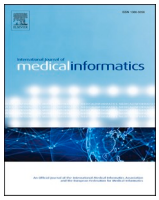
International Journal of Medical Informatics 190 (2024) 105548

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Evaluation of risk factors and survival rates of patients with early-stage breast cancer with machine learning and traditional methods

Emrah Gökay Özgür [a,*], Ayse Ulgen [b], Sinan Uzun [c], Gülnaz Nural Bekiroğlu [a]

[a] Marmara University, School of Medicine, Department of Biostatistics, Turkiye
[b] Department of Mathematics and Physics. School of Science and Technology. Nottingham Trent University. United Kingdom. Girne American University, Faculty of Medicine, Department of Biostatistics, Cyprus
[c] Marmara University, Institute of Health Sciences, Department of Biostatistics, Turkiye

ARTICLE INFO

Keywords:
Breast cancer
Machine learning
Early stage
Risk factor

ABSTRACT

Background: This article is aimed to make predictions in terms of prognostic factors and compare prediction methods by using Cox proportional hazards regression analysis (CPH), some machine learning techniques and Accelerated Failure Time (AFT) model for post-treatment survival probabilities according to clinical presentations and pathological information of early-stage breast cancer patients.
Material and methods: The study was carried out in three stages. In the first stage, the CPH method was applied. In the second stage, the AFT model and in the last stage, machine learning methods were applied. The data set consists of 697 breast cancer patients who applied to Marmara University Hospital oncology clinic between 01.01.1994 and 31.12.2009. The models obtained by using various parameters of the patients were compared according to the C index, 5-year survival rate and 10-year survival rate.
Results and conclusion: According to the models obtained as a result of the analyses applied, MetLN and age were obtained as a significant risk factor as a result of CPH method and AFT methods, while MetLN, age, tumor size, LV1 and extracapsular involvement were obtained as risk factors in machine learning methods. In addition, when the c-index values of the handheld models are examined, it is obtained as 69.8 for the CPH model, 70.36 for the AFT model, 72.1 for the random survival forest and 72.8 for the gradient boosting machine. In conclusion, the study highlights the potential of comparing conventional statistical methods and machine-learning algorithms to improve the precision of risk factor determination in early-stage breast cancer prognosis. Additionally, efforts should be made to enhance the interpretability of machine-learning models, ensuring that the results obtained can be effectively communicated and utilized by clinical practitioners. This would enable more informed decision-making and personalized care in the treatment and follow-up processes for early-stage breast cancer patients.

1. Background

Breast cancer is a type of cancer that affects the cells in the breast. It is the most common type of cancer in women, but it can also occur rarely in men. [1]. Breast cancer constitutes 12.5% of new annual cancer cases worldwide [2]. One in 8 or 10 women in the world will be diagnosed with breast cancer in their lifetime The death rate from breast cancer among these women is 2.5%. [3]. Worldwide, almost 2.3 million women were diagnosed with breast cancer and 685,000 deaths globally in 2020; [4]. Breast cancer is considered to be mainly influenced by lifestyle risk factors. However, genetic and heritability studies have shown that there is also a genetic component that is considered significant [5,6]. Early-stage breast cancer refers to cancer that is in the early stages of development and has not yet spread beyond the breast. Early diagnosis of the disease and an effective treatment method reduce the mortality rate [7]. Several early-stage breast cancer types include ductal carcinoma in situ (DCIS) and invasive breast cancer [8]. DCIS is a non-invasive type of breast cancer that is confined to the milk ducts and has not spread to the surrounding breast tissue [9] Invasive breast cancer has spread beyond the milk ducts and into the surrounding breast tissue [10]. (Both types of early-stage breast cancer can be treated with surgery, radiation, and/or chemotherapy [11].

