OPEN ACCESS

Saudi Journal of Biomedical Research
Abbreviated Key Title: Saudi J Biomed Res
ISSN 2518-3214 (Print) | ISSN 2518-3222 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Feature Extraction and Classification Analysis of High-Dimensional Biological Data Based on Dimensionality Reduction Fusion Method

Yankun Li[1#], Yulong Liu[1#], Ziyu Shang[1], Zhiyu Zheng[2], Mengting Ran[1], Zhimin Wang[3*]

[1]Department of Environmental Science and Engineering, North China Electric Power University, Baoding, 071003, PR China
[2]Department of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, PR China
[3]Affiliated Hospital of Hebei University, Department of Rheumatology and Immunology, Baoding, 071000, PR China
#indicates that the two authors contributed equally to this work

DOI: 10.36348/sjbr.2024.v09i01.001 | **Received:** 27.11.2023 | **Accepted:** 31.12.2023 | **Published:** 03.01.2024

***Corresponding author:** Zhimin Wang
Affiliated Hospital of Hebei University, Department of Rheumatology and Immunology, Baoding, 071000, PR China

Abstract

Identification and extraction of characterized information from complex high-dimensional biological data is a very meaningful issue. The dimensionality reduction fusion method based on random forest, feature extraction and neural network is proposed to recognize and classify two datasets of mRNA and lncRNA. It is shown that the proposed fusion method achieved accurate identification/classification of cancer and non-cancer groups, and simultaneously selected identity variables that have biological relevance to lung cancer (tumor) as potential biomarkers from a large number of variables. It is considered as an effective tool and theoretical support for lung cancer identification in clinical application, and it can be extended to other kinds of cancer or biological data. Ultimately, an advanced method for feature extraction and classification analysis of high-dimensional data is provided.

Keywords: High-dimensional data, Classification, Dimensionality reduction, Feature selection, Tumor biomarker.

Copyright © 2024 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)** which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

1. INTRODUCTION

Currently, with the rapid development of biological measure technology, more and more gene expression values have been determined. Nevertheless, due to the relatively small sample size and the high dimension of the data, it is difficult to find out the correlations and differences between data variables by simple observation or statistical analysis. Moreover, due to the constraints of experimental conditions, systematic errors, and possible artificial errors during the measurement process, interference and redundant information will inevitably be introduced in the acquired data. The above aspects pose a great challenge to extract potential biological information from complex high-dimensional data.

To solve the above troubles, with the help of machine learning [1-5], we can deduce the substance's essential properties from the measured data, and then classify or cluster the substances of different types. Cancer is a disease caused by somatic cell mutation, and some genetic mutations are associated with the pathogenesis of cancer [6-9]. The core analysis of cancer gene expression data mainly focuses on the screening of important genes and identification of cancer (sub-type), which can find the genes that have a strong correlation with the occurrence, progression, and metastasis of cancer [10]. Besides, from thousands of variables, selecting a few variables for experimental design can also save experimental costs and improve the efficiency of data analysis.

Since the successful realization of cancer identification based on the gene expression profile of leukemia [11], machine learning has been increasingly applied in cancer identification [12]. Random test-partial least squares discriminant analysis (RT-PLSDA) was applied to the identification of different types of cancer [13]. Data of LncRNA-miRNA-mRNA combined with support vector machine (SVM) was used for the detection of cancer and non-cancer groups [14-18]. A combinational feature selection method in conjunction with ensemble neural networks was proposed to classify tumor microarray datasets and extract latent biomarkers [19]. A co-classification approach based on the global component model and cancer component model was used for cancer recognition including leukemia and breast cancer etc. [20], Li *et al.*, [21], combined the locally linear embedding method and Fisher discriminant criterion to classify five different types of cancer. A two-

Stage Hybrid gene selection strategy by combining genetic algorithm (GA) and mutual information (MI) was applied to colon cancer, lung cancer, and ovarian cancer datasets [22]. Based on the above studies, nevertheless, further studies are needed to improve the accuracy and robustness of the cancer classification method and to search for more new tumor biomarkers with potential clinical values. Some new ideas of machine learning are gradually being applied to the field of data dimensionality reduction.

