Biophysics

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

AIMS Biophysics, 10(1): 67–89.
DOI: 10.3934/biophy.2023006
Received: 26 September 2022
Revised: 02 February 2023
Accepted: 08 February 2023
Published: 17 February 2023

Research article

Predicting factors and top gene identification for survival data 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: For high-throughput research with biological data-sets generated sequentially or by transcriptional micro-arrays, proteomics or other means, analytic techniques that address their high dimensional aspects remain desirable. The computation part basically predicts the tendency towards mortality due to breast cancer (BC) by using several classification methods, i.e., Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Decision Tree (DT), and compared the models' performances. We proceed with the RF method since it provides better results than any other underlying models based on accuracy. We have also demonstrated some traditional and competing risk models, illustrated the models with real data analysis, depicted their curves' natures and also compared their fits using prediction error curves and the concordance index. Furthermore, two different survival splitting rules are used by using separate Random Survival Forest (RSF) methods and also constructing the ranking of risk factors due to breast cancer. The results show that high-level grade and diameter are the most important predictors for mortality progression in the presence of competing events of death, and lymph nodes, age and angiography are other vital criteria for this purpose. We have also implemented RSF backward selection criteria, which enables top gene selection related to mortality progression due to breast cancer. This method identifies c-MYB, CDCA7, NUSAP1, BIRC5, ANGPTL4, JAG1, IL6ST, and remaining genes that are mainly responsible for mortality progression due to breast cancer. In this work, *R* software is used to obtain and evaluate the results.

Keywords: breast cancer; random forest; accuracy; brier score; minimal depth; variable importance

1. Introduction

Cancer is a disease that arises from cells that leave the cell cycle, start to proliferate in an uncontrolled manner and spread into surrounding tissues. This proliferation could be induced by hormones that are impinging on the breast. Generally, most breast cancers begin in the ducts or lobules. 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. Genetic micro-array analysis of the 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 the therapeutic point of view. It gives detailed information about the expression levels of thousands of genes in BC and depicts molecular portraits of BC. A high risk of getting breast cancer may depend on a family history of breast cancer or inherited changes in BRCA1 and BRCA2 genes. There are several applications of a recently developed mathematical field called topological data analysis (TDA). The two most important methods of TDA are (i) the Progression Analysis of Disease and (ii) the analysis of Betti numbers [1]. 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 helps to identify Estrogen Receptor-positive (ER+) cells, which form 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 negative evolution. There is no other way to distinguish between healthy and victimized people who belong to this group. This group has an understandable, distinct and also statistically significant molecular signature. It can reveal coherent biology but conceal for cluster analysis and fail for fitting into the classification (which is accepted) of Normal-like subtypes of Estrogen Receptor-positive BC and also in the case of Luminal A/B. This group is known as c-MYB+ BC [1]. When high dimensional data has been considered, gene expression data gives various proposals and aspects [2]. Based on the sequential forward selection, an algorithm is developed which is used for regression and several classification purposes for selecting DNA methylation probes that are very important with respect to the expression of their corresponding genes [2].

In the previous few years, BC has been analyzed substantially, so the prognosis rate is increasing, and the death rate is decreasing. However, more research is still required for a full understanding of its mechanism and corresponding systematic treatment. Conventionally, doctors mainly rely on biological techniques for diagnosing cancer, which are as follows: (i) Ultrasonography, (ii) B-Scan, and (iii) Fine Needle Aspiration Cytology (FNAC) [3].

Our purpose is to gain significant potential information for breast cancer transcriptional genomic data which have a great influence on gene expression and have a significant role in mortality due to breast cancer. So, we make comparisons, we have demonstrated by approaching logistic regression (LR), random forests (RF) and support vector machine (SVM), along with linear discriminant analysis (LDA) and decision tree (DT). Among all, the random forests (RF) method performs best in this work based on accuracy. So, we proceed with the RSF method for clinical data purposes and the backward RSF method for gene selection purposes. This study seeks to investigate the most appropriate model to examine the most important risk factors of clinical data and genes that significantly influence the mortality rate. In section 2, we have mentioned materials and basic statistical tools. Then, the comparison of the performances of the underlying models is discussed in section 3. Next, in section 4, some traditional and competing risk models have been demonstrated. In section 5, we have applied the proposed models with real data examples and depicted their curves' natures, and we also compared their fits using prediction error curves and the concordance index. Additionally, two different survival splitting rules are used by using separate RSF methods in section 6, and we also find the ranking of risk factors due to BC. In section 7, we have implemented RSF backward selection criteria, which enables top gene selection related to breast cancer progression. Finally, the last section consists of the general discussions and conclusions of this work.

2. Methods and materials

2.1. Data description

The data on Breast Cancer was provided by NKI Breast Cancer [4]. The data had a sample of 272 breast cancer patients (as rows), with 1570 columns. The data set consists of 272 breast cancer patients and 1567 attributes (1554 gene attributes, 10 clinical attributes, and 3 patient general attributes).

2.2. Preliminaries of logistic regression

Logistic regression (LR), applied on a binary dependent variable, is uttermost important for ordered categorical response [5]. Binomial regression is a regression analysis procedure where the dependent variable (sometimes referred to as Y) is a series of Bernoulli trials, or it may be the result of a series of one of two possible disjoint outcomes (conventionally with "success" denoted as 1 and "failure" denoted as 0) in statistics. The log-binomial model is a model of the binomial generalized linear model (i.e., GLM) together with a log link function, which is widespread in epidemiological and bio-statistical fields. For a binary regression, the dependent variable is denoted as Y. Let X be the independent variable, and let $\Phi(x) = P(Y = 1|X = x)$. The logistic regression is as follows:

$$\Phi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{2.1}$$

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

$$\text{logit}[\Phi(x)] = \log \frac{\Phi(x)}{1 - \Phi(x)} = \alpha + \beta x \tag{2.2}$$

The expression (2.2) equates the logit function to a 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 (2.2), which gives a basic interpretation for the magnitude of β. In this work, we have taken the alive condition as response variables whose values are 1 and 0, according to alive and dead, respectively.

2.3. Random forests (RF) algorithm

Random forests (RF), a decision tree, is a classification approach, suitable for both (i) parametric and (ii) non-parametric purposes. This method also establishes a multitude of decision trees using

Bootstrap. 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 reduces both (i) test error and (ii) out-of-bag (OOB) error. Random forests overcome several problems which are generated by another tree-based method [6, 7].

2.4. Feature selection using support vector machine (SVM)

Support vector machine (SVM) algorithm is constructed specifically for binary classification, and also it can be extended to multi-class classification. Sometimes data are obtained from more than two classes, and then it is known as multi-class. For classification purposes, SVM deposits data points into n dimensional space (where n is the number of attributes). It makes two regions by separating the space with the help of hyperplanes. The formula is as follows:

$$\min_{w,b} \frac{||\omega||}{2} + \zeta \sum_{i=1}^{N} \eta^{(i)} \tag{2.3}$$

$$\text{subject to: } y^{(i)} \left(\omega^T \Phi(x^{(i)}) + b \right) \geq 1 - \eta^{(i)} \tag{2.4}$$

where $\eta^{(i)} \geq 0$ for all $i \in \{1, 2, \ldots, N\}$ be a slack variable, and ζ is the penalty of the error term and where ω is the normal vector to the hyperplane. The kernel function Φ is involved in transforming the data set from the input space to a higher dimensional output space (where the data can be linearly separated) whenever the data set is not linearly separable in the input space.