With the development of technology, it becomes easier to determine the risk factors of diseases. We can obtain risk factors of diseases with

* Corresponding author at: Marmara University, School of Medicine, Department of Biostatistics, Başıbüyük mah, Başıbüyük yolu sok, Maltepe, İstanbul, Turkiye.
E-mail addresses: emrahgokayozgur@gmail.com, emrah.ozgur@marmara.edu.tr (E.G. Özgür).

https://doi.org/10.1016/j.ijmedinf.2024.105548
Received 29 May 2024; Received in revised form 4 July 2024; Accepted 9 July 2024
Available online 11 July 2024

machine learning methods. Machine learning is a type of artificial intelligence that involves the use of algorithms to analyse and learn from data [12]. It has the potential to be used in various aspects of breast cancer diagnosis and treatment, including early-stage breast cancer. One potential application of machine learning in early-stage breast cancer is in the development of predictive models. These models use data from past patients to identify patterns and predict the likelihood of certain outcomes, such as the likelihood of cancer returning after treatment [13]. This can help doctors tailor treatment plans to each individual patient and improve outcomes. Machine learning algorithms can also be used to analyse medical images, such as mammograms, to identify abnormalities and potentially cancerous lesions [14–16]. This can improve the accuracy and speed of diagnosis, particularly in cases where it is difficult for a radiologist to identify a lesion on an image.

This article is aimed to make predictions in terms of prognostic factors and compare prediction methods by using Cox proportional hazards (CPH) regression analysis, some machine learning techniques and Accelerated Failure Time model (AFT) for post-treatment survival probabilities according to clinical presentations and pathological information of early-stage breast cancer patients.

2. Material and methods

Cox regression, also known as the Cox proportional hazards model, is used to analyse the relationship between survival time and predictor variables, commonly in medical research. It models the hazard function, estimating the likelihood of an event occurring. The AFT model estimates the effect of covariates on event time. Unlike the Cox proportional hazards model, which models the hazard rate as a function of time and covariates, the AFT model models the logarithm of the time to event as a linear function of the covariates. It models the logarithm of time as a linear function of covariates, allowing comparison between different groups or conditions. AFT models use distributions like Weibull, Log-normal, Gompertz, and so on. Machine learning is an AI technique where algorithms learn from data to make predictions or decisions without explicit programming. It's used in various fields, solving complex problems efficiently. Random forest is an ensemble algorithm for classification and regression. It combines predictions from multiple decision trees. Random forests are accurate, robust, resistant to overfitting, and widely used. Random survival forests (RSF) extend random forests to analyze survival data [17–19]. RSF predicts event risk and identifies important predictors [20]. Gradient boosting is another ensemble algorithm for classification and regression [21]. It combines weak learners to create a strong learner, focusing on areas where previous learners performed poorly. Gradient boosting is a powerful and effective machine-learning algorithm used in various fields.

In summary, both traditional methods and machine learning methods are frequently used in survival data. These methods have differences from each other. The assumption of proportionality is important in Cox regression analysis. When the proportionality assumption is not met, the reliability of Cox regression results decreases. When the proportionality assumption is not met in survival data, rather than other parametric methods, especially The AFT model is used as an alternative to Cox regression. The AFT model is a survival analysis model that relates an individual's survival time to covariates through a log-linear relationship. However, if the distribution used in the AFT model does not fit the data, biased results may occur. Unlike traditional methods, assumptions such as linearity and proportionality are not taken into account in machine learning methods. Therefore, machine learning methods can be easily applied to survival data. The success of machine learning methods in large and complex data sets makes their use important.

In the study, patients with a diagnosis of breast cancer, who were over 18 years old, without distant organ metastases, and who applied to the Marmara University Medical Oncology outpatient clinic between 01.01.1994 and 31.12.2009 were included and the demographic information and pathology data of the patients were retrospectively scanned. 697 patients with complete required data were included in the study.

The variables used in our study; age at diagnosis, menopausal status, histology status, tumour size in the right or left breast, tumour diameter, number of metastatic lymph nodes, oestrogen, and progesterone receptor status, multifocal status, Her-2 status, lymphatic invasion and vascular invasion status, neural invasion status, extra capsulation status, tumour grade, follow-up period, and whether the patient survived during this period.

This study comprises three stages. In the first stage, the conventional Kaplan-Meier method was used to calculate survival probability according to the follow-up time of early-stage breast cancer patients with the variables determined. In order to determine the significant independent variables that affect survival; univariate statistical analyses were conducted applying the Log-rank test, followed by multivariate analyses using CPH model.

In the second stage, when the proportionality assumption was not provided in the CPH model, the preferred parametric AFT model was applied to the data set and results were obtained.

