AIMS Biophysics

http://www.aimspress.com/journal/biophysics/

AIMS Biophysics, 8(4): 372–384.
DOI: 10.3934/biophy.2021029
Received: 20 September 2021
Accepted: 23 November 2021
Published: 01 December 2021

Research article

Feature selection and classification approaches in gene expression of breast cancer

Sarada Ghosh[1], **Guruprasad Samanta**[2] **and Manuel De la Sen**[3,*]

[1] Department of Statistics, Gurudas College, Phool Bagan, Kolkata-700054, India
[2] Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711103, India
[3] Institute of Research and Development of Processes, University of the Basque Country, 48940 Leioa, Bizkaia, Spain

* **Correspondence:** Email: manuel.delasen@ehu.eus.

Abstract: DNA microarray technology with biological data-set can monitor the expression levels of thousands of genes simultaneously. Microarray data analysis is important in phenotype classification of diseases. In this work, the computational part basically predicts the tendency towards mortality using different classification techniques by identifying features from the high dimensional dataset. We have analyzed the breast cancer transcriptional genomic data of 1554 transcripts captured over from 272 samples. This work presents effective methods for gene classification using Logistic Regression (LR), Random Forest (RF), Decision Tree (DT) and constructs a classifier with an upgraded rate of accuracy than all features together. The performance of these underlying methods are also compared with dimension reduction method, namely, Principal Component Analysis (PCA). The methods of feature reduction with RF, LR and decision tree (DT) provide better performance than PCA. It is observed that both techniques LR and RF identify TYMP, ERS1, C-MYB and TUBA1a genes. But some features corresponding to the genes such as ARID4B, DNMT3A, TOX3, RGS17 and PNLIP are uniquely pointed out by LR method which are leading to a significant role in breast cancer. The simulation is based on *R*-software.

Keywords: breast cancer; random forest; decision tree; principal component analysis

1. Introduction

Cancer is a disease which arises from cells that leave the cell cycle and start to proliferate in an uncontrolled manner and spread into surrounding tissues [1]. This proliferation could be induced by hormones that are impinging on the breast. There are various types of breast cancer (BC). The kind of

breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin in different parts of breast. A breast is made up of three main parts (lobules, ducts, and connective tissue). The lobules are the glands which produce milk. The ducts are tubes which carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together. Generally, most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasis. The risk for breast cancer is due to a combination of different factors. The main factors that influence the risk include being a woman and getting older. Most breast cancers are found in women who are 50 years old or older [2]. A genetic micro-array analysis of genetic transcriptional difference between normal and malignant cells is provided in gene expression profiling (GEP) which has come into clinical use in recent years and is beneficial from therapeutic point of view. It gives the detail information about the expression levels of thousands of genes in BC and depicts molecular portraits of BC. The experts of BC have intensively studied GEP and their findings have begun to be embraced by clinical oncologists in their day-to-day practice. It is very important to learn the technique successfully and its potential applications [3]. A high risk of getting breast cancer may depend on family history of breast cancer or inherited changes in BRCA1 and BRCA2 genes [4]. There are several applications of a recently developed mathematical field called topological data analysis (TDA). The most two important methods of TDA be (i) Progression Analysis of Disease and (ii) the analysis of Betti numbers [5]. These techniques are applied to a set of microarrays from tissue donated by women undergoing mammoplasty surgery. The results are obtained from breast cancer research, under varying experimental conditions. Progression Analysis of Disease (PAD) highlights genes that are significantly differentially expressed even if it is just for a small number of patients. PAD has envisioned some needful implementations which provide a clear visualization that can help for further exploration. It can be also used for a large range of high throughput data, which is a useful example for analyzing of data of breast cancer. PAD helps to identify Estrogen Receptor-positive (ER+) which is a unique subgroup. This subgroup can demonstrate high levels of c-MYB and low levels of IIG (innate inflammatory genes). So, 100% survival is exhibited by patients and there is no evolution. There is no other way for distinction between healthy and victimized people who belong to this group. This group has a understandable, distinct and also statistically significant molecular signature. It can reveal coherent biology but conceal for cluster analysis and fail for fitting into classification (which is accepted) of Normal-like subtypes of Estrogen Receptor-positive BC and also for in case of Luminal A/B. This group is known as c-MYB+ BC [5].

