

Development of prognostic model and multivariate analysis for breast cancer survival patients using SEER database

Nihar Ranjan Panda^{1*} Kamal Lochan Mahanta¹ Jitendra Kumar Pati² Soumya Subhashree Satapathy³ Ruchi Bhuyan⁴

¹Department of Mathematics, CV Raman Global University, Bhubaneswar, India.

²KIIT International School, KIIT University, Bhubaneswar, India.

³Center for Biotechnology, School of Pharmaceutical Science, SOA deemed to be University, Bhubaneswar, India.

⁴Department of Medical Research, IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, India.

ARTICLE INFO

Article history:

Received 1 August 2023

Accepted as revised 30 October 2023

Available online 2 November 2023

Keywords:

Breast cancer, prognostic model, risk factors, regression.

ABSTRACT

Background: Many studies employed machine learning (ML) to forecast the prognosis of breast cancer (BC) patients and discovered that the ML model showed high individualized forecasting ability. Breast cancer is the most frequent kind of carcinoma in women globally and ranks as the leading cause of death in women.

Objectives: This study intends to use the Surveillance, Epidemiology, and End Results dataset to categorize breast carcinoma cases' alive and dead conditions. Deep learning and machine learning have been extensively utilized in clinical studies to address various categorization problems due to their ability to manage massive data sets in an organized manner. Pre-processing the data allows it to be visualized and analyzed for making critical choices. This study describes a realistic machine learning-based strategy for categorizing the SEER breast cancer dataset.

Materials and methods: We employed classification and machine learning algorithms to classify breast cancer mortality. Four well-known classification ML algorithms were employed in this study. To identify risk factors, we employed multivariate analysis using the data set.

Results: The decision tree performed the best accuracy (0.914) among all the models. T4 stage ($\beta=1.4$, $p<0.001$, OR=4.22, 95% CI (2.06-8.64), N2 stage ($\beta=0.39$, $p=0.008$, OR= 1.49, 95% CI (1.111-1.997) found to be major risk factors for breast cancer mortality using multivariate analysis.

Conclusion: The significant prognostic variables affecting the breast carcinoma survival rates reported in the current research are relevant and might be turned into decision support systems in the medical realm.

Introduction

Breast cancer is widely recognized as the prevailing type of carcinoma affecting women globally. It has emerged as the primary cause of cancer. Epidemiological research has discovered that the global incidence of breast cancer is estimated to surpass nearly 2 million cases by the year 2030.¹ According to recent data from GlobeCan in 2020, in India, breast cancer accounted for 13.5% of all cancer cases and 10.6% of all cancer-related mortalities in India. Breast cancer is categorized as a form of tissue cancer primarily affecting the inner layer of milk glands or lobules as well as the ducts, which are small tubes responsible for transporting milk. Several crucial factors contribute

* Corresponding contributor.

Author's Address: Department of Mathematics,
CV Raman Global University, Bhubaneswar,
India.

E-mail address: niharpanda1994@gmail.com

doi: 10.12982/JAMS.2024.008

E-ISSN: 2539-6056

to the development of breast cancer, such as age, race, a deficit of iodine in the diet, and elevated hormone levels^{2,3} Breast cancer is characterized by the combined effect of genetic predisposition and environmental influence. This interaction gives rise to the gradual accumulation of genetic and epigenetic alterations within breast cancer cells, ultimately contributing to ailment development.²

In the past, various methods have been employed for the detection of breast cancer, such as biopsies, ultrasonography, and mammography. These diagnostic approaches have proven valuable in facilitating the initial identification of breast cancer. However, it is important to note that these detection methods do not always provide precise results and can be financially burdensome. Chemotherapy serves as a treatment approach for breast cancer. Still, it exerts a significant impact on the quality of life of patients, profoundly affecting their physical and mental social being.⁴ In contemporary times, advanced detection and treatment methods have emerged, offering a promising solution to overcome the drawbacks associated with conventional methods. Machine learning, the branch of artificial intelligence, has gained considerable attention to significantly enhance accuracy and precision detection, as well as providing innovative treatment strategies, thereby addressing the drawbacks previously encountered.⁵

Materials and methods

The data used for this research were obtained from SEER database, which was extracted from Kaggle website.³¹ The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Programme <https://seer.cancer.gov/about/overview.html> was an authoritative source of information on cancer incidence and survival in the United States, covering roughly 48.0% of the US population. This dataset of breast cancer patients was taken from the NCI's SEER Program's November 2017 update, which offers information on population-based cancer statistics.

