EasyChair Preprint

№ 8421

Classification of Cancer Subtypes Based on Multi-Granularity Cascade Forest

Chunxiao Jiang and Hua Duan

July 10, 2022

MDPI

Article

Classification of cancer subtypes based on multi-granularity cascade forest

Chunxiao Jiang [1,‡], Hua Duan [1,‡,*]

[1] School of Mathematics and Systems Science, Shandong University of Science and Technology 266000, China,1479817921@qq.com(C.J);huaduan59@163.com(H.D)
* Author to whom correspondence should be addressed.
‡ These authors contributed equally to this work.

Abstract: Treatment options are different for different cancer subtypes.It is of great significance for cancer patients and medical field to determine the types of cancer subtypes in time.For some redundant genes and noisy genes in the sample data of cancer subtypes,decision trees were used for feature selection,which effectively improved the classification performance of the classification model.In order to solve the problem of over-fitting caused by traditional machine learning methods in classification of cancer subtypes,multi-granularity Cascade forest (gcForest),an algorithm combined with machine learning and deep neural network,was applied.Comparing gcForest with support vector machine,logistic regression,random forest and K-nearest Neighbor method,the experimental results show that gcForest has better performance than other traditional machine learning algorithms.

Keywords: cancer subtype data;gcForest;decision tree;multi-classification

Citation: Lastname, F.; Lastname, F.; Lastname, F. Classification of cancer subtypes based on multi-granularity cascade forest. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

0. Introduction

With the improvement of human material conditions,all kinds of diseases,especially cancer, also pose a great threat to people's survival and health.Cancer is one of the main diseases leading to human death.Cancer is a common disease especially for middle-aged and elderly people.The number of cancer cases and deaths worldwide has exploded,witlh about 14.1 million new cancer patients and 8.2 million new deaths occurred in 2012 alone[1].According to the cancer statistics report,about 1.6 million people are newly diagnosed with cancer and 1.3 million people die of cancer every year in China[2].

From a microscopic perspective, the occurrence of cancer is the result of uncontrolled proliferation of cells, the essential cause of which is genetic variation and epigenetic variation[3][4][5].Cell proliferation and apoptosis are important processes to maintain normal operation of the body.Cells multiply by dividing, and at the same time, the chromosomes that carry the genetic information replicate themselves.Apoptosis is an initiative programmed death process.These two processes maintain the balance of cell number and ensure the normal expression of body function.However, when mutations occur, the balance between cell proliferation and apoptosis is broken, and cells begin to proliferate, spread and spread, forming cancer.So cancer is fundamentally a genetic disease.Genes are deoxyribonucleic acid (DNA) fragments carrying genetic information on chromosomes, which are diverse. Under the influence of external factors, gene mutations produce different combinations, which will cause diseases such as cancer and seriously threaten human life[6][7].

Cancers that appear in the same category are subdivided into different subtypes according to the different inducing genes,and different cancer subtypes have distinct prognostic responses and treatment outcomes to treatment regimens.The discovery and identification of cancer subtypes is of great importance in the treatment of cancer,and it is a key basis for providing personalized and precise treatment for cancer patients. Since Golub[8] et al first used genomic data to classify cancer in 1999, research on various cancers

by using genomic data has gradually attracted attention. By using genomic sequencing technology to obtain cancer genomic data,researchers can classify cancer subtypes at the molecular level.