When high dimensional data has been considered, gene expression data gives various proposal in case of feature selection [6–8]. Random Forest (RF) is one of the popular algorithm which is applied for feature selection [9]. In this work, our purpose is to achieve significantly potential information for breast cancer transcriptional genomic data which make a great influence for the gene expression and lead a significant role for mortality due to breast cancer. So, we make comparisons for selection of features by approaching Logistic Regression, Random Forest and Decision Tree (DT) along with Principal Component Analysis (which helps to reduce dimension). We have gained a significant influence in prediction accuracy for features selection by using LR, RF and DT than PCA. LR ,RF and DT have also been compared among themselves during feature selection and it has been observed that RF and DT cannot able to identify some features whereas LR can capable for identifying these. The piecemeal method has also been analyzed (by using LR method) for identifications of some features.

This study seeks to investigate the most appropriate model to examine the uttermost important genes that significantly influence the mortality rate.

In Section 2, we have discussed about basic statistical tools and model derivation. Then the comparison among the performance of the underlying models are discussed in the Section 3 and 4. Next, in Section 5 we have been interpreted the results related to Breast Cancer. Finally, the last section consists of the general discussions and conclusions of this work.

2. Basic statistical tools and model derivation

2.1. Data description

The data of Breast Cancer is provided by NKI Breast Cancer Data [10]. A sample of 272 breast cancer patients (as rows), 1570 columns. Network is built using only gene expressions. Meta-data includes patient information, treatment, and survival. In this work, we have considered only varying gene expressions of 272 samples consisting of 195 alive and 77 death samples.

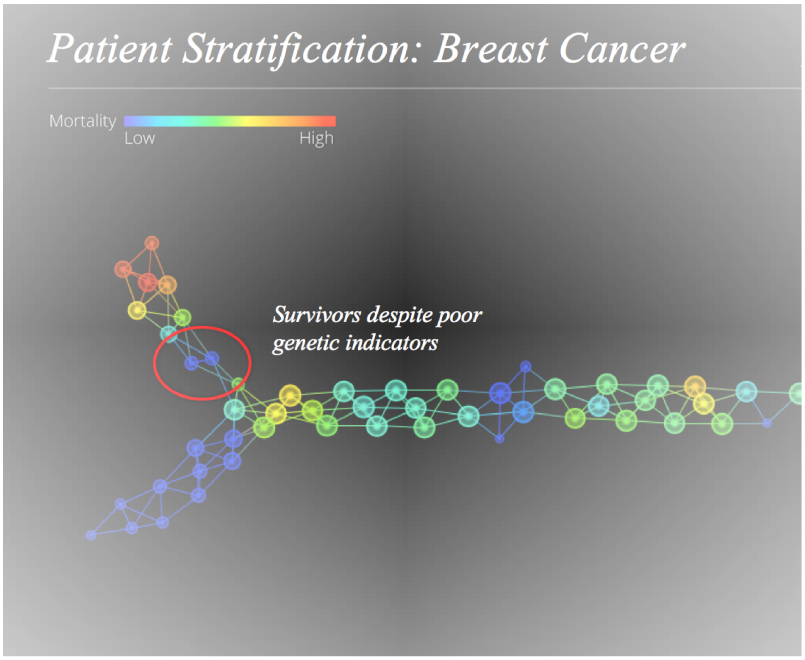

Figure 1. Patient stratification: breast cancer.

GEO (Gene Expression Omnibus) data set where the column list is limited to the top varying genes. Each node is a group of patients similar to each other. Flares (left) represent sub-populations that are distinct from the larger population. One differentiating factor between two flares is estrogen expression (low = top flare, high = bottom flare). Bottom flare is a group of patients with 100% survival. Top flare shows a range of survival very poor towards the tip (red), and very good near the base (circled) in the data-set.

2.2. Preliminaries of logistic regression

The statistical model (logistic regression) is generally applied on a binary dependent variable. This model is uttermost important for response data which are ordered categorical [1]. It is also used gradually in a large variety of applications in social science research but also speedily used in biomedical studies and marketing in the last 25 years. Besides, in genetics purpose it is also used widely. Binomial regression is a regression analysis procedure where the dependent variable (sometimes referred as Y) is a series of Bernoulli trials, or it may be result of a series of one of two possible disjoint outcomes (conventionally "success" denoted as 1, and "failure" denoted as 0) in statistics. The log-binomial model is a model of binomial generalized linear model (i.e., GLM) together with a log link function which is basically famous in epidemiological and bio-statistical fields. For a binary regression, dependent variable is denoted as Y and let X be independent variable, and let $\Phi(x) = P(Y = 1|X = x)$. The logistic regression is as follows:

$$\Phi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{1}$$

The log odds is said to be logit function, which is as follows:

$$\text{logit}[\Phi(x)] = \log \frac{\Phi(x)}{1 - \Phi(x)} = \alpha + \beta x \tag{2}$$