2.5. Linear discriminant analysis (LDA) for feature selection

Discriminant analysis is very useful and well-known to select features, and it can be used successfully for many classification purposes. Linear discriminant analysis (LDA) is based on the covariance matrix of training data. At first, this approach was described in the case of two-class, and then it can be swiftly extended to multi-class cases via multiple discriminant analysis. This analysis is very useful for multi-class classification purposes. In the case of LDA, the assumption is $P(x|y = 0)$ and $P(x|y = 1)$ (i.e., conditional probability density function) both are normally distributed, with (μ_0, Σ_0) and (μ_1, Σ_1) respectively, where μ_0 and μ_1 are the corresponding means, and Σ_0 and Σ_1 are the corresponding covariance matrices. In this work, we have classified one point as distinct from another one if the following condition is satisfied:

$$(x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) + \ln |\Sigma_0| + (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln |\Sigma_1| > \tau \tag{2.5}$$

where τ (two classes without specified weights) is 0.5 in this work.

2.6. Decision tree (DT)

Decision tree (DT), a tree-based approach for representing every decision, leads to the last regression or classification result. A decision 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 according to a certain parameter. The tree-based methods generally handle high-dimension datasets. 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.

3. Comparison for models' performances based on accuracy

3.1. Fundamental evaluation measures

For classification purposes, generally, the classifier is evaluated by a confusion matrix. For a binary classification problem, a matrix is a square of 2 × 2, as shown in Table 1. In this table, the column represents the classifier prediction, while the row is the real value of the class labels. Accuracy, the most common metric for classifier evaluation, assesses the overall effectiveness of the algorithm by estimating the probability of the true value of the class label. Accuracy is defined as:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{31}$$

The abbreviations TP, FN, FP and TN of the confusion matrix cells are defined as follows:

TP: true positive (the number of positive cases that are correctly identified as positive),
FN: false negative (the number of positive cases that are misclassified as negative cases),
FP: false positive (the number of negative cases that are incorrectly identified as positive cases),
TN: true negative (the number of negative cases that are correctly identified as negative cases).

Table 1. 2 × 2 Contingency Table for Accuracy.

	Predicted Positive	Predicted Negative
Actual positive	TP	FN
Actual negative	FP	TN

For comparing the importance of features that are chosen by LR, RF, SVM, etc., firstly, we have to construct the model of classification for every feature and make a comparison among the predicting accuracies. In this work, the underlying models are fitted for mortality due to BC, and every classification is used with only clinical data, only gene data and a combination of both. In terms of accuracy for the clinical-based, gene-based and combine models, there is a clear vision that the overall percentage of RF is the best among all except in the case of model performance. When we predict clinical-based modeling in terms of accuracy, the method RF is the best model among all underlying models. In this work, DT is the second best among all. The method LDA is the worst among all the performing models (from Table 2). Next, the results have been provided for gene-based models in terms of accuracy, and it is predicted that the RF method is also the best among all. Whenever we are predicting genes, there is not present any clear view for choosing the second best among the rest of the underlying models. Here, DT provides the worst results, shown in Table 2. In combined-based modeling, the best performing method is RF, along with SVM (though it is never the best in clinical-based and gene-based modeling).

Finally, it is noted that LDA is worst among all in combined cases with respect to accuracy, shown in Table 2. As per the obtained results, RF is the best among all, since its accuracy is higher than any other classification shown in Table 2. So, we proceed with an improved random forest-based rule extraction method for breast cancer diagnosis for mortality rate due to BC and demonstrate various analyses, and we also depict several methods from statistical points of view.

Table 2. Accuracy results based on several classifications.

Classification	Overall	Gene	Clinical
Logistic	68.276	68.276	72.069
RF	73.931	71.379	72.759
SVM	73.483	68.704	72.062
LDA	67.241	67.241	71.034
DT	70.345	70.000	70.690

4. Models

4.1. Cox proportional-hazards model

The most popular regression model for event time data is the proportional hazards model introduced by Cox (CPH) [8]. In this model, it is assumed that hazard ratios are constant over time and that each covariate under consideration has a linear effect on the logarithm of the hazard rate, whenever the other covariates are given. The Cox regression model is as follows:

$$\lambda(t|\mathbf{x}) = \lambda_\mathbf{0}(\mathbf{t}) \exp(\beta^\mathbf{t}\mathbf{x}) \tag{4.1}$$

with the unspecified baseline hazard rate $\lambda_0(t)$ for a (possibly fictitious) individual with a covariate vector of zeros, the P-dimensional vector of covariates $\mathbf{x}$ and the vector of regression coefficients β. The hazard ratio between two individuals i and j can be computed as follows:

$$\frac{\lambda(t|\mathbf{x_j})}{\lambda(t|\mathbf{x_i})} = \frac{\lambda_0(t)\exp(\beta^\mathbf{t}\mathbf{x_j})}{\lambda_0(t)\exp(\beta^\mathbf{t}\mathbf{x_i})} = \exp(\beta^\mathbf{t}(\mathbf{x_i} - \mathbf{x_j})) \tag{4.2}$$

Generally, survival analysis examines the relationship of the survival distribution to covariates. Cox proportional hazard regression can investigate the effect of different variables on the time (a specified event takes to happen). This model has established an association between the survival time of patients and one or more predictor variables.

4.2. Competing risk forests

4.2.1. Competing risk

In this section, the breast cancer clinical-based dataset process is modeled by the statistical model developed for competing risk data. When an individual is at risk of failing from K distinct types of events, these different event types are called competing risks, which are broadly covered in the statistical literature. An alternative approach to competing risks is the consideration of a bivariate random variable (T, D), where T is a random variable for the event time, and D is a random variable for the event type. For each individual $i = 1, ..., n$, a couple of event times or last time is known to be free of any event; t_i and a status variable indicating the type of event $d_i \in \{1, \ldots, K\}$ or a censored event time $(d_i = 0)$ are observed.

4.2.2. Splitting rule

The splitting rule is used to grow competing risk trees. For notational convenience, the rules for the root node are described properly, but the idea extends obviously to any tree node and to bootstrap data [9]. Let us consider $(T_i, \delta_i)_{1\leq i\leq n}$ to be the survival times and event indicators, where n is the number of individuals within a node, and let $t_1 < t_2 < ... < t_m$ be the distinct and ordered event times from $(T_{1\leq i\leq n})$. Suppose that the proposed split for the root node is of the form $x \leq c$ and $x > c$ for a continuous predictor x; c is a split value for predictor x. Such a split forms two daughter nodes containing two new sets of competing risk data. At first, B bootstrap samples are drawn from the original data, and a survival tree on each of the $b = 1, ..., B$ bootstrap samples is grown. At every node, p (where p is equal to the square root of the total number of predictor variables) predictor variables will be randomly selected to be split in such a way that the splitting value maximizes the difference in the objective function. In other words, the best split for node h is the one in which the predictor and split value maximize the difference in survival between the two daughter nodes for all x and c.

4.2.3. Random survival forests

Definition: A random survival forest (RSF) is an assembly of trees method for analysis of right-censored time-to-event data. It is an extension of Breiman's random forest method [9].