1. Related work

Cancer subtype data are characterized by high dimension, small sample size and sample imbalance. In order to improve the learning ability of the model, an effective method is to remove the redundant features in the data and select the most effective features to represent the original data for training, that is, feature selection on the data set.Feature selection methods can be divided into three categories: filtering method, wrapping method and embedding method. Filtering method refers to the use of a standard to measure the relationship between each feature and the sample category, and the first K features in the last order are selected as the feature set[9], which is a very common method. Golub[8] et al proposed Signal to Noise Ratio method for feature gene selection for the first time using gene expression profiles. Zhu[10] et al used t-test method to identify genes with significant differences. Liang[11] et al combined the Two methods of Person correlation coefficient and Signal to Noise Ratio to select mixed features. Guyon[12] et al used mutual information to judge the relationship between features for feature selection. The filtering method has the advantages of simple principle and simple calculation, but ignores the relationship between genes, resulting in the existence of redundant genes, which makes the computational complexity very large, and even affects the accuracy of classification. The winding method uses classifiers to estimate the performance of feature subsets and adjusts the feature subsets according to the results. Peng[13] proposed a feature selection method based on hybrid genetic algorithm and support vector machine. Abualigah[14] combined clustering with particle swarm optimization for feature selection. Diao[15] combined fuzzy rough set and harmony search algorithm to achieve feature selection.Dash[16] uses rough sets and nonlinear analysis for feature selection. Winding method takes the classification accuracy of selected gene subset as the judgment standard, so the classification accuracy of this method is relatively high. However, this method calls classifiers for many times, resulting in high computational cost, and its biomedical significance is unclear. Embedding method is to embed the feature selection process into the training process of classifier. Li[17] et al adopted the integration method of recursive classification tree for feature selection.Uğuz[18] used principal component analysis to achieve feature selection.Guyon[19] proposed a recursive feature gene elimination method based on support vector machine (SVM-RFE).RamÓn[20] proposed the feature selection method of random forest integration.The advantages of this method are that feature selection and classifier are embedded, and the computation is reduced and the classification accuracy is improved.However, this method is highly dependent on the classifier and the results of gene selection are not universal.The decision tree algorithm adopts this kind of model structure.The experimental results show that the classification model with the feature selection method has better prediction performance than that without the feature selection method.

In recent years, machine learning and cancer genome mapping have been widely used in cancer research[21][22]. Many traditional classification models, such as decision tree method (DT), support vector machine (SVM), K nearest neighbor method (KNN), logistic regression (LR), and random forest (RF), have been used to classify cancer subtypes based on gene expression data.With the development and application of deep learning, neural network and other deep learning methods are increasingly used in cancer subtype classification. However, due to the complexity of neural network, its hyperparameters are difficult to adjust, easy to fall into overfitting, which seriously affects the classification performance of the model.Due to the characteristics of cancer subtype samples, the application of deep neural network in cancer subtype classification still has certain limitations.

In order to solve the above problems of neural network and avoid the risk of overfitting due to small sample size, Zhou and Feng[23] proposed a new decision tree integration method–gcForest model.This model utilizes multi-layer learning of deep learning to avoid

over-fitting due to small sample size. Similar to deep neural network, gcForest has a multi-layer linked structure, and each layer contains many random forests. gcForest is composed of two parts.The first part is multi-granularity scanning, which adopts sliding window structure to scan sample data from top to bottom and input them into different random forests.The second part is the cascade forest, according to the data to automatically determine the number of cascading layers. Comparing the gcForest model with other traditional models, the experiment shows that the gcForest model has better performance than other traditional models.

2. Problem description

Firstly, the data is preprocessed to remove outliers and normalize the data.The processed data is then input into the model, which consists of a first-stage feature selection process and a second-stage classification process. In the first stage, feature selection is carried out by decision tree. In the second stage, the output of the first stage is classified by gcForest,which is a combination of machine learning and neural network. Finally, the model is compared with SVM, LR, KNN and RF classification models. in Figure 1 shows the overall workflow.