In the third phase of this study, survival probabilities were calculated for the same patients with the RSF model and GBM models, which are machine-learning algorithms. In machine-learning algorithm applications, datasets are divided into two parts. In the implementation of machine learning algorithms, training, and test datasets are created from the patients involved in the study to train the algorithms and measure their performance. In machine-learning use, the dataset is usually split into as 90–10, 80–20, and 70–30 ratios for training and testing, then analyses were conducted. The robust results were obtained with an 80–20 split for this study, which was therefore used. Additionally, hyperparameter optimization was performed and the values that gave the most appropriate results were used. All risk (hazard) factors included in our study and affecting survival; survival models were performed using RSF and GBM algorithms. Then, in line with the algorithm flow, these hazard factors are ranked according to the degree of importance listed.

In the study the predictions performances of the models were evaluated using 5 and 10-year survival rates and C-index fit statistics for the machine-learning generated survival model algorithms (RSF, GBM) and AFT method, CPH method.

3. Results

In the first stage of the study, the 5- and 10-year survival probabilities of early-stage breast cancer patients were calculated using the Kaplan-Meier method. In determining the significant risk factors that affect survival; Log-Rank test was used to detect significant risk factors in univariate analyses and CPH regression method was used in multiple analyses. In the second stage of the study, Since the most appropriate hazard function distribution of overall survival was found to be in accordance with the log-logistic distribution, the log-logistic AFT regression model was applied to the same data set. In the last stage; 5- and 10-year survival probabilities were calculated by applying machine-learning algorithms RSF and GBM algorithms to the same data set, and the risk factors affecting survival were listed according to their importance. In addition, significant risk factors of the models were determined and the model with the most appropriate and correct prediction was tried to be determined from a biostatistical and clinical point of view.

Categorical variables used to investigate the significance of risk factors; menopausal status, histology status, tumour in the right or left breast, grade (grade of tumour), multifocality, lymph invasion or visin invasion (LV1) status, neuron invasion status, extracapsular extension status, oestrogen, and progesterone receptor status, and Her-2. The log-rank test was applied to test the significance of these variables to survival. As seen in Table 1, Menopause status($p = 0.011$), tumour grade($p = 0.033$), multifocality ($p = 0.031$), lymph and blood vessel invasion

Table 1
Results of Logrank Test of Categorical Variable.

Log Rank(Mantel −Cox)	Chi-Square	df	p
Menopause status	6.4	1	**0.011**
Histology Status	9.6	10	0.483
Right Left Breast Condition	0.7	1	0.411
Tumor Grade	6.8	2	**0.033**
Multifocality	4.7	1	**0.031**
Lymph and Blood Vessel Invasion Status	14.1	2	**0.008**
Neuroinvasion	11.5	2	**0.003**
Extra Capsular Involvement Status	29.3	2	**<0.001**
Estrogen and Progesterone Receptor Status	2.7	5	0.744
HER2 Status	4.5	4	0.367

status ($p = 0.008$), neuro-invasion ($p = 0.003$) and extracapsular involvement status ($p < 0.001$) were found to be statistically significant according to Log-Rank test results. Univariate CPH regression analysis was used to test the significance of age, tumour diameter, and the number of metastatic lymph nodes on survival for continuous variables. As shown in Table 2, age ($p < 0.001$), tumour size ($p = 0.002$) and the number of metastatic lymph nodes ($p < 0.001$) were found to be statistically significant. Although there are variables such as extracapsular, tumour diameter, and neuro-invasion that violated the proportional hazard (PH) assumption, the results of the CPH analysis with significant variables are shown in Table 3.

Since the PH assumption is forced in the CPH model, relying on the results of the model in question will decrease, and the necessity to search for an alternative method has arisen. Therefore, in the second stage, the AFT model, which interprets the risk factors as time ratio, was run with Log-logistics, the optimal hazard function distribution for overall survival applying among other hazard distributions such as log-normal, Gompertz, Weibull. According to the AFT analysis results given in Table 4, age and MetLN (Metastatic Lymph node) variables were found to be statistically significant in the multiple models.

In the last stage, survival models were evaluated by using machine-learning algorithms. During model evaluation, 557 (80 %) of 697 patients were determined as training data and 140 (20 %) as test data. The performance of the models was compared with the C-index fit statistic. With the evaluated models, 5- and 10-year survival mean and medians were calculated over the test data, and the risk factors were listed according to the degree of importance.