RSF is introduced for extending RF to the setting of right-censored survival data [10]. Implementation of RSF follows the same general principles as RF: (a) Survival trees are grown by using bootstrapped data. (b) Random feature selection is used when splitting tree nodes. (c) Trees are generally grown deeply. (d) The survival forest ensemble is calculated by averaging terminal node statistics (TNS). In this work, we have approached the competing risks which build on the framework of random survival forests (RSF). The performance of these models has been confirmed in different areas [9]. Among them, random survival forests (RSF), a non-parametric tree-based ensemble method, can automatically handle the difficulties of the Cox model [11]. It can be also used effectively in high-dimensional datasets. Generally, the RF method has been used for classification and regression purposes, but it can be extended to censored lifetime datasets. An advantage of this approach is that it is fully non-parametric, including the effects of the treatments and predictor variables. In traditional methods, the assumption was a distribution for the lifetimes or, in the case of the Cox regression, a linear-exponential form for the treatment effects. In the RSF method, the splitting rules are used to grow the tree, and the estimated values calculated within the terminal nodes are used to define the ensemble procedure. Currently, RSF has four distinct methods that can be used to maximize a splitting value c for a variable x. The first method, the one that is used in the survival analysis below, is the log-rank splitting method, which, as the name suggests, uses a multitude of log-rank tests to measure the severity of node separation at a value c for a predictor x. The value of the log-rank test is given by the following formula:

$$L(x,c) = \frac{\sum_{i=1}^{N}(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i})}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}(1 - \frac{Y_{i,1}}{Y_i})\frac{Y_i - d_i}{Y_i - 1}d_i}} \tag{4.3}$$

where $Y_{i,j}$ is the individuals who are at risk (alive) or who had an event (death), and $d_{i,j}$ is the number of events at time t_i in daughter node j, where $j \in \{1,2\}$. The second method is an approximation to the previous one and is therefore named approximate log-rank splitting. The goal is to find the x and c which give the largest magnitude of the log-rank test. That is, we wish to find a predictor $\overline{x}$ and split value $\overline{c}$ such that $|L(\overline{x},\overline{c})| \geq |L(x,c)|$ for every x and c. This process is repeated at every node until the terminal node is reached. In order to approximate the numerator of $L(x,c)$, a revision is done using the Nelson-Aalen cumulative hazard estimator for the parent node. The Nelson-Aalen estimator [12] is as follows:

$$\hat{H}(t) = \sum_{t_i \leq t}\frac{d_i}{Y_i} \tag{4.4}$$

So, we can rewrite the current numerator of the $L(x,c)$ as follows:

$$\sum_{i=1}^{N}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right) = D_j - \sum_{l=1}^{n} I\{x_l \leq c\}\hat{H}(T_l) \tag{4.5}$$

where $D_j = \sum_{i=1}^{N} d_{i,j}$; $j = 1,2$. Furthermore, it can be simplified by considering $D = \sum_{i=1}^{N} d_i$, and therefore the log-rank test is as follows:

$$\frac{D^{\frac{1}{2}}(D_j - \sum_{l=1}^{n} I\{x_l \leq c\}\hat{H}(T_l))}{\sqrt{\{\sum_{l=1}^{n} I\{x_l \leq c\}\hat{H}(T_l)\}\{D - \sum_{l=1}^{n} I\{x_l \leq c\}\hat{H}(T_l)\}}} \tag{4.6}$$

4.3. Conditional inference forests

Conditional inference forests (CIForests) for another fully non-parametric (tree-based) method used in survival analysis like random survival forests. It is based on Breiman's random forests. A conditional inference tree is constructed as follows [13]:

(I) For each predictor variable, test the null hypothesis that there is independence between the response variable(s) and the predictor variable. If we fail to reject the null, stop; otherwise, choose the predictor variable which has the strongest association with the response. The association strength is assessed using the p-values from all partial null hypotheses of a single predictor and the response(s). A split only occurs when the p-value is smaller than a specified value.

(II) Divide the observations of the selected predictor variable using a binary split. This splitting criterion is based on multiplicity-adjusted p-values (Bonferroni or Monte Carlo), univariate p-values (Univariate), or, on values of the test statistic. When the criterion, specified by the option min criterion, is exceeded, a split is made. This method allows a tree to be grown to the correct size without the need for pruning.

(III) Repeat the previous steps until a terminal node is reached (no additional predictor variables can be split). The ensemble survival function is:

$$\hat{S}^{CIForests}(t \mid x_i) = \prod_{t_{l,h} \leq t} \left(1 - \frac{\sum_{b=1}^{B} d_{l,h}}{\sum_{b=1}^{B} Y_{l,h}} \right) \tag{4.7}$$

4.4. Prediction error curves

4.4.1. Definition

Integrated Brier Score: Integrated Brier Score (IBS) is an overall measure for the prediction of the model at all times.

Prediction error curves are increasingly used to assess and compare predictions in survival analysis. These curves are obtained when the Brier score is followed over time. The technique based on bootstrap re-sampling or bootstrap sub-sampling can be applied to assess and compare the predictive power of various regression modeling strategies on the same set of data. Instead, the three models have been compared using two newer methods, namely, prediction error curves and the concordance index (c-index). The prediction error is accessed by an expected time-dependent Brier score. For right censored data, the squared residual of a subject at t time point is weighted by using the inverse probability of censoring weights. This censoring weight is as follows:

$$\hat{W}_i(t) = \frac{(1 - \overline{Y}_i(t))\Delta_i}{\hat{G}(T \mid X_i)} + \frac{\overline{Y}_i(t)}{\hat{G}(t \mid X_i)} \tag{4.8}$$

where $\overline{Y}_i(t) = I(\overline{T} > t)$ is the observed status of an individual i at time t, and $\hat{G}(t \mid x) \approx P(C_i > t \mid X_i = x)$ is the estimate of the conditional survival function of the censoring times. For a new observation, or, if a test dataset $\overline{D}_M$ is available, the expected Brier score is estimated by

$$BS(t, \hat{S}) = \frac{1}{M} \sum_{i \in \overline{D}_M} \hat{W}_i(t)(\overline{Y}_i(t) - \hat{S}(t \mid x_i)) \tag{4.9}$$

where M is the number of subjects in $\overline{D}_M$, and $\hat{S}$ is the predicted survival probability for a subject i at time t based on a training dataset. In order to protect against overfitting, ten-fold cross-validation iterated five times was used for each of the three survival methods on each of the three datasets.

4.5. Concordance index

Unlike prediction error curves and other measures of survival performance, the c-index, is not dependent on a fixed time point for the evaluation of a model and takes into account the censoring status of an individual. Two observations are said to be concordant if the observation that fails first is predicted to have a worse outcome. The process of obtaining the c-index is as follows:
I. Over the entire data, form all possible pairs of observations.
II. If, within a pair, the observation with the shorter survival time is censored, or both observations have the same survival time, but at least one is not an event, omit the pair. All of the remaining pairs are considered permissible.
III. Scoring
(i) A permissible pair receives a value of 1 if any of the followings holds: (a) Their survival times are not equal, and the shorter survival time is predicted to have a worse outcome. (b) Their survival times

are equal, and their predicted outcomes are also equal. (c) Their survival times are equal, not both events, and the predicted outcome is worse for the observation with the observed event.
(ii) A permissible pair receives a value of 0.5 if any of the followings holds: (a) Their survival times are not equal, but their predicted outcomes are equal. (b) Their survival times are equal, but their predicted outcomes are not equal. (c) Their survival times are equal, not both events, and the predicted outcome is not worse for the observation with the observed event.
IV. The c-index, C, is given by C = Concordance/Permissible. V. The error rate is given by Error = $1 - C$, where $0 \leq \text{Error} \leq 1$. $C = 1$ or Error = 0 indicates perfect prediction, whereas C = Error= 0.5 indicates doing no better than guessing.

