narra j

Short Communication

A comparative assessment on gene expression classification methods of RNA-seq data generated using next-generation sequencing (NGS)

Setia Pramana[1*], I Komang Y. Hardiyanta[2], Farhan Y. Hidayat[2] and Siti Mariyah[1,3]

[1]Politeknik Statistika STIS, Jakarta, Indonesia; [2]BPS' Statistics Indonesia, Jakarta, Indonesia; [3]School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

*Corresponding author: setia.pramana@stis.ac.id

Abstract

Next-generation sequencing or massively parallel sequencing have revolutionized genomic research. RNA sequencing (RNA-Seq) can profile the gene-expression used for molecular diagnosis, disease classification and providing potential markers of diseases. For classification of gene expressions, several methods that have been proposed are based on microarray data which is a continuous scale or require a normal distribution assumption. As the RNA-Seq data do not meet those requirements, these methods cannot be applied directly. In this study, we compare several classifiers including Logistic Regression, Support Vector Machine, Classification and Regression Trees and Random Forest. A simulation study with different parameters such as over dispersion, differential expression rate is conducted and the results are compared with two mRNA experimental datasets. To measure predictive accuracy six performance indicators are used: Percentage Correctly Classified, Area Under Receiver Operating Characteristic (ROC) Curve, Kolmogorov Smirnov Statistics, Partial Gini Index, H-measure and Brier Score. The result shows that Random Forest outperforms the other classification algorithms.

Keywords: Microarray data, gene expression, support vector machine, classification, random forest

Introduction

Transcriptome is a collection of all gene readouts that are present in a cell. It is from the instruction that are carried out of the DNA and were needed to be read by transcribing into an RNA. Reading the sequence and quantities of mRNA which has a vital role of making proteins can determine when and where each gene will be turned on and turned off in cells and tissues of an organism. Transcriptome analysis (transcriptomics) is aimed to understand the expression of genome at the transcription level and provides the information on the respective gene structure, regulation of gene expression, gene product function and genome dynamics used for molecular diagnosis, disease classification and providing potential markers of diseases. Studies revealed several driver genes in breast cancer patient survival prognosis [1, 2]. Studies also found that a panel of genes used to predict the overall survival of prostate cancer patients [3, 4]. Furthermore, there are plenty applications in targeted therapy in pharmacogenomics [5, 6].

Currently, there are two types of technologies commonly used to measure the expression of thousands of genes simultaneously: microarray and next-generation sequencing (NGS) with the

https://narraj.org/

Received: November 13, 2021 - Accepted: March 22, 2022 - Published online: April 1, 2022

RNA sequencing approach. Microarray refers to a hybridization-based technology which has been around for over a couple of decades. It utilizes short oligonucleotide probes representing genomic DNA immobilized on solid surface such as glass or silicon slides, then uses fluorescent labeling as a method of quantification [7]. The technology has been mainly used for gene expression analysis, where RNAs extracted from cell or tissue samples could either be directly labeled, or converted to cDNA or cRNA first.

DNA sequencing is the method to get the exact order of nucleotides (composition) in a DNA. The first DNA sequencing technique is Sanger sequencing developed by Frederick Sanger and colleagues in 1977. The next technique is NGS technology known also as massive parallel sequencing, which breaks the limitations of the traditional Sanger sequencing. One of the NGS application is RNA sequencing (RNA-Seq) which sequence the mRNA and measure the gene expression. The biggest advantage RNA-Seq has over microarray is that it directly accesses the sequence without hybridization, allowing the differential expression analysis of organisms without reference genome [8]. It is also more accurate, as it can suggest precisely the location of transcription boundaries to a single-base resolution [9]. Currently between the two, microarray is still generally less costly compared to RNA-Seq. Additionally, microarray is the most mature technology for high-throughput screening, which means that both the hardware and analytical tools for it have been refined over the years.

Pathology of tumor could be better understood by investigating the comparison of gene expression levels between samples from diseased patients and those from healthy individuals [10, 11, 12, 13]. In addition, classification of gene expression based on RNA-seq can detect severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and other common RNA respiratory viruses [14] and then to be used to identify biomarkers and therapeutic targets in managing coronavirus disease 2019 (COVID-19) [15].

