Laboratory Investigation (2021) 101:430–441
https://doi.org/10.1038/s41374-020-00525-x

ARTICLE

Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data

Fei Deng[1] · Jibing Huang[1] · Xiaoling Yuan[2] · Chao Cheng[3,4] · Lanjing Zhang[5,6,7,8]

Received: 31 July 2020 / Revised: 20 October 2020 / Accepted: 2 December 2020 / Published online: 11 February 2021

Abstract

Most biomedical datasets, including those of 'omics, population studies, and surveys, are rectangular in shape and have few missing data. Recently, their sample sizes have grown significantly. Rigorous analyses on these large datasets demand considerably more efficient and more accurate algorithms. Machine learning (ML) algorithms have been used to classify outcomes in biomedical datasets, including random forests (RF), decision tree (DT), artificial neural networks (ANN), and support vector machine (SVM). However, their performance and efficiency in classifying multi-category outcomes of rectangular data are poorly understood. Therefore, we compared these metrics among the 4 ML algorithms. As an example, we created a large rectangular dataset using the female breast cancers in the surveillance, epidemiology, and end results-18 database, which were diagnosed in 2004 and followed up until December 2016. The outcome was the five-category cause of death, namely alive, non-breast cancer, breast cancer, cardiovascular disease, and other cause. We analyzed the 54 dichotomized features from ~45,000 patients using MatLab (version 2018a) and the tenfold cross-validation approach. The accuracy in classifying five-category cause of death with DT, RF, ANN, and SVM was 69.21%, 70.23%, 70.16%, and 69.06%, respectively, which was higher than the accuracy of 68.12% with multinomial logistic regression. Based on the features' information entropy, we optimized dimension reduction (i.e., reduce the number of features in models). We found 32 or more features were required to maintain similar accuracy, while the running time decreased from 55.57 s for 54 features to 25.99 s for 32 features in RF, from 12.92 s to 10.48 s in ANN, and from 175.50 s to 67.81 s in SVM. In summary, we here show that RF, DT, ANN, and SVM had similar accuracy for classifying multi-category outcomes in this large rectangular dataset. Dimension reduction based on information gain will increase the model's efficiency while maintaining classification accuracy.

Introduction

Most of the biomedical data are rectangular in shape, including those of 'omics, large cohorts, population studies, and surveys. Few missing data were present in these datasets. An increasing number of human genomic and survey data have been produced in recent years [1]. Rigorous analyses on these large datasets demand considerably more efficient and more accurate algorithms, which are poorly understood.

Machine learning (ML) algorithms are aimed to produce a model that can be used to perform classification, prediction, estimation, or any other similar task [2, 3]. The

Supplementary information The online version of this article (https://doi.org/10.1038/s41374-020-00525-x) contains supplementary material, which is available to authorized users.

✉ Lanjing Zhang
lanjing.zhang@rutgers.edu

1 School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, China

2 Department of Infectious Disease, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine Shanghai, Shanghai, China

3 Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA

4 The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX 77030, USA

5 Department of Pathology, Princeton Medical Center, Plainsboro, NJ, USA

6 Department of Biological Sciences, Rutgers University, Newark, NJ, USA

7 Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA

8 Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA

unknown dependencies/associations are estimated based on a given dataset and later can be used to predict the output of a new system or dataset [2–4]. Therefore, ML algorithms have been used to analyze large biomedical datasets [5–7]. Studies have compared the accuracy of several ML algorithms for classifying microarray or genomic data [8–10], and show a superior performance of random forests (RF). However, only a few studies accessed the accuracy of ML algorithms in classifying multi-category outcomes, which are important for the more-in-depth understanding of biological and clinical processes. Only when we can accurately predict or identify the factors linked to multicategory outcomes, we can provide patients with more targeted prevention and treatment for the most likely outcome. Hence, we aimed to understand the performance and efficiency of RF, decision tree (DT), artificial neural networks (ANN), and support vector machine (SVM) algorithms in classifying multi-category outcomes of rectangular-shaped biomedical datasets.

As an example, we used a large population-based breast cancer dataset with long-term follow-up data to create a large rectangular dataset. The reasons were: (1) cancer is one of the most common diseases [11]. Knowledge gained by studying cancer can be easily generalized to other biomedical fields; (2) breast cancer is the second most common cancer in U.S. women [11, 12] and would provide a sufficiently large sample size ($n = 45{,}085$); (3) breast cancers diagnosed in 2004 had 12 years of follow-up on average and very few missing outcomes/follow-ups were anticipated; (4) breast cancers diagnosed in 2004 would have a moderately good prognosis and their outcomes will be rather diversified (i.e., many patients might not die of breast cancer). Therefore, this study is designed to systematically compare the performance and efficiency of DT, RF, SVM, and ANN algorithms in classifying multi-category causes of death (COD) in a large biomedical dataset (breast cancers).

Methods

Data analysis

We obtained individual-level data from the surveillance, epidemiology, and end results-18 (SEER-18) (www.seer.cancer.gov) SEER*Stat database with treatment data using SEER*Stat software (Surveillance Research Program, National Cancer Institute SEER*Stat software (seer.cancer.gov/seerstat) version <8.3.6>) as we did before [13–15]. SEER-18 is the largest SEER database including cases from 18 states and covering near 30% of the U.S. population. The SEER data were de-identified and publicly available. Therefore, this was an exempt study (category 4) and did not require an institutional review board (IRB) review. All incidental invasive breast cancers of SEER-18 diagnosed in 2004 were included and had the follow-up to December 2016. The individual deaths were verified via certified death certificates (2019 data-release). We chose the diagnosis year of 2004 with the consideration of the implementation of the sixth edition of the tumor, node, and metastasis staging manual (TNM6) of the American joint commission on cancer (AJCC) in 2004. Moreover, we only included the primary-cancer cases that had a survival time >1 month, age of 20+ years, and known COD.