5. Analysis

Table 3. *Comparison with integrated brier scores of the models.*

Model	IBS
CPH	0.195
RSF	0.161
CIForests	0.165

For evaluating the performance of RSF in this work, we have used the method of comparing the RSF with Cox proportional hazard regression (CPH) and conditional inference forests or CIForests models. The underlying models are compared by applying the Integrated Brier score (IBS) criterion. For comparing the performance of the used RSF methods with traditional counterparts of CPH, CIForests method is used. The data set is randomly divided into a training set (i.e., 70%) and a test set (i.e., 30%), and the process is repeated 100 times. The result of such a computation is shown in Table 3. In both cases of CPH and CIForests, the RSF counterpart provides better performance than the others. In the case of RSF, it gives 0.161 IBS, which is lower than any other model (shown in Table 3). The first technique for comparing survival analysis models makes use of the concordance index (c-index) of each model over time. The c-index gives the probability of concordance between the predicted and the observed survivals. A c-index of 1 refers to the model making a perfect predictions and a c-index of 0.5 means the model did no better than guessing. In Figure 1, it is shown that RSF performs better than other underlying models in the prediction performance of survival. Conditional inference forests tend to be better than Cox proportional hazards but still far inferior to RSF. It is also very interesting that at first Cox seems to be slightly convex. Then, it becomes almost parallel to the x-axis. RSF and CIForests are almost parallel to each other, but RSF performs better than any other underlying model.

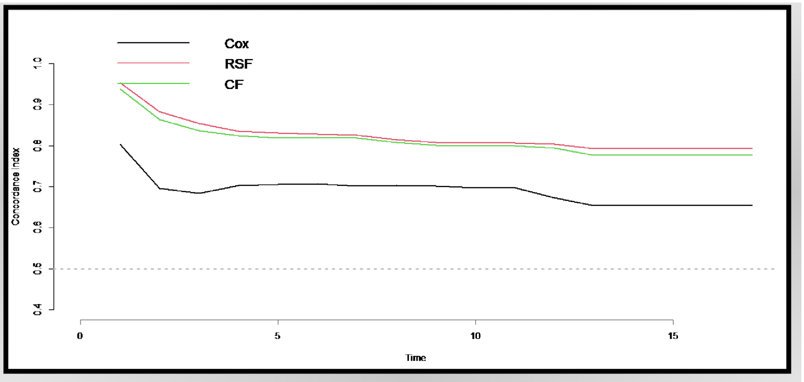


Figure 1. *Comparison of c-indexes for breast cancer data.*

Table 4. *Comparison for prediction of error rates of the models in different times.*

Time	Risk	Reference	CPH	RSF	CF
3.121	151	0.180	0.158	0.137	0.143
6.319	101	0.221	0.188	0.163	0.167
9.982	49	0.236	0.221	0.189	0.190
18.341	0	0.140	0.129	0.109	0.110

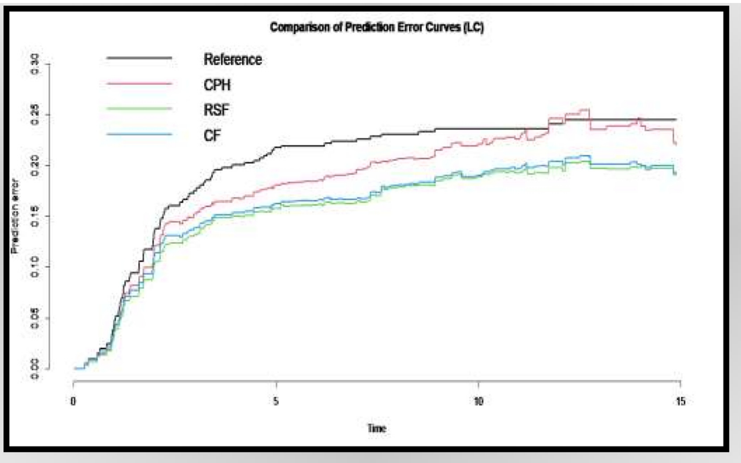


Figure 2. *Comparison of prediction error curves for breast cancer data.*

For comparison purposes, it is shown in Table 4 how prediction errors for each model purpose fluctuate as time goes on. At first, for the initial purpose, the prediction error for each model is zero whenever the number of subjects at risk at time t_0 (where t_0 is the initial time) is 201. As time goes on, the error rate is gradually increasing, and after a certain time, it falls. However, for each case, Table 4 provides that the errors for the RSF model are consistently lower than other underlying models, so the RSF model is the best among all. The prediction error is accessed by an expected time-dependent

Brier score which is already shown in Table 3. The second technique for comparing models is shown by how prediction error for each model purpose fluctuates as time goes on. For the underlying dataset of interest, the prediction error curves for reference, CPH, RSF and CIForests are plotted and compared for finding a suitable model. The approach for comparing models is to see how the prediction error of each model fluctuates with a time period. For the underlying dataset, the prediction error curves for various models have been plotted in Figure 2. Random survival forests seem to consistently have a lower prediction error than the other models over time, as shown in Figure 2.

6. Methodological issues

6.1. Measure of variable importance: Ranking

Random forests can contribute a swiftly computable internal measure of variable importance (i.e., VIMP) that is used for ranking variables. This method is especially very beneficial in the case of high-dimensional genomic data. Permutation importance (which measures the predictive value of a feature) and node impurity indices (example: Gini index, abbreviated G.I.) are two important measures for evaluation [9]. Permutation importance has been applied to the measurement in the case of RF. For the tree data purpose, the underlying covariates are permuted randomly in the out-of-bag (OOB) data (i.e., the original data left out from the bootstrap sample which is used for growing the tree, approximately $1 - 0.632 = 0.368$ of the original sample) for calculating a permutation, and then permuted out-of-bag data are dropped down the tree. Then, prediction error is calculated using the estimation of out-of-bag, i.e., OOB (i.e., the average error for each calculated using predictions from the trees that do not contain the respective bootstrap sample). The difference between the out-of-bag error without permutation and the estimate is known as the VIMP of the variable.

In this work, for measurement purposes, modified VIMP has been used for high-dimensional genomic data. For example, the use of sub-sampling without replacement in place of bootstrapping is suggested by [14] in the case where variables vary in their numbers of categories or scales of measurement. They have also constructed a conditional permutation VIMP that can correct the bias in the case of correlated variables. There exist many valuable applications using permutation importance. However, a ranked based method is harder than the problem of variable selection that simply seeks to select a group of variables without imposing a ranking structure. Due to the complexity of several biological systems, lists of the ranked genes, based on random forests or random survival forests (considering correlation and interaction effects), are more beneficial than univariate ranked gene lists which are based on Cox proportional model by using one variable at a time.

6.2. Minimal depth (MD) to select variable

Ishwaran [15] proposed a new approach to select tree-based forest variables known as minimal depth (MD). With forests, the splitting variables (which are close to the root node) have a very strong effect on the case prediction accuracy. So, such an effect on the VIMP method can be used without any problems. To calculate VIMP, noising up test data leads to poor prediction and large VIMP for this purpose since terminal node assignments maintain distance from their original values for such purpose. The variables which can split higher in the tree have much less impact, since terminal node assignments are not so perturbed. Many advantages are present whenever we have considered minimal depth. This method

(MD) is not dependent on the prediction error. This measurement is used for assessing performance in the case of the MD method for avoiding controversial issues. In the case of survival analysis, a controversy is present regarding whether the concordance index (which is a ranked-based method) is preferable for measurement based on the Brier score. For classification purposes, it is conceded that error due to misclassification may be sub-optimal in the case of RF algorithm with analyses involving unbalanced samples [16], shown in a case of common occurrence for many genomic data purposes. For comparing the performance of the model, [17] has described a comprehensive review of approaches. Apart from this, there exists another advantage (unlike VIMP), that the MD method can be worked out in the closed form. So, a rigorous threshold value can be calculated efficiently in the case of high-dimension chosen variables. Especially, the mean of the MD method under the null of no association with the outcome can be easily computed. Table 5 shows the variable importance (VIMP) and minimal depth (MD) values for all used covariates due to mortality for BC that can be used to rank variables for the underlying method. In this work, the assumptions are as follows: (i) If the value of VIMP is greater than −0.0022, treat it as an effective variable, and (ii) less than −0.0022 values provide treatment effects. Such underlying treatment effects (i.e., chemotherapy, hormonal, amputation) can almost cure breast cancer and cease the mortality rate due to BC. Furthermore, the smaller values of the MD method point out better predictiveness of the underlying variables. In this work, we demonstrate two different procedures viz. VIMP and MD to rank all included variables. Larger values of the VIMP are related to the variables with a better rank. Using the generalized log-rank splitting rule, the variables used in RSF for mortality purposes due to BC are ranked here as displayed in the VIMP column. The largest VIMP value for the event of interest belongs to the cancer-grade criterion, according to Table 5. So, it is the first top-rank variable in mortality progression due to BC. Angiography, age at diagnosis, diameter, and nodes of the cancer tissue have also VIMP greater than −0.0022. So, these underlying factors lead an important role in predicting death progression in BC. Lymph and hist-type criteria are the other risk factors due to BC. Additionally, based on the MD values, the first most important variable is the diameter of the cancer cell, while the second most important variable for death progression by age and then nodes, grade, angiography, etc, respectively, which are responsible factors for mortality progression due to breast cancer.

