Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2021;13(3):236-51

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2021-81105

Performance Comparison of Support Vector Machines, Random Forest and Artificial Neural Networks in Binary Classification: Descriptive Comparison Study

İkili Sınıflandırmada Destek Vektör Makineleri, Rastgele Orman ve Yapay Sinir Ağlarının Performans Karşılaştırması: Tanımlayıcı Kıyaslama Çalışması

Emre DİRİCAN[a], Zeki AKKUŞ[b]

[a]Department of Biostatistics, Hatay Mustafa Kemal University Faculty of Medicine, Hatay, TURKEY
[b]Department of Biostatistics, Dicle University Faculty of Medicine, Diyarbakır, TURKEY

ABSTRACT Objective: In this study, it was aimed to find the method with high classification success among the methods used in the study by comparing the supervised machine learning methods according to the classification performance. **Material and Methods:** In our study, both the real data set obtained from 302 patients with invasive ductal carcinoma and 24 different data sets obtained by simulation were used to compare the classification performance of support vector machines, random forest and artificial neural networks. The success of classifications of the methods used was compared according to the general accuracy, F-measure, Matthews correlation coefficient, area under the curve (AUC) and discriminant power in breast cancer data. In addition, the difference in training-test accuracy in the simulation data and the significance of this difference were also evaluated. **Results:** The highest survival classification accuracy (80%) for the test set of stage III patients with invasive ductal carcinoma was obtained from support vector machines (SVM) with the radial-based kernel. The highest values in other performance metrics (F-measure=0.87, Matthews correlation coefficient=0.22, AUC=0.89 and discriminant power=0.52), and the most successful results in simulation data were generally obtained from SVM. **Conclusion:** SVM had higher accuracy in both the real data set and simulation data than random forest and artificial neural networks.

Keywords: Machine learning; classification; breast cancer; support vector machines; random forest

ÖZET Amaç: Bu çalışmada, danışmanlı makine öğrenimi yöntemleri sınıflama performansına göre kıyaslanarak, çalışmada kullanılan yöntemlerin içerisinden sınıflama başarısı yüksek olan yöntemin bulunması amaçlandı. **Gereç ve Yöntemler:** Çalışmamızda, destek vektör makineleri, rastgele orman ve yapay sinir ağları yöntemlerinin sınıflama performanslarını kıyaslamak için hem invaziv duktal karsinomlu 302 hastadan elde edilen gerçek veri seti hem de simülasyonla elde edilen 24 farklı veri seti kullanıldı. Kullanılan yöntemlerin sınıflama başarıları meme kanseri verilerinde genel doğruluk, F-ölçütü, Matthews korelasyon katsayısı, eğri altında kalan alan [area under the curve (AUC)] ve ayırsama gücüne göre kıyaslandı. Ayrıca simülasyon verilerinde eğitim-test doğrulukları farkı ve bu farkın anlamlılığı da değerlendirildi. **Bulgular:** İnvaziv duktal karsinomlu evre III hastalarının test seti için en yüksek sağkalım sınıflama doğruluğu (%80), radyal tabanlı çekirdek ile destek vektör makinelerinden [support vector machines (SVM)] elde edildi. Diğer performans ölçütlerindeki (F-ölçütü=0,87; Matthews korelasyon katsayısı=0,22; AUC=0,89 ve ayırsama gücü=0,52) en yüksek değerler ve simülasyon verilerinde en başarılı sonuçlar, genel olarak SVM'den elde edilmiştir. **Sonuç:** SVM, hem gerçek veri setinde hem de simülasyon verilerinde, rastgele orman ve yapay sinir ağlarına göre daha yüksek doğruluk oranına sahiptir.

Anahtar kelimeler: Makine öğrenimi; sınıflama; meme kanseri; destek vektör makineleri; rastgele orman

Machine learning (ML) is one of the data mining technologies. It is a scientific discipline that focuses on how computers learn from data for various computational methods such as classification and clustering.[1]

Statistical learning theory is the main theorem underlying ML. In this theory, the model form is unknown, and it is tried to find the model that gives the optimum result among the models considered to be

Correspondence: Emre DİRİCAN
Department of Biostatistics, Hatay Mustafa Kemal University Faculty of Medicine, Hatay, TURKEY/TÜRKİYE
E-mail: emredir44@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 08 Jan 2021 ***Received in revised form:*** 10 Jun 2021 ***Accepted:*** 06 Jul 2021 ***Available online:*** 09 Sep 2021

correct.[2] Performance metrics are needed both to evaluate the performance of the determined model and to examine its validity. Although the correct classification percentage is generally considered among the aforementioned metrics, many criteria such as sensitivity, specificity, F-measure (F), Matthews correlation coefficient (MCC), discriminant power (DP) etc. can be used in binary classifiers.[3] In order to the determined ML method to be considered as the highest performance method, it should be compared with other ML techniques with its success in performance metrics. However, training and then optimizing a technique in the supervised learning are processes that require experience and time.[4] Therefore, determining the model with high performance in classification for many types of data sets is useful for the researcher to pass time-consuming processes more quickly.

The data sets used to predict the performance of ML methods can be of many types. The first of these is the data set consisting of real-world data, and in our study, data of breast cancer patients were used for this type of data set. In breast cancer studies, many histopathological subtypes [in situ lobular carcinoma, invasive ductal carcinoma (IDC), invasive lobular carcinoma etc.] are generally evaluated together. In our study, however, only patients with histopathological subtype IDC stage III were evaluated. Another type of data set is obtained by simulation. Simulations are used as the third mode of science that complementing theory and experiments. Simulations are very useful in situations that are too complex to be analyzed analytically or too expensive to study experimentally and also impractical.[5] In our research, it has been shown that the ML method, which has a high performance in classifying the data obtained by simulation, gives similar results for real data. In addition, after obtaining a successful survival classification result in IDC stage III breast cancer patients, it was studied to determine the important variables. In this study, we aimed to evaluate the performance of three ML methods in classifying many independent variables to a binary dependent variable with both real data set and simulation data.