The 5 most important variables in the RSF method are METLN, age, tumour size, LV1 extracapsular involvement, in the GBM, age, METLN, extracapsular involvement, tumour size, LV1.

4. Discussion

The primary goal of this study was to identify significant risk factors affecting the prognosis of early-stage breast cancer patients by using both conventional statistical methods, including the CPH model, AFT model, and machine-learning algorithms, and to compare their performance of the applied models from both biostatistical and clinical point of view.

Conventional statistical methods, such as Kaplan-Meier, Log-Rank test, CPH regression model, were applied to identify significant risk factors and predict survival probabilities. While these methods have been widely used in medical research, the limitations of methods, such as the proportional hazard assumption in the CPH model, led to finding alternative methods. One of these alternative methods can be counted as the AFT model. The AFT model, which interprets risk factors as time ratios, was found to be an effective alternative to the CPH model, providing valuable insights into the impact of risk factors on survival time.

In this context, machine-learning algorithms, including RSF and GBM, were used to perform survival models. These algorithms offer several advantages, such as the ability to handle complex interactions between variables, the capacity to deal with non-linear relationships, and the capability to rank risk factors according to their importance in predicting survival probabilities.

The results of the study indicate that conventional methods determined only two important risk factors such as MetLN and age while machine-learning algorithms were determining several important risk factors such as age, tumour size, number of metastatic lymph nodes, and extracapsular involvement. In addition, machine-learning algorithms demonstrated their potential by providing additional insights into the relative importance of these risk factors in determining the prognosis of patients with early-stage breast cancer. In the RSF model, MetLN, age, tumour size, LV, and extracapsular involvement were obtained in the order of importance, while in the GBM model, Age, MetLN, extracapsular involvement, tumour size, and LV were obtained in the order of importance.

When it comes to the c-index values, that show the concordance between two variables, the highest c-index value is valuable therefore it is obtained from the GBM model, then RSF, AFT and CPH models, respectively as shown in Table 5 and Fig. 1. Considering the 5 and 10-year survival probabilities, the highest survival probability in 5-year survival rates were obtained from the AFT model with 94.30 %. Survival probability was 93.40 % in the CPH model, 93.00 % in the GBM model, and 92.90 % in the RSF model. The same ranking was still continued in 10-year survival probability. While the highest 10-year survival probability was obtained from the AFT model with 82.10 %, it was obtained as 78.80 % from the CPH model, 78.60 % from the GBM model, and 78.50 % from the RSF model.

Table 2
Univariate COH method results for continuous variables.

Cox Regression (COH)	B	SE	Wald	df	p	Exp (B)
Age	0.028	0.006	19.27	1	**<0.001**	1.293
Tumor Size	0.136	0.040	9.48	1	**0.002**	1.146
Number of Metastatic Lymph Node	0.092	0.011	48.11	1	**<0.001**	1.097

Table 3
Results of Multiple COH regression model.

	B	SE	z	p	Exp (B)
Age	0.029	0.008	3463	**<0.001**	1.030
Menopause status	−0.076	0.219	−0.349	0.726	0.926
Tumor Size	0.057	0.047	1218	0.223	1.059
Number of Metastatic Lymph Node	0.072	0.013	5.55	**<0.001**	1.075
Tumor Grade	0.066	0.135	0.488	0.625	1.068
Multifocality	0.162	0.186	0.867	0.385	1.175
Lymph and Blood Vessel Invasion Status	0.254	0.160	1585	0.112	1.290
Extra Capsular Involvement Status	0.097	0.119	0.815	0.415	1.102
Neuroinvasion	−0.017	0.103	−0.173	0.862	0.982

Table 4
Results of Multiple AFT regression model.

	B	SE	z	p
Age	−0.018	0.005	−3.17	**0.001**
Menopause status	0.070	0.151	0.46	0.642
Tumor Size	−0.058	0.035	−1.65	0.098
Number of Metastatic Lymph Node	−0.056	0.011	−5.08	**<0.001**
Tumor Grade	−0.077	0.052	−0.81	0.418
Multifocality	−0.181	0.139	−1.30	0.194
Lymph and Blood Vessel Invasion Status	−0.174	0.112	−1.56	0.118
Extra Capsular Involvement Status	−0.055	0.079	−0.71	0.480
Neuroinvasion	0.008	0.068	0.13	0.897

Table 5
Results of C- index and survival probabilities of COH Model, Machine learning algorithms and AFT model.