Table 5. VIMP and MD of the variables used in RSF for breast cancer data.

Variable	VIMP	Variable's rank	Minimal depth	Variable's rank
Chemo	-0.0115	3	0.91	3
Hormonal	-0.0022	2	0.96	1
Amputation	-0.0031	1	0.92	2
Posnodes	-0.0010	4	0.69	8
Grade	0.0764	10	0.76	7
Angioniv	0.0211	9	0.84	6
Lymphinfill	0.0008	5	0.65	10
Age	0.0187	8	0.89	5
Histtype	0.0001	6	0.90	4
Diam	0.0161	7	0.68	9

7. Gene selection

7.1. Stepwise procedures for variable selection

In this work, at first we used the RF method for choosing the top genes selected with the help of stepwise procedures. RF is capable of modeling for a large number of predictors and also can achieve good performance for prediction purposes. However, we have to find a small number of variables with equivalent or better prediction ability, which is required not only for interpretation purposes but also for use in practical situations. Optimal parameters for a random forests model are already found for all response variables. Since RF is the best among all the underlying models, as shown in section 3, we have proceeded with the improved random forest-based rule extraction method for breast cancer diagnosis method and fitted using the entire data. [18] have demonstrated a backward elimination method by using RF to select genes from microarray data. This procedure consists of the following steps: (i) First, we have to fit data using the RF method, and then, according to permutation VIMP, rank all genes responsible for BC. (ii) Then, we have to fit RF iteratively, and at every iteration, we have to discard a proportion of genes from the bottom of the list of genes ranked according to importance (default 10%). (iii) Next, we have to choose a group of genes whenever RF technique reaches the smallest OOB error-rate. (iv) Lastly, the error rate of prediction has been estimated to mitigate selection bias by using the bootstrap method. Then, the RSF backward method is used on the data consisting of all genes to calculate a final RSF with the smallest prediction error rate. With RSF there is no need for standardization of data [19]. So, the crude data set is used for simulation purpose. In this work, the RF method provides insight for selecting top genes. Generally, trees are built from root to leaves, and the closer a variable is to the root. By choosing 'important' parameters with the help of stepwise selection criteria, top genes can be discovered for every dependent variable. In this work, next, we systematically removed noise genes by implementing the stepwise RSF backward method, which is as follows: (i) First we have to calculate an RSF by using a dataset of genes to be tested. (ii) Then, we have to rank the genes by MD method and also discard the gene with the worst MD method from the dataset. (iii) Next, with the remaining data, we have to calculate a new RSF. (iv) Repeat the underlying steps (ii) and (iii) until only one gene remains. Lastly, (v) select the set of genes with the smallest prediction error rate.

7.2. Interpretations of the results related to breast cancer

In this work, the RF method points out some features of gene expressions. It can be concluded that the corresponding genes of these features are correlated with breast cancer disease, whose dysregulation (abnormality) may be assisted in the progression of mortality. RF method chooses the top 40 genes, and next, the implemented RSF backward selection algorithm enables gene selection for responsibility for BC. The application of this process provides the result that almost 10 genes associated with breast cancer disease have slightly improved risk prediction compared with RSF on all genes (Shown in Figure 3). The analysis has revealed that some of the genes are expressed only in breast cancer, and we have explained the significance of such genes. We have found in this work that the high expression of c-MYB is related to breast tumors in humans. This gene 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. It inhibits ESR1 function by selectively competing with coactivator NCOA3 for binding to ESR1 in ESR1-positive breast cancer cells. Thus, the gene expression of the c-MYB is correlated with estrogen receptors (ERs) expression in the case of breast tumors, i.e., a c-MYB gene can increase the mortality rate due to breast cancer [1]. Apart from this, in this work, RF classification is uniquely identified for the CDCA7, NUSAP1, BIRC5, ANGPTL4, JAG1 and IL6ST genes, which are mainly responsible for breast cancer development and progression purpose. CDCA7 (cell division cycle associated family of genes) are involved in embryonic development and dysregulated in various types of human cancer. However, the biological role and molecular mechanism of CDCA7 in Triple-negative breast cancer (TNBC) have not been defined. This gene is preferentially and markedly expressed in TNBC cell lines and tissues. High expression of CDCA7 is associated with metastatic relapse status and predicted poorer disease-free survival in patients with TNBC. CDCA7 silencing in TNBC cell lines effectively impairs cell proliferation, invasion, and migration in vitro. Importantly, depletion of CDCA7 strongly reduces the tumorigenicity and distant colonization capacities of TNBC cells in vivo. Additionally, CDCA7 can increase the expression of EZH2, a marker of aggressive breast cancer which is involved in tumor progression, by enhancing the transcriptional activity of its promoter. This increase in EZH2 expression is essential for the CDCA7-mediated effects on TNBC progression. It is revealed by immunohistochemical analysis that the CDCA7/EZH2 axis is clinically relevant, which suggests that CDCA7 plays an important role in TNBC progression by transcriptionally upregulating EZH2 and might be a potential prognostic factor and therapeutic target in TNBC [20].

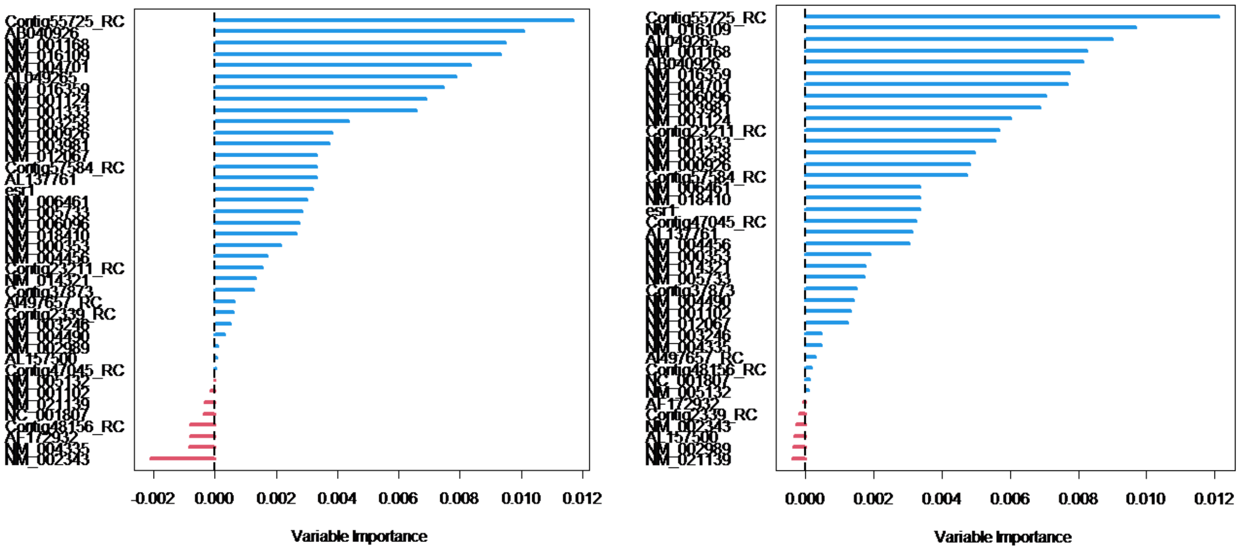


Figure 3. Gene selection by backward RSF method, responsible for breast cancer.