Statistical analysis

All the statistical analysis was done by R programming language (version 4.3.1) and JASP Statistical software. We summarized the categorical variables using frequencies and percentages. The association between breast cancer survival and other predictors was evaluated by using a chi-square test or Fisher exact test. Similarly, in continuous variables, data were summarized in terms of mean (sd) or median (iqr). Comparison between continuous variables concerning breast cancer survival status was done by students t-test or man Whitney test. The multivariate logistic regression model observed the adjusted association between survival status and other predictors.

Data processing

In the context of machine learning, processing information refers to the modification, transformation, and preparation of raw data to prepare it for analysis and training machine learning algorithms.^{6,7} It entails several strategies and processes to clean, organize, and

prepare work data before it is fed into a machine learning algorithm. The SEER Data was collected and checked for missing values and outliers using various methods. Once the data were clean and ready for analysis, we built various machine-learning algorithms to predict breast cancer mortality.

Machine learning

Machine learning techniques may be utilized for forecasting and diagnosing carcinoma of the breast by analyzing data sources such as medical imaging, patient demographics, and clinical records.^{8,9} Keep in mind that establishing a trustworthy breast cancer prediction model needs access to high-quality data, working with healthcare experts, and thorough testing to assure its safety and efficacy. Naive Bayes (NB), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbor (KNN), SVM, classification and regression trees (CART), Random forest (RF), multivariate adaptive regression splines (MARS), and logistic regression are examples of common ML methods. (LR) and extreme gradient boosting (XGBoost) can be used to predict breast cancer.

The Random Forest (RF) technique is frequently employed in clinical investigations because of its higher performance than other methodologies.^{10,11} It has been proven to attain excellent accuracy and to be especially successful in dealing with extremely non-linear data and datasets with many characteristics. Furthermore, RF is more adaptable to data noise and easier to tune than other ensemble learning techniques. The technique includes a method for predicting missing values, the usage of a Weighted Random Forest (WRF) for balancing mistakes in unbalanced data, and the capacity to assess the relevance of the parameters used for classification.

In the current investigation, logistic regression describes the connection between the independent variables (the 23 factors influencing survival status) and the dependent variable (survival status). The odds ratio is determined using the Gaussian distribution. Because logistic regression is best suited for binary outcomes, patients' condition for survival (alive/dead) is recorded as 1/0 in the dataset. Because the dependent variable is a binary survival status, logistic regression was chosen as one of the evaluation techniques to measure accuracy using binary values.¹²⁻¹⁴

The logistic equation is given by

$$(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1x_1)}} = \frac{e^{(\beta_0+\beta_1x_1)}}{1+e^{(\beta_0+\beta_1x_1)}} \dots (1)$$

When there are several predictors in the data, the equation becomes:

$$P(y) = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\dots+\beta_ix_i)}} \dots (2)$$

$$= \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\dots+\beta_ix_i)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\dots+\beta_ix_i)}} \dots (3)$$