MATERIAL AND METHODS

Our study is retrospective descriptive for the department where the data of breast cancer patients is used, and descriptive for the simulation data. We mentioned that there were two different approaches to performance evaluations. In the first approach, the real data set was obtained from Inonu University Turgut Özal Medical Center Medical Pathology Department. Data on 302 patients with IDC were obtained from 1,470 breast cancer cases by examining each of 75,000 biopsy reports between 2008 and 2016. In the study, survivability (alive-exitus) classification was made using a total of 15 independent variables together with prognostic factors for breast cancer [*tumour size, tumour necrosis, hormone receptors estrogen (ER) & progesterone (PR), Her-2 (CerbB-2), number of lymph nodes, number of metastatic lymph nodes, lymphovascular invasion (LVI), ki-67 score*].

Tumor size is an independent prognostic factor in the tumor-node-metastasis (TNM) staging system and, it shows a good correlation with nodal metastasis incidence and survival.[6] Necrosis is a form of cell death caused by factors external to the cell, such as hypoxia, and is often associated with rapidly growing, aggressive forms of cancer in the breast, colon, etc. Identification of tumor necrosis in breast cancer can provide prognostic knowledge indicating death.[7] Ki-67 score is the pith protein expressed at G1, S, G2 and M phases of tumor cells and a solid tumor proliferation marker being related with the prognosis of breast carcinoma and its response to neoadjuvant chemotherapy.[8]

Chemotherapy response and progression of molecular subtypes differ. In Luminal A; ER and/or PR is positive, HER-2 is negative, and the proliferation index is low. In Luminal B; tumors are high grade, and may be PR+ or PR-, or HER2+ or HER2-. Typically, triple negative breast carcinoma (TNBC) is the kind lacking ER and PR with overexpression of HER2. Compared to other subtypes, TNBC tumors are usually larger and they are related with 2.5-fold more metastasis within five years after diagnosis.[9,10] LVI is associated with increased lymph node metastasis and the risk of progression to systemic disease. It is a poor prognostic factor for relapse and survival in node-negative patients.[11,12]

Age is a prognostic factor in IDC and varies by geographical region or demographics. In regions with young populations such Africa, and Turkey, breast carcinomas are more frequent under the age of 40, and these tumors are found at further stages compared to the Western societies.[13] The presence of axillary lymph node is one of the most important factors in prognosis estimation for the patients. Metastatic axillary lymph node ratio is known as an important prognostic factor in breast cancer.[14] Obesity is related with increased invasive breast cancer risk in postmenopausal women. Especially, a body mass index of 35.0 or higher was strongly associated with risk for ER receptor-positive and PR receptor-positive breast cancers (hazard ratio, 1.86; 95% confidence interval, 1.60-2.17).[15]

For the real data set; individuals with histopathological type of IDC, stage 3 according to TNM staging (any of 3A, 3B and 3C), followed for at least 30 months were included in the study. Patients whose histological type and stage could not be determined exactly and metastatic tumors were excluded from the study. In addition, patients who applied for confirmation outside of the designated center were not included in the study. This study was approved by Dicle University Medical Faculty Ethics Committee for Non-interventional Studies (date: 23.6.2017, number: 2017/137).

In the second approach, the simulation study, analyzes were performed with different data sets by changing the incidence of the risky situation in the dependent variable, the number of observations and the number of independent variables. In these data sets, the data obtained from the binomial distribution according to the observation rates of 0.2-0.4-0.6 and 0.8 for the dependent variable were included in the analysis. Independent variables were created from datasets derived from multivariate normal distribution with zero mean and one standard deviation. By combining the data sets created by using different numbers of independent variables (15, 25 and 35) with the dependent variable (labels: 0-1) obtained from the binomial distribution, 24 different data sets with 100 repetitions were obtained. The performance of ML methods was evaluated according to performance metrics calculated separately for each repetition.

R 3.3.3 (R programming languages/Project) and RStudio interface were used for analysis. Support vector machine (SVM), random forest (RF) and artificial neural network (ANN) models were used in the analysis of the data sets.[17] The optimal hyperparameters for these techniques were determined using the cross-validation (5 repeats of 10-fold) method and testing different hyperparameters. In all ML methods, 70% of the data set was used to train the model (training set) and 30% was used to test the learned model (test set).

Support vector machine: It aims to create a decision boundary between two classes that enables the prediction of labels from one or more feature vectors. This decision boundary, known as the hyperplane, is orientated in such a way that it is as far as possible from the closest data points from each of the classes. These closest points are called support vectors.[18] An alternative use for SVM is the kernel method, which can be calculated directly in the original low-dimensional space without having to represent the mapping of the sample data in the high-dimensional space, which can effectively avoid the complex computation in the high-dimensional space.[19]

Random forest: The purpose of the RF method is to get more efficient results with more than one decision-maker as in other ensemble methods. The variables selected for branch splitting at any fork in any tree is not selected from the full set of possible descriptors but from a randomly selected subset of predetermined size.[20] The predetermined number of variables is suggested as $\sqrt{p} = m_{try}$. Although there are studies that allow other values (p/2, 0.1p etc.) on this part, the results that give optimum performance by classifying with different values should be taken into account in determining the appropriate "m_{try}" number.[21]

Artificial neural networks: ANN designed according to the neural structure of the brain is very useful in predictive modelling due to its ability to model nonlinear relationships in a high dimensional data set.[22] There are two types of ANN, feed-forward and feed-back, depending on the connection design of the neurons. In feed-forward, the data moves in only one direction forward from the input layer to the output layer. In addition, multi-layer perceptrons (MLPs) are ANNs consisting of the input layer, the output layer and one

or more hidden layers between these two layers.[23] Feedforward-backpropagation MLP has been used in our study as in many ANN analysis.[24] Explanations on the sub-units of ML techniques are in Table 1.

TABLE 1: Explanations on machine learning techniques.

Machine learning methods	Units		Optimization technique
SVM	Kernels	Linear	**Cross-validation (5 repeats of 10-fold)** **70% training set, 30% test set**
		Polynomial	
		Sigmoid	
		Radial basis	
RF	mtry=3	DT=100	
	mtry=3	DT=500	
	mtry=4	DT=100	
	mtry=4	DT=500	
ANN	Activation Functions	Sigmoid	
		Hyperbolic Tangent	
		Relu	
		Linear	

SVM: Support vector machine; RF: Random forest; ANN: Artificial neural network; DT: Number of decision trees.

In our study, general accuracy (D), MCC, F, DP, and area under the curve (AUC) were used to both evaluate the performance of the classifiers and to examine the validity of the model. The confusion matrix from which model performance metrics are obtained is given in Table 2.