The expression (2) equates the logit function to linear predictor, where the intercept α is biased. Y is independent of X if $\beta = 0$ and since logistic density is symmetric, the function $\Phi(x)$ approaches 1 at the same rate that it approaches 0. The odds become an exponential function of x whenever we exponentiate both sides of (2) which gives a basic interpretation for the magnitude of β. In this work, we have taken the alive condition as responsible variables whose values are 1 and 0 according to alive and death respectively.

2.3. Decision tree (DT) for selecting feature

Decision tree (DT) is one of the popular and easiest techniques which builds regression or classification model in the form of a tree structure. A decision tree (also known as a classification tree or a reduction tree) is a predictive model which is a mapping from observations about an item to conclusions about its target value. It is used as an iterative logarithm in decision analysis where the data is continuously split up according to a certain parameter. The tree based methods generally handle high-dimensions dataset. Decision trees can be drawn by hand or created with a graphics program or specialized software. Informally, decision trees are very useful for focusing discussion when a group must make a decision.

2.4. Random forest (RF) for selecting feature

Random forest is a decision tree based regression and classification technique which is suitable for both parametric and non-parametric cases [9] and this algorithm also establishes multitude decision trees using Bootstrap. For building these decision trees, each time a split in a tree is considered. RF is performed for feature selection where every feature is assigned with random values and make a

comparison for decreasing in the model's performance accuracy which permits to rank the variables according of their importance [11]. In this work, we have performed RF classification for feature selection and then also pointed out the top features for constructing the classification model. Then, Bootstrap validation is also executed to measure the accuracy. RF algorithm leads to reduction in both test error and out-of-bag (OOB) error [12]. Random forest overcomes problems generated by other tree based techniques. It does so by forcing each split to consider only a subset of the predictors.

2.5. Dimension reduction technique by using principal component analysis (PCA)

Principal Component Analysis, a statistical tool, is used for dimension reduction (i.e., high dimensional data reduces into low dimensions) in such a manner that the high dimensional data projects onto a new subspace with lower dimension than the original data. These new components, getting by PCA, are applied for predictive analysis and also for construction of model. In this work, we have executed the dimension reduction method by PCA and Bootstrap validation on the breast cancer transcriptional genomic data.

2.6. Random sampling technique (RST)

Random sampling is a technique for sampling from a population in which (i) the selection of a sample unit is based on chance and (ii) every element of the population has a known non-zero probability of being selected. The advantages of RST is to cull a small sample from a large population and used it to investigate for making generalizations about the information of the desirable larger group. In this work a random sampling of 205 features has been chosen to observe the effects of the random selection of features on the overall model performance.

3. Comparison among the several model-performance

For comparing the importance of features which are chosen by LR, DT and RF, firstly, we have to construct the model of classification for every feature and then make a comparison among the predicting accuracies. In this work, we have prepared total 6 different feature sets. In the present data-set, there are 1554 features when the full model (FM) is considered. We have performed the individual across all features and LR model contains 192 features. Next, we have performed DT classification algorithm which suggests 200 features and also compared with other classifications. In RF classification, top 200 features are chosen for comparing with the features. Lastly, in this work, we have performed the dimension reduction method by PCA and also have comprised 168 principal components which reveal 95% variation of the data-set. A classification model is built for each data-set by using RF and then the performance is also measured using bootstrap method. The reduction of error rate is compared with each of the proposed models. In *R* script '*caret*' package is used to construct the model from different classification techniques.

Here, the predictors are the breast cancer genomic data for 1554 transcripts that are measured from 272 samples. LR method is applied for the response variable with each of predictors separately. It is also measured for the performance of the model by ROC curve using estimation of the Area Under Curve (AUC). AUC value provides us to rank for each feature. The high value of AUC indicates how strongly any predictor is capable for classification between two different classes of the response variable. Features with higher AUC are selected and overall performance of these features have

been estimated by constructing a model using RF. In the work of Rakotomamonjy (2004) [13], it is observed that LR classification can maximize both AUC and accuracy. The AUC of gene expression "NM_003247" is shown in Figure 2 and similar AUC are being observed for others variants of THBS gene. For the purpose of measuring accuracy, Bootstrap validation is executed. Then, a comparison is performed between the estimated accuracies among the proposed models with dimension reduction by using PCA, DT and RF (Table 1).