The features (i.e., variables) were dichotomized for more efficiency and slightly better performance [7], while the actual values were also tuned for and analyzed using RF models. A total of 54 features were included (Supplementary Table 1). We conducted correlation analyses and produced a correlation matrix (Supplementary Fig. 1) to identify the closely correlated factors. The outcomes of the classification models were the patient's five-category COD. The COD was originally classified using SEER's recodes of the causes of death, which were collected through death certificates of deceased patients (https://seer.cancer.gov/). We simplified the SEER COD into five categories based on the prevalence of COD [16–18], including alive, non-breast cancer, breast cancer, cardiovascular disease (CVS), and other cause.

The most common task of ML techniques in the learning process is classification [19]. The tenfold cross-validation approach was used to tune all models, which is also termed as model optimization (see Supplementary methods) as described before [20, 21]. This approach is an approximation of leave-p-out cross-validation has the advantage of repeat the process to reduce variance and being less sensitive to the partitioning of the data than the holdout method. Briefly, the samples were randomly divided into ten same-size subsamples, among which one subsample remained as the validation set and the other nine as the test set. The (cross-)validation would be performed ten times as the test subsample was shuffled among the ten subsamples. The mean of the cross-validation's performances would be calculated and used as the performance of the model. Multinomial logistic regression (MLR) and the ML analyses were carried out using MATLAB (version 2018a, MathWorks, Natick, MA).

Model tuning

The detailed model tuning process is described in Supplementary material. Several DT methods, such as CHAID, CART, and exclusive CHAID are available with MATLAB [22]. We used classification and regression tree(CART) to predict the categories, using the Gini index as the split criterion and 100 iterations for each run (Supplementary

Fig. 2). There are no default RF packages/toolboxes in MATLAB's own toolbox. We thus used the Randomforest-Matlab open-source toolbox developed by Jaiantilal et al. [23, 24].

In tuning RF models, the parameter nTrees, which was to set how many trees in a random forest, and may have an impact on the classification results. We set the value of nTrees from 1 to 600 separately, the results show that if this parameter is not too few (greater than ten), the accuracy of recognition can reach 69–70% (Supplementary Fig. 3). Therefore, we set this parameter to 136 in RF-based analyses. The value of Mtry node for the best model performance was identified by setting the parameter value from 1 to 20 with 1 as the interval. We found 5 was the value, which was indeed consistent with the default value generated using the MATLAB model's default (i.e., Mtry = floor (sqrt(size(P_train,2)))).

Among different training algorithms of ANN, we used the Trainscg algorithm because it is the only conjugate gradient method that did not require a linear search. The number of input layer nodes is 54 and the number of output layer nodes is 6. To tune the ANN model, we conducted experiments of either single hidden or double hidden layers, with the node numbers ranged from 5 to 100. According to the tuning results of accuracy and mean squared error (MSE), we would set the model to double hidden layers, and the number of layers for the highest accuracy and lowest MSE.

We used the multi-class error-correcting output codes (ECOC) model the SVM modeling which allows classification in more than two classes; and the MATLAB fitcecoc function that creates and adjusts the template for SVM [25]. The Kernel functions considered in the SVM were: linear, radial basis function, Gaussian, and polynomial.

Performance analysis

We analyzed the performance metrics of each proposed model, including accuracy, recall, precision, F1 score, and specificity [26, 27]. They were defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{All}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

True positive (TP) and true negative (TN) were defined as the number of samples that are classified correctly. False positive (FP) and false negative (FN) were defined as the number of samples that are misclassified into the other mutational classes [26, 27]. The specificity or true negative rate (TNR) is defined as the percentage of mutations that are correctly identified:

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

The receiver operating curve (ROC) is a graph where recall is plotted as a function of 1-specificity. It can more objectively measure the performance of the model itself [25]. The model performance was also evaluated using the area under the ROC, which is denoted the area under the curve (AUC). An AUC value close to 1 highlight a high-performance model, while an AUC value close to 0 demonstrate a low-performance model [28, 29]. AUC is independent of the class prior distribution, class misclassification cost, and classification threshold, which can more stably reflect the model's ability to sort samples and characterize the overall performance of the classification algorithm. The formula used to determine the AUC can be written as follows [29, 30]:

$$\text{AUC} = \frac{\sum \text{TP} + \sum \text{TN}}{\text{P} + \text{N}} \tag{6}$$

where P is the total number of positive class and N is the total number of negative class.

Dimension reduction based on the information entropy and information gain

Information entropy is an indicator to measure the purity of the sample set. The formula is as follows:

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i \tag{7}$$

The measure of information gain is to see how much information a feature can bring to the classification system. The more information it brings, the more important the feature. Information gain IG(A) is the compute of the difference in entropy from start to end the set S is split on an attribute A, the information gain is defined as follows [31]:

$$\text{IG}(A, S) = H(S) - \sum_{t \in T} p(t) H(t) \tag{8}$$

where $H(S)$ is the entropy of set S, T is the subsets created from splitting set S by attribute A such that $S = U_{t \in T} t$, $p(t)$ is the proportion of the number of elements in t to the number of elements inset S, and $H(t)$ is the entropy of subset T.

Table 1 Baseline characteristics of the women with breast cancer diagnosed in 2004 that were included in this study.

