, true positive rate) against the fraction of false positives out of the negatives (*i.e.*, false positive rate) at various thresholds. Generally, the area under the curve (AUC) is estimated as an important measurement for model comparison, which reflects the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Because of the imbalance of the samples we used, the precision-recall (PR) curve is required to deal with the highly skewed data. The area under the PR curve is typically used to measure the relationship between the precision and recall and the performance of a classifier. A large area represents both high precision and high recall.

The predictive accuracy results are shown in Table 3. In the table, we describe the recognition rate of our proposed ensemble method compared with five single classifiers and the majority voting algorithm for each cancer data set. From the table, it is clear that the integration strategy of multiple models significantly outperforms classification models using in isolation and results in more stable performance. In addition, our deep learning-based multi-model ensemble method obtains better predictions than the majority voting, raising the accuracy to 99.20%, 98.78% and 98.41%, for the LUAD, STAD and BRCA data sets, respectively.

The three ROC curves are shown in Fig. 5 for the three cancer data sets, respectively. According to the results in the figures, the integration strategy gets higher AUC scores by combining multiple different classifiers, thus obtains better classification performance than each classifier operating alone. Furthermore, the AUC scores of the proposed deep learning-based ensemble method are higher than that of the majority voting for all three data sets, owing to its ability to learn and discover hidden structure automatically.

The three corresponding PR curves are shown in Fig. 6. In the figure, we observe that the proposed ensemble

ACCEPTED MANUSCRIPT

method obtains an area that is larger than or the same as that of each single classifier and the majority voting. We also observe that the proposed ensemble method deals well with skewed data, which reflects the imbalance of clinical samples.

4 Discussion

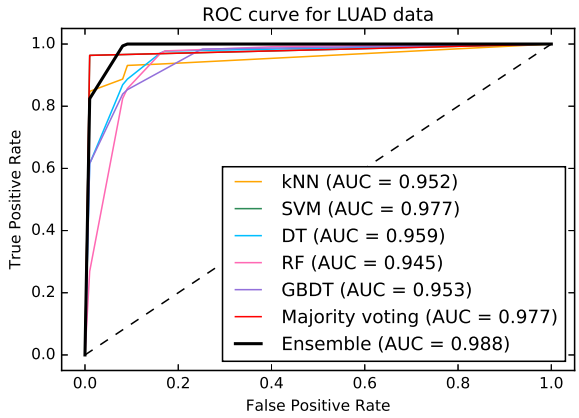
Based on the results, we observe that the proposed deep learning-based multi-model ensemble method yields satisfactory results that are superior to single classifiers and the majority voting algorithm in cancer prediction. Due to the complexity and high mortality of cancer, timely and accurate diagnosis is critical. Thus, improving the prediction accuracy by applying computer-aided techniques is of great help to cancer treatment.

In the study, we made a comparison between the multi-model ensemble method we have proposed in this paper and five different classification models acting solo. The five classifiers are classical and advanced ones that have been widely used in cancer prediction. According to the observations in a previous study [6], for SVMs and RFs, each classifier may outperform the other on different data sets. The same situation may happen for other classifiers, which indicates that each method has its own shortcomings relative to others. It is this observation that has motivated us to propose the strategy of integrating different classifiers in order to obtain a more accurate and unbiased classification model. Our results on three data sets show that the multi-model ensemble method leads to higher accuracy than all the five classifiers acting solo on all the data sets. In addition, the ROC curves indicate that a single classifier exhibits unstable prediction performance for different data sets. This is probably a consequence of different sensitivities of classifiers to different data distributions, sample sizes and redundant features. However, by going a step further to ensemble the outputs of the five classifiers, our proposed method continues to train the weight of each classifier. In this procedure, classifiers with higher accuracy have a greater role to play and interference information of the classifiers with lower accuracy is excluded. Therefore, the advantages of each classifier are fully considered and utilized, and better prediction performance is obtained.