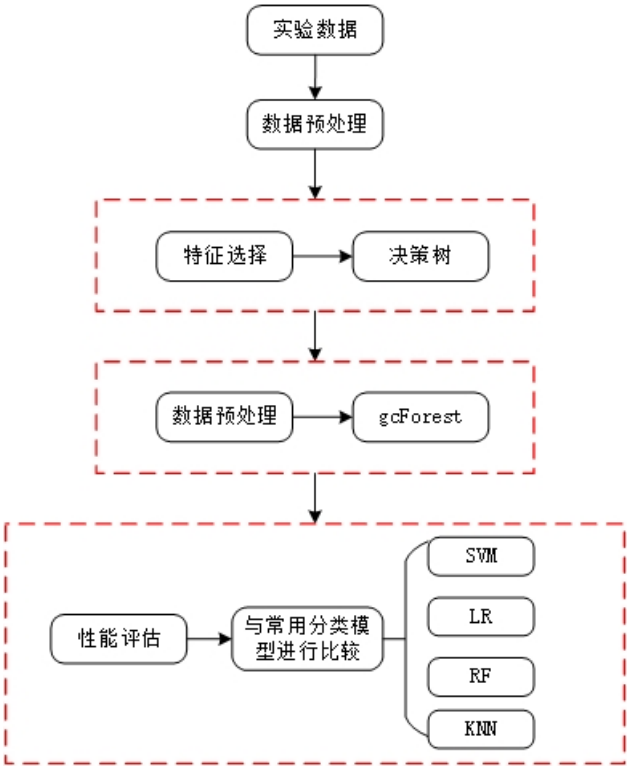

Figure 1. Flowchart of classification of cancer subtypes.

3. Model building

3.1. Feature selection

Information entropy is the most commonly used index to measure the purity of sample set.Assuming that the proportion of the k-th sample in the current set of sample D is $P_k = (1, 2, \cdots, |y|)$, the information entropy of D is defined as:

$$Ent(D) = -\sum_{1}^{|y|} p_k \log_2^{p_k} \tag{1}$$

The lower the value of $Ent(D)$, the higher the purity of D.

Information theory is a branch of science that deals with information quantitatively. The change of information before and after dividing data sets is called information gain. Assuming that discrete attribute a has V possible values $\{a^1, a^2, \cdots, a^v\}$, if a is used to divide the sample set D, v branch nodes will be generated, where the v branch node contains all the samples in D whose value is a^v on attribute a, namely D^v.The information entropy of D^v can be calculated according to the information entropy formula. Considering that different branch nodes contain different number of samples, weight $|D^v|/|D|$ is assigned to each branch node, that is, the branch node with more samples has greater influence. Thus, the information gain obtained by dividing sample set D with attribute a can be calculated:

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{D} Ent(D^v) \tag{2}$$

Generally speaking, the larger the information gain is, the greater the "purity improvement" obtained by using attribute a for partitioning. Therefore, the information gain can be used for attribute partitioning of the decision tree. In fact, it is to select the attribute with the largest information gain.

Decision tree is a tree structure, in which each internal node represents a judgment on an attribute, each branch represents the output of a judgment result, and finally each leaf node represents a classification result, which is essentially a tree composed of multiple judgment nodes. Decision tree algorithm is a recursive method to select the optimal feature and segment the training data according to the optimal feature, so that each sub-data set can have a best classification process.This process corresponds to the division of feature space and the construction of decision tree.

In the process of decision tree construction, the most important thing is to divide data sets. The biggest principle of dividing data sets is to make disordered data more orderly, which depends on information gain in information theory. The feature with the highest information gain is the optimal choice.Therefore, decision trees can be applied to feature selection.

3.2. Classification model

The gcForest model is composed of two parts. The first part is multi-granularity scanning. The original features generate feature vectors through sliding windows and input these feature vectors into different types of random forest for training and output class vectors.Then connect the class vector of the output, that is, the output of the multi-granularity scan.As shown in Figure 2, it is assumed that there are 300 original features, the sliding window size is 100 dimensions, and the sliding step size is 1. Sliding a feature window generates a 100-dimension feature vector, and a total of 201 feature vectors are generated.If the original features have spatial relations, as shown in Figure 3, assume that the image feature size is 20×20, the sliding window size is 10×10, and the sliding step size is 1, 121 matrices of size 10×10 can be obtained through the sliding feature window.