	Alive, *n* (%)	Breast cancer, *n* (%)	CVS, *n* (%)	Non-breast cancer, *n* (%)	Other cause, *n* (%)
Total	28,323 (62.82)	7752 (17.19)	3120 (6.92)	1874 (4.16)	4016 (8.91)
Age 65+ years					
No	21,382 (75)	4727 (61)	417 (13)	695 (37)	849 (21)
Yes	6941 (25)	3025 (39)	2703 (87)	1179 (63)	3167 (79)
Grade 2tier					
Low	16,939 (65)	2686 (41)	1933 (69)	1057 (64)	2489 (69)
High	9265 (35)	3801 (59)	849 (31)	586 (36)	1103 (31)
T Category					
T1–2	25,813 (91)	4530 (58)	2645 (85)	1556 (83)	3443 (86)
T3–4	1450 (5)	2181 (28)	283 (9)	156 (8)	324 (8)
Unknown	1060 (4)	1041 (13)	192 (6)	162 (9)	249 (6)
N category					
0	19,410 (69)	2399 (31)	2128 (68)	1221 (65)	2744 (68)
1	6398 (23)	2253 (29)	544 (17)	338 (18)	764 (19)
2	1445 (5)	1123 (14)	174 (6)	119 (6)	184 (5)
3	518 (2)	962 (12)	74 (2)	60 (3)	94 (2)
Unknown	552 (2)	1015 (13)	200 (6)	136 (7)	230 (6)
M category					
0	27,595 (97)	5551 (72)	2876 (92)	1684 (90)	3718 (93)
1	174 (1)	1636 (21)	79 (3)	105 (6)	115 (3)
Unknown	554 (2)	565 (7)	165 (5)	85 (5)	183 (5)
Summary stage 2000 (1998+)					
Blank(s)	14,498 (51)	947 (12)	1434 (46)	881 (47)	1940 (48)
Distant	9929 (35)	2158 (28)	984 (32)	546 (29)	1246 (31)
Localized	2453 (9)	2170 (28)	341 (11)	213 (11)	400 (10)
Regional	178 (1)	1637 (21)	80 (3)	111 (6)	118 (3)
Unknown/unstaged	1265 (4)	840 (11)	281 (9)	123 (7)	312 (8)
Diagnosis confirmation					
Microscopic diagnosis	28,283 (100)	7539 (97)	3084 (99)	1852 (99)	3980 (99)
Radiologic and clinical diagnosis	33 (0)	134 (2)	30 (1)	16 (1)	33 (1)
Other	<10	79 (1)	<10	<10	<10
Histology type					
IDC	20,014 (71)	5113 (66)	2064 (66)	1220 (65)	2585 (64)
ILC	1930 (7)	590 (8)	266 (9)	136 (7)	383 (10)
MDLC	2472 (9)	592 (8)	218 (7)	135 (7)	317 (8)
IDC with mixed feature	157 (1)	38 (0)	17 (1)	6 (0)	19 (0)
ILC with mixed feature	860 (3)	153 (2)	111 (4)	58 (3)	142 (4)
Others	2890 (10)	1266 (16)	444 (14)	319 (17)	570 (14)
ER/PR receptor status					
ER− PR−	5234 (18)	2180 (28)	431 (14)	332 (18)	590 (15)
ER+ PR−	3094 (11)	1004 (13)	370 (12)	187 (10)	486 (12)
ER− PR+	393 (1)	114 (1)	20 (1)	23 (1)	26 (1)
ER+ PR+	16,630 (59)	3118 (40)	1816 (58)	1053 (56)	2316 (58)
Other/unknown	2972 (10)	1336 (17)	483 (15)	279 (15)	598 (15)

Table 1 (continued)

	Alive, *n* (%)	Breast cancer, *n* (%)	CVS, *n* (%)	Non-breast cancer, *n* (%)	Other cause, *n* (%)
Laterality					
Missing	14 (0)	45 (1)	<10	<10	<10
Left	14,162 (50)	3871 (50)	1617 (52)	929 (50)	2036 (51)
Unknown site	23 (0)	165 (2)	13 (0)	30 (2)	10 (0)
Right	14,124 (50)	3671 (47)	1487 (48)	910 (49)	1967 (49)
Surgery					
Lumpectomy	17,179 (61)	2349 (30)	1599 (51)	1000 (53)	2083 (52)
Mastectomy	10,593 (37)	3650 (47)	1297 (42)	732 (39)	1630 (41)
Other/unknown	551 (2)	1753 (23)	224 (7)	142 (8)	303 (8)
Radiotherapy					
No	12,538 (44)	4556 (59)	2001 (64)	1005 (54)	2475 (62)
Yes	15,785 (56)	3196 (41)	1119 (36)	869 (46)	1541 (38)
Chemotherapy					
No	15,240 (54)	3451 (45)	2641 (85)	1297 (69)	3238 (81)
Yes	13,083 (46)	4301 (55)	479 (15)	577 (31)	778 (19)
Percent of high school education attainment, quartile[a]					
Q1	7412 (26)	1694 (22)	717 (23)	442 (24)	931 (23)
Q2	7266 (26)	1800 (23)	773 (25)	492 (26)	1049 (26)
Q3	6719 (24)	2077 (27)	808 (26)	481 (26)	1023 (25)
Q4	6926 (24)	2181 (28)	822 (26)	459 (24)	1013 (25)
Percent of persons in poverty, quartile[a]					
Q1	7541 (27)	1706 (22)	720 (23)	454 (24)	903 (22)
Q2	7179 (25)	1854 (24)	708 (23)	487 (26)	1004 (25)
Q3	6687 (24)	2005 (26)	845 (27)	478 (26)	1089 (27)
Q4	6916 (24)	2187 (28)	847 (27)	455 (24)	1020 (25)
Percent of foreign-born residents, quartile[a]					
Q1	6695 (24)	2066 (27)	893 (29)	516 (28)	1177 (29)
Q2	7006 (25)	1950 (25)	790 (25)	458 (24)	1042 (26)
Q3	7264 (26)	1814 (23)	742 (24)	485 (26)	968 (24)
Q4	7358 (26)	1922 (25)	695 (22)	415 (22)	829 (21)
Rural urban continuum 2003					
Metro	25,411 (90)	6831 (88)	2705 (87)	1642 (88)	3452 (86)
Non-metro	2912 (10)	921 (12)	415 (13)	232 (12)	564 (14)

The cells with case number fewer than ten were statistically suppressed to protect patient privacy. T, N and M categories were classified according to the AJCC 6 TNM staging manual.