For classification of gene expressions, several methods that have been proposed are based on microarray data [16] which is a continuous scale or needs a normal distribution assumption. Since the RNA-Seq data do not meet those requirements, these methods cannot be applied directly. In addition, as the RNA-Seq creates data with over-dispersion (the variance exceeds the mean), it should be taken into consideration. Otherwise, it would influence the model performances. Many researchers have proposed alternative approaches. Anders and Huber [17] proposed a technique on the basis of negative binomial distribution, with variance and mean linked by local regression for differential expression analysis in count data. Furthermore, Tan et al. [18] discussed the application of, and modifications to, LR, Principal Components Analysis, Linear Discriminant Analysis, Support Vector Machine (SVM), and Partial Least Squares in the RNA-Seq data. Zararsiz et al [19] recently discussed the log transformation and Random Forest (RF) and SVM approach are good choice for classification for RNA-Seq data.

Herein, we compared the performance four classification algorithms, LR, SVM, RF and Classification and Regression Trees (CART), on different data situations. A simulation study with different parameters such as overdispersion dan differential-expression rate were conducted and compared. To measure predictive accuracy, 10-fold cross validation and five performance indicators were used.

Methods

Simulation setup

A simulation study adopted from Zararsiz *et al.* [19] was carried out to evaluate the influence of some parameters. The simulation of gene-expression datasets was conducted under 216 different scenarios using the following negative binomial model;

$$X_{ij}|y_{ij} = k \sim NB(s_i g_j d_{kj}, \varphi) \quad (1)$$

Where g_j represents the total number of mapped counts per gene (i.e, gene total), φ – the dispersion parameter, si – the number of mapped counts per sample, and d_{kj} – the differential-expression parameter of the j-th gene between classes k.

The study focused only on binary classification (number of classes was 2). The simulated datasets consist all feasible combinations of: (1) different dispersion situations: slightly over dispersed, substantially over dispersed and highly over dispersed; (2) number of observations (n): 20, 40, 60, 80; (3) number of genes (p): 50, 100, 500; and (4) differentially expressed (DE) gene rates as (d_{kj}) 1% and 50%.

The simulation was repeated 10 times and performed on PoiClaClu [20] package of R software with CountDataSet function. The logarithmic transformation approach (rlog) was then implemented to transform the data into smaller skewed distribution and extreme values [21]. Note that in this study we focus on binary classification and include number of genes in the scenario of simulation.

Implementation of classifiers

The classification of the simulated data is carried out using several algorithms: LR, SVM [22], CART [23] and RF [24]. LR is one of mathematical model approaches for classification where outcome variable is a binary variable. The purpose was to evaluate the influences of multiple explanatory variables (categorical or numeric) on the outcome variable. LR models the probability of an event by a linear combination of predictor or independent variables.

SVM was originally reported by Vapnik [22], in which this approach has been used in multiple classification problems. SVM is one of popular classification methods according to the statistical learning theory. SVM is known for its effective mathematical background, good generalization ability, wide range of application, and learning capability. In addition, SVM is has the ability to perform linear or nonlinear classification and engage with high-dimensional data.

CART introduced in 1984 [23] and is one of the widely used tree classifiers implemented in broad-spectrum applications. The tree classifier is also known as Recursive Partitioning and Tree-Based Technique. The principle of these techniques is partitioning the space and identify some representative centroids. Classification trees partitioning the space in hierarchal manner. It is initiated with the entire space and recursively divided it into less regions, then each region has a class label assigned. The set of splitting rules are used to segment the predictor space can be summarized in a tree, known as decision tree methods. To overcome offer fitting the maximally grown tree was pruned. A cross-validation approach was implemented to obtain the optimal tree that has the lowest rate of error classification. The assignment of each terminal node to a class was performed by choosing the class that lower the feasibility of error classification.