The second part is the cascade forest structure. The class vector output from the first part is input into the cascade forest.Figure 4 shows the specific structure of the cascaded forest. Each cascade layer is composed of decision tree forests.Suppose there are three classes, the multi-granularity scanning part each forest generates three-dimensional class vector, each layer contains two random forests and two completely random forests, each forest contains 500 trees.Each layer receives the features processed by the previous layer, and takes the output results of the features processed by each layer and the original features as the input of the next layer for layer by layer processing.In order to reduce the risk of over-fitting, k-fold cross validation is used to verify the whole cascade structure. When the accuracy rate does not increase, the number of layers will not increase and the training process is terminated. Therefore, the number of layers of the cascade structure can be determined automatically.Finally, all the output class vectors are averaged, and the class with the highest probability value is the predicted result.

Figure 5 shows the overall structure of gcForest.Suppose there are 300 original features, the samples are divided into three categories, and the two sliding Windows are 100 and 200 in size, respectively, and the step size is 1.When the size of the sliding window is 100, the original features are trained by random forests and completely random forests, and 1206 dimensional feature vectors are output.When the size of the sliding window is 200, the original features are trained by random forests and completely random forests to output 606-dimensional feature vectors, and then the 1206-dimensional feature vectors and 606-dimensional feature vectors are successively input into the cascaded forest part for training.

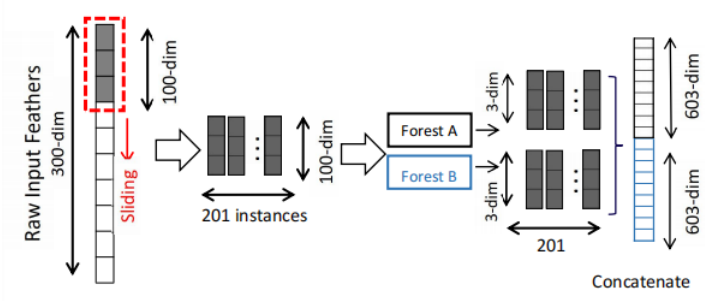

Figure 2. Sequence feature multi-granularity scanning.

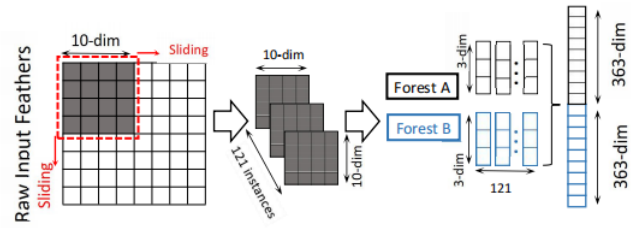

Figure 3. Image feature multi-granularity scanning.

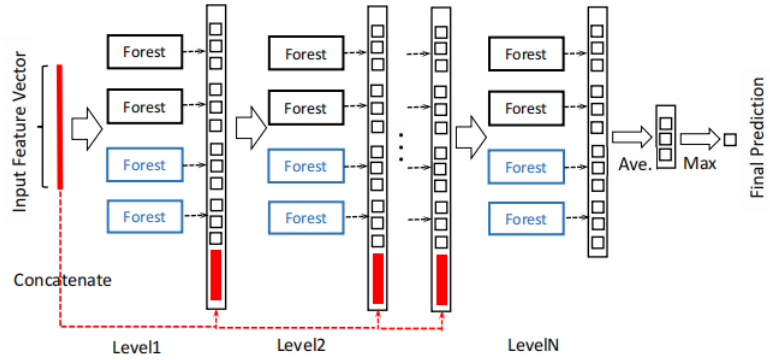

Figure 4. Cascade forest section.