IDC invasive ductal carcinoma, *ILC* invasive lobular carcinoma, *MDLC* mixed invasive ductal and lobular carcinoma, *ER* estrogen receptor, *PR* progesterone receptor.

[a]County attributes of the year 2000 (from the U.S. Census Bureau); education attainment defined as the percent of residents with less than high-school graduate in the county; person in poverty defined as the percent of residents with income below 200% of poverty in the county.

The information gain of a feature can indicate how much information it brings to the classification system and can be used as a feature weight. When the model uses more features the classification time will be longer. Arbitrarily reducing the characteristics will likely reduce classification accuracy. Therefore, we screened the features based on the calculated information gains to achieve a balance between run time and classification accuracy. We then step-wise deleted the features of the least information of gain, which were likely the least important. Because DT and RF were both ensemble-based algorithms and had similar performances, we conducted dimension reduction with RF, ANN, and SVM models and expect similar results with DT models.

Results

Dataset characteristics and the model tuning

Of the 52,818 samples, we step-wise excluded the 5294 (~10%) cases missing tumor-grade data (_grade), 1770 (3.4%) missing survival-time and 352 (0.6%) missing laterality data (lat_bi), and 317 (0.6%) missing data of TNM6 N category. Overall, we included 45,085 cases of breast cancer diagnosed in 2004 that were included in the SEER-18 and had no missing data in all features (Table 1). The variables were dichotomized into 54 features. Likely due to the unique nature of the pathological and socio-economic factors, the correlation analyses showed only a few features had a correlation coefficient >0.95 or <−0.95 (Supplementary Fig. 1).

For DT models, the minimum number of leaf nodes with the best performance of the DT range 20–300 and peak at 93 (Supplementary Fig. 2). Therefore, the minimum number of samples contained in the leaf node was 93 for optimization. The overall classification accuracy could reach 67.3%, which was about 4% higher than the original DT, and the cross-validation error has also decreased. However, due to the uneven distribution of the data samples, some categories such as non-breast cancer and other cause were pruned, causing the loss of some data information.

For RF models, the parameter nTrees set how many trees in a RF and may have an impact on the classification results. The Mtry is the number of features randomly sampled as candidates at each split and was optimized (Supplementary methods). After tuning the models, we set the nTrees parameter to 136 in RF-based analyses with the best Mtry node value of 6 (Supplementary Fig. 3).

According to the tuning results of accuracy and MSE in ANN models, when the number of layers was greater than 20, the models' performance appeared stabilized (Table 2). Therefore, we set the model to double hidden layers, and the number of layers is 40.

For the SVM models of linear, radial basis function, Gaussian or polynomial function, we found the linear kernel function had the highest accuracy (69.06%) and shortest run-time (175.50 s, Table 3), with a one-vs-one approach.

Performance analysis results

Based on the confusion matrices (Fig. 1), the 4 ML models appeared to have similar performance (Table 4) and all were more accurate than the MLR. The best classification accuracy of DT, RF, ANN, and SVM models in this study was 69.21%, 70.23%, 70.16%, and 69.06%, respectively, and higher than 68.12% of a conventional statistical algorithm (MLR). However, to evaluate the pros and cons of a model, it is not enough to just look at the accuracy. The values of recall, TNR, and F1 can specifically reflect the classification of each category. Further comparing the metrics among the four models show that the precision-recall and F1 in classifying breast cancer by RF model were better than the DT, ANN, and SVM models. This is mainly because the RF model is an integrated learning algorithm; through the voting mechanism, one can balance a certain error. Compared with ML algorithms, MLR had lower overall accuracy and a much lower recall (1.13% vs ~50%) and a lower specificity (Table 4) for predicting the death of breast cancer.

Table 2 Performance of artificial neural network models with various numbers of layers.

Hidden layer settings	Neural network algorithm	Total accuracy (%)	MSE
Single hidden layer			
5	Trainscg	69.61	0.0852
10	Trainscg	69.34	0.0857
20	Trainscg	69.88	0.0846
40	Trainscg	70.10	0.0852
50	Trainscg	70.12	0.0852
60	Trainscg	69.45	0.0861
70	Trainscg	69.54	0.0860
80	Trainscg	69.37	0.0867
90	Trainscg	68.88	0.0876
100	Trainscg	69.10	0.0865
Double hidden layers			
[5,5]	Trainscg	69.63	0.0857
[10]	Trainscg	69.72	0.0853
[20]	Trainscg	69.83	0.0851
[40]	Trainscg	70.16	0.0852
[50]	Trainscg	69.88	0.0854
[60]	Trainscg	65.06	0.1398
[70]	Trainscg	69.70	0.0863
[80]	Trainscg	69.48	0.0885
[90]	Trainscg	69.21	0.0882
[100,100]	Trainscg	69.28	0.0867

MSE mean squared error.

ROC analysis results

After a comparative analysis of the above performance metrics, we found that the RF model was superior to the DT, ANN, and SVM models. However, in the classification of other causes, the ANN model has a higher recognition rate than the RF model, and the F1 values of non-breast cancer cannot be calculated. Therefore, we used the ROC curve to further analyze non-breast cancer and other cause in RF and ANN models.

Table 3 Classification accuracies and efficiencies of support vector machine models by the kernel function.