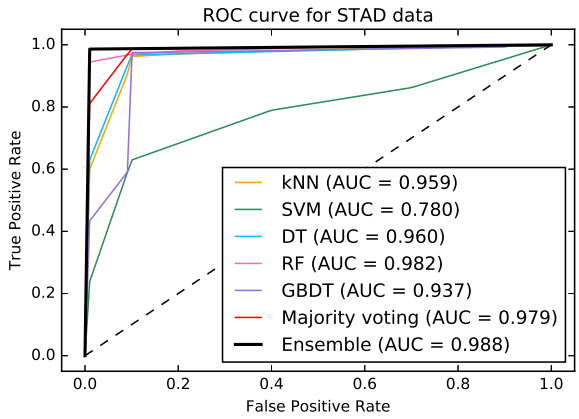
Additional tests were performed to compare the prediction accuracy of the proposed deep learning-based ensemble method with the majority voting algorithm. As a commonly used ensemble approach in various fields, the majority voting algorithm is also employed in cancer prediction. Cho and Won [11] observed a better classification performance of the majority voting than SVMs and kNN on cancer data sets, which is also confirmed by our results. Furthermore, we observed that our deep learning-based ensemble method obtains a higher accuracy and AUC score than the majority voting algorithm. The results may be attributed to the fact that the majority voting algorithm does not regard the weights of different classifiers and only considers linear relationships. As compared to the majority voting algorithm, the deep learning used in the ensemble stage in our proposed method automatically learns hidden intricate structures, including nonlinear structures. Through the training of deep learning, the unknown relationships among classifiers and the label of samples are discovered and fitted to the best, so that both the outputs of different classifiers and the relationships among them are fully taken into account. Consequently, higher accuracy of cancer prediction is achieved. Our results also confirm that deep learning, with the ability to fit complex relationships, especially nonlinear relationships, and with very little engineering by hand, will be of great use in taking advantages of increases in the amount of information and data. We believe that the application of deep learning in the field of disease diagnosis will be very promising with a broad development space.

We have to point out that the deep learning-based multi-model ensemble method incurs a higher computational cost. To overcome this limitation to a certain extent, we applied the feature selection technique in the data preprocessing phase, which greatly reduces the running time and improves the prediction accuracy in the same time. With the rapid increase in the amount of gene expression data and the variety of features, feature selection

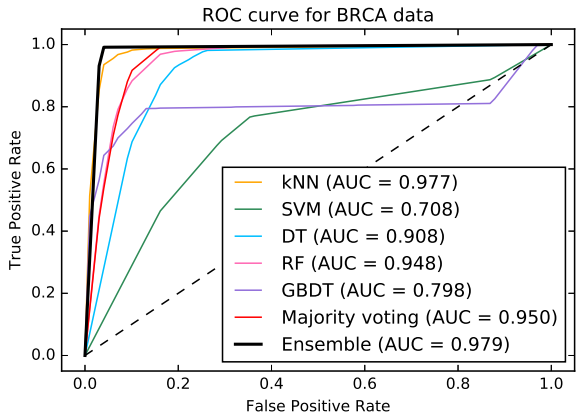
ACCEPTED MANUSCRIPT



(a) LUAD



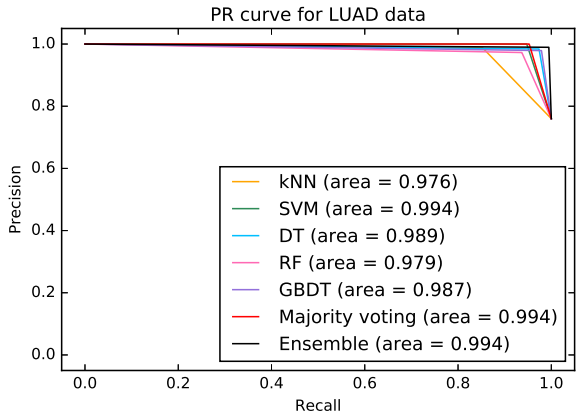
(b) STAD



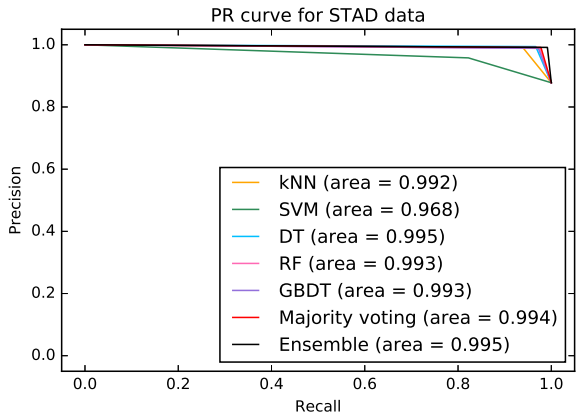
(c) BRCA

Figure 5: The ROC curves for individual and ensemble methods on three data sets: (a) LUAD; (b) STAD; (c) BRCA.