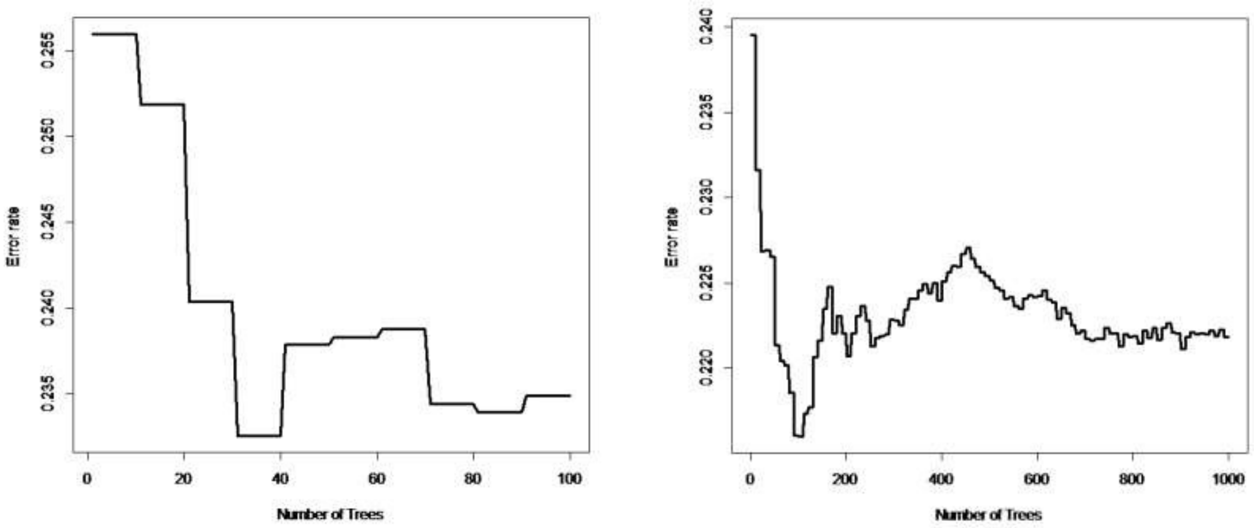
NUSAP1 has been implicated in an elevated risk of breast cancer. It has been reported to function in mitotic spindle assembly, chromosome segregation, and regulation of cytokinesis. It has hitherto unknown functions in the key BRCA1-regulated pathways of double-strand DNA break repair and centrosome duplication. Depletion of NUSAP1 from cells led to the suppression of double-strand DNA break repair via the homologous recombination and single-strand annealing pathways. This gene plays a most important role in these processes through the regulation of BRCA1 protein levels, and BRCA1 over-expression from a plasmid mitigates the defective phenotypes seen upon NUSAP1 depletion. It is revealed that there exists a novel association between BRCA1 and NUSAP1 [21]. 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 in repairing damaged DNA. If estrogen receptors, progesterone receptors, or large amounts of HER2/neu protein are not present in the cells of breast cancer, then the mutation of harmful BRCA1 increases the level. In the general population, sometimes about 12% of women will develop BC in their lifetime. In contrast, recent work has concluded that at age of 80, around 72% of women (who inherit a harmful BRCA1 mutation) will develop breast cancer. The harmful BRCA1 mutations also have a high risk of developing a new primary cancer in the opposite breast in the years following a breast cancer diagnosis in the case of women. It has been disclosed that by 20 years after a first breast cancer diagnosis, about 40% of women (who inherit a harmful BRCA1 mutation) will develop cancer in their other breast [22]. BIRC5 (also known as Survivin) is a member of the inhibitor of apoptosis (IAP) gene family, which encodes negative regulatory proteins that prevent apoptotic cell death. It plays dual roles in promoting cell proliferation and preventing apoptosis. It is recognized to act as an important regulator of the localization of the chromosome passage protein complex during mitosis and cytokinesis. BIRC5 has potential involvement in the case of breast cancer. This gene is also associated with the age of onset in patients of breast cancer [23]. The copy number of BIRC5 has been pointed out to high progress in tumor tissues, and it has the potential to be a marker for the detection and prognosis of BC at an early age. There exists a great association between the expression levels of ANGPTL4 and the prognosis of breast cancer. ANGPTL4 serves an important role in tumor-associated activities, such as tumor cell motility and invasiveness, cell migration, endothelial cell function, vascular leakage, neoangiogenesis and cell adhesion and motility, by interacting with matrix proteins in a variety of solid tumors. Its expression is higher in invasive ductal carcinoma (IDC) (near about 63.4%) compared with normal breast tissues, and the levels of ANGPTL4 mRNA are higher in human breast cancer and in breast cancer cell lines. ANGPTL4 is an independent prognostic factor for BC [24]. It is positively associated with malignant progression and poor prognosis of BC. Next, JAG1 seems to play a central role in linking various pathways, involving well-established cancer-related molecules. In breast cancer, high levels of JAG1 promote stem cell self-renewal and potentiate mammosphere formation in vitro. The involvement of this gene in breast cancer stem cells (CSC) has also been confirmed by mouse models in which mammary-specific deletion of Lfng. An N-acetylglucosamine transferase that prevents Notch activation (a procedure to connect cells and cells that line patent stable blood vessels through direct interaction with the Notch ligand) by Jagged ligands, induces basal-like BC with higher JAG1 activity and enhanced CSC proliferation. It has also been involved in CSC biology in other tumor types. For example, JAG1 can be expressed by both tumor and endothelial cells and play a most important role in glioma/glioblastoma-initiating cells. NOTCH promotes breast CSC survival and self-renewal, and over-expression of NOTCH1 and the NOTCH ligand JAG1 predicts poor outcomes. Approximately 15%–20% of breast cancers are HER2-positive [25]. In HER2-positive breast cancer tissue, higher Jagged1 membrane staining, but not cytoplasmic or perinuclear Jagged1 expression, predicts poor overall survival for women with primary, invasive HER2-positive breast cancer. IL6ST gene is also responsible and plays a central role in TNBC, which is already mentioned in the case of CDCA7 gene purpose. Higher expression of IL6ST shows a significant association with longer overall survival in TNBC patients. IL6ST is the signal transducer for interleukin 6 (IL6), ciliary neurotrophic factor (CNTF), leukemia inhibitory factor (LIF) and oncostatin M (OSM). In general, IL6ST is lower in TNBC when compared to non-TNBCs. It is shown by (name) that high expression of IL6ST has been shown to be a good prognostic factor in breast cancer, as it increases patients' overall survival, which supports our finding in TNBC where higher expression of IL6ST is shown to be associated with significantly increased survival. Multiple studies identified IL6ST as being positively associated with estrogen receptor alpha (ER-α) expression in breast cancer, which confirms the decreased levels of this gene in TNBC patients [26]. Additionally, the presence of elevated levels of ER-α in benign breast epithelium appears to point out an increased risk of BC. Worldwide, 60-70% BC patients are estrogen receptor alpha positive. Apart from this, some genes are also responsible for breast cancer disease. High expression of Sperm-associated antigen 5 (SPAG5) has been detected in BC. The biological function and regulatory mechanism of SPAG5 in breast cancer remain unclear. [27] have revealed the potential biological function of SPAG5 in BC cells. The mRNA and protein expression of SPAG5 both are significantly up-regulated in BC cell lines. The silencing of SPAG5 inhibited the proliferation and invasion of breast cancer cells, as has been shown by functional experiments, while the overexpression of SPAG5 promoted the proliferation and invasion of BC cells. In addition, SPAG5 promoted the expression of Wnt3 and β-catenin and increased the activation of β-catenin transcriptional activity investigated by the mechanistic procedure. The gene SPAG5, which promotes the proliferation and invasion of breast cancer cells by activating β-catenin, has a familiar role in the progression of BC. The THBS gene is a member of the thrombospondin family. It is a disulfide-linked homotrimeric glycoprotein that 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 thrombospondin-1 type 1 repeats that can activate transforming growth factor beta 1 (TGFβ1 which play a major role in breast cancer progression [28]. TK1, Tubulin-1-alpha and TYMP genes (TYMP genes are identified only by LR and SVM) are majorly correlated with the activity of breast cancer. The 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 the TK1 gene plays a significant role in breast cancer [29]. A common genetic spectrum for breast cancer at any age is supported by the PFKM gene, which is also known as a novel breast cancer gene [30]. The association between the gene expression of PFKM and a high risk of breast cancer disease is plausible for several reasons, which are given below:

(i) The PFKM gene is expressed in cell lines of BC [31].

(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 [32].

(iii) A relationship between gene expression of PFKM and the risk of BC is consistent, with observations that suggest that due to aberrant glucose metabolism, a large amount of glucose can be consumed by tumor cells, through a glycolytic pathway which produces lactate [32].

Since the biology of gene expression of PFKM and its modulators together with inhibitors is well characterized, finally, TP53 (known as tumor suppressor protein) has been shown for suppress the gene expression of PFKM in the system of model [33]. For breast cancer prevention and treatment purposes, the identification of the PFKM gene region has potential translational implications as a breast cancer susceptibility locus. There exists a great and potential role for the GNG4 (G protein gamma-4 subunit/guanine nucleotide-binding protein-4) gene for BC tumorigenesis and metastasis. Apart from this, ARF1 (ADP-ribosylation factor 1) is another gene in the ARF gene family that leads an important role in breast cancer progression, and high-level amplification of ARF1 is associated with increased mRNA expression and poor outcomes in patients with breast cancer. In this work, some other genes, TOX3, TYMP, DSC2, PNLIP and RGS17, are also responsible for breast cancer progression [21, 34]. Moreover, we also attempt to improve the interpretation of backward RSF analyses by increasing the number of trees. In this work, we have considered $n = 100$ and 1000, where n denotes the number of trees of the underlying method (shown in Figure 4). These results show that the error rate is stable whenever the number of trees is increasing.



(1.a) **n = 100 (left)**

(1.b) **n = 1000 (right)**

Figure 4. Error rate vs number of trees.

The RSF method is beneficial to identify the variables which are associated with disease in the case of a complex data set considering time to event as an outcome. The analysis of the RSF method also is providing comparable findings which are used in Cox regression in general. It also can address the problem of having multicollinearity, and the method is very useful whenever the data set is highly dimensional.

8. Concluding remarks

Classical treatment selection due to cancers mostly depends on the judgment of doctors and clinical observations, but in general, it is very hard to predict most outcomes. This work has shown its comprehensive understanding of a cancer genomic data-set. The results not only help to predict BC survivability but also predict the genes which are mainly responsible for breast cancer progression. The approach which is used in this work can be also applied in disease learning, especially in the case of high-dimensional data such as genomics. This methodology not only achieves better prediction for clinical purpose but also points out the significant attributes of the patients, which is beneficial in the epidemiological field. This work demonstrates a useful application of the fundamental idea of high-dimensional data in epidemiological research. Furthermore, in the case of several diseases (even those beyond cancer scope), such methodology of modeling and exploration can be appealed for treatment optimization and gene selection systems.

In recent years, similar efforts have been made for developing customized cancer treatments. The idea is based on the gene of a breast cancer patient and also predicts how it is correlated with breast cancer. Also, some therapies are used and target the genes specific to certain patients. Nevertheless, some limitations are present in this technique. The scheme mainly requires investigation and meticulous lab work, which faces difficulties due to time consumption and expense. Although we have analyzed pathways related to BC, cancer genomics is not fully understandable in many difficult cases.

In this work, we have demonstrated several classification methods, i.e., Logistic Regression, Random Forest, Support Vector Machine, Linear Discriminant Analysis and Decision Tree, to predict basically the tendency towards mortality due to breast cancer and also compared their performances. In previous work, some classifications had been evaluated using the features selected method in the case of breast cancer [35]. However, we have not only evaluated and compared the performances of different algorithms but also proceeded with the Random Forest method since it provides better results than any other underlying models based on accuracy. Many different machine learning methods [36] have already been applied for microarray data analysis, like Support Vector Machines [37, 38] or Random forest [39]. Furthermore, in the last few years, these Algorithms have been used for solving both problems of feature selection and classification in gene expression data analysis. Genetic Algorithms [40] have been employed for building selectors where each allele of the representation corresponds to one gene. However, we have demonstrated some traditional and competing risk models, illustrated the models with real data analysis, depicted their curves' nature and also compared their fits using prediction error curves and the concordance index. Furthermore, two different survival splitting rules have been implemented by using separate Random Survival Forest (RSF) methods and also constructing the rank of risk factors due to breast cancer. The results show that high-level grade and diameter are the most important predictors for mortality progression in the presence of competing events of death. Lymph nodes, age, and angiography are other vital criteria for this purpose. We have also implemented RSF backward selection criterion, which enables us to make top gene selection related to mortality progression due to breast cancer and identifies some important genes responsible for mortality progression due to breast cancer.

The classification performance of each model is shown in Table 2 and compared among all based on accuracy. For classification purposes, the best model is associated with the random forests method on each data group: clinical, gene and combined. SVM technique stands in the second position for gene data, and combining data and logistic regression is the second most important method for clinical data purposes. Since RF is the best among all for individual data, we proceed with mainly two different approaches: the RSF method for clinical data purposes and the backward RSF method for gene selection purposes. The focus of the present work is to identify the important prognostic factors which affect the duration of time from breast cancer infection for mortality progression in the presence of a competing event of death. This RSF method benefits from many useful properties and has various important features. RSF points out that high-level grade and diameter are the most important predictors for mortality progression in the presence of competing events of death providing results. Meanwhile, the second most important variables are angiography and age, respectively. The criterion “high-level grade” is also statistically significant in Cox-regression and CIForests method of mortality progression due to BC using traditional models.

Nevertheless, using classical models, age is the only significant variable for the use of the Cox proportional-hazards model and CIForests model. These comparisons (shown in Figure 1 and 2) show a relative consistency between the results of the traditional model strategies and the RSF. We have also compared the performances of the methods in this work. According to the results, RSF outperforms classical models in terms of lower prediction error (shown in Table 4). This can be attributed to the property of considering all complex relationships between variables by the RSF model. Ishwaran et al. [9] in their study also showed that their proposed RSF model outperformed traditional models in competing risk cases. Epidemiologists are motivated for considering the underlying RSF backward selection as a sensible complement to conventional regression-based selection methods for selecting suitable variables whenever analyzing complex survival data which are highly correlated. RSF method has various advantages, compared with regression approaches. This method is completely data-set driven so it is independent of hypothesis testing. This technique does not test the goodness-of-fit of the data set in the case of hypothesis but seeks a model that provides the best explanation of the data set. It is a very suitable and useful technique for exploratory analysis of survival data where previous knowledge is still limited. The RSF backward technique is particularly suitable for selecting variables whenever complex survival data are highly correlated. The underlying method has been demonstrated to point out the unknown covariates related to BC. RSF backward method can be easily implemented and applied for reducing the dimension of the data set and also can improve the interpretability. At the current stage in the case of cancer research, this technique inevitably establishes optimization of error rate. For investigating the direction and potential non-linearity of individual gene associations, partial plots are the first step. “The translation and verification of RSF technique findings into clinically understandable relation measures” can be extended for future research purposes.