Kernel function	Alive (%)	Non-breast cancer (%)	Breast cancer (%)	CVS (%)	Other cause (%)	Total accuracy (%)	Run-time (s)
Linear	96.22	0.00	39.43	0.00	0.00	69.06	175.50
Radial basis function	95.60	1.26	6.23	1.29	6.27	63.86	286.31
Gaussian	95.60	1.26	6.23	1.29	6.27	63.86	290.28
Polynomial	27.75	3.14	32.79	13.83	36.51	27.46	3710.50

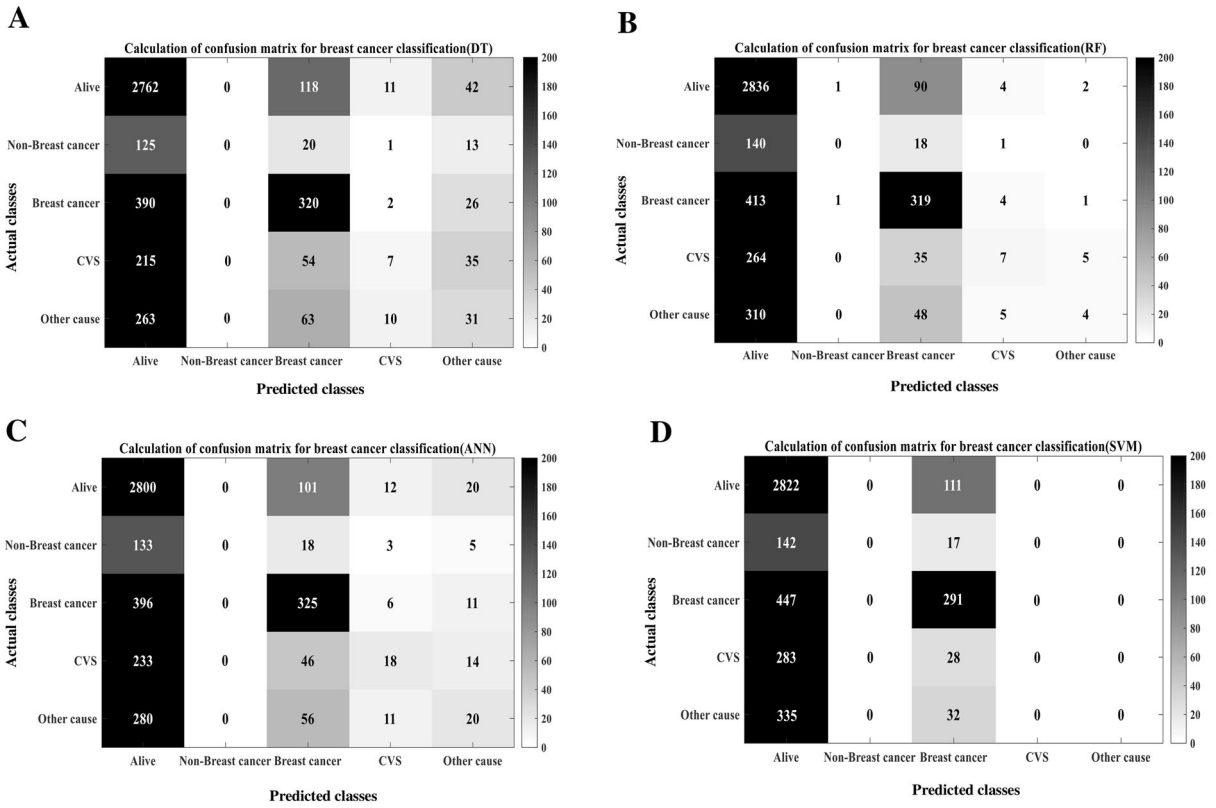

Fig. 1 Confusion matrices of the 4 tuned machine learning models. There were similar performance metrics of decision tree (**A**), random forest (**B**), artificial neural networks (**C**), and support vector machine (**D**) models.

The AUC of the 5 COD in the RF model was overall lower than those in the ANN model (Fig. 2). Because the AUC index can measure the performance of the model more objectively, the results show that the overall performance of the ANN model is similar or better than that of the RF model.

Dimension reduction based on the information entropy and information gain

This dataset had 54 categorical/binary features. Based on the training datasets, the information entropy and information gain of 54 features were calculated (Supplementary Table 3). Using information entropy and information gain, we obtained the following important features: age >65; surgery; TNM6 metastasis subgroup1; AJCC stage 4; TNM6 metastasis subgroup2; surgery–other; TNM6 tumor subgroup1; TNM6 tumor subgroup2; AJCC stage 1; surgery-lumpectomy; TNM6 lymph node subgroup5; and TNM6 lymph node subgroup1. Then we characterized the key features in a step-wise fashion (Supplementary Fig. 4).

We successfully reduced the data dimension based on information gain and shortened the run times in RF, ANN, and SVM models, while maintaining the overall classification accuracy (Table 5 and Supplementary Tables 3–5). Removal of features with low information gain (0.0000–0.0005) in RF models led to a slight increase in the alive class and the overall accuracy rates, while no accuracy changes in CVS and breast cancer classes. The classification of CVS was always 100%; accuracy rate of alive class and the overall had a slight improvement, respectively about 0.5% and 0.3%; breast cancer and CVS had no significant

Table 4 The performance of the decision tree, random forests, artificial neural networks, and support vector machine models.

Statistical Measure	Alive (%)	Non-breast cancer (%)	Breast cancer (%)	CVS (%)	Other cause (%)
Decision tree					
Accuracy	69.21				
Precision	73.56	NA	55.65	22.58	21.09
Recall	94.17	0.00	43.36	2.25	8.45
Specificity (TNR)	26.50	100.00	91.65	99.23	96.38
F1	82.60	NA	48.74	4.09	12.06
Random forest					
Accuracy	70.23				
Precision	71.56	0.00	62.55	33.33	33.33
Recall	96.69	0.00	43.22	2.25	1.09
Specificity (TNR)	22.65	99.94	93.71	99.56	99.75
F1	82.25	NA	51.12	4.22	2.11
Artificial neural networks					
Accuracy	70.16				
Precision	72.88	NA	59.52	36.00	28.57
Recall	95.47	0.00	44.04	5.79	5.45
Specificity (TNR)	25.84	100.00	92.78	98.99	98.43
F1	82.66	NA	50.62	9.97	9.15
Support vector machine					
Accuracy	69.06				
Precision	70.04	NA	60.75	NA	NA
Recall	96.22	0.00	39.43	0.00	0.00
Specificity (TNR)	19.43	100.00	93.75	100.00	100.00
F1	81.07	NA	47.82	NA	NA
Multinomial logistic regression					
Accuracy			68.12		
Precision	69.71	61.10	13.73	50.00	22.73
Recall	96.38	42.66	1.13	0.54	1.24
Specificity (TNR)	29.33	73.98	85.71	90.41	99.59
F1	80.90	50.25	2.10	1.07	2.35

CVS cardiovascular disease, *NA* not applicable, *TNR* true negative rate.

changes; non-breast cancer accuracy was slightly reduced, and the classification effect is unstable. Therefore, the features with low information gain (0.0000–0.0005) may be considered as redundant features, and deleted in the models, while the running times were scientifically reduced. We also found similar changes in ANN and SVM models.