TABLE 2: Confusion matrix for the calculation of model performance metrics.

		Real (Actual) Classification		
		Disease (+)	**Non-disease (-)**	
Predicted Classification	Positive (+)	True Positive TP	False Positive FP	Positive Predictive Value $\frac{TP}{TP+FP}$
	Negative (-)	False Negative FN	True Negative TN	Negative Predictive Value $\frac{TN}{FN+TN}$
		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{TN+FP}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

MCC is the correlation coefficient between observed and predicted classification. The MCC value ranges from -1 to +1 with a value of +1 representing a perfect prediction, 0 as no better than random prediction and -1 the worst possible prediction.[25]

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The F expresses the balance between precision and sensitivity. F is 0 when either the precision or the sensitivity is 0.[26]

$$F - measure = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision}$$

The area under the ROC curve (AUC) is a univariate depiction of the ROC curve. It ranges from 0.5 to 1, with larger values representing higher ML performance. The AUC is equal to the probability that the decision value assigned to a randomly-drawn positive sample is larger than the value assigned to a randomly-drawn negative sample.[27]

DP evaluates how well the classifier distinguishes between positive and negative cases by summarizing sensitivity and specificity. The test is a poor discriminant DP<1 and limited DP<2 and fair DP<3, good in other cases.[28]

$$DP = \frac{\sqrt{3}}{\pi}(\log\left(\frac{sensitivity}{1 - sensitivity}\right) + \log\left(\frac{specificity}{1 - specificity}\right))$$

STATISTICAL ANALYSIS

In our study, categorical variables were expressed with frequency and percentage, and continuous variables with mean and standard deviation. Normality assumption was analyzed by Kolmogorov-Smirnov test. Chi-square tests and Student t-test were used for univariate analyzes in the relationship of independent variables with the primary output variable. SPSS v21 (Armonk, NY: IBM Corp.) program was used for these analyzes. After removing the noisy data in the real data set, imputation was performed for a small number (4%) of missing data. The data were prepared for analysis with the standardization process. ML methods (SVM, RF, ANN) were used to classify the survival with determined independent variables. R 3.3.3 (R programming languages/Project packages "e1071", "randomForest", "MASS", "caret", "caTools", "dplyr", "pROC", "mxnet", "ggplot2") and Rstudio interface were used for this analysis. The performance of the methods was determined according to the D, positive predictive value, MCC, F, DP, and AUC metrics. These metrics were also used when comparing ML techniques on artificial data.

RESULTS

In the first approach, the mean age (years) of 302 patients with breast carcinoma was calculated as 50.54±12.38 and the median survival time was 39 (interquartile range=28.0-64.0) months. It was determined that 222 (73.5%) of the patients were alive and 80 (26.5%) patients died due to breast cancer.

All independent variables (features) and primary output (target) variable that expressing survival status were shown in Table 3. The values given in this table indicate the frequency and significance of independent variables in survival status. A small number of male breast cancer cases were also included to preserve the originality of the data set. In addition, the ki-67 score was expressed as a percentage in the reports. However, since it was accepted that 14% value could be a critical value according to the surgical approach, ki-67 was categorized (≤%14 and >%14) based on this value.[28] The independent variables and univariate analysis results related to the real data set obtained from breast cancer patients were shown in Table 3. Although the relationship between the number of lymph nodes (p=0.549), chemotherapy (p=0.631), radiotherapy (p=0.488) and survival was found to be insignificant, they were included in the classification because it was thought to contribute to the performance of ML methods.

The highest survival classification accuracy (80%) in the test set for stage III patients with IDC was obtained from SVM with radial-basis kernel. For other performance metrics (F=0.87, MCC=0.22, AUC=0.89 and DP=0.52), the most successful results were obtained from SVM. Performance metrics other than classification accuracy were calculated for numerical comparison of ML model with other methods. Performance

metrics were observed to be high in the training set for the highest performance SVM model, but low in the test set. This is due to the fact that the number of observations used as a training set (70%) is higher than the number of observations used as a test set (30%). Especially in the SVM method, if all breast cancer data are used as test data, it is thought that the values to be obtained for all metrics will be high (Table 4). In addition, it has been revealed by the classification that the variables with order of significance (Figure 1) are effective in the survival of IDC stage III patients.

As a result of the analysis of 24 data sets obtained by simulation, training-test set accuracies, accuracy differences (ΔD) and its significance test (H0: ΔD is non-zero hypothesis, one sample t-test) positive predictor and F were given. The means of the above metrics calculated separately for 100 repetitions from the data sets were given in Table 5. However, the significance test data were the values of accuracy differences obtained from the results calculated from each repetition.

TABLE 3: Variables used in invasive ductal carcinoma data set.

		Survival		p value
		Alive	Ex	
		222 (73.5)	80 (26.5)	
Age		49.54±12.16	53.32±12.64	0.019**
Gender	Female	216 (75.3)	71 (24.7)	0.005****
	Male	6 (40)	9 (60)	
BMI	Normal weight	35 (67.3)	17 (32.7)	0.035*
	Overweight	100 (81.3)	23 (18.7)	
	Obesity	82 (67.8)	39 (32.2)	
Tumour size	T1	35 (85.4)	6 (14.6)	0.001*
	T2	123 (80.4)	30 (19.6)	
	T3	56 (63.6)	32 (36.4)	
	T4	8 (40)	12 (60)	
Tumour necrosis	No	154 (81.9)	34 (18.1)	0.001*
	Yes	68 (59.6)	46 (40.4)	
ER	Negative	54 (57.4)	40 (42.6)	0.001*
	Positive	168 (80.8)	40 (19.2)	
PR	Negative	68 (64.8)	37 (35.2)	0.012*
	Positive	154 (78.2)	43 (21.8)	
C-erb B2 (Her2)	Negative	111 (79.3)	29 (20.7)	0.034*
	Positive	111 (68.5)	51 (31.5)	
Number of lymph nodes		18.37 ± 7.61	17.77 ± 7.78	0.549**
Number of metastatic lymph nodes	0-3 LNm	67 (84.8)	12 (15.2)	0.009*
	4-9 LNm	94 (73.4)	34 (26.6)	
	10+ LNm	61 (64.2)	34 (35.8)	
LVI	LVI (-)	51 (87.9)	7 (12.1)	0.009***
	LVI (+)	171 (70.1)	73 (29.9)	
Ki-67 Proliferation	Ki-67 ≤14%	80 (80.8)	19 (19.2)	0.045*
	Ki-67 >14%	142 (70)	61 (30)	
Chemotherapy	No	14 (66.7)	7 (33.3)	0.631***
	Yes	208 (74)	73 (26)	
Radiotherapy	No	63 (70.8)	26 (29.2)	0.488*
	Yes	159 (74.6)	54 (25.4)	
Hormonotherapy	No	68 (57.1)	51 (42.9)	0.001*
	Yes	154 (84.2)	29 (15.8)	

*Pearson chi-square test; **Independent Student t-test; ***Yates continuity correction; ****Fisher exact chi-square test; BMI: Body mass index; ER: Estrogen; PR: Progesterone; LVI: Lymphovascular invasion.