Table 1. Estimation of mean error rates.

Model	Error rate
FM	31.756
LR	23.329
DT	26.534
RF	22.782
PCA	29.337
RST	33.332

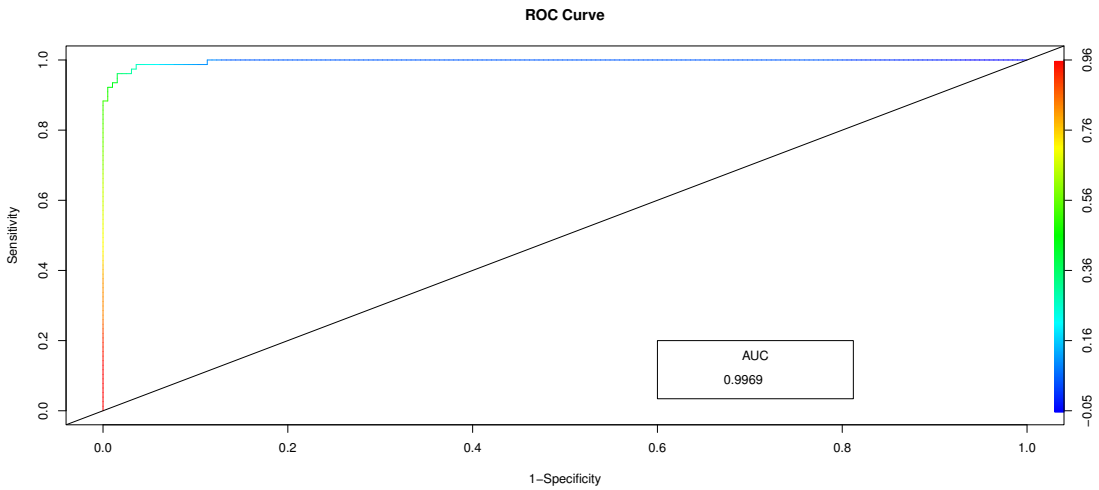

Figure 2. ROC curve.

4. Discussion

The AUC for the data-set having 1554 features are computed, with at most 0.996 (= value of AUC) for THBS gene. The selection procedure is made when transcripts having AUC value greater than 0.9 provides the result of 192 features in the horizontal line of Figure 3.

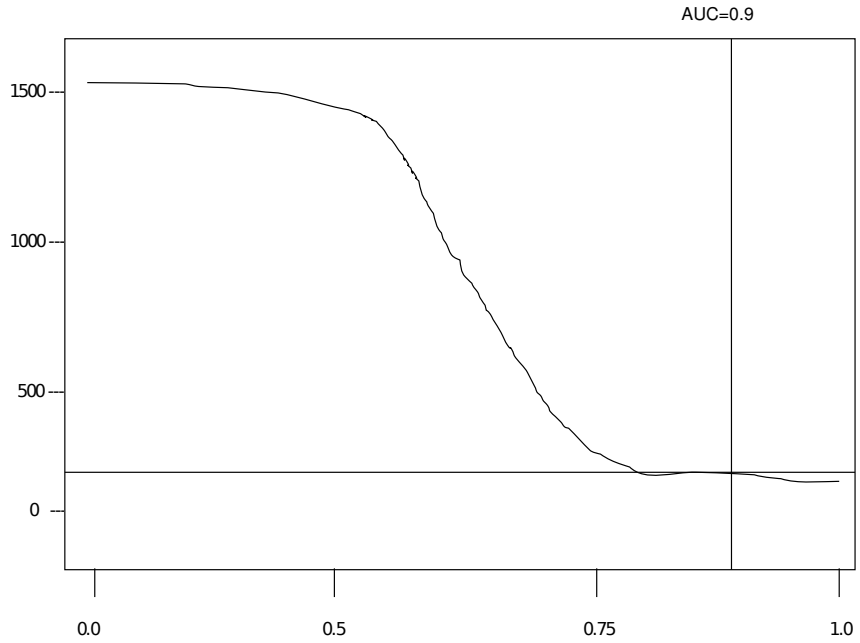

Figure 3. The number of feature at different AUC.