Feature importance in RF models using the data of convention encoding or one-hot encoding

Our previous works have shown that one-hot encoding (dichotomization of features) led to a slight increase in prediction accuracy of the RF model on prostate cancer using Stata [7]. Consistent with that finding, our tuned RF model on actual values had a prediction accuracy of 69.70%, which was slightly lower than the accuracy of 70.23% on dichotomized features (Supplementary Table 6). However, the top-five important features were different in the two models, except age 65+ years (vs <65 years, Supplementary Figs. 5 and 6).

Discussion

We here compared the performance and efficacy of DT, RF, ANN, and SVM in classifying five-category outcomes of a large rectangular database (54 features and ~45,000 samples). The accuracy in classifying five-category COD with DT, RF, ANN, and SVM was 69.21%, 70.23%, 70.16%, and 69.06%, respectively. It is noteworthy that the accuracy

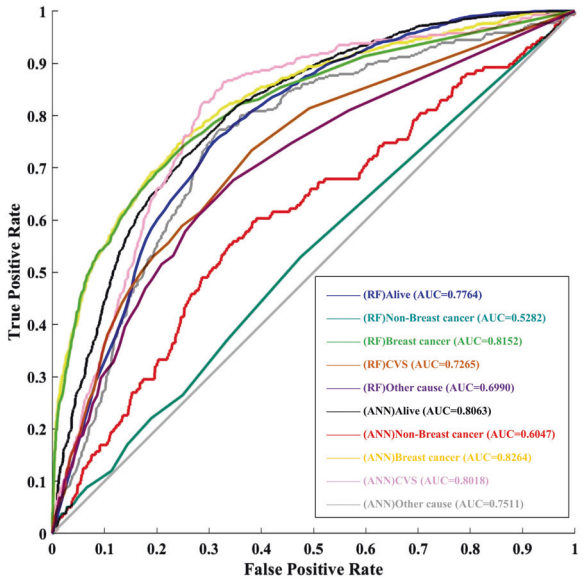

Fig. 2 The receiver operator curve of the tuned random forests (RF) and artificial neural networks (ANN) models by 5 causes of death. The areas under curve (AUCs) in the RF model were overall lower than those in the ANN model.

in classifying five-category outcomes is much more difficult than classifying binary-category outcomes since it depends on four sequential classification processes in one-over-all or one-over-rest algorithms (i.e., final accuracy is the accuracy multiplication for four times). Moreover, based on the information entropy and information gain of feature values, we could reduce the feature number in a model to 32 and maintained a similar accuracy, while the running time decreased from 55.57 s for 54 features to 25.99 s for 32 features in RF, from 12.92 s to 10.48 s in ANN and from 175.50 s to 67.81 s in SVM. The DT algorithm was not tested after dimension reduction for its lower performance than RF and its theoretical framework (ensemble-based) like RF.

Few studies to our knowledge investigated the dimension reduction of multicategory classification. A two-stage approach has been shown to effectively and efficiently select features for balanced datasets [32], but no specific reduction in run time was reported. We here show that dimension reduction and efficiency improvement can be achieved by removing features of low to medium information gain (<0.0005) in RF, ANN, and SVM models, which apparently have little effect on the overall classification performance. Such a strategy may be applied to other ML models in classifying unbalanced large rectangular datasets, while caution should be used when classifying outcomes in a balanced dataset.

Consistent with a previous report [7], one-hot encoding (i.e., dichotomization of the features) produced slightly better prediction accuracy in our data than conventional encoding and also has the advantage of no need for normalization. The top-five important features in the models based on the dichotomized and actual-value features differed considerably except age. This finding further confirms the important role of age in modeling five-category COD. However, five of the top ten important features shared in these two models, including advanced age, pathologic staging, surgery status, N category, and histology type. These features thus should be carefully examined, recorded, and considered for predicting or preventing five-category COD in breast cancer patients.

The four included ML algorithms each have their own theoretical frameworks. DT is a logical-based ML approach [33]. The structure of the DT is similar to a flowchart. Using top-down recursion, the classification tree produces the category output. Starting from the root node of the tree, test and compare property values on its internal node, then determine the corresponding branch, and finally reach a conclusion in the leaf node of the DT. This process is repeated at each node of the tree by selecting the optimal splitting features until the cut-off value is reached [34]. A leafy tree tends to overtrain, and its test accuracy is often far less than its training accuracy. By contrast, a shallow tree can be more robust and be easy to interpret [35].

DT works by learning simple decision rules extracted from the data features. But RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently, and with the same distribution for all trees in the forest [36]. When RF used for a classification algorithm, the deeper the tree is, the more complex the decision rules and the fitter the model [36, 37]. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Random decision forests overcome the problem of overfitting of the DTs and are more robust with respect to noise.

ANN technique is one of the artificial intelligence tools. It is a mathematical model that imitates the behavior characteristics of animal neural networks [38]. This kind of network carries on distributed parallel information processing by adjusting the connection between a large number of internal nodes, so as to achieve the purpose of processing information. After repeated learning and training, network parameters corresponding to the minimum error are determined, and the ANN model classifies the output automatically from the dataset.