TABLE 4: Model performance metrics for the real data set.

Methods		Optimum approaches		D[1]	D[2]	F[1]	F[2]	MCC[1]	MCC[2]	AUC[1]	AUC[2]	DP[1]	DP[2]
SVM	Radial Basis	c^1=4. γ^1=0.029		0.93		0.96		0.83		0.98		1.49	
			c^2=2. γ^2=0.034		0.8		0.87		0.22		0.89		0.52
RF	m_{try}=3.			0.82		0.88		0.48		0.83		0.65	
	ntree=100				0.77		0.86		0.14		0.7		0.45
ANN	Hyperbolic Tangent	S^1=6. De^1=0.4		0.75		0.85		0.23		0.82		0.39	
			S^2=4. De^2=0.4		0.77		0.86		0.18		0.84		0.4

SVM: Support vector machine; RF: Random forest; ANN: Artificial neural network; AUC: Area under the curve; [1]Training set; [2]Test set; D: Accuracy; F: F-measure; MCC: Matthews correlation coefficient; DP: Discriminant power; c, γ: Hyperparameters for SVM; S: Size; De: Decay.

According to the various risk rates in the dependent (target) variable (0.2-0.4-0.6-0.8), mostly the best performance results were obtained from the SVM model between 60% and 80%. It was observed that the performance of SVM was higher in the artificial data set where structures of variable were very similar (all variables comprise 0-1). Therefore, successful result was obtained from the SVM method for all types of independent variables, regardless of the rate in the dependent variable. It was seen that successful results were obtained from RF method only when the rate of the risky situation was very high and low (0.80-0.20), and when the number of observation and variables increased. All the results of the simulation data are given in Appendix Table 1, Table 2, Table 3, Table 4. When the results obtained from both the real data set and the simulation datasets were evaluated together, it was understood that the SVM model was a very successful ML technique for binary classifiers. The reason for testing the difference in training-test sets accuracy (ΔD) was that this difference can provide information about the performance of the model. In other words, as seen in Table 5, the smaller the (ΔD) value or the greater the p value obtained from the test result, the higher the performance of the model. However, it was thought that more ML techniques and more simulations were needed to be able to say this clearly.

The rank order of importance of prognostic factors affecting survival was determined by the SVM method, which was the most performing model (Figure 1). According to these results, it was understood that the prognostic factors that had the most impact on survival were tumour size, ER hormone receptor and age (considering that hormonotherapy is the treatment method). The variables that contribute the least to the classification were gender, number of lymph nodes and chemotherapy. The fact that almost every patient received chemotherapy, which was a routine treatment, negatively affected its importance.

TABLE 5: Methods with the highest performance according to different rates in simulation data.

Sample size	Number of variable	ML methods	Hyperparameters	D[1]	D[2]	ΔD	p	PPV[1]	PPV[2]	F[1]	F[2]
The rate of the risky situation in the dependent variable was about 20%											
N=500	*p=15*	**SVM**	c[1]=0.25.	0.8		0.004	0.861	0.8		0.89	
			γ^1=0.040								
			c[2]=0.25.		0.8				0.8		0.89
			γ^2=0.043								
	p=25	**SVM**	c[1]=2.	0.8		-0.0019	0.001	0.8		0.89	
			γ^1=0.022								
			c[2]=0.25.		0.8				0.8		0.89
			γ^2=0.021								
	p=35	**SVM**	c[1]=0.25.	0.79		-0.001	0.809	0.79		0.89	
			γ^1=0.015								
			c[2]=0.25.		0.8				0.8		0.89
			γ^2=0.015								
N=1000	*p=15*	**SVM**	c[1]=0.25.	0.8		-0.0011	0.006	0.8		0.89	
			γ^1=0.040								
			c[2]=0.25.		0.8				0.8		0.89
			γ^2=0.041								
	p=25	**SVM**	c[1]=2.	0.8		-0.0013	0.012	0.8		0.89	
			γ^1=0.022								
			c[2]=0.25.		0.8				0.8		0.89
			γ^2=0.021								
	p=35	**RF**	mtry=3	0.8		0.0002	0.428	1		0.89	
			ntree=100		0.8				0.99		0.48
The rate of the risky situation in the dependent variable was about 40%											
N=500	***p=15***	**SVM**	c[1]=0.25.	0.63		0.0347	0.008	0.62		0.76	
			γ^1=0.039								
			c[2]=0.1.		0.59				0.6		0.74
			γ^2=0.040								
	p=25	**SVM**	c[1]=0.25.	0.62		0.0314	0.018	0.62		0.76	
			γ^1=0.023								
			c2=0.25.		0.59				0.6		0.74
			γ^2=0.022								
	p=35	**SVM**	c[1]=0.25.	0.65		0.0479	0.014	0.64		0.78	
			γ^1=0.015								
			c[2]=0.25.		0.6				0.6		0.74
			γ^2=0.015								
N=1000	***p=15***	**SVM**	c[1]=0.25.	0.61		0.009	0.049	0.6		0.75	
			γ^1=0.040								
			c[2]=4.		0.6				0.6		0.75
			γ^2=0.037								
	p=25	**SVM**	c[1]=0.25.	0.62		0.0148	0.007	0.61		0.76	
			γ^1=0.022								
			c[2]=0.25.		0.6				0.6		0.75
			γ^2=0.023								
	p=35	**SVM**	c[1]=0.25.	0.61		0.0107	0.045	0.6		0.75	
			γ^1=0.015								
			c[2]=0.25.		0.59				0.6		0.75
			γ^2=0.014								