RF is proposed by Breiman [24] to combine numerous weak classifiers to generate a significantly improved as well as strong classifier. The main idea of this algorithm is to build a larger number of unpruned decision trees then combine them by averaging the predictions of individual trees in the forest. The technique is expected to be less affected by the noise and highly efficient on large data. The first step of RF is to “grow” many classification trees using ensemble approach. Then each tree gives a classification and taken as a vote towards the foregoing tree. The classification is based on the classification possessing the most votes over all the trees in the forest.

To deal with overfitting and validate each classifier model, the 10-fold cross-validation was performed. The best parameters for each classifier were then selected based on the 10-fold cross-validation result.

Evaluation process

As many as 70% of the simulated data were allocated into training, and the rest (30%) – test sets. The entire model building processes were performed on training sets, where model performances were evaluated in test sets. Most of the studies rely only on single performance measurement. Here, we used six performance indicators which assessed different performances: Percentage Correctly Classified (PCC), Kolmogorov-Smirnov Statistic (KS) [25], Area Under ROC Curve (AUC), Partial Gini Index (PG-Index) [26], H-measure (H-m) [27], and Brier Score (BS) [28]. The PCC along with KS assess the correctness of categorical predictions [29]. AUC and PG-Index evaluate the distinguishing ability of the algorithm, whereas the H-

measure (H-m) is considered to overcome the shortcomings of AUC. BS evaluates the accuracy of the probability predictions.

Since we have six different performance indicators to combine all indicators, for each performance indicator, the classification algorithms were ranked so that the best algorithm score highest. The average of the ranks for each algorithm were calculated and compared. The better the performance, the higher the rank score.

Results

Effect of sample size, number of gene and over-dispersion level on Percentage Correctly Classified (PCC)

The most popular classification performance index is accuracy or also known as PCC. The results of performance of each algorithm based on PCC parameter for different sample sizes (20, 40 and 60) and over dispersion situations (high, medium and low) are shown in **Figure 1**. It shows that when the data is highly over dispersed (high), the accuracy tends to be smaller as compared to low and medium over dispersion. It shows that most of the classification algorithms perform well when the data is less dispersed and relatively homogeneous.

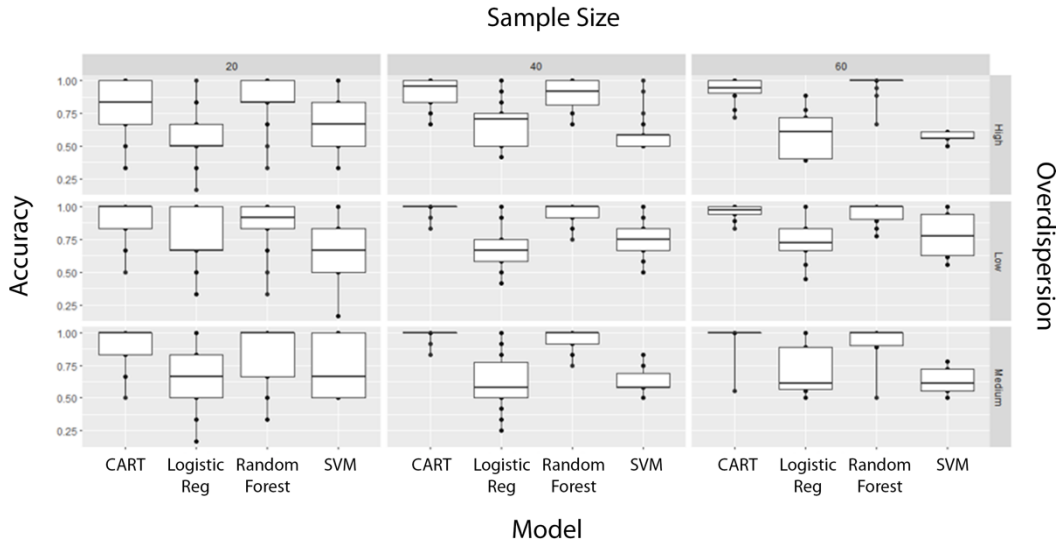

Figure 1. Accuracy of four algorithms based on Percentage Correctly Classified (PCC) parameter on different sample sizes and over-dispersion levels.