In this work, the dimensionality reduction fusion method based on random forest (RF), feature extraction and neural network (NN) was proposed and applied to two biological datasets of mRNA and lncRNA. The results show that the proposed method can accurately identify cancer and non-cancer samples, and identify key variables as potential lung cancer biomarkers from a large number of variables, which has practical value and application prospect in cancer identification. So, an effective cancer identification way and new potential lung cancer biomarkers are available. Ultimately, an advanced method for feature extraction and classification analysis of high-dimensional data is provided.

2. Theory and Algorithm

The main idea of the dimensionality reduction fusion method (called as "fusion method" for short) is as follows: Firstly, the first dimensionality reduction is conducted on the original data by processing it through RF model. Then, the second dimensionality reduction is conducted on the original data by the combination of feature extraction and NN model. Finally, the final informative/feature variables are obtained by intersecting the selected variables from the above steps.

The according specific process of the fusion method is as follows: The raw data was preprocessed using min-max scaling method, and then the preprocessed data flowed into two branches.

1. In the first branch, the data was divided into training set and testing set using a ratio of 7:3, and then fed into RF model to acquire the identification of important variables and recognition accuracy.
2. In the second branch, the data underwent Pearson correlation coefficient analysis (PCCA) to select a set of feature variables. Then the selected variables were further processed using principal component analysis (PCA). The PCA-transformed data was also split into training set and testing set using a ratio of 7:3, and inputted into the NN model, generating the recognition accuracy of the two groups.
3. Finally, the feature variables were obtained by taking the intersection of the selected variables from the above two branches.

2.1 Min-Max Scaling

To reduce the impact of large variables' expression levels on small variables' expression levels, data were normalized according to the following formula.

$$x'_{ij} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min} + \delta}\ (i = 1,2,\ldots,\mathrm{N}, j = 1,2,\ldots,\mathrm{M}) \qquad (2\text{-}1)$$

Where i is the sample number, j is the variable number, $x_j^{\min}$ and $x_j^{\max}$ represent the minimum and maximum values of the j-th variable. The parameter δ is a very small value to avoid having a denominator of 0, as in cancer dataset, $x_j^{\min}$ and $x_j^{\max}$ both may be equal to 0.

It is obvious that after the min-max scaling process, the expression level of each variable falls into 0 ~ 1. It only performs linear transformations on the data, without changing the relative size relationship of the data. By compressing the data into a fixed range, differences in data dimensions are eliminated and the impact of outliers on the data is reduced.

2.2 Pearson Correlation Coefficient Analysis (PCCA)

The calculation formula of the Pearson correlation coefficient (PCC) is:

$$r(f_i, f_j) = \frac{\mathrm{Cov}(f_i, f_j)}{\sqrt{\mathrm{Var}(f_i)\mathrm{Var}(f_j)}} \qquad (2\text{-}2)$$

Where f_i and f_j respectively represent two sets of variables, and PCC is the covariance (Cov) of the two sets of variables divided by the product of the standard deviations (Var) of the two sets of variables [23]. It can be seen that PCCA can only be calculated for two sets of variables with equal volume. For the datasets of unequal volume, the larger dataset needs to be changed with the same volume as the smaller dataset.

The higher the score of the PCC, the stronger the correlation between the variables. An extremely strong correlation between the variables is generally regarded when the coefficient is between 0.8-1.0; A strong correlation (0.6-0.8), a moderate correlation (0.4-0.6), a weak correlation (0.2-0.4) and no correlation (0.0-0.2) are generally considered. By using the above threshold and selecting the severity of filtering variables, relevant feature variables could be selected.

2.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is used to retain as much data information as possible while reducing the dimensionality of the data matrix. The goal of PCA is to determine the most meaningful principal axis that is expressed by the dataset and filters out the noise, revealing the hidden structure of the data [24].

The steps of PCA can be described as follows [25]. Each row of matrix **X** is centered by subtracting the mean value of that row to obtain zero mean normalization. Next, the covariance matrix is calculated along with its eigenvalues and corresponding eigenvectors. The eigenvalues and eigenvectors are arranged in descending order to create a matrix P. By selecting the top *k* columns of matrix P, a new matrix Y is formed through Y = XP.

2.4 Random Forest Model

RF [26-28], is randomly built in a forest, which is composed by many decision trees, and each decision tree is not associated with others. Whenever a new sample is encountered, a separate judgment is made by each decision tree in the forest to determine the category to which the sample belongs, and voting is used to select the final classification result based on the most selected samples. Firstly, the most suitable number of trees and leaf nodes with the smallest error need to be found. In this study, they were selected by calculating the mean squared error (MSE) under different number of trees and leaf nodes. The formula for the MSE is:

$$\mathrm{MSE} = \frac{1}{n} * \sum (y_i - \hat{y}_i)^2 \qquad (2\text{-}3)$$

Where MSE is the value obtained by the Loss function, *n* is the total number of samples, y_i represents the true value of the *i*-th sample, and $\hat{y}_i$ represents the predicted value of the *i*-th sample.