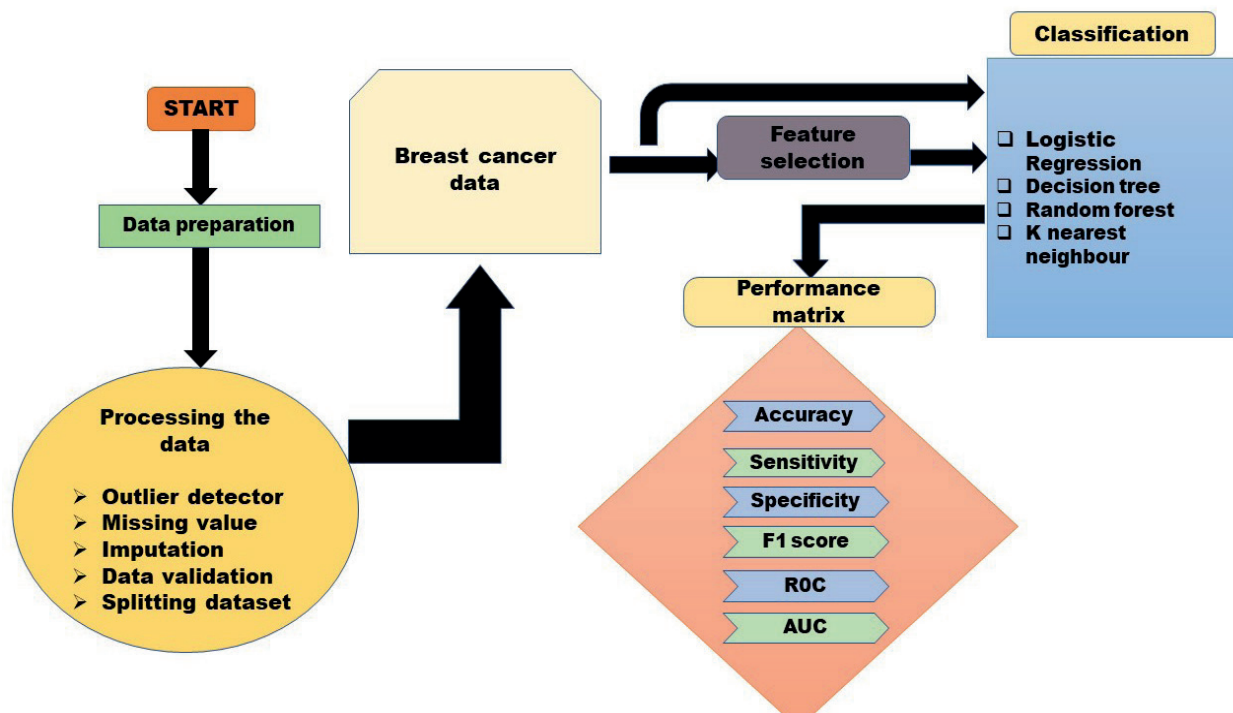


Figure 1. Road map from data processing to model building.

K-Nearest Neighbours (KNN) is a well-known machine learning technique that may be used for both classification and regression applications.^{15 16} It represents a straightforward yet effective method that is frequently used as a starting point for comparison with more complicated models. KNN may be employed for carrying out tasks related to classification, such as assigning a label from a class to a specified input data point. It operates by locating the k nearest neighbours in the feature space and giving a majority class to the query point among them. KNN may also be employed to solve regression problems when the aim is to forecast a continuous value. rather than using the overwhelming vote in regression, KNN computes the average (or weighted average) of the target values of the k nearest neighbors as the forecast value for the query point. KNN utilizes a non-parametric method, which means it makes no assumptions about the data distribution. It can deal with complicated and nonlinear interactions between

characteristics and target variables without requiring a specialized model structure.^{17,18}

Decision trees represent a clear and understandable depiction of the decision-making procedure.¹⁹⁻²¹ The learned tree framework may be visualized and comprehended, making the logic behind the predictions easier to convey. Developing predictions for new cases is quick when a Decision Tree has been trained. The method traverses the tree structure by assessing the feature conditions at each internal node until it reaches the expected outcome at a leaf node. Decision Overfitting is common in trees, especially when the tree depth is not restricted.²² They can build too complicated trees that remember noise in the training data, resulting in poor generalization on unseen data. Pruning and establishing maximum depth or minimum samples per leaf might assist in reducing fitting.^{23,24}