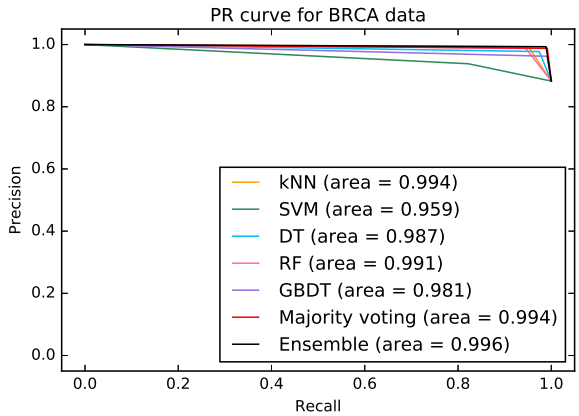
ACCEPTED MANUSCRIPT



(a) LUAD



(b) STAD



(c) BRCA

Figure 6: The precision-recall curves for individual and ensemble methods on three data sets: (a) LUAD; (b) STAD; (c) BRCA.

ACCEPTED MANUSCRIPT

is very important and necessary. Overall, feature selection in the discovery of important genes and in the study of pathology deserves more attention.

5 Conclusions

Cancer is a major health problem worldwide. Although the machine learning methods have been more and more widely used in cancer prediction, no one method outperforms all the others. In this paper, we presented a deep learning-based multi-model ensemble approach to the prediction of cancer. Specifically, we analyzed gene expression data obtained from three kinds of tissues, lung, stomach and breast. In order to avoid over-fitting in classification, we identified differentially expressed gene data between normal and tumor phenotypes with the DESeq technique. The results show that differential expression analysis is necessary to reduce the dimensionality of data and to select effective information, thus increasing the accuracy of the prediction and reducing the computational time to a large extent. The multi-model ensemble method then utilizes the predictions of multiple different models as inputs to a deep neural network, which is trained to combine the model predictions to form an optimal final prediction. The majority voting algorithm combines the predictions from different classifiers as a contrast. We analyzed the three kinds of cancer data on five classifiers separately as well as on the majority voting method and our proposed multi-model ensemble method. The results show that the proposed ensemble model outperforms every other classifier as well as the majority voting in various evaluation metrics. The deep learning-based multi-model ensemble method reduces the generation error and obtains more information by using the first-stage predictions as features than it is trained in isolation. Moreover, by using deep learning, the intricate relationships among the classifiers are learned automatically, thus enabling the ensemble method to achieve better prediction.

Conflict of interest

The authors do not have financial and personal relationships with other people or organizations that could inappropriately influence (bias) their work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant No. 31270210. The authors would like to thank the reviewers in advance for their comments and suggestions.

References

[1] About cancer. National Cancer Institute, February 2015.

[2] GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. http://globocan.iarc.fr/.

[3] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

ACCEPTED MANUSCRIPT

[5] E. Sayed, A. Wahed, I. A. Emam, and A. Badr. Feature selection for cancer classification: An SVM based approach. *International Journal of Computer Applications*, 46(8):20–26, 2012.

[6] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1):1–10, 2008.

[7] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of International Conference on Machine Learning*, volume 96, pages 148–156, 1996.

[9] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[10] A. C. Tan and D. Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2003.

[11] S. B. Cho and H. H. Won. Machine learning in dna microarray analysis for cancer classification. In *Asia-Pacific Bioinformatics Conference*, pages 189–198, 2003.

[12] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov. Applications of deep learning in biomedicine. *Molecular Pharmaceutics*, 2016.

[13] H. Hijazi and C. Chan. A classification framework applied to cancer gene expression profiles. *Journal of Healthcare Engineering*, 4(4):255–84, 2012.

[14] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):1–12, 2010.

[15] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 1.3. Springer Science+Business Media, 2006.

[16] L. Breiman. Random forest. *Machine Learning*, 45:5–32, 2001.

[17] http://scikit-learn.org/stable/.

[18] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. *Journal of Machine Learning Research*, 15, 2010.

[19] The TCGA Database. http://cancergenome.nih.gov/.

[20] Cancer Genome Atlas Research Network. *Nature*, 513(7517):202–209, 2014.

[21] J. Wu, X. Zhao, Z. Lin, and Z. Shao. A system level analysis of gastric cancer across tumor stages with RNA-seq data. *Molecular BioSystems*, 11(7):1925–1932, 2015.

ACCEPTED MANUSCRIPT