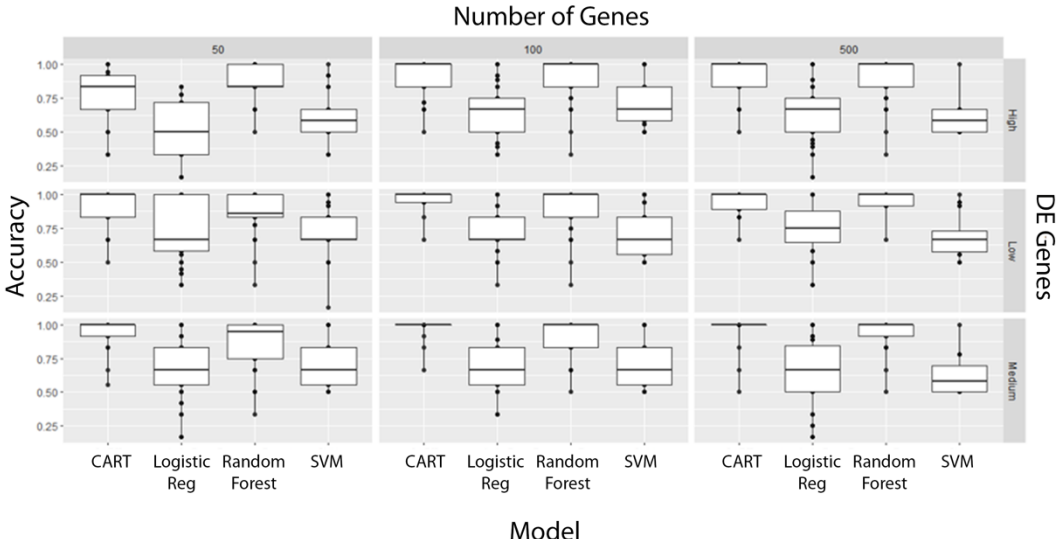

Figure 2. Accuracy of four algorithms based on Percentage Correctly Classified (PCC) on different number of genes and differentially expressed (DE) genes.

The impact of number of genes (variables) on the classification accuracy (PCC) is presented in **Figure 2**. It can be seen that large number of genes (variables) do not seem to affect the accuracy. Accuracy is expected to increase when more variables are into account, with overfitting as the prices. The cross validation has worked well to prevent overfitting due to large number of variables (**Figure 2**).

There are limitations of accuracy as performance indicator, e.g., unbalanced cases. Hence it is better to use other indicators.

The effect of sample size and overdispersion on average of the ranks of six indicators

The performance of the classification algorithms based on the average of the ranks of six indicators (PCC, KS, AUC, PG-Index, H-m and BS) are presented in **Figure 3**. The average of the ranks for each algorithm is plotted for different classification approaches, sample sizes and over dispersion situations.