The rate of the risky situation in the dependent variable was about 60%											
N=500	*p=15*	SVM	c^1=0.25.	0.66		0.0712	0.001	0.94		0.5	
			γ^1=0.040								
			c^2=0.5.		0.59				0.31		0.22
			γ^2=0.039								
	p=25	SVM	c^1=0.25.	0.61		0.0047	0.394	0.98		0.2	
			γ^1=0.022								
			c^2=0.25.		0.6				0.42		0.23
			γ^2=0.021								
	p=35	SVM	c^1=0.25.	0.63		0.0279	0.07	0.98		0.54	
			γ^1=0.015								
			c^2=0.25.		0.6				0.39		0.3
			γ^2=0.015								
N=1000	*p=15*	SVM	c^1=2.	0.6		0.0112	0.036	0.9		0.2	
			γ^1=0.039								
			c^2=0.25.		0.59				0.21		0.04
			γ^2=0.040								
	p=25	SVM	c^1=0.25.	0.6		0.0065	0.227	0.97		0.22	
			γ^1=0.022								
			c^2=1.		0.6				0.37		0.23
			γ^2=0.021								
	p=35	SVM	c^1=0.25.	0.62		0.0224	0.168	0.98		0.45	
			γ^1=0.014								
			c^2=0.25.		0.6				0.22		0.39
			γ^2=0.015								
The rate of the risky situation in the dependent variable was about 80%											
N=500	*p=15*	SVM	c^1=2.	0.8		0	0.985	1		0.37	
			γ^1=0.040								
			c^2=0.25.		0.8				0		!
			γ^2=0.040								
	p=25	SVM	c^1=0.25.	0.8		-0.0006	0.777	1		0.44	
			γ^1=0.023								
			c^2=0.25.		0.8				!		!
			γ^2=0.021								
	p=35	SVM	c^1=0.25.	0.8		-0.0058	0.063	1		0.4	
			γ^1=0.015								
			c^2=0.25.		0.81				!		!
			γ^2=0.015								
N=1000	*p=15*	SVM	c^1=2.	0.8		-0.0011	0.001	1		0.0676	
			γ^1=0.023								
			c^2=0.25.		0.8				!		!
			γ^2=0.039								
	p=25	SVM	c^1=0.25.	0.8		-0.0016	0.001	!		!	
			γ^1=0.042								
			c^2=0.25.		0.8				!		!
			γ^2=0.022								
	p=35	RF	mtry=3	0.8		0.0033	0.415	!		0.0137	
			ntree=500		0.8				!		!

ML: Machine learning; SVM: Support vector machine; RF: Random forest; [1]Training set; [2]Test set; D: Accuracy; *p*: Number of independent variables; ΔD: Mean of the accuracy difference; p: One sample t-test significance, PPV: Positive predictive value; F: F-measure; !: Non-calculable value; c, γ: Hyperparameters for SVM.

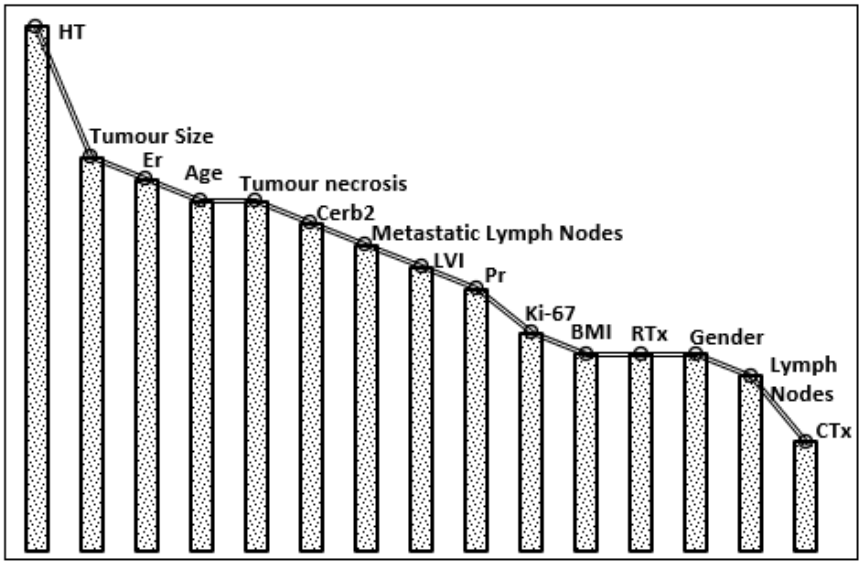

FIGURE 1: Rank order of importance of variables in support vector machine.

HT: Hormonotherapy; ER: Estrogen; LVI: Lymphovascular invasion; PR: Progesterone; BMI: Body mass index; RTx: Radiotherapy; CTx: Chemotherapy.

DISCUSSION

In this study, three ML methods divided into 12 sub-units were compared with five metrics for the actual data set and four metrics (with ΔD) for the simulation data. The highest classification accuracy (80%) for test data in the real data set was obtained from SVM with radial-basis. In the simulation data, while the highest classification accuracy was obtained from SVM in general, successful results were rarely seen in the RF method.

Chao et al. used seven independent variables in the survival classification for breast cancer in their study.[29] In the breast cancer survival classification study; tumor size, age, radiotherapy, number of lymph nodes, metastatic lymph nodes, pathological staging and chemotherapy variables were used. They stated that the most important prognostic factors among these variables are tumor size, metastatic lymph nodes and pathological staging. In our study, tumor size was found to be one of the most important variables for survival, and the significance of metastatic lymph nodes was moderate. Pathological staging variable was not included in our study. In the study, they used SVM, logistic regression and decision trees, and presented the results in the form of training and test set accuracy. According to the analysis results, it was seen that high classification accuracy was obtained from SVM. In another survival study in which patients with IDC were used as a data set, logistic regression and alternating decision trees (ADTree) were used. Methods were compared according to accuracy and AUC metric.[30]

Park et al. demonstrated the classifiability of survival with prognostic factors in breast cancer patients using ANN, SVM and semi-supervised learning methods with 16 independent variables. They suggested using a model that bridges the domains of learning theory and medicine to aid medical specialists in independent variables selection when building a predictive model for medical domain.[31] In another study, Ganggayah et al. demonstrated the classifiability of survival of breast cancer patients by prognostic factors using seven ML methods and 23 independent variables (highest accuracy obtained with SVM). In their studies, excluding cancer stage, the four most important variables determined by the RF method were tumor size, number of lymph nodes, treatment type and lymph node status (metastatic lymph nodes). Similarities were found with our study in terms of the rank order of importance of the variables.[32]

APPENDIX TABLE 1: The rate of the risky situation in the dependent variable was about 20%.