Then, the importance index of each variable was obtained through the training model of RF. According to the order of importance index, feature variables were retained to conduct a new training process to obtain new importance index, and then the new feature variables were retained. During this cycle, the number of variables gradually decreased, ultimately resulting in the selected variables while maintaining the highest recognition accuracy.

2.5 Neural Network Model

The bionics-inspired concept is adopted by the NN model, and the modeling is achieved through simulation of the structure and function of biological neural networks [29]. Concerning the network structure's design, it is determined based on accumulated experience and a large number of experimental results. The design of the network structure becomes relatively flexible, allowing for the design of a single hidden layer NN to accomplish the classification task.

During the NN model training, MSE loss function is utilized. MSE loss function quantifies the average error between the predicted and actual values, imposing higher penalties for more significant errors. Throughout the training, model's parameters are fine-tuned by minimizing the MSE, enhancing the overall performance. The vertical axis of the MSE loss function graph portrays the discrepancy between the predicted and actual values. A smaller vertical axis signifies predictions that are closer to the true values, implying lower loss. As the vertical axis diminishes, it indicates that the model is reducing errors and enhancing accuracy.

3. DATA AND SAMPLING

3.1 Lung Cancer Dataset 1

The data is provided by Arindam Bhattacharjee *et al.*, [30], and downloaded from https://www.pnas.org/content/98/24/13790. It contains 139 lung adenocarcinoma samples and 17 non-cancer samples, and each sample contains 12600 mRNA.

3.2 Lung Cancer Dataset 2

LncRNA expression profiles of 27 early-stage lung adenocarcinoma and adjacent nonmalignant lung samples derived from patients were used for identification. It was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113852. Each sample containing 39186 lncRNAs in raw data, and 6259 lncRNAs with $p<0.001$ (by *t*-test method) were selected for calculation.

4. RESULTS AND DISCUSSION

4.1 RF Model

The raw data was min-max scaled, and the optimal number of trees and leaf nodes in the RF model were determined. The range of tree selection depends on the size and complexity of the dataset. In this study, the number of trees was searched within the range of 0~500. For processing large dataset, a smaller number of leaf nodes is usually preferred to ensure stable prediction with a larger number of samples per leaf node. However, a smaller number of leaf nodes may lead to overfitting, so a suitable balance point needs to be found. After considering various factors, the number of leaf nodes was searched within the range of 5~10. The MSEs under different number of trees and leaf nodes are given in Figure 1. It can be seen that initially as the number of trees increase, MSE decreases intensely. When the number of trees reaches and exceeds 200, MSE decreases slightly and tends to stabilize, so the number of 200 trees is selected. Also, the optimization of leaf nodes number is given in Figure 1. Clearly, after the number of 200 trees, based on the yellow line with the lower and constant MSE, 10 leaf nodes are selected in the two datasets.