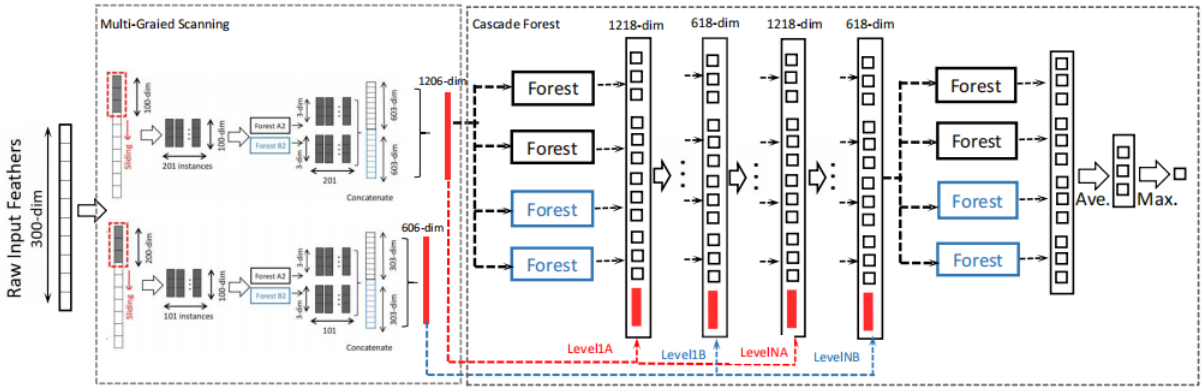

Figure 5. The overall structure of gcForest.

4. Experiment

4.1. Data sets and parameters

The Cancer Genome Atlas (TCGA) was developed by The National Cancer Institute,(NCI) and the National Human Genome Research Institute (NHGRI) launched a joint project in 2006 to document clinical data on various Human cancers, including subtypes of tumors,These data include genome variation, message Ribonucleic Acid(mRNA) expression,micro Ribonucleic Acid (miRNA) expression and methylation, and are important sources of cancer data.Data sets of Serous Cystadenocarcinoma, NOS and other 19 cancer subtypes were downloaded from TCGA, including 4240 samples in total. The data sets of 19 cancer subtypes are shown in Table 1.Since the data of cancer subtypes contained some noisy genes and redundant genes, the data were preprocessed firstly, including outlier deletion and data normalization, and then feature selection was carried out using decision trees.

Table 1. Information on a dataset of 19 cancer subtypes.

Cancer subtypes	Sample size
UCEC-Serous cystadenocarcinoma, NOS	128
UCEC-Endometrioid adenocarcinoma, NOS	377
THCA-Papillary carcinoma, follicular variant	104
THCA-Papillary adenocarcinoma, NOS	352
PRAD-Adenocarcinoma, NOS	369
OV-Serous cystadenocarcinoma, NOS	261
LIHC-Hepatocellular carcinoma, NOS	282
LGG-Oligodendroglioma, NOS	100
LGG-Oligodendroglioma, anaplastic	67
LGG-Mixed glioma	110
LGG-Astrocytoma, NOS	54
LGG-Astrocytoma, anaplastic	108
LAML-Acute myeloid leukemia, NOS	173
KIRP-Papillary adenocarcinoma, NOS	222
KIRC-Clear cell adenocarcinoma, NOS	515
HNSC-Squamous cell carcinoma, NOS	419
GBM-Glioblastoma	153
COAD-Mucinous adenocarcinoma	68
COAD-Adenocarcinoma, NOS	378

According to the characteristics of the samples, two random forests and two completely random forests are adopted in each layer of the cascaded forest, and each forest is composed of 110 trees, so as to ensure the diversity of the integration.Parameter Settings of gcForest classification model are shown in Table 2.

Table 2. gcForest parameter settings.

Parameter	Value
n_mgsRFtree	30
tolerance	0
stride	1
cascade_test_size	0.3
n_cascadeRF	4
n_cascadeRFtree	110
min_samples_mgs	10
min_samples_cascade	7
cascade_layer	np.inf
n_jobs	1

4.2. Experimental results and analysis

Confusion matrix is a case analysis table that summarizes the prediction results of classification model in machine learning. In the form of matrix, the situation in the data set is summarized according to the real category and the category predicted by classification model.The obfuscation matrix can better observe the performance of the model on each category and make the category more distinctive.Figure 6 is the confusion matrix of gcForest model in each category.