	Number of variable		D[1]	D[2]	ΔD	p	PPV[1]	PPV[2]	F[1]	F[2]
N=500	*p=15*	SVM	0.8004	0.8000	0.0040	0.861	0.8000	0.8000	0.8886	0.8882
		RF	0.7933	0.7899	0.0034	0.116	0.9914	0.9842	0.8844	0.8819
		ANN	0.9064	0.7319	0.1745	0.001	0.9061	0.7982	0.9460	0.8399
	p=25	SVM	0.7981	0.7998	-0.0019	0.001	0.7981	0.7998	0.8876	0.8887
		RF	0.7931	0.7906	0.0025	0.014	0.9974	0.9937	0.8845	0.8828
		ANN	0.9221	0.7410	0.1811	0.001	0.9200	0.8024	0.9558	0.8456
	p=35	SVM	0.7943	0.7954	-0.0010	0.809	0.7943	0.7954	0.8853	0.8859
		RF	0.7913	0.7930	-0.0160	0.517	0.9979	0.9950	0.8833	0.8843
		ANN	0.9304	0.7477	0.1826	0.001	0.9330	0.8085	0.9602	0.8507
N=1000	*p=15*	SVM	0.7999	0.8003	-0.0011	0.006	0.7997	0.8003	0.8886	0.8890
		RF	0.7931	0.7880	0.0051	0.001	0.9942	0.9880	0.8845	0.8813
		ANN	0.8542	0.7489	0.1053	0.001	0.8580	0.8045	0.9159	0.8513
	p=25	SVM	0.8026	0.8040	-0.0013	0.012	0.8026	0.8040	0.8904	0.8912
		RF	0.8052	0.8012	0.0039	0.239	0.9992	0.9980	0.8920	0.8894
		ANN	0.9061	0.7237	0.1823	0.001	0.9079	0.8015	0.9446	0.8335
	p=35	SVM	0.7966	0.7936	0.0030	0.346	0.7992	0.7995	0.8883	0.8884
		RF	0.8007	0.8041	0.0002	0.428	0.9999	0.9866	0.8893	0.4811
		ANN	0.9309	0.7265	0.2043	0.001	0.9417	0.8025	0.9589	0.8359

SVM: Support vector machine; RF: Random forest; ANN: Artificial neural network; [1]Training set; [2]Test set; D: Accuracy, *p*: Number of independent variables, ΔD: Mean of the accuracy difference, p: One sample t-test significance, PPV: Positive predictive value, F: F-measure.

APPENDIX TABLE 2: The rate of the risky situation in the dependent variable was about 40%.

	Number of variable		D[1]	D[2]	ΔD	p	PPV[1]	PPV[2]	F[1]	F[2]
N=500	p=15	SVM	0.6265	0.5917	0.0347	0.008	0.6223	0.5988	0.7642	0.7407
		RF	0.5459	0.5476	-0.0016	0.867	0.7767	0.7599	0.6671	0.6630
		ANN	0.7778	0.5372	0.2405	0.001	0.7843	0.5982	0.8322	0.6317
	p=25	SVM	0.6237	0.5923	0.0314	0.018	0.6166	0.5969	0.7605	0.7391
		RF	0.5610	0.5585	0.0025	0.745	0.8213	0.7946	0.6865	0.6769
		ANN	0.8731	0.5215	0.3516	0.001	0.8796	0.5945	0.9003	0.6155
	p=35	SVM	0.6456	0.5977	0.0479	0.014	0.6434	0.5999	0.7794	0.7434
		RF	0.5654	0.5547	0.1060	0.155	0.8449	0.8232	0.6973	0.6875
		ANN	0.9312	0.5188	0.4124	0.001	0.9384	0.5979	0.9435	0.5980
N=1000	p=15	SVM	0.6062	0.5971	0.0090	0.049	0.6024	0.5987	0.7508	0.7453
		RF	0.5616	0.5461	0.0154	0.001	0.8335	0.7846	0.6945	0.6746
		ANN	0.7276	0.5419	0.1857	0.001	0.8335	0.7846	0.7936	0.6472
	p=25	SVM	0.6158	0.6009	0.0148	0.007	0.6109	0.6016	0.7580	0.7495
		RF	0.5692	0.5625	0.0066	0.049	0.8751	0.8400	0.7074	0.6950
		ANN	0.8378	0.5173	0.3205	0.001	0.8397	0.5934	0.8720	0.6081
	p=35	SVM	0.6056	0.5949	0.0107	0.045	0.6019	0.5957	0.7512	0.7452
		RF	0.5857	0.5831	0.0025	0.380	0.9111	0.8881	0.7257	0.7193
		ANN	0.8670	0.5324	0.3345	0.001	0.8785	0.6135	0.8916	0.6223

SVM: Support vector machine; RF: Random forest; ANN: Artificial neural network; [1]Training set; [2]Test set; D: Accuracy; ΔD: Mean of the accuracy difference; p: One sample t-test significance; PPV: Positive predictive value; F: F-measure.

APPENDIX TABLE 3: The rate of the risky situation in the dependent variable was about 60%.