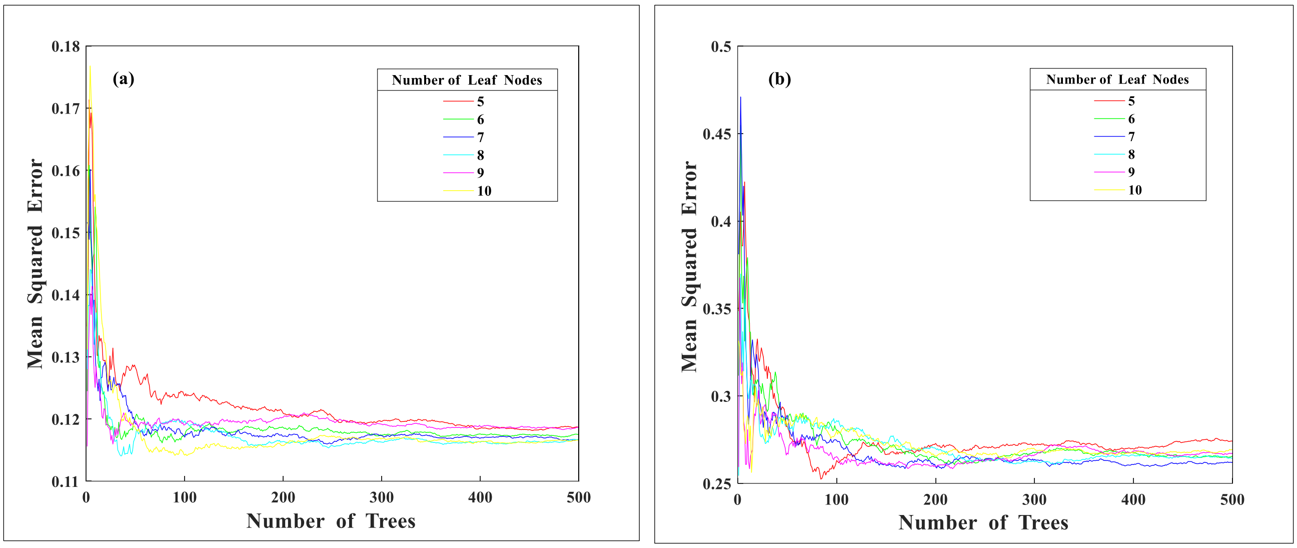


Figure 1: MSE under different number of trees and leaf nodes (a) Dataset 1; (b) Dataset 2

The RF model was trained 100 times to determine the variable importance and model accuracy. The importance values of 12600 mRNA variables and 6259 lncRNA are presented in Figure 2. An initial classification accuracy of 95.47% was achieved with all mRNA variables. All mRNA variables were gradually filtered based on their importance, and at the beginning no change in classification accuracy was observed. When the number of mRNA variables was less than 7600, a decrease in classification accuracy was observed. Therefore, these 7600 mRNA variables were selected and retained by the RF model. According to the above steps, lncRNA variables were also screened, and ultimately 2000 lncRNA variables were selected and retained by the RF model.

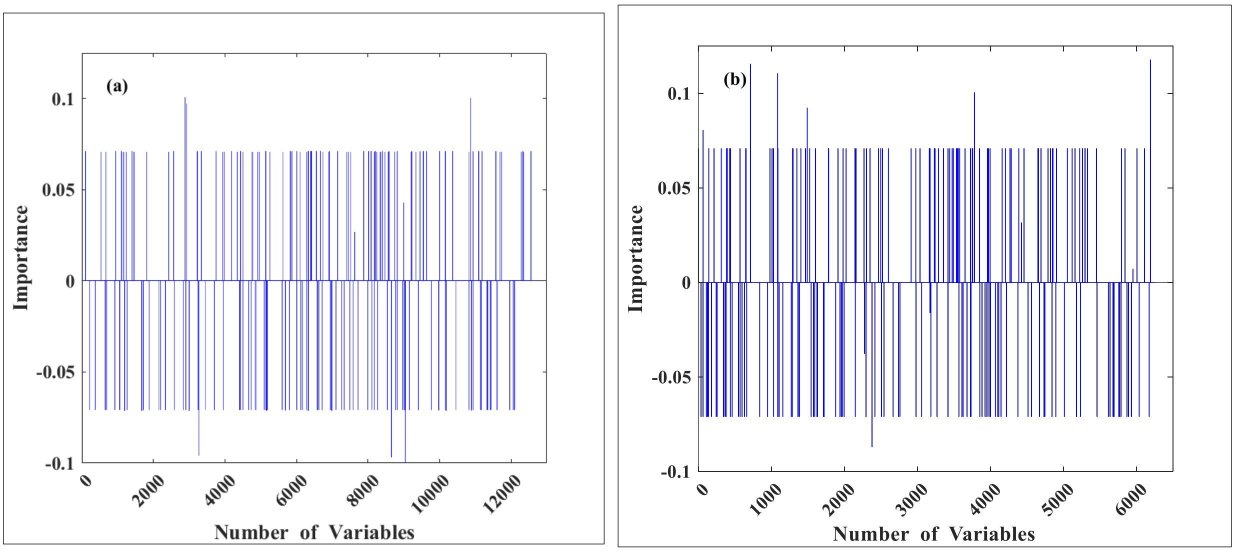


Figure 2: Importance of variables in (a) Dataset 1; (b) Dataset 2

4.2 Feature Extraction

Seventeen random samples taken from all cancer samples with seventeen non-cancer samples were subjected to PCCA. PCC were set a threshold and relevant feature variables could be selected. In this study, PCCA has been completed from the group level (step 1) and the variable level (step 2).