Model	Importance Risk Factor	C-index (se)(%)	5 year Survival Rate	10 year Survival Rate
CPH	MetLN Age	69.8(2.3)	93.40 %	78.80 %
RSF	MetLN Age Tumor Size LV1 Extracapsular Involvement	72.1(0.6)	92.90 %	78.50 %
GBM	Age MetLN Extracapsular Involvement Tumor Size LV1	72.8(0.4)	93.00 %	78.60 %
AFT	MetLN Age	70.36 (2.2)	94.30 %	82.10 %

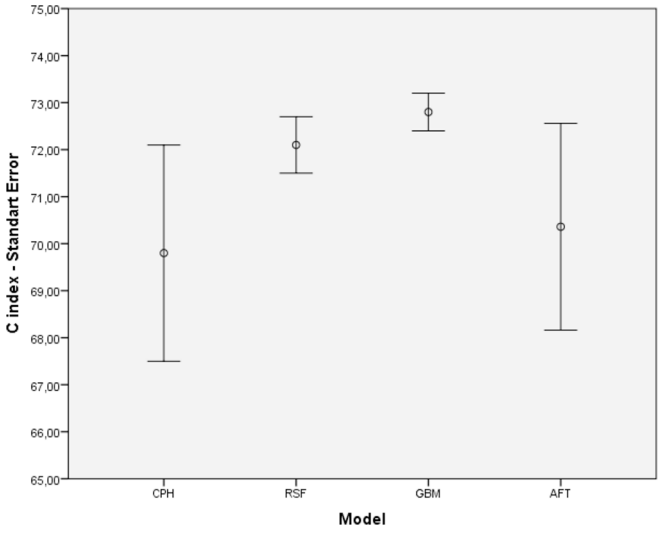


Fig. 1. C index and Standart Error for Models.

Although RSF and GBM models have provided promising results, it is important to be aware of certain limitations of these models. For instance, the performance of machine-learning algorithms can be affected by the quality of the input data and the selection of appropriate hyperparameters. Moreover, the interpretability of the results from machine-learning models can be challenging, especially for clinical practitioners who may not be familiar with these techniques.

Studies using machine-learning methods on early-stage breast cancer are available in the literature. Some of these studies estimate the survival probabilities of patients, while others classify patients. Nicolo et al. used machine-learning methods to determine the risk factors affecting early-stage breast cancer in their study in 2023. According to their study result, they found the prediction performances of machine learning methods and conventional methods to be similar. Orozco et al., in their study in 2022, concluded that the models obtained from machine-learning methods have higher discrimination. In the study of Xiong and et al. in 2022, they worked on a prediction model for breast cancer. In that study, they used logistic regression and three machine learning methods. As a result, they stated that the best prediction model was obtained using the GBM method as our study. In the study of Ganggayah et al. in 2019, the factors affecting the survival of breast cancer patients were investigated. They used logistic regression and six machine learning methods in the study. In general, although the results obtained from these methods are similar to each other, they concluded that the model obtained by the random forest method is the best model.

In conclusion, the study highlights the potential of comparing conventional statistical methods and machine-learning algorithms to improve the precision of risk factor determination in early-stage breast cancer prognosis. The results suggest that age, tumour size, number of metastatic lymph nodes, and extracapsular involvement are important factors in predicting survival probabilities. Further research is needed to validate these findings and explore the potential of other machine-learning algorithms in predicting the prognosis of breast cancer patients. The results may differ due to the different ways in which machine learning algorithms and traditional methods process and analyse data. However, if the data is robust and of high quality, utilizing diverse analytical methods does not constitute a risk but rather an opportunity to identify intriguing and unexpected relationships. Additionally, efforts should be made to enhance the interpretability of machine-learning models, ensuring that the results can be effectively communicated and utilized by clinical practitioners. This would enable more informed decision-making and personalized care in the treatment and follow-up processes for early-stage breast cancer patients.