SVM is another popular ML tool based on statistical learning theory, which was first proposed by Vapnik and his colleagues [39, 40]. Unlike traditional learning methods, SVMs are approximate implementations of structural risk minimization methods. The input vector is mapped to a high-dimensional feature space through some kind of nonlinear mapping which was selected in advance. An optimal

Table 5 Information-gain-based dimension reduction in the three machine learning models.

Accuracy	Alive (%)	Non-breast cancer (%)	Breast cancer (%)	CVS (%)	Other cause (%)	Total accuracy (%)	Time, mean (s)
Feature change							
Random forests							
Raw data (54 features)	96.69	0.00	43.22	2.25	1.09	70.23	55.57
Delete IG < 0.0002(46 features)	96.66	0.00	42.14	2.25	0.54	69.99	46.63
Delete IG < 0.0005(40 features)	96.52	0.00	42.14	1.93	0.27	69.85	37.03
Delete IG < 0.001(32 features)	96.52	0.00	41.87	1.93	0.27	69.81	25.99
Delete IG < 0.005(21 features)	96.22	0.00	41.19	2.89	0.27	69.57	14.59
Delete IG < 0.01(15 features)	96.59	0.00	40.11	2.57	1.09	69.68	10.15
Delete IG < 0.02(10 features)	97.89	0.00	33.33	0.64	1.09	69.28	5.94
Delete top three (7 features)	97.75	0.00	31.84	0.00	0.00	68.81	3.85
Artificial neural networks							
Raw data (54 features)	95.47	0.00	44.04	5.79	5.45	70.16	12.92
Delete IG < 0.0002(46 features)	95.29	0.00	43.36	0.64	4.09	69.48	8.92
Delete IG < 0.0005(40 features)	95.36	0.00	44.17	4.50	7.36	70.19	10.32
Delete IG < 0.001(32 features)	95.74	0.00	42.41	4.50	5.18	69.96	10.48
Delete IG < 0.005(21 features)	95.77	0.00	41.60	3.54	2.18	69.54	12.15
Delete IG < 0.01(15 features)	96.73	0.00	39.70	2.57	0.54	69.65	8.83
Delete IG < 0.02(10 features)	97.82	0.00	33.88	0.00	0.82	69.25	9.66
Delete top three (7 features)	97.92	0.00	29.95	0.00	0.00	68.61	11.30
Support vector machine							
Raw data (54 features)	96.22	0.00	39.43	0.00	0.00	69.06	175.50
Delete IG < 0.0002(46 features)	96.22	0.00	39.43	0.00	0.00	69.06	135.55
Delete IG < 0.0005(40 features)	96.18	0.00	39.43	0.00	0.27	69.06	100.38
Delete IG < 0.001(32 features)	96.22	0.00	39.43	0.00	0.27	69.08	67.81
Delete IG < 0.005(21 features)	96.28	0.00	39.02	0.00	0.00	69.03	45.18
Delete IG < 0.01(15 features)	96.15	0.00	38.75	0.00	2.18	69.08	32.37
Delete IG < 0.02(10 features)	97.68	0.00	31.71	0.00	0.00	68.74	28.04
Delete top three (7 features)	97.89	0.00	31.44	0.00	0.00	68.83	18.26

IG information gain.

classification hyperplane is constructed in this feature space, to maximize the separation boundary between the positive and negative examples [39, 40]. Support vectors are the data points closest to the decision plane, and they determine the location of the optimal classification hyperplane.

The 4 ML algorithms had different strengths and weaknesses, while all outperformed conventional statistical algorithm (MLR). The RF algorithm in our study seems to have the best overall performance for its lack of being unable to classify some CODs and the best overall accuracy. Despite the similar classification accuracy (~70%), the DT algorithm could not accurately classify the non-breast cancer group. Given the similar theoretical framework, we did not access its performance after dimension reduction. The ANN algorithm in our study is most efficient before and after dimension reduction. Surprisingly, we also notice a small increase in accuracy after dimension reduction which warrants further investigation. The SVM algorithm in our study appears very sensitive to the subgroup size (i.e., number of samples) and was not able to classify two of the five COD, although it also acceptable classification accuracy.

The study's limitations should be noted when applying our findings to other databases. First, this type of rectangular database is typical in survey and population-study, but not so in computational biology. The major difference is the large p in 'omics datasets vs the large n in epidemiological datasets, which referred to feature number and sample number, respectively. Second, some of the outcomes were not accurately classified. It is likely owing to the unbalanced outcome distribution. On the other hand, such an undesired situation reflexes real-world evidence/experience. Further studies are needed to improve the classification accuracy in the classes of fewer samples. Third, our works were exclusively based on the MATLAB platform and may not be applicable to other platforms such as R or python. Fourth, a molecular subtype of the cancer was not available due to the lack of human epidermal growth factor receptor 2 (Her2) data. Fifth, the prediction accuracy of the ML algorithms was moderately acceptable (~70%), although ML had much better recall and sensitivity than MLR. The challenge in predicting multi-category outcomes remains outstanding, despite the use of tuned ML algorithms. Thus, future works are needed to improve the prediction performance. Finally, ideally, we should use a large database to validate our models, but it is very difficult to curate and apply the tuned models to another large database that is similar to the SEER database.

In summary, we here show that RF, DT, ANN, and SVM algorithms had similar accuracy, but outperformed MLR, for classifying multi-category outcomes in this large rectangular dataset. Dimension reduction based on information gain will significantly increase the model's efficiency while maintaining classification accuracy.

Acknowledgements We thank Lingling Han at Shenzhen Horb Technology Corporate, Ltd. for invaluable discussions and comments.