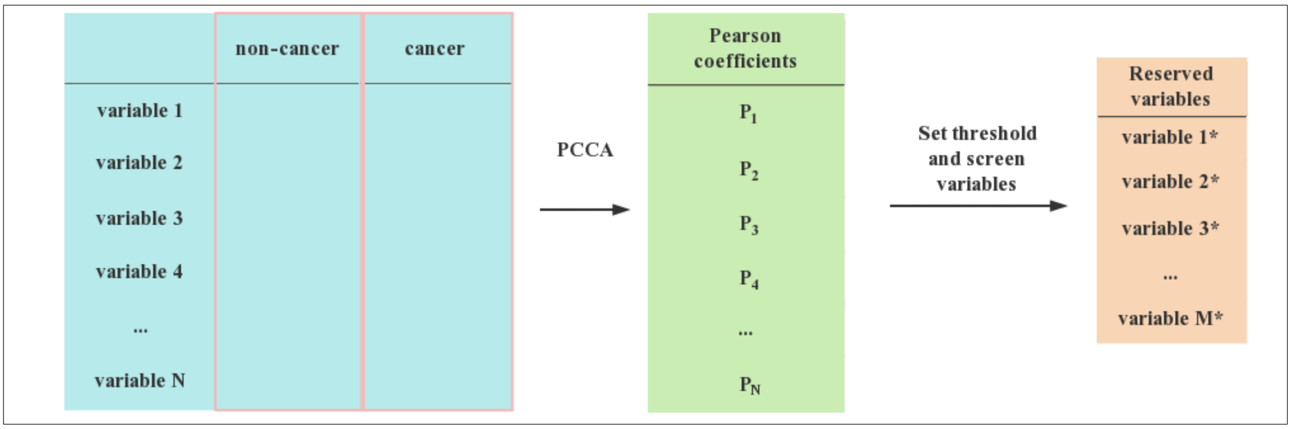


Figure 3: The flowchart of step 1

The flowchart of step 1 is shown in Figure 3. Assuming there were initially N variables, the PCC (P_1, P_2, ..., P_N) between the non-cancer and cancer groups was calculated for each variable. A proper threshold was set and the calculated PCC was compared with the threshold, then the variables with PCC greater than this threshold were considered to be related to cancer and retained. Certainly, the variables with PCC less than this threshold were considered to have no relation to cancer and deleted. Thus, M variables were selected (variable 1^*, variable 2^*, ..., variable M^*) and arranged in descending order according to PCC values.

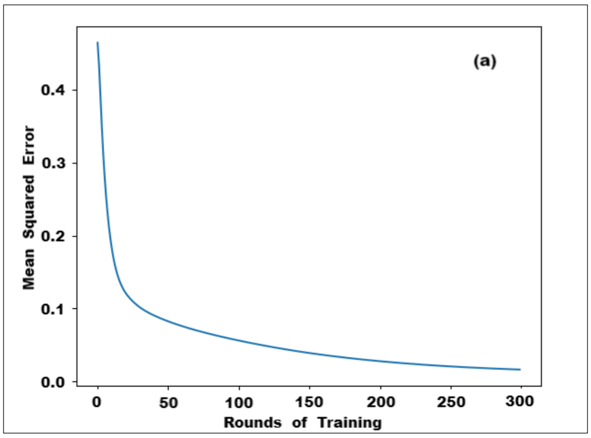
Due to the redundancy between the data variables, redundant variables with strong correlation need to be removed (step 2). The PCCs between two variables ($P^*_{(1,2)}$, $P^*_{(1,3)}$, $P^*_{(2,3)}$, ..., $P^*_{(M-1,M)}$) from the selected M variables were calculated in sequence from step 1. A proper threshold was set, and if the PCC of two variables was greater than the threshold, which indicated a strong correlation between the two variables. Then in the two variables, the variable with front order in step 1 (having a stronger correlation with cancer) was reserved, and finally X selected variables (variable 1^{**}, variable 2^{**}, ..., variable X^{**}) were obtained.

For dataset 1, a threshold of 0.6 at the group level and 0.5 at the variable level was chosen to achieve higher classification accuracy. This process was repeated 20 times, ultimately 342 mRNA feature variables were selected. PCA analysis was then conducted on these variables and when the accumulated explained variance ratio reached 90.19%, the dimension of the PCA-transformed data was reduced to 156*60 (156 represents 156 samples).