CRediT authorship contribution statement

Emrah Gökay Özgür: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Ayşe Ülgen:** Writing – original draft, Methodology, Investigation. **Sinan Uzun:** Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Nural Bekiroğlu:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

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

References

[1] American cancer society: cancer facts figure 209, Nicola C, https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2019.html.
[2] Breast cancer Facts and Statistics, Jan, 03. breastcancer.org/facts-statistics.
[3] H. Nadia, G. Micael, Breast cancer, Lancet 389 (10074) (2017) 1134–1150, https://doi.org/10.1016/S0140-6736(16)31891-8.
[4] World Health Organisation. Breast Cancer, March, 2021. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
[5] K. McPherson, C.M. Steel, J.M. Dixon, Breast cancer-epidemiology, risk factors, and genetics, BMJ 321 (7261) (2000) 624–628.
[6] S. Möller, L.A. Mucci, J.R. Harris, et al., The heritability of breast cancer among women in the Nordic twin study of cancer, Cancer Epidemiol. Biomarkers Prev. 25 (1) (2016) 145–150.
[7] American cancer society. cancer.org.
[8] M. Aslaug, R. Anders, W. Fredrik, et al., Frequent aberrant DNA methylation of ABCB1, FOXC1, PPP2R2B and PTEN in ductal carcinoma in situ and early invasive breast cancer, Breast Cancer Res. 12 (1) (2010) R3, 10.116/bcr2466.
[9] F. Lesley, M. Lucy, F. Adele, et al., Low-grade Ductal Carcinoma in situ (DCIS): how best to describe it? Breast 23 (5) (2014 Oct) 693–696, https://doi.org/10.1016/j.breast.2014.06.013.
[10] N.S. Ganesh, D. Rahul, S. Jyotsana, et al., Various types and management of breast cancer: an overview, J. Adv. Pharm. Technol. Res. 1 (2) (2010 Apr) 109–126.
[11] A.D. Kristine, B.J. Paul, A. Michael, et al., Course of fatigue in women receiving chemotherapy and/or radiotherapy for early-stage breast cancer, J. Pain Symptom Manage. 28 (4) (2004 Oct) 373–380, https://doi.org/10.1016/j.jpainsymman.2004.01.012.
[12] A. Arushi, S. Purushottam, A. Mohammed, et al., Classification model for accuracy and intrusion detection using machine learning approach, PeerJ Comput. Sci. 7 (7) (2021 Apr) e437, https://doi.org/10.7717/peerj-cs.437.eCollection2021.
[13] D. Dursun, O. Asil, J.K. Zhenyu, A machine learning-based approach to prognostic analysis of thoracic transplantations, Artif. Intell. Med. 49 (1) (2010 May) 33–42, https://doi.org/10.1016/j.artmed.2010.01.002.
[14] E. Yusuf, H.F. Yuk, L.N. Jing, et al., Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms, Sci. Rep. 12 (1) (2022 Mar 10) 3883, https://doi.org/10.1038/s41598-022-07693-4.

[15] B. Saskya, R.M. Lydia, M. Alejandro, et al., Machine Learning and Artificial Intelligence for Surgical Decision Making, Surg. Infect. (Larchmt) 22 (6) (2021 Aug) 626–634, https://doi.org/10.1089/sur.2021.007.
[16] R. Nithya, B. Santhi, R. Masoumeh, Computer Vision System for Mango Fruit Defect Detection Using Deep Convolutional Neural Network, Foods 11 (21) (2022 Nov 2) 3483, https://doi.org/10.3390/foods11213483.
[17] D. Arwinder, S. Ashima, eBreCaP: extreme learning-based model for breast cancer survival prediction, IET Syst. Biol. 14 (3) (2020 Jun) 160–169, https://doi.org/10.1049/iet-syb.2019.0087.
[18] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[19] I. Hemant, B.K. Udaya, H.B. Eugene, et al., Random Survival Forest, Ann. Appl. Stat. 2 (3) (September 2008) 841–860, https://doi.org/10.1214/08-AOAS169.
[20] A.V. Bharath, Y. Xiaoying, O.W. Colin, et al., Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis, Circ. Res. 121 (9) (2017 Oct 13) 1092–1101, https://doi.org/10.1161/CIRCRESAHA.117.311312.
[21] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobot. 7 (2013) 21, https://doi.org/10.3389/fnbot.2013.00021.eCollection2013.