	Number of variable		D[1]	D[2]	ΔD	p	PPV[1]	PPV[2]	F[1]	F[2]
N=500	p=15	SVM	0.6616	0.5904	0.0712	0.001	0.9420	0.3105	0.5002	0.2221
		RF	0.5624	0.5638	-0.0014	0.926	0.2230	0.2396	0.2930	0.2970
		ANN	0.7939	0.5249	0.2689	0.001	0.7742	0.3885	0.7071	0.3570
	p=25	SVM	0.6080	0.6034	0.0047	0.394	0.9792	0.4167	0.1998	0.2326
		RF	0.5693	0.5643	0.0050	0.480	0.1660	0.1726	0.2261	0.2311
		ANN	0.9162	0.5042	0.4120	0.001	0.9167	0.3851	0.8993	0.3696
	p=35	SVM	0.6285	0.6007	0.0279	0.070	0.9804	0.3900	0.5390	0.2980
		RF	0.5632	0.5620	0.0012	0.882	0.1391	0.1651	0.1975	0.2235
		ANN	0.9394	0.5249	0.4145	0.001	0.9413	0.4014	0.9188	0.3914
N=1000	p=15	SVM	0.6004	0.5892	0.0112	0.036	0.8956	0.2071	0.2037	0.0426
		RF	0.5631	0.5587	0.0044	0.560	0.1486	0.2033	0.2101	0.2654
		ANN	0.7276	0.5405	0.1872	0.001	0.6972	0.3978	0.5869	0.3295
	p=25	SVM	0.6040	0.5975	0.0065	0.227	0.9691	0.3704	0.2211	0.2260
		RF	0.5735	0.5772	-0.0037	0.518	0.1166	0.1483	0.1766	0.2110
		ANN	0.8838	0.5192	0.3646	0.001	0.8884	0.3815	0.8557	0.3653
	p=35	SVM	0.6218	0.5994	0.0224	0.168	0.9828	0.2238	0.4539	0.3947
		RF	0.5752	0.5667	0.0084	0.022	0.1051	0.1373	0.1638	0.2006
		ANN	0.8881	0.5432	0.3449	0.001	0.8837	0.4217	0.8546	0.4086

SVM: Support vector machine; RF: Random forest; ANN: Artificial neural network; [1]Training set; [2]Test set; D: Accuracy; ΔD: Mean of the accuracy difference; p: One sample t-test significance; PPV: Positive predictive value; F: F-measure.

APPENDIX TABLE 4: The rate of the risky situation in the dependent variable was about 80%.

	Number of variable		D[1]	D[2]	ΔD	p	PPV[1]	PPV[2]	F[1]	F[2]
N=500	p=15	SVM	0.8002	0.8002	0.0000	0.985	1.0000	0.0000	0.3664	!
		RF	0.7961	0.7945	0.0016	0.693	0.0121	0.0053	0.0499	0.0785
		ANN	0.8863	0.7423	0.1439	0.001	0.8809	0.1946	0.6544	0.1347
	p=25	SVM	0.8022	0.8028	-0.0006	0.777	1.0000	!	0.4444	!
		RF	0.8004	0.8012	-0.0007	0.431	0.0009	!	0.0286	!
		ANN	0.9388	0.7416	0.1972	0.001	0.9830	0.1894	0.9082	0.1368
	p=35	SVM	0.8004	0.8063	-0.0058	0.063	1.0000	!	0.3902	!
		RF	0.8024	0.8023	0.0001	0.910	0.0019	!	0.0273	!
		ANN	0.9448	0.7472	0.1976	0.001	0.9983	0.2027	0.9633	0.1697
N=1000	p=15	SVM	0.8018	0.8029	-0.0011	0.001	1.0000	!	0.0676	!
		RF	0.7940	0.7894	0.0047	0.002	0.0049	0.0083	0.0242	0.0438
		ANN	0.8497	0.7426	0.1071	0.001	0.8513	0.1456	0.5137	0.1156
	p=25	SVM	0.8003	0.8019	-0.0016	0.001	!	!	!	!
		RF	0.8004	0.8002	0.0003	0.592	0.0010	0.0012	0.0142	0.0351
		ANN	0.9047	0.7198	0.1849	0.001	0.8949	0.1906	0.7497	0.1663
	p=35	SVM	0.7978	0.7991	-0.0012	0.001	!	!	!	!
		RF	0.8027	0.7994	0.0033	0.415	!	!	0.0137	!
		ANN	0.9259	0.7272	0.1988	0.001	0.9659	0.1982	0.8857	0.1697

SVM: Support vector machine; RF: Random forest; ANN: Artificial neural network; [1]Training set; [2]Test set; D: Accuracy; ΔD: Mean of the accuracy difference; p: One sample t-test significance; PPV: Positive predictive value; F: F-measure; !: Non-calculable value.

Khondoker et al. used RF, SVM, Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (kNN) methods both in the simulation of the real data set and in the fully simulated data sets (obtained using the poisson distribution).[33] They compared the methods according to classification error, sensitivity and specificity metrics. As a general result, they suggested LDA when $p/n<0.5$ (where p indicates the number of variables and n was the sample size) in the highly correlated data set. However, when $p/n\geq0.5$, they suggested radial basis SVM without mentioning the correlation. They reported that all four methods in the study (excluding RF) performed better at high correlation.

Kate and Nadig used 16 independent variables for survival classification in their study.[34] In the study, they used ML methods such as naive bayes, logistic regression and decision trees, and evaluated the results according to the AUC metric. They stated that the most successful performance was obtained from the naive bayes method.

Engelhardt et al. obtained a larger data set with multivariate normal distribution by simulating a small data set.[35] After this, they compared these data sets with LDA, Quadratic Linear Discriminant Analysis, Penalized Discriminant Analysis (PDA), RF, classification and regression trees, ANN and kNN methods, according to the misclassification (error rate) rate. They reached the conclusion that the lowest misclassification rate is in PDA.

CONCLUSION

In the real data set, it was observed that most of the determined independent variables (prognostic factors and treatment methods) were effective in the classification. Therefore, it was concluded that these variables are important for the survival of IDC stage III patients. However, it is thought that higher performance results will be achieved when more variables that are thought to affect survival are classified using SVM method. In addition, SVM was found to be a high performance model in artificial data sets designed for different rates and number of observations. It was concluded that SVM has higher classification accuracy and performance metrics than other methods (RF and ANN) in binary classification of both real and artificial data sets.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Emre Dirican, Zeki Akkuş;* ***Design:*** *Emre Dirican, Zeki Akkuş;* ***Control/Supervision:*** *Emre Dirican, Zeki Akkuş;* ***Data Collection and/or Processing:*** *Emre Dirican;* ***Analysis and/or Interpretation:*** *Emre Dirican, Zeki Akkuş;* ***Literature Review:*** *Emre Dirican;* ***Writing the Article:*** *Emre Dirican;* ***Critical Review:*** *Emre Dirican, Zeki Akkuş;* ***References and Fundings:*** *Emre Dirican, Zeki Akkuş;* ***Materials:*** *Emre Dirican.*