For dataset 2, a threshold of 0.4 at the group level and 0.5 at the variable level was chosen to achieve higher classification accuracy. Ultimately, 344 lncRNA feature variables were selected. When the PCA accumulated explained variance ratio reached 97.19%, the dimension of the PCA-transformed data was reduced to 54*40 (55 represents 54 samples).

4.3 NN Model

Then the PCA-transformed data were introduced to the NN model, which had 50 neurons in the middle-hidden layer and a learning rate of 0.05. After 300 rounds of training, the MSE loss function image is shown in Figure 4. As the rounds of training increases, MSE is gradually decreased and eventually stabilized. To ensure the reliability of the classification accuracy, ten models were separately trained on two datasets, and the obtained average classification accuracy were 93.33% and 94.99%, which demonstrates the model can be used for identifying cancer and non-cancer groups.



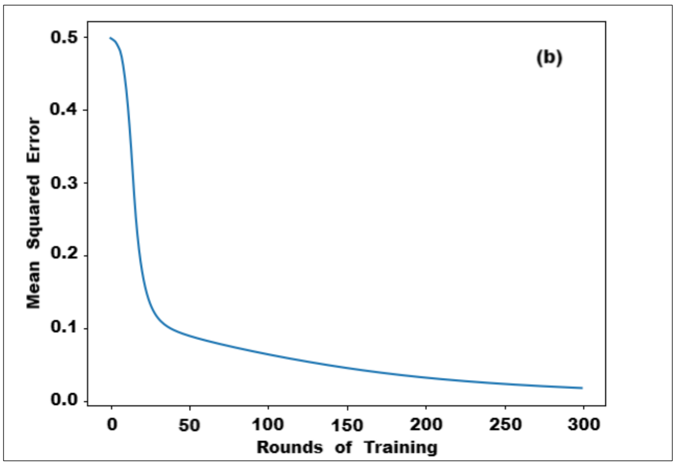


Figure 4: Loss function image of (a) Dataset 1; (b) Dataset

Finally, through the dimensionality reduction fusion method analysis, the 7600 selected variables in the RF model were intersected with the 342 selected variables in the PCCA to obtain 195 mRNA feature variables in dataset 1. The 2000 selected variables in the RF model were intersected with the 344 selected variables in the PCCA to obtain 134 lncRNA feature variables in dataset 2. The selected feature variables could be used as potential biomarkers in lung cancer screening, whose physiological significance can be further researched.

5. CONCLUSION

The dimensionality reduction fusion method based on random forest, feature extraction, and neural network is proved to be a valuable approach for analysis of high-dimensional biological data. The fusion method successfully distinguished cancer and non-cancer groups with high classification accuracy. Additionally, it identified specific variables that are biologically relevant to lung cancer, potentially serving as tumor biomarkers. Based on this, further researches can be carried out on other types of cancer or biological data.

Funding: This study is supported by National College Students Innovation and Entrepreneurship Training Program (X2023-097).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Sanuade, O. A., Hassan, A. M., Akanji, A. O., Olaojo, A. A., Oladunjoye, M. A., & Abdulraheem, A. (2020). New empirical equation to estimate the soil moisture content based on thermal properties using machine learning techniques. *Arabian journal of geosciences*, *13*, 1-14.
2. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504-507.
3. Ramamurthy, M., Robinson, Y. H., Vimal, S., & Suresh, A. (2020). Auto encoder based dimensionality reduction and classification using convolutional neural networks for hyperspectral images. *Microprocessors and Microsystems*, *79*, 103280.
4. Li, Y., & Zeng, X. (2016). Serum SELDI-TOF MS analysis model applied to benign and malignant ovarian tumor identification. *Analytical Methods*, *8*(1), 183-188.
5. Li, Y. K., Dong, R. N., Zhang, J., Huang, K. N., & Mao, Z. Y. (2021). Variable Selection Methods in Spectral Data Analysis. *Spectroscopy and Spectral Analysis*, *41*(11), 3331-3338.
6. Liu, X., Wang, J., Duan, L., Zhang, Y., & Yang, D. (2022). lncRNAs have special significance in diagnosis and therapy for cancer and inflammation. *Cell Biology and Toxicology*, *38*(6), 923-925.
7. Yang, J., Qu, T., Li, Y., Ma, J., & Yu, H. (2022). Biological role of long non-coding RNA FTX in cancer progression. *Biomedicine & Pharmacotherapy*, *153*, 113446.
8. McCullough, K. B., Hobbs, M. A., Abeykoon, J. P., & Kapoor, P. (2018). Common adverse effects of novel therapies for multiple myeloma (MM) and their management strategies. *Current hematologic malignancy reports*, *13*, 114-124.
9. Zheng, S., Zheng, D., Dong, C., Jiang, J., Xie, J., Sun, Y., & Chen, H. (2017). Development of a novel prognostic signature of long non-coding RNAs in lung adenocarcinoma. *Journal of cancer research and clinical oncology*, *143*, 1649-1657.
10. Zhang, Y., Tao, Y., Ji, H., Li, W., Guo, X., Ng, D. M., ... & Liao, Q. (2019). Genome-wide identification of the essential protein-coding genes and long non-coding RNAs for human pan-cancer. *Bioinformatics*, *35*(21), 4344-4349.
11. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class