The criterion permits us for the selection of 192 features for model construction and the performance of the model is also compared with other models. The rates of errors for all different data-sets are being measured. The level of the significance is also observed in the difference in the rate of error of mean and corresponding p-value estimated using two sample t-test. Every pair of the data-set provides us the information whether the rates of errors between them are significantly differ or not.

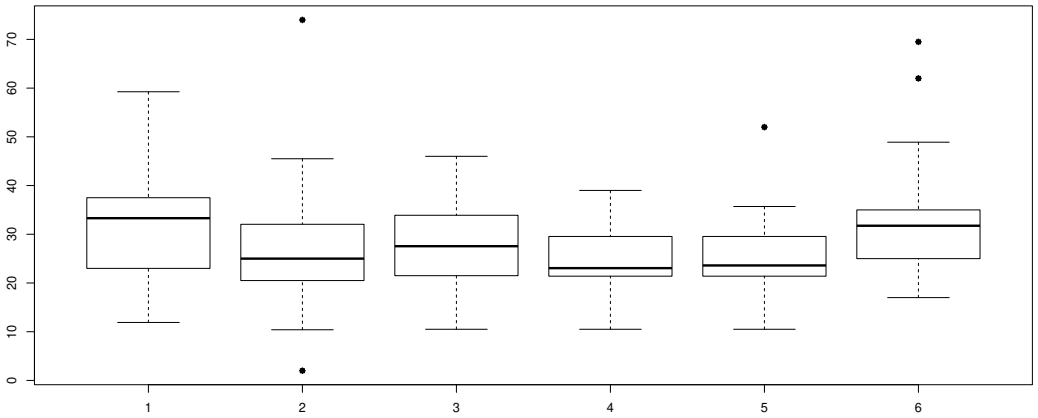

Figure 4. Boxplot for the rate of error.

We have constructed a random sampling data-set and then compare with other models, considering that there should be a significant increase in the accuracy between feature selection by random sampling and feature selection by others. The random sampling has been performed by setting the seed value as 100 in R. Then we make a comparison among the features which are obtained from the underlying models (i.e., LR, DT and RF) and it is also found that many of the features do not overlap. In Figure 4, the box-plot for the rates of errors are obtained using bootstrap method of five distinct data-sets.

The methods: LR (23.329%) and RF (22.782%) give the minimum rates of errors among all other methods. Clinical trials are important for discovering new treatments for diseases, as well as new ways to detect, diagnose, and reduce the chance of developing the disease. It has so many benefits that provide a scientific basis for advising and treating patients. It can access to promising new treatments often not available outside the clinical-trial setting. It plays a more active role in health care purpose and treatment may be more effective than the standard approach. Researchers can able to provide close monitoring, advice, medical care and more frequent health check-ups as part of treatment. It makes opportunities for helping others to get a better treatment for health problems in the future. It plays an active role in purpose of health care and gains a better understanding of disease or condition. So, it helps the society by contributing to medical research. Even if people do not get any beneficiary from the results of the clinical trial, they can take part for gathering the information that can help others and adds to scientific knowledge. People who take part in clinical trials are vital to the process of improving medical care [14].

5. Interpretations of the results related to breast cancer

In this work, the LR and RF method exceptionally points out some features of gene expressions. But LR classification algorithm uniquely identifies some features of gene expression more than RF method. It can be concluded that the corresponding genes of these features are correlated with breast cancer disease, whose dysregulation (abnormality) may be assisted for the progression in mortality. With zinc finger motifs and multiple homeodomains, Estrogen receptor (ESR1 chosen by LR and RF algorithm) gene encodes a transcription factor and also regulates the myogenic and neuronal differentiation [15]. The encoded protein suppresses expression of the alpha-fetoprotein gene by binding to an AT-rich enhancer motif. The protein has also been shown to negatively regulate c-MYB, and transactivate the cell cycle inhibitor cyclin-dependent kinase inhibitor 1A (also known as p21CIP1) [16]. This gene is persisted to function as a tumor suppressor in different types of cancers and an association with atrial fibrillation is also created by the sequence variants of this gene. Multiple transcript variants expressed from alternate promoters and encoding different isoforms have been found for this gene [17]. Inhibits ESR1 function by selectively competing with coactivator NCOA3 for binding to ESR1 in ESR1-positive breast cancer cells [18]. Thus the gene expression of the c-MYB is correlated with estrogen receptors (ERs) expression in the case of breast tumors, i.e., c-MYB gene can increase the mortality rate due to breast cancer [19].