Data availability statement

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

Conflict of interest

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

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. They are also thankful to Nirapada Santra, SRF and Bijoy Kumar Das, SRF, and Protyusha Dutta, JRF of the IIEST, Shibpur for helping during the preparation of this manuscript. 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 IT1555-22.

References

1. Nicolau M, Levine AJ, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, *Proceedings of the National Academy of Sciences of the United States of America.* 108: 7265–7270. https://doi.org/10.1073/pnas.1102826108

2. Baur B, Bozdag S (2016) A feature selection algorithm to compute gene centric methylation from probe level methylation data. *PLoS One* 11: e0148977. https://doi.org/10.1371/journal.pone.0148977

3. Trop I, Dugas A, David J, et al. (2011) Breast abscesses: evidence-based algorithms for diagnosis, management, and follow-up. *Radiographics* 31: 1683–1699. https://doi.org/10.1148/rg.316115521

4. NKI Breast Cancer Data, Data World, 2016. Available from: https://data.world/deviramanan2016/nki-breast-cancer-data.

5. Ghosh S, Samanta GP (2019) Statistical modeling for cancer mortality. *Lett Biomath* https://doi.org/10.1080/23737867.2019.1581104

6. Breiman L (2001) Random forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324

7. Livingston F (2005) Implementation of Breiman's random forest machine learning algorithm. *ECE591Q Mach Learn J Pap* 1–13.

8. Cox DR (1972) Regression models and life-tables. *J Roy Stat Soc: Ser B (Meth)* 34: 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

9. Ishwaran H, Gerds TA, Kogalur UB,et al. (2014) Random survival forests for competing risks. *Biostatistics* 15: 757–773. https://doi.org/10.1093/biostatistics/kxu010

10. Ishwaran H, Kogalur UB, Blackstone EH, et al. (2008) Random survival forests. *Ann Appl Stat* 2: 841–860. https://doi.org/10.1214/08-AOAS169

11. Ishwaran H, Kogalur UB, Gorodeski EZ, et al. (2010) High-dimensional variable selection for survival data. *J Am Stat Assoc* 105: 205–217. https://doi.org/10.1198/jasa.2009.tm08622

12. Borgan Ø (2005) Nelson–Aalen Estimator, *Encyclopedia of Biostatistics.* https://doi.org/10.1002/0470011815.b2a11054

13. Mogensen UB, Ishwaran H, Gerds TA (2012) Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw* 50: 1–20. https://doi.org/10.18637/jss.v050.i11

14. Strobl C, Boulesteix AL, Zeileis A, et al. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8: 25. https://doi.org/10.1186/1471-2105-8-25

15. Ishwaran H, Kogalur UB, Chen X, et al. (2010) Random survival forests for high-dimensional data. *Stat Anal Data Min ASA Data Sci J* 4: 115–132. https://doi.org/10.1002/sam.10103

16. Calle ML, Urrea V, Boulesteix AL, et al. (2011) Auc-rf: a new strategy for genomic profiling with random forest. *Hum Hered* 72: 121–132. https://doi.org/10.1159/000330778

17. Steyerberg EW, Vickers AJ, Cook NR, et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21: 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

18. Diaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7: 3. https://doi.org/10.1186/1471-2105-7-3

19. Dietrich S, Floegel A, Troll M, et al. (2016) Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol* 45: 1406–1420. https://doi.org/10.1093/ije/dyw145

20. Ye L, Li F, Song Y, et al. (2018) Overexpression of CDCA7 predicts poor prognosis and induces EZH2-mediated progression of triple-negative breast cancer. *Int J Cancer* 143: 2602–2613. https://doi.org/10.1002/ijc.31766

21. Chen L, Yang L, Qiao F, et al. (2015) High levels of nucleolar spindle-associated protein and reduced levels of BRCA1 expression predict poor prognosis in triple-negative breast cancer. *PLoS One* 10: e0140572. https://doi.org/10.1371/journal.pone.0140572

22. 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. https://doi.org/10.1001/jama.2017.7112

23. Sušac I, Ozretić P, Gregorić M, et al. (2019) Polymorphisms in Survivin (BIRC5 Gene) are associated with age of onset in breast cancer patients. *J Oncol* 3483192. https://doi.org/10.1155/2019/3483192

24. Cai YC, Yang H, Wang KF, et al. (2020) ANGPTL4 overexpression inhibits tumor cell adhesion and migration and predicts favorable prognosis of triple-negative breast cancer. *BMC Cancer* 20: 878. https://doi.org/10.1186/s12885-020-07343-w

25. Wang J, Xu B (2019) Targeted therapeutic options and future perspectives for HER2-positive breast cancer. *Signal Transduct Tar* 4: 34. https://doi.org/10.1038/s41392-019-0069-2

26. Wilson BJ, Giguère V (2008) Meta-analysis of human cancer microarrays reveals GATA3 is integral to the estrogen receptor alpha pathway. *Mol Cancer* 7: 49. https://doi.org/10.1186/1476-4598-7-49

27. Jiang J, Wang J, He X, et al. (2019) High expression of SPAG5 sustains the malignant growth and invasion of breast cancer cells through the activation of Wnt/β-catenin signalling. *Clin Exp Pharmacol Physiol* 46: 597–606. https://doi.org/10.1111/1440-1681.13082

28. Weng TY, Wang CY, Hung YH, 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. https://doi.org/10.1371/journal.pone.0161007.

29. 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. https://doi.org/10.1186/s12935-018-0633-9

30. 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 Epidem Biomar* 23: 658–669. https://doi.org/10.1158/1055-9965.EPI-13-0340

31. 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. https://doi.org/10.1016/j.ymgme.2010.04.006

32. Smerc A, Sodja E, Legisa M (2011) Posttranslational modification of 6-phosphofructo-1-kinase as an important feature of cancer metabolism. *PloS One* 6: e19645. https://doi.org/10.1371/journal.pone.0019645

33. Danilova N, Kumagai A, Lin J (2010) p53 upregulation is a frequent response to deficiency of cellessential genes. *PloS One* 5: e15938. https://doi.org/10.1371/journal.pone.0015938

34. 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. https://doi.org/10.1158/1078-0432.CCR-17-3490

35. Wu J, Hicks C (2021) Breast cancer type classification using machine learning. *J Pers Med* 11: 61. https://doi.org/10.3390/jpm11020061

36. Lu Y, Han J (2003) Cancer classification using gene expression data. *Inf Syst* 28: 243–268. https://doi.org/10.1016/S0306-4379(02)00072-8

37. Guyon I, Weston J, Barnhill S, et al. (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46: 389–422. https://doi.org/10.1023/A:1012487302797

38. Hernandez Hernandez JC, Duval B, Hao JK (2007) A genetic embedded approach for gene selection and classification of microarray data, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 5th European Conference*, Springer Berlin Heidelberg, 90–101.

39. Dai B, Chen RC, Zhu SZ, et al. (2018) Using random forest algorithm for breast cancer diagnosis, 2018 International Symposium on Computer, Consumer and Control (IS3C). IEEE, 449–452. https://doi.org/10.1109/IS3C.2018.00119

40. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning, Addison-Wesley, 36.

AIMS Press