The confusion matrix can more intuitively reflect the quality of the model. All the correct prediction results are on the diagonal, and all the wrong prediction results are outside the diagonal, so it can be intuitively seen where there are mistakes.As can be seen from Figure 6, the number of samples outside the diagonal is small, that is to say, there are few wrong prediction results, indicating that the classification performance of the model is good.

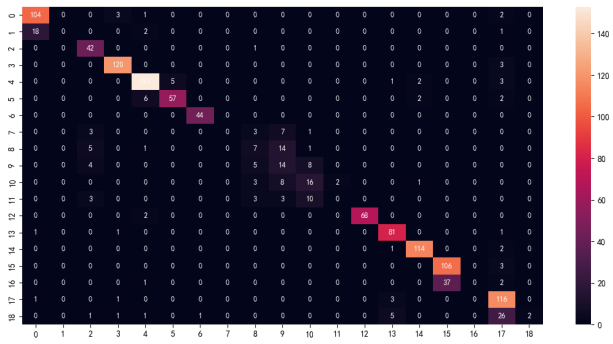

Figure 6. The overall structure of gcForest.

gcForest was compared with four traditional machine learning algorithms (SVM, LR, RF and KNN), and the commonly used evaluation indexes were considered: accuracy, precision, recall, F1 score, etc. To evaluate the performance of the algorithm, 70% of the samples were randomly selected for training and 30% for testing.

As can be seen from Table 3,the accuracy of gcForest model is the highest at 0.836, and RF (0.804) is the highest among other traditional methods.gcForest is 0.022 higher than RF,indicating that THE performance of gcForest is superior to other models.The precision , recall and F1 score of each method were calculated.As can be seen from Table 3, the gcForest classifier has achieved better results compared with SVM, RF, KNN and LR classifiers, so the overall performance of gcForest is better than other models and has better classification performance.

Table 3. Performance comparison of different classification algorithms.

Classification algorithm	Accuracy	F1-score	Precision	Recall
SVM	0.718	0.51	0.529	0.534
RF	0.804	0.641	0.657	0.639
KNN	0.771	0.583	0.599	0.59
LR	0.711	0.495	0.509	0.511
gcForest	0.836	0.612	0.684	0.64

In addition, in order to more comprehensively evaluate the overall performance of SVM, RF, KNN, LR and gcForest classifier,Receiver Operating Characteristic(ROC) curves were used to compare the robustness of different classification models. Figure 7 shows the average ROC curves of different models on TCGA dataset after training respectively.The more convex ROC curve is, the better the classification performance is.However, if the ROC curves of two classifiers are very similar, it will be difficult to compare their performance,so Area Under The Curve(AUC) value is needed for comparison. AUC is the area under ROC curve[24]. The higher the AUC value, the better the classification performance.

In the multi-classification problem, there are two calculation methods for AUC value, micro and macro respectively.Micro method is to divide m classification problems into M dichotomies and obtain M confusion matrices.Then, the True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR) and False Negative Rate (FNR) values corresponding to m confusion matrices were added and the mean values were calculated respectively.Finally, calculation accuracy rate, accuracy rate, recall rate, etc.Macro method first divides m classification problems into m dichotomies and obtains M confusion matrices. However, different from micro method, Macro method calculates the accuracy, accuracy and recall rate of m confusion matrices, and finally calculates the average value.This paper studies a multi-classification problem, and the horizontal and vertical coordinates of ROC curve are FPR and TPR,so the Micro method is adopted to calculate the AUC value of the evaluation model.As can be seen from Figure 7, gcForest has a higher AUC value and a better classification effect.