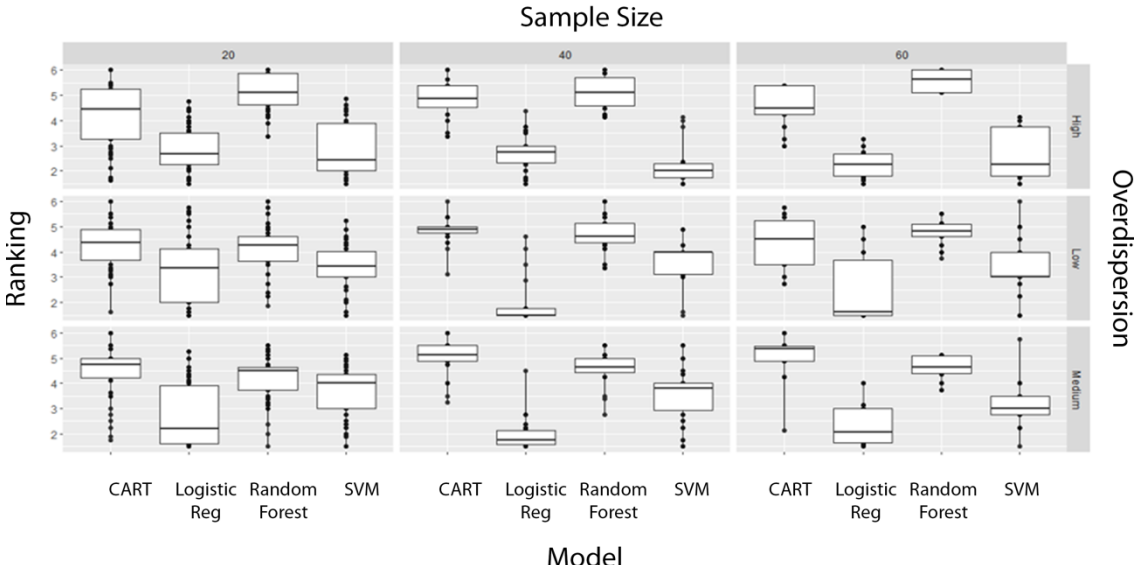

Figure 3. Average ranks of six indicators (PCC, KS, AUC, PG-Index, H-m and BS) of four classification approaches with different sample sizes and levels of over-dispersion.

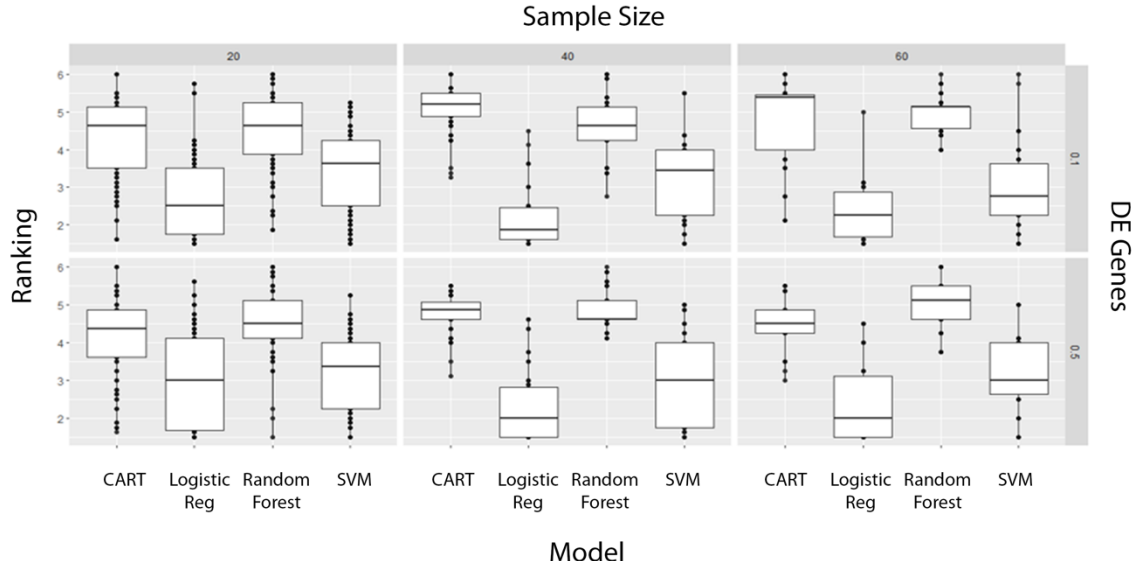

Figure 4. Average ranks of six indicators (PCC, KS, AUC, PG-Index, H-m and BS) of four classification approaches with different sample sizes and percentage of differentially expressed genes.

The effect of sample size and percentage of differentially expressed (DE) gene on average of the ranks of six indicators

The effect of sample size and percentage of DE gene on average ranks of six indicators (PCC, KS, AUC, PG-Index, H-m, and BS) of four classification algorithms are presented in **Figure 4**. Two percentages of DE gene (1% and 5%) were assessed. The results suggested that the higher percentage of DE genes leaded to an increase in the classification performance (**Figure 4**).