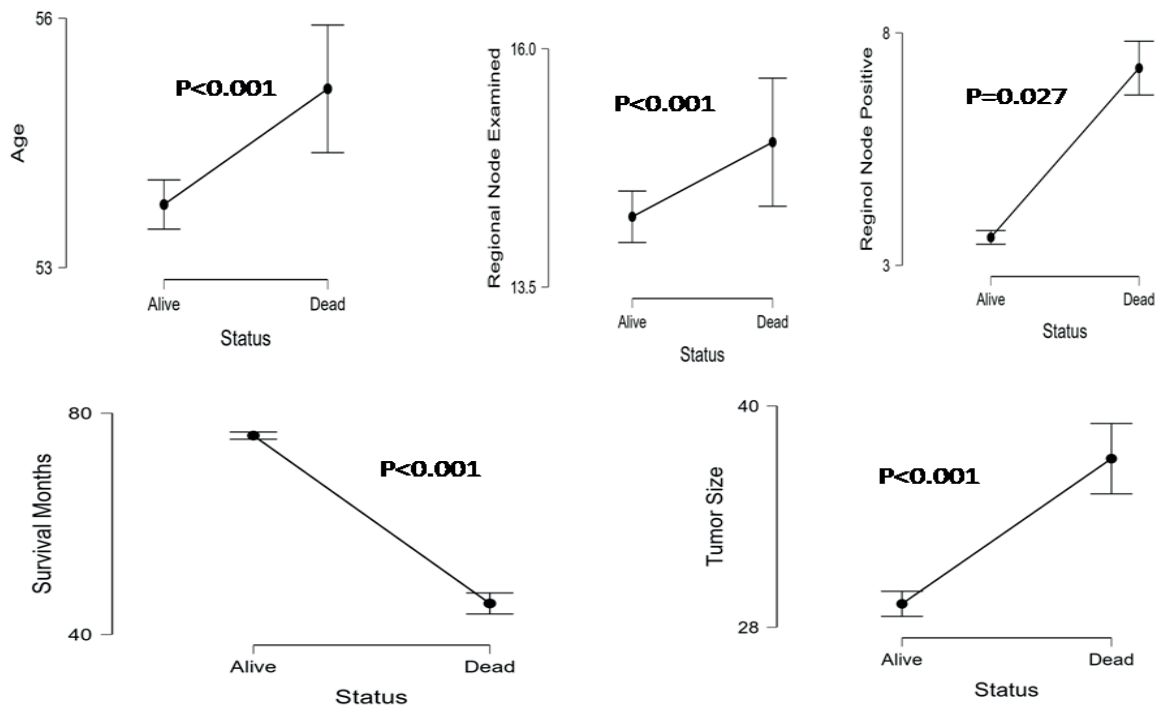


Figure 2. Mean plot showing the difference between alive and dead breast cancer patients.

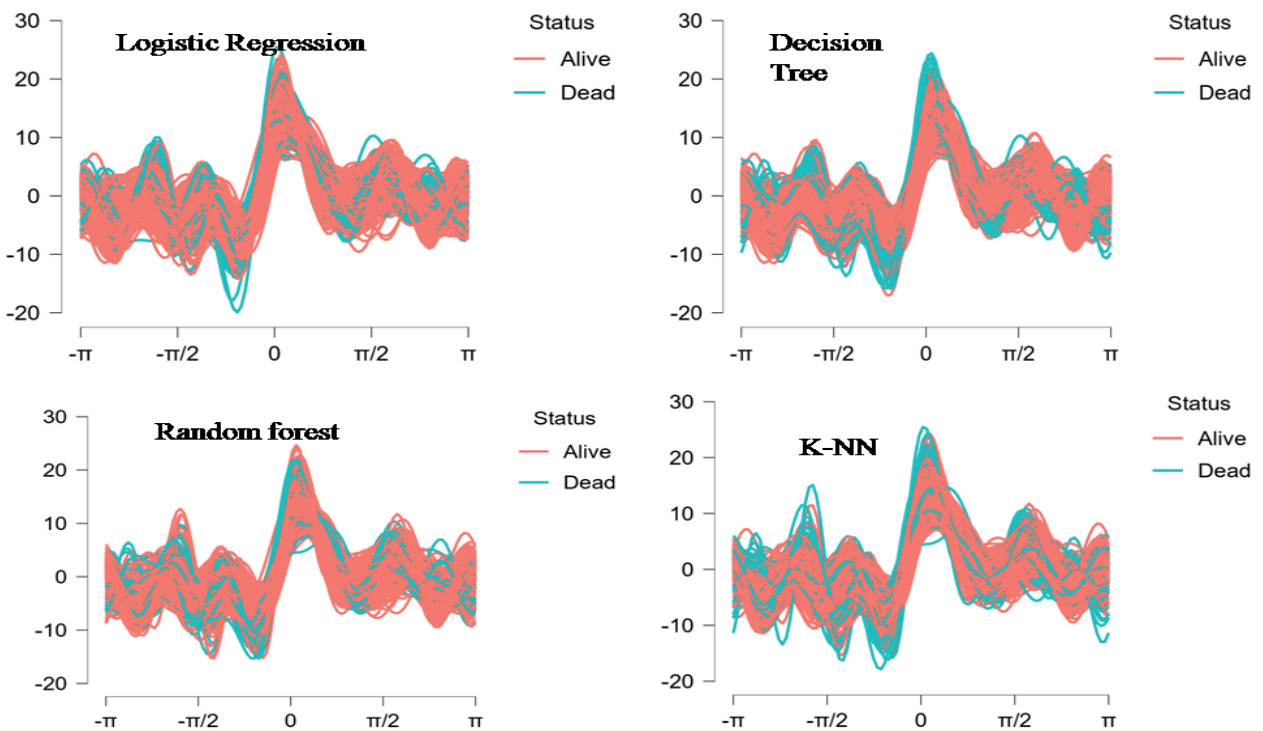


Figure 3. Andrews plot describing the classification of breast cancer using machine learning.