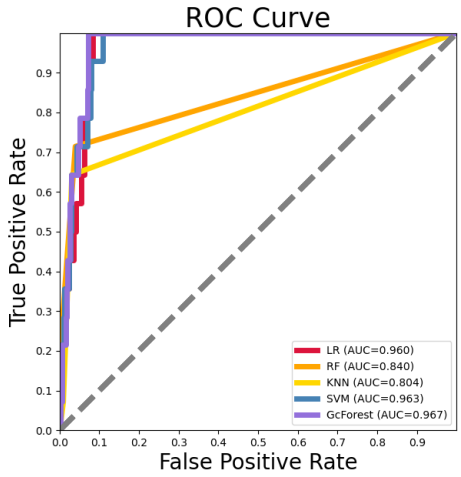

Figure 7. The overall structure of gcForest.

5. Conclusion

gcForest algorithm is a combination of traditional machine learning algorithm and deep learning, and it is a decision tree integration method. Its hyperparameters are much less than those of deep neural network, and its model complexity depends on automatic data determination.According to the characteristics of the sample data, the sample is preprocessed, the outliers are removed, and the data is normalized.Then the decision tree

method is applied to feature selection of sample data.As can be seen from Figure 6, gcForest has good classification performance.It can be seen from Table 4 that the performance of classification model using decision tree for feature selection is better than that without feature selection.Therefore, the application of decision tree in feature selection of cancer subtype data can effectively improve the classification performance of the model.

Table 4. Comparison of classification performance of gcForest with and without feature selection.

Classification algorithm	Accuracy	F1-score	Precision	Recall
gcForest without decision tree	0.83	0.604	0.58	0.636
gcForest	0.836	0.612	0.684	0.64

In order to verify that the performance of gcForest classification model has better robustness, gcForest is compared with KNN, SVM, LR and RF.As can be seen from Table 3, the accuracy of gcForest is 11.8% higher than SVM, 12.5% higher than LR, 3.2% higher than RF, and 6.5% higher than KNN. The performance of gcForest classification model has better robustness.In addition, their accuracy rate, recall rate and F1 score were compared, and their ROC curves were drawn. It can be seen from Table 3 and Figure 7 that gcForest classification model has better classification performance than other classification models.

Classification of cancer subtypes is crucial for cancer treatment and diagnosis.In this paper, we focus on the gcForest classification model and evaluate the gcForset classification model based on the TCGA sample dataset of 4240 cancer patients.As the cancer subtype samples are characterized by high dimension, small sample size and unbalanced data, the cancer subtype samples are selected by decision tree, which effectively improves the classification performance of gcForest classification model.Moreover, compared with other traditional machine learning algorithms, gcForest algorithm is obviously superior to other traditional machine learning algorithms.However, there are still some limitations and problems to be solved. For example, in some extreme class imbalance and high-dimensional small-scale data sets, gcForest needs to be further improved to improve its stability.In addition, it has been proved in recent years that the fusion of multiple omics data can help improve the classification performance of cancer subtypes[25][26].In this study, we focused only on the classification of cancer subtypes based on gene expression data, which is also one of the problems.In the future, the gcForest classification model will be applied to multi-omics data classification tasks.

6. Patents

Funding: Acknowledgement This work was supported in part by NSFC (61702306), Sci. & Tech. Development Fund of Shandong Province of China (2016ZDJS02A11, ZR2017BF015 and ZR2017MF027), SDUST Research Fund (2015TDJH102 and 2019KJN024), and Shandong Chongqing Science and Technology Cooperation Project (cstc2020jscx-lyjsAX0008).

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in https://cancergenome.nih.gov/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stewart B W;Wild C P.World Cancer Report.*J* **2014**.
2. Zhi wei D; You lin Q, Lian di L I. Report of Chinese cancer control strategy.*J*, Bulletin of Chinese Cancer, **2002**, *5*, 4-14.
3. Sting J; Caldas C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis.*J*, Nature Reviews Cancer, **2007**, *7(10)*, 791-799.
4. Bianchini G; Iwamoto T; Qi Y.Prognostic and therapeutic implications of distinct kinaseexpression patterns in different subtypes of breast cancer.*J*, Cancer research, **2010**, *70(21)*, 8852-8862.