Discussions

The results shows that in general, larger sample size increases the classification accuracy. Providing more information (i.e., number of samples) would improve the power to detect differences. On the other hand, having more variables (i.e., genes) would result on overfitting and complexity thus it impacts the classification accuracy and machine learning efficiency. Furthermore, the result shows that overfitting occurs due to the addition of large number of genes can be solved by performing k-fold cross validation. This finding suggests that researchers need to perform feature selection before the modeling to decrease the quantity of genes leading the elevation of the discrimination power [31]. Although the gene filtering and feature selection have been carried out, normally there are plenty number of genes remaining. In this case, performing k-fold cross validation in the modeling would increase the classification performance.

A common complication in the data generated from RNA-Seq is overdispersion, where the observed variation exceeds that predicted from the binomial distribution. The results show overdispersion influence the classification performances. The higher the overdispersion, the lower the classification accuracy. More biological replicates would increase the sparseness of the data (i.e., increasing overdispersion) which reduce the classification power [19,30].

Comparing the performance of four classification algorithms in all conditions suggest that the algorithm RF has higher accuracy and average ranks of six performance indicators. The larger average ranks of performance indicators show the better classification predictions.

In addition, RF can handle the overdispersion relatively better than the other algorithms. RF is one of ensemble methods which aggregates multiple machine learning models with the aim of decreasing both bias and variance [32,33]. Hence, the result from an ensemble method such as RF will be better than any of individual machine learning model. The result also shows that adding and combining more significant genes on class conditions is similar to combining their abilities to forecast can increase in the classification performance. In general, in all simulated datasets, the classification performance of RF outperforms the CART, LR and SVM. It is in line with the previous studies [31, 32,33]. RF algorithm had been previously shown to perform outstanding in several bioinformatics tasks [34].

Overall, the accuracy of classifiers on RNA-Seq data depends on the number of samples, number genes used and the variation of the data (overdispersion). High variability of gene expression is expected due to the differences of sample and library preparation, type of sequencers and biological samples variation [31]. To obtain the best classification performance, researcher must set standard or protocols of wet lab preparation to minimize the variation, then performing feature selection, gene filtering before implementing the classification algorithms. Implementing RF with K-fold cross validation would provide best classification accuracy.

Conclusions

Several situations of RNA-Sequencing data and compared several algorithms to perform binary classification was simulated. The classification performance is measured by several indicators, measuring different type of classification performances. Our simulations show that data with less overdispersion, large sample size and have more differentially expressed genes could improve the discriminatory ability and the accuracy of the classifiers. In overall assessment, the RF algorithm works better on classification than CART, LR and SVM.

Declarations

Ethics approval

Not required.

Acknowledgments

This work has been supported by the research funding of Politeknik Statistika STIS.

Conflict of interest

All data in this paper has been composed and that the work has not be submitted for any other event. All in this work submitted by the authors, except where work which has formed part of jointly-authored publications has been included. The authors declare that there is no conflict of interest in make this paper.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Underlying data

Derived data supporting the findings of this study are available from the first author on request.

How to cite

Pramana S, Hardiyanta IKY, Hidayat FY, Mariyah S. A comparative assessment of classification methods for RNA-sequencing data. Narra J 2022; 2(1): e60. http://doi.org/10.52225/narra.v2i1.60.

References

1. Suo C, Hrydziuszko O, Lee D, *et al.* Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. Bioinformatics 2015; 31(16):2607-2613.
2. Vu TN, Pramana S, Calza S, *et al.* Comprehensive landscape of subtype-specific coding and non-coding RNA transcripts in breast cancer. Oncotarget 2016; 7(42):68851-68863.
3. Peng Z, Andersson K, Lindholm J, *et al.* Operator dependent choice of prostate cancer biopsy has limited impact on a gene signature analysis for the highly expressed genes IGFBP3 and F3 in prostate cancer epithelial cells. PLoS One 2014; 9(10):e109610.
4. Peng Z, Andersson K, Lindholm J, *et al.* Improving the Prediction of Prostate Cancer Overall Survival by Supplementing Readily Available Clinical Data with Gene Expression Levels of IGFBP3 and F3 in Formalin-Fixed Paraffin Embedded Core Needle Biopsy Material. PloS one 2016; 11(1):e0145545.
5. Lin D, Shkedy Z, Yekutieli D, *et al.* Modeling dose-response microarray data in early drug development experiments using R: order-restricted analysis of microarray data: Springer; 2012.
6. Rabbani B, Nakaoka H, Akhondzadeh S, *et al.* Next generation sequencing: implications in personalized medicine and pharmacogenomics. Mol Biosyst 2016; 12(6):1818-1830.
7. Wu J, Kim D. Transcriptomics and gene regulation: Springer; 2016.
8. Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 2011; 9:34.
9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10(1):57-63.
10. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003;2:S75–83.