Author contributions FD, CC, and LZ designed the study, FD and JH conducted the study and drafted the manuscript, all authors discussed, revised, and edited the manuscript, and LZ supervised the work.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Liu DD, Zhang L. Trends in the characteristics of human functional genomic data on the gene expression omnibus, 2001–2017. Lab Investig. 2019;99:118–27.
2. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inf. 2007;2:59–77.
3. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.
4. Bishop CM. Pattern recognition and machine learning. New York, NY, USA: Springer; 2006.
5. Chow ZL, Thike AA, Li HH, Nasir NDM, Yeong JPS, Tan PH. Counting mitoses with digital pathology in breast phyllodes tumors. Arch Pathol Lab Med. 2020;144:1397–400.
6. Koo J, Zhang J, Chaterji S. Tiresias: context-sensitive approach to decipher the presence and strength of MicroRNA regulatory interactions. Theranostics. 2018;8:277–91.
7. Wang J, Deng F, Zeng F, Shanahan AJ, Li WV, Zhang L. Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model. Am J Cancer Res. 2020;10:1344–55.
8. Maniruzzaman M, Jahanur Rahman M, Ahammed B, Abedin MM, Suri HS, Biswas M, et al. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. Comput Methods Programs Biomed. 2019;176:173–93.
9. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genom. 2008;9:S13.
10. Haibe-Kains B, Desmedt C, Sotiriou C, Bontempi G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics. 2008;24:2200–8.
11. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70:7–30.
12. Goetz MP, Gradishar WJ, Anderson BO, Abraham J, Aft R, Allison KH, et al. NCCN guidelines insights: breast cancer, version 3.2018. J Natl Compr Canc Netw. 2019;17:118–26.
13. Chavali LB, Llanos AAM, Yun JP, Hill SM, Tan XL, Zhang L. Radiotherapy for patients with resected tumor deposit-positive colorectal cancer: a surveillance, epidemiology, and end results-based population study. Arch Pathol Lab Med. 2018;142:721–9.
14. Yang M, Bao W, Zhang X, Kang Y, Haffty B, Zhang L. Short-term and long-term clinical outcomes of uncommon types of invasive breast cancer. Histopathology. 2017;71:874–86.
15. Mayo E, Llanos AA, Yi X, Duan SZ, Zhang L. Prognostic value of tumour deposit and perineural invasion status in colorectal

cancer patients: a SEER-based population study. Histopathology. 2016;69:230–8.
16. Bevers TB, Helvie M, Bonaccio E, Calhoun KE, Daly MB, Farrar WB, et al. Breast cancer screening and diagnosis, version 3.2018. J Natl Compr Cancer Netw. 2018;16:1362–89.
17. Afifi AM, Saad AM, Al-Husseini MJ, Elmehrath AO, Northfelt DW, Sonbol MB. Causes of death after breast cancer diagnosis: a US population-based analysis. Cancer. 2020;126:1559–67.
18. Clough-Gorr KM, Thwin SS, Stuck AE, Silliman RA. Examining five- and ten-year survival in older women with breast cancer using cancer-specific geriatric assessment. Eur J Cancer. 2012;48:805–12.
19. Amrane M, Oukid S, Gagaoua I, Ensari T. Breast cancer classification using machine learning. stanbul: Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT); 2018. p. 1–4. https://doi.org/10.1109/EBBT.2018.8391453.
20. Mao Y, Fu Z, Dong L, Zheng Y, Dong J, Li X. Identification of a 26-lncRNAs risk model for predicting overall survival of cervical squamous cell carcinoma based on integrated bioinformatics analysis. DNA Cell Biol. 2019;38:322–32.
21. Dong RZ, Yang X, Zhang XY, Gao PT, Ke AW, Sun HC, et al. Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. J Cell Mol Med. 2019;23:3369–74.
22. Grzesiak W, Zaborski D. Examples of the use of data mining methods in animal breeding. Adem Karahoca, editor. Data mining applications in engineering and medicine. London, UK: IntechOpen Limited; 2012; 303–24.
23. Wang XC, Shi F, Yu L, Li Y. Cases analysis of MATLAB neural network. Beijing: Beijing University of Aeronautics and Astronautics. 2009. p. 59–62.
24. Jaiantilal A. Classification and regression by randomforest-matlab. (2009, 2012). https://code.google.com/archive/p/randomforest-matlab/ Accessed 22 July 2020.
25. Gonçalves CB, Leles ACQ, Oliveira LE, Guimaraes G, Cunha JR, Fernandes H. Machine learning and infrared thermography for breast cancer detection. Multidiscipl Digit Publish Inst Proc. 2019;27:45.
26. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Australasian joint conference on artificial intelligence. 2006; 1015–21.
27. Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput Sci Inf Technol. 2011;2:37–45.
28. Youssef AM, Pradhan B, Jebur MN, El-Harbi HM. Landslide susceptibility mapping using ensemble bivariate and multivariate statistical models in Fayfa area, Saudi Arabia. Environ Earth Sci. 2015;73:3745–61.
29. Costache R, Hong H, Wang Y. Identification of torrential valleys using GIS and a novel hybrid integration of artificial intelligence, machine learning and bivariate statistics. Catena. 2019;183:104179.
30. Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, et al. Landslide susceptibility mapping using J48 decision tree with AdaBoost, bagging and rotation forest ensembles in the Guangchang area (China). Catena. 2018;163:399–413.
31. Anyanwu MN, Shiva SG. Comparative analysis of serial decision tree classification algorithms. Int J Comput Sci Secur. 2009;3:230–40.
32. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol. 2010;9: Article17. https://www.degruyter.com/document/doi/10.2202/1544-6115.1492/html.
33. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21:660–74.
34. Lan T, Hu H, Jiang C, Yang G, Zhao Z. A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. Adv Space Res. 2020;65:2052–61.
35. Garcia Leiv R, Fernandez AnA, Mancus V, Casari P. A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. IEEE Access. 2019;7:99978–87.
36. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
37. Nguyen C, Wang Y, Nguyen HN. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J Biomed Sci Eng. 2013;6:551–60.
38. Jain AK, Jianchang M, Mohiuddin KM. Artificial neural networks: a tutorial. Computer. 1996;29:31–44.
39. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
40. Fradkin D, Schneider D, Muchnik I. Machine learning methods in the analysis of lung cancer survival data. DIMACS technical report 2005–35. 2006.