5. Heiser L M; Sadanandam A; Kuo W L; Benz S C. Subtype and pathway specific responses toanticancer compounds in breast cancer.*J*, Proceedings of the National Academy of Sciences, **2012**, *109(8)*, 2724-2729.
6. Prat A; Parker J S; Karginova O.Phenotypic and molecular characterization of the claudin-lowintrinsic subtype of breast cancer.*J*, Breast Cancer Research, **2010**, *12(5)*, 1-18.
7. Jahid M J; Huang T H; A Personalized Committee Classification A pproach to ImprovingPrediction of Breast Cancer Metastasis.*J*,Bioinformatics, **2014**, *30(13)*, 1858-1866.
8. Golub TR; Slonim D K; Tamayo P.Molecular Classification of Cancer: Class Discovery and ClassPrediction by Gene Expression Monitoring.*J*, science, **1999**, *286(5439)*, 531-537.
9. Chan W H; Mohamad M S; Deris S.Identification of informative genes and pathways using animproved penalized support vector machine with a wei ghting scheme.*J*, Computers in Biology &Medicine, **2016**, *77(C)*,102-115.
10. Zhu S, Wang D, Yu K.Feature Selection for Gene Expression Using Model-Based Entropy.*J*, IEEE/ACM Transactions on Computational Biology &Bioinformatics, **2010**, *7(1)*, 25-36.
11. Goh L; Song Q; Kasabov N. A novel feature selection method to improve classification of gene expression data.*C 2004*, 161-166.
12. Guyon I; Elisseeff A. An introduction to variable and feature selection[J]. Journal ofMachine Learning Research.*J* **2003**, *3(6)*, 1157-1182.
13. Peng S; Xu Q; Ling X B. Molecular classification of cancer types from microarray data using thecombination of genetic algorithmms and support vector machines.*J*, FEBS Letters, **2003**, *555*,358-362.
14. Abualigah L M; Khader A T; Hanandeh E S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm.*J*, Journal of Computational Science, **2018**, *25*,456-466.
15. Diao R D R; Shen Q S Q.Two new approaches to feature selection with harmony search.*C* **2010**,1-7.
16. Dash S; Patra B.Rough set aided gene selection for cancer classification.*C* **2012**,290-294.
17. Xia L I; Harbin. An N ovel Ensemble Method of Feature Gene Selection Based on Recursive Partition-Tree.*J*, Chinese Journal of Computers, **2004**, *27(5)*,675-682.
18. Uğuz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm.*J*, Knowledge-Based Systems, **2011**, *24(7)*,1024-1032.
19. Guyon I; Weston J; Barnhill S. Gene Selection for Cancer Classification using Support VectorMachines.*J*, Machine Learning,,**2002**, *46(1-3)*,389-422.
20. Ramón Díaz-Uriarte; Sara Alvarez de Andrés. Gene selection and classific ation of mi croarray data usingrandomforest.*J*, Bmc Bioinformatics,,**2006**,*7(1)*,3-10.
21. Sun D; Wang M; Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multidimensional data.*J*, IEEE/ACM transactions on computational biology and bioinformatics,**2018**, *16(3)*,841-850.
22. Becker A S; Marcon M; Ghafoor S. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer*J*, Investigative radiology,**2017**,*52(7)*,434-440.
23. Zhou Z H; Feng J. Deep forest.*J*, National Science Review,**2019**,*6(1)*,74-86.
24. Zhu Q; Pan M; Liu L. An ensemble feature selection method based on deep forest for microbiome-wide association studies.*C* **2018**,248-253.
25. Bhattacharyya M; Nath J; Bandyopadhyay S. MicroRNA signatures highlight new breast cancer subtypes.*J*,Gene,**2015**,*556(2)*,192-198.
26. Cantini L; Isella C; Petti C. MicroRNA–mRNA interactions underlying colorectal cancer molecular subtypes.*J*, Nature communications, **2015**,*6(1)*,1-12.