11. Ge SG, Xia J, Sha W, Zheng CH. Cancer subtype discovery based on integrative model of multigenomic data. IEEE/ACM Trans Comput Biol Bioinform. 2016;14:1–1.
12. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
13, Peng Z, Andersson K, Lindholm J, *et al.* Improving the prediction of prostate cancer overall survival by supplementing readily available clinical data with gene expression levels of IGFBP3 and F3 in formalin-fixed paraffin embedded core needle biopsy material. PLoS ONE 2016;11(1): e0145545.
14. Acera Mateos, P., Balboa, R.F., Easteal, S. *et al.* PACIFIC: a lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses. Sci Rep 2021; 11: 3209.
15. Soremekun OS, Omolabi KF, Soliman MES. Identification and classification of differentially expressed genes reveal potential molecular signature associated with SARS-CoV-2 infection in lung adenocarcinomal cells. Inform Med Unlocked 2020;20:100384.
16. Gohlmann H, Talloen W. Gene expression studies using Affymetrix microarrays: CRC Press; 2009.
17. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010; 11(10):R106.
18. Tan KM, Petersen A, Witten D. Classification of RNA-seq data. In: Statistical analysis of next generation sequencing data. edn.: Springer; 2014: 219-246.
19. Zararsiz G, Goksuluk D, Korkmaz S*, et al*. A comprehensive simulation study on classification of RNA-Seq data. PLoS One 2017; 12(8):e0182507.
20. Witten DM. Classification and clustering of sequencing data using a Poisson model. Ann Appl Stat 2011; 5(4):2493-2518.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15(12):550.
22. Vapnik V. The nature of statistical learning theory: Springer; 1999.
23. Breiman L, Friedman J, Olshen R. Classification and Regression Trees. Belmont, California: Wadsworth; 1984.
24. Breiman L. Random forests. Machine learning 2001; 45(1):5-32.
25. Thomas LC, Edelman DB, Crook JN. Credit Scoring and its Applications: SIAM monographs on mathematical modeling and computation. Philadelphia: University City Science Center, SIAM; 2002.
26. Pundir S, Seshadri R. A novel concept of partial lorenz curve and partial gini index. Int J Eng Science Innov Technol2012; 1(2):296-301.
27. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 2009; 77(1):103-123.
28. Hernández-Orallo J, Flach PA, Ramirez CF. Brier curves: a new cost-based visualisation of classifier performance. In: 2011 2011; 2011.
29. Lessmann S, Baesens B, Seow H-V*, et al*. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 2015; 247(1):124-136.
30. Nagalakshmi U, Wang Z, Waern K, *et al*.The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008, 320(5881):1344–1349.
31. Johnson NT, Dhroso A, Hughes KJ, Korkin D. Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? RNA. 2018 Sep;24(9):1119-1132.
32. Abu El Qumsan M. Assessment of supervised classification methods for the analysis of RNA-seq data. Theses. Aix-Marseille Université; 2019.
33. Ramroach S., John M., Joshi A. The efficacy of various machine learning models for multi-class classification of RNA-seq expression data. In: Arai K., Bhatia R., Kapoor S. (eds) Intelligent Computing. CompCom 2019. Advances in Intelligent Systems and Computing, Vol 997 Springer; 2019.
34. Anne-Laure B, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2; 2012.