Apart from this, LR classification is uniquely identified the Arid4b, DNMT3A, TOX3, PNLIP and RGS17 genes. There exists a great and potential role of Arid4b gene for BC tumorigenesis and metastasis. Arid4b gene is a component of the mSin3a histone deacetylase complex that has a familiar role in the progression of BC and is a candidate gene underlying a metastasis QTL peak which have been identified on mouse chromosome 13 [20]. Besides, the gene DNMT3A also associated with breast cancer disease, can provide the instructions for making an enzyme called DNA methyltransferase 3 alpha which play an important in many cellular functions. This gene is related with breast cancer cells line and make a significant role [21, 22]. TOX3 gene which encoded protein is glutamine-rich due to CAG repeats in the coding sequence. A minor allele of this gene has been implicated in an elevated risk of breast cancer. The high level of RGS17 gene expression is noticed in diverse human cancers and associates with tumor progression and it is also suggest that RGS17 leads an important role in

breast cancer progression [23]. PNLIP, a member of the lipase family of proteins is another important gene which play a significant role in progression of breast cancer disease [24]. The mutations in this PNLIP gene cause congenital pancreatic lipase deficiency, a rare disorder which is characterized by steatorrhea. The underlying models (i.e., LR, DT and RF) commonly point out certain features corresponding to the genes THBS, BRCA, TUBA1A, PFKM, DSC2 and KRT86. THBS gene is a member of the thrombospondin family. It is a disulfide-linked homotrimeric glycoprotein which mediates cell-to-cell and cell-to-matrix interactions. It is correlated with a potent inhibitor of tumor growth and angiogenesis. It is a multi-domain matrix glycoprotein that has been shown to be a natural inhibitor of neovascularization and tumorigenesis in healthy tissue. The TSP-1 3TSR which is a recombinant version of the THBS1 antiangiogenic domain containing all three thrombosopondin-1 type 1 repeats that can activate transforming growth factor beta 1 (TGFβ1 which play a major role in breast cancer progression [25]. The BRCA1 gene instructs for making a protein that acts as a tumor suppressor. Tumor suppressor proteins help prevent cells from growing and dividing rapidly or in an uncontrolled manner. The BRCA1 protein is involved for repairing damaged DNA. If estrogen receptors, progesterone receptors, or large amounts of HER2/neu protein is not present in the cells of breast cancer, then the mutation of harmful BRCA1 increase the level. In the general population, sometimes about 12% of women will develop BC of their lifetime. By contrast, a recent work concluded that at age of 80, around 72% of women (inherit a harmful BRCA1 mutation) and around 69% of women (inherit a harmful BRCA2 mutation) will develop their breast cancer. The harmful BRCA1 or BRCA2 mutations also have a high risk for developing a new primary cancer in the opposite breast in the years following a breast cancer diagnosis in case of women. It has been disclosed that by 20 years after a first breast cancer diagnosis, about 40% of women (inherit a harmful BRCA1 mutation) and about 26% of women (inherit a harmful BRCA2 mutation) will develop cancer in their other breast [26, 27]. TK1, Tubulin-1-alpha and TYMP genes (TYMP genes are identified only by LR and RF) are majorly correlated with activity of breast cancer. TK1 gene(i.e., thymidine kinase 1) which is highly associated with breast cancer disease, catalyzes the addition of a gamma-phosphate group to thymidine and TK1 gene plays a sinificant role for breast cancer [28]. Tubulin isoforms in breast cancer to explore any correlation between tubulin alterations and taxane resistance [29]. TYMP (thymidylate phosphorylase, previously known as ECGF1) is a large panel for breast cancer disease [30]. A common genetic spectrum for breast cancer at any age is supported by PFKM gene which is also known as a novel breast cancer gene [31]. The association between the gene expression of PFKM with high risk of breast cancer disease is plausible for several reasons which are given below:

(i) PFKM gene is expressed in cell lines of BC [32].

(ii) There is an association between the variants in the gene with the post-translational modifications which have been depicted to alter the metabolism and the growth of cancer cells have been promoted [33].