REFERENCES

1. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med. 2018;56(4):516-24. [Crossref] [PubMed]
2. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988-99. [Crossref] [PubMed]
3. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. [Crossref] [PubMed] [PMC]
4. Revathi S, Malathi A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research & Technology (IJERT). 2013;2(12):1848-53. [Link]
5. Shao Y, Liu Y, Ye X, Zhang S. A machine learning based global simulation data mining approach for efficient design changes. Adv Eng Softw. 2018;124:22-41. [Crossref]
6. Böcker W. WHO-Klassifikation der Tumoren der Mamma und des weiblichen Genitale: Pathologie und Genetik [WHO classification of breast tumors and tumors of the female genital organs: pathology and genetics]. Verh Dtsch Ges Pathol. 2002;86:116-9. German. [PubMed]
7. Tata A, Woolman M, Ventura M, Bernards N, Ganguly M, Gribble A, et al. Rapid detection of necrosis in breast cancer with desorption electrospray ionization mass spectrometry. Sci Rep. 2016;6:35374. [Crossref] [PubMed] [PMC]
8. Jones RL, Salter J, A'Hern R, Nerurkar A, Parton M, Reis-Filho JS, et al. The prognostic significance of Ki67 before and after neoadjuvant chemotherapy in breast cancer. Breast Cancer Res Treat. 2009;116(1):53-68. [Crossref] [PubMed]
9. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121(7):2750-67. [Crossref] [PubMed] [PMC]
10. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res. 2007;13(15 Pt 1):4429-34. [Crossref] [PubMed]
11. Wong JS, O'Neill A, Recht A, Schnitt SJ, Connolly JL, Silver B, et al. The relationship between lymphatic vessell invasion, tumor size, and pathologic nodal status: can we predict who can avoid a third field in the absence of axillary dissection? Int J Radiat Oncol Biol Phys. 2000;48(1):133-7. [Crossref] [PubMed]
12. Lee AH, Pinder SE, Macmillan RD, Mitchell M, Ellis IO, Elston CW, et al. Prognostic value of lymphovascular invasion in women with lymph node negative invasive breast carcinoma. Eur J Cancer. 2006;42(3):357-62. [Crossref] [PubMed]
13. Agarwal G, Pradeep PV, Aggarwal V, Yip CH, Cheung PS. Spectrum of breast cancer in Asian women. World J Surg. 2007;31(5):1031-40. [Crossref] [PubMed]
14. Woodward WA, Vinh-Hung V, Ueno NT, Cheng YC, Royce M, Tai P, et al. Prognostic value of nodal ratios in node-positive breast cancer. J Clin Oncol. 2006;24(18):2910-6. [Crossref] [PubMed]
15. Neuhouser ML, Aragaki AK, Prentice RL, Manson JE, Chlebowski R, Carty CL, et al. Overweight, obesity, and postmenopausal invasive breast cancer risk: a secondary analysis of the women's health initiative randomized clinical trials. JAMA Oncol. 2015;1(5):611-21. [Crossref] [PubMed] [PMC]
16. R Core Team. R: A language and environment for statistical computing. 2013. [Accessing Date: 05 September 2019]. Accessing Link: [Link]
17. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-7. [Crossref] [PubMed]
18. Li M, Chen F, Lei M, Li C. [Near-infrared spectrum of coal origin identification based on LVQ with SVM algorithm]. Guang Pu Xue Yu Guang Pu Fen Xi. 2016;36(9):2793-7. Chinese. [PubMed]
19. Palmer DS, O'Boyle NM, Glen RC, Mitchell JB. Random forest models to predict aqueous solubility. J Chem Inf Model. 2007;47(1):150-8. [Crossref] [PubMed]
20. Breiman L. Random forests. Machine Learning. 2001;45:5-32. [Crossref]
21. Renganathan V. Overview of artificial neural network models in the biomedical domain. Bratisl Lek Listy. 2019;120(7):536-40. [Crossref] [PubMed]
22. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019;29(7):R231-6. [Crossref] [PubMed]
23. Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting. 1998;14(1):35-62. [Crossref]
24. Akosa JS. Predictive accuracy: A misleading performance measure for highly imbalanced data. Proceedings of the SAS Global Forum. 2017:1-12. [Link]
25. Murat Yilmaz C, Kose C, Hatipoglu B. A Quasi-probabilistic distribution model for EEG Signal classification by using 2-D signal representation. Comput Methods Programs Biomed. 2018;162:187-96. [Crossref] [PubMed]
26. Wang S, Li D, Petrick N, Sahiner B, Linguraru MG, Summers RM. Optimizing area under the ROC curve using semi-supervised learning. Pattern Recognit. 2015;48(1):276-87. [Crossref] [PubMed] [PMC]
27. Nellore SB. Various performance measures in Binary classification--An Overview of ROC study. IJISET-International Journal of Innovative Science, Engineering & Technology. 2015;2(9):596-605. [Link]
28. Üncel M, Aköz G, Yıldırım Z, Pişkin G, Değirmenci M, Solakoğlu Kahraman D, et al. Meme kanserinin klinikopatolojik özelliklerinin moleküler alt tipe göre değerlendirilmesi [Evaluation of clinicopathological features of breast cancer according to the molecular subtypes]. Tepecik Eğit ve Araşt Hast Dergisi. 2015;25(3):151-6. [Link]
29. Chao CM, Yu YW, Cheng BW, Kuo YL. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst. 2014;38(10):106. [Crossref] [PubMed]
30. Horiguchi K, Toi M, Horiguchi S, Sugimoto M, Naito Y, Hayashi Y, et al. Predictive value of CD24 and CD44 for neoadjuvant chemotherapy response and prognosis in primary breast cancer patients. J Med Dent Sci. 2010;57(2):165-75. [PubMed]
31. Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Engineering Applications of Artificial Intelligence. 2013;26(9):2194-205. [Crossref]
32. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak. 2019;19(1):48. [Crossref] [PubMed] [PMC]
33. Khondoker M, Dobson R, Skirrow C, Simmons A, Stahl D. A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. Stat Methods Med Res. 2016;25(5):1804-23. [Crossref] [PubMed] [PMC]
34. Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. Int J Med Inform. 2017;97:304-11. [Crossref] [PubMed]
35. Engelhardt A, Kanawade R, Knipfer C, Schmid M, Stelzle F, Adler W. Comparing classification methods for diffuse reflectance spectra to improve tissue specific laser surgery. BMC Med Res Methodol. 2014;14:91. [Crossref] [PubMed] [PMC]