discovery and class prediction by gene expression monitoring. *science*, *286*(5439), 531-537.

12. Rau, A., Flister, M., Rui, H., & Auer, P. L. (2013). Exploring drivers of gene expression in the Cancer Genome Atlas. *Bioinformatics. 35*, 62-68.
13. Mao, Z., Cai, W., & Shao, X. (2013). Selecting significant genes by randomization test for cancer classification using gene expression data. *Journal of biomedical informatics*, *46*(4), 594-601.
14. Li, S., Chen, X., Liu, X., Yu, Y., Pan, H., Haak, R., ... & Schmalz, G. (2017). Complex integrated analysis of lncRNAs-miRNAs-mRNAs in oral squamous cell carcinoma. *Oral oncology*, *73*, 1-9.
15. Schneider, H. W., Raiol, T., Brigido, M. M., Walter, M. E. M., & Stadler, P. F. (2017). A support vector machine based method to distinguish long non-coding RNAs from protein coding transcripts. *BMC genomics*, *18*(1), 1-14.
16. Sun, L., Liu, H., Zhang, L., & Meng, J. (2017). lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine. *Plos One. 10*, e0139654.
17. He, Y., Ma, J., Wang, A., Wang, W., Luo, S., Liu, Y., & Ye, X. (2018). A support vector machine and a random forest classifier indicates a 15-miRNA set related to osteosarcoma recurrence. *OncoTargets and therapy*, 253-269.
18. Zhao, J., Cheng, W., He, X., Liu, Y., Li, J., Sun, J., ... & Gao, Y. (2018). Construction of a specific SVM classifier and identification of molecular markers for lung adenocarcinoma based on lncRNA-miRNA-mRNA network. *OncoTargets and therapy*, 3129-3140.
19. Liu, B., Cui, Q., Jiang, T., & Ma, S. (2004). A combinational feature selection and ensemble neural network method for classification of gene expression data. *BMC bioinformatics*, *5*, 1-12.
20. Lu, X. G., Chen, D., Du, J. Y., & Zhou, J. (2010). Using Co-classification approach to detect the type of Cancer. *Comput. Sci. 37*, 232-236.
21. Li, B., Tian, B. B., Zhang, X. L., & Zhang, X. P. (2014). Locally linear representation Fisher criterion based tumor gene expressive data classification. *Computers in Biology and Medicine*, *53*, 48-54.
22. Jansi Rani, M., & Devaraj, D. (2019). Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification. *Journal of medical systems*, *43*, 1-11.
23. Benesty, J., Chen, J. D., Huang, Y. T., & Cohen, I. (2009). Pearson correlation coefficient, *Noise reduction in speech processing*. *2*, 1-4.
24. Kurita, T. (2019). Principal component analysis (PCA). *Computer Vision: A Reference Guide*, 1-4.
25. Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An overview of principal component analysis. *Journal of Signal and Information Processing*, *4*(3B), 173.
26. Gao, W., & Zhou, Z. H. (2020). Towards convergence rate analysis of random forests for classification. *Advances in neural information processing systems*, *33*, 9300-9311.
27. Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied energy*, *262*, 114566.
28. Alam, M. Z., Rahman, M. S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, *15*, 100180.
29. Wu, Y. C., & Feng, J. W. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, *102*, 1645-1656.
30. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., ... & Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, *98*(24), 13790-13795.