(iii) An relationship between gene expression of PFKM and the risk of BC is consistent with observations which suggests that due to aberrant glucose metabolism, a large amounts of glucose can be consumed by tumor cells, through a glycolytic pathway which produces lactate [33].

In this work, it has been observed that the biology of gene expression of PFKM and its modulators together with inhibitors are well characterized. Also, TP53 (known as tumor suppressor protein) has been shown to suppress the gene expression of PFKM in the system of models [34]. For the purpose of breast cancer prevention and treatment, the identification of the PFKM gene region has potential

translational implications as a breast cancer susceptibility locus [35, 36]. The two other genes (which can be identified by the methods LR, DT and RF) DSC2 (Desmocollin 2) and gene KRT86 (Keratin86) are associated with breast cancer and also play a lead role for increasing the risk of many diseases including breast cancer [37, 38].

6. Conclusions

In this article, we have demonstrated various classification techniques. The classification algorithm LR uniquely points out some important features but it is not the best chosen for constructing a model with large number of features because LR tends to over-fitting the model whenever the number of features is increasing. After plotting ROC curves, the strength of association between independent and dependent variables can be measured from estimation of the prediction performance. In this work, the performance for the features selection is carried out by selecting models which are best fitted and then built the models by using RF. The maximization of ROC-based criterion is also performed for judging the capability of LR. The model performance is also compared with several methods of features selection and also made a comparison with the method of dimension reduction, i.e., PCA. RST and PCA are also accomplished and compared with other models. There is no similarity among the identification of features by using LR and random forest and sometimes in case of decision tree and logistic regression. Because it is observed that the significant difference between the predicted accuracies by the sets of features is not present (i.e., p-value is greater than 0.05). Some of the features identified using LR method are omitted by DT and RF methods and play a significant role for accelerating breast cancer disease.

Data availability statement

The data used to support the findings of this study are included in the references within the article.

Acknowledgments

The authors are grateful to the learned reviewers and Editors for their careful reading, valuable comments and helpful suggestions, which have helped them to improve the presentation of this work significantly. The third author (Manuel De la Sen) is grateful to the Spanish Government for its support through grant RTI2018-094336-B-I00 (MCIU/AEI/FEDER, UE) and to the Basque Government for its support through grant IT1207-19.

Conflict of interest

The authors declare that they have no conflict of interest regarding this work.

References

1. Ghosh S, Samanta GP (2019) Statistical modelling for cancer mortality. *Lett Biomath* 6: 1–12 .
2. Centers for Disease Control and Prevention, Breast Cancer in Young Women, 2020. Available from: https://www.cdc.gov/cancer/breast.

3. Bao T, Davidson NE (2008) Gene expression profiling of breast cancer. *Adv Surg* 42: 249–260.
4. Centers for Disease Control and Prevention, Family Health History and the BRCA1 and BRCA2 genes, 2020. Available from: https://www.cdc.gov/genomics.
5. Nicolau M, Levine AJ, Carlssonn G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *P Natl Acad Sci USA* 108: 7265–7270.
6. Everson TM, Lyons G, Zhang H, et al. (2015) DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. *Genome Med* 7: 89.
7. Baur B, Bozdag S (2016) A feature selection algorithm to compute gene centric methylation from probe level methylation data. *PLoS One* 11: e0148977.
8. Mallik S, Bhadra T, Maulik U (2017) Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data. *IEEE T Nanobiosci* 16: 3–10.
9. Breiman L (2001) Random forests. *Mach Learn* 45: 5–32.
10. Ramanan D, NKI Breast Cancer Data, Data World, 2016. Available from: https://data.world/deviramanan2016/nki-breast-cancer-data.
11. Livingston F (2005) Implementation of Breiman's random forest machine learning algorithm. *ECE591Q Mach Learn* 1–13.
12. Gareth J, Daniela W, Trevor H, et al. (2013) *An Introduction to Statistical Learning with Applications in R*, New York: Springer.
13. Rakotomamonjy A (2004) Support Vector Machines and Area Under ROC curve, *Citeseer^X*. Available from: http://citeseerx.ist.psu.edu/.
14. Lang T, Siribaddana S (2012) Clinical trials have gone global: Is this a good thing?. *PLOS Med* 9: e1001228.
15. Gurdon JB, Javed K, Vodnal M, et al. (2020) Long-term association of a transcription factor with its chromatin binding site can stabilize gene expression and cell fate commitment. *P Natl Acad Sci USA* 117: 15075–15084.
16. GTR: Genetic Testing Registry, National Center for Biotechnology Information, 2009. Available from: https://www.ncbi.nlm.nih.gov.
17. Sun JW, Collins JM, Ling D, et al. (2019) Highly variable expression of ESR1 splice variants in human liver: Implication in the liver gene expression regulation and inter-person variability in drug metabolism and liver related diseases. *J Mol Genet Med* 13: 434.
18. Gupta A, Hossain MM, Miller N, et al. (2016) NCOA3 coactivator is a transcriptional target of XBP1 and regulates PERK–eIF2α–ATF4 signalling in breast cancer. *Oncogene* 35: 5860–5871.
19. Quintana AM, Liu F, O'Rourke JP, et al. (2011) Identification and regulation of c-Myb target genes in MCF-7 cells. *BMC Cancer* 11: 30.
20. Winter SF, Lukes L, Hunter KW (2010) Abstract 2371: Arid4b is a potential breast cancer progression modifier gene. *Cancer Res* 70: 2371.

21. Jahangiri R, Jamialahmadi K, Gharib M, et al. (2019) Expression and clinicopathological significance of DNA methyltransferase 1, 3A and 3B in tamoxifen-treated breast cancer patients. *Gene* 685: 24–31.
22. Khazayel S, Mokarram P, Mohammadi Z, et al. (2018) Derivative of stevioside; CPUK02; restores ESR1 gene methylation in MDA-MB 231. *Asian Pac J Cancer P* 19: 2117–2123.
23. Li Y, Li L, Lin J, et al. (2015) Deregulation of RGS17 expression promotes breast cancer progression. *J Cancer* 6: 767–775.
24. Zhang G, He P, Tan H, et al. (2013) Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clin Cancer Res* 19: 4983–4993.
25. Weng T, Wang C, Hung Y, et al. (2016) Differential expression pattern of THBS1 and THBS2 in lung cancer: Clinical outcome and a systematic-analysis of microarray databases. *PLoS One* 11: e0161007.
26. Howlader N, Noone AM, Krapcho M, et al. (2017) SEER Cancer Statistics Review, 1975–2014. National Cancer Institute, Bethesda.
27. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. (2017) Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* 317: 2402–2416.
28. Weagel EG, Burrup W, Kovtun R, et al. (2018) Membrane expression of thymidine kinase 1 and potential clinical relevance in lung, breast, and colorectal malignancies. *Cancer Cell Int* 18: 135.
29. Nami B, Wang Z (2018) Genetics and expression profile of the tubulin gene superfamily in breast cancer subtypes and its relation to taxane resistance. *Cancers* 10: 274.
30. Marangoni E, Laurent C, Coussy F, et al. (2018) Capecitabine efficacy is correlated with TYMP and RB1 expression in PDX established from triple-negative breast cancers. *Clin Cancer Res* 24: 2605–2615.
31. Ahsan H, Halpern J, Kibriya MG, et al. (2014) A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. *Cancer Epidemiol Biomark Prev* 23: 658–669.
32. Zancan P, Sola-Penna M, Furtado CM, et al. (2010) Differential expression of phosphofructokinase-1 isoforms correlates with the glycolytic efficiency of breast cancer cells. *Mol Genet Metab* 100: 372–378.
33. Šmerc A, Sodja E, Legiša M (2011) Posttranslational modification of 6-phosphofructo-1-kinase as an important feature of cancer metabolism. *PloS One* 6: e19645.
34. Danilova N, Kumagai A, Lin J (2010) p53 upregulation is a frequent response to deficiency of cellessential genes. *PloS One* 5: e15938.
35. Deng H, Yu F, Chen J, et al. (2008) Phosphorylation of Bad at Thr-201 by JNK1 promotes glycolysis through activation of phosphofructokinase-1. *J Biol Chem* 283: 20754–20760.
36. Usenik A, Legiša M (2010) Evolution of allosteric citrate binding sites on 6-phosphofructo-1-kinase. *PloS One* 5: e15447.
37. Landemaine T, Jackson A, Bellahcène A, et al. (2008) A six-gene signature predicting breast cancer lung metastasis. *Cancer Research* 68: 6092–6099.

38. Notas G, Pelekanou V, Kampa M (2015) Tamoxifen induces a pluripotency signature in breast cancer cells and human tumors. *Mol Oncol* 9: 1744–1759.