Results

Table 1 represents the demographic and clinical features of breast cancer variables. The data were separated into two groups for comparing the number and percentage. A total of 4024 patients' data were analyzed. The mortality percentage was 15.3. Table 2 describes the factors associated with breast cancer mortality. The analysis observed that age ($\beta=0.001$, $p<0.001$, OR=1.028, 95% CI (1.015-1.041), T4 stage ($\beta=1.4$, $p<0.001$, OR=4.22, 95% CI

(2.06-8.64), N2 stage ($\beta=0.39$, $p=0.008$, OR=1.49, 95% CI (1.111-1.997), regional node examined ($\beta=-0.03$, $p<0.001$, OR=0.96, 95% CI (0.954-0.985), regional node positive ($\beta=0.078$, $p<0.001$, OR=1.08, 95% CI (1.04-1.12), survival month ($\beta=-0.06$, $p<0.001$, OR=0.94, 95% CI (0.935-0.946), Grade 1 ($\beta=-2.272$, $p=0.03$, OR=0.18, 95% CI (0.03-0.86), progesterone status ($\beta=-0.5174$, $p<0.001$, OR=0.596, 95% CI (0.442-0.803) were significantly associated with breast cancer mortality.

Table 1. Demographic and clinical parameters for breast cancer patients.

Variables		Dead (N=616)	Alive (N=3408)
Race	Black	73 (11.8)	218 (6.3)
	Other	33 (5.3)	284 (8.4)
	White	510 (82.7)	2,903 (85.1)
Marital Status	Divorced	90 (14.6)	396 (11.6)
	Married	358 (58.11)	2,285 (67.04)
	Separated	15 (2.43)	30 (0.8)
	Single	104 (16.88)	511 (14.9)
	Widowed	49 (7.95)	186 (5.45)
T Stage	T1	157 (25.48)	1,446 (42.43)
	T2	303 (49.18)	1,483 (43.51)
	T3	116 (18.83)	417 (12.23)
	T4	40 (6.49)	62 (1.81)
N Stage	N1	270 (43.83)	2,462 (72.24)
	N2	165 (26.78)	655 (19.21)
	N3	181 (29.38)	291 (8.53)
6 th Stage	IIA	96 (15.88)	1,209 (35.47)
	IIB	135 (21.91)	995 (29.19)
	IIIA	184 (29.87)	866 (25.41)
	IIIB	20 (3.24)	47 (1.39)
	IIIC	181 (29.38)	291 (8.53)
Differentiate	Moderately differentiated	305 (49.1)	2,046 (60.03)
	Poorly differentiated	263 (42.69)	848 (24.88)
	Undifferentiated	9 (1.46)	10 (0.29)
	Well differentiated	39 (6.3)	504 (14.78)
Grade	1	39 (6.33)	504 (14.78)
	2	305 (49.55)	2,046 (60.03)
	3	263 (42.69)	848 (24.88)
	4	9 (1.46)	10 (0.29)
A Stage	Distant	35 (5.68)	57 (1.67)
	Regional	581 (94.31)	3,351 (58.32)
Estrogen status	Negative	108 (17.53)	161 (4.72)
	Positive	508 (82.46)	3,247 (95.27)
Progesterone status	Negative	204 (33.11)	449 (14.49)
	Positive	412 (66.68)	2,914 (85.50)
Age		55.15±9.6	53.75±8.8
Tumor size		37.14±24.11	29.26±20.3
Regional node examined		15.01±8.4	14.23±8.02
Regional node positive		7.24±7.3	3.6±2.3
Survival months		45.61±23.96	75.94±19.8

Table 2 Multivariate analysis for independent risk factors associated with breast cancer.

Predictor	Estimate	SE	Z	p	Odds ratio	95% Confidence Interval	
						Lower	Upper
Intercept	2.94326	0.97116	3.031	0.002	18.978	2.8287	127.32
Age	0.02763	0.0065	4.253	<0.001	1.028	1.015	1.041
Race:	Black Ref						
Other	-0.90096	0.29066	-3.1	0.002	0.406	0.2298	0.718
White	-0.45649	0.18989	-2.404	0.016	0.634	0.4366	0.919
Marital Status:	DivRef						
Married	-0.16482	0.1663	-0.991	0.322	0.848	0.6122	1.175
Separated	0.4864	0.48814	0.996	0.319	1.626	0.6248	4.234
Single	-0.08819	0.20602	-0.428	0.669	0.916	0.6114	1.371
Widowed	0.04388	0.26033	0.169	0.866	1.045	0.6273	1.74
T Stage:	T1 Ref						
T2	0.38822	0.15162	2.56	0.01	1.474	1.0953	1.985
T3	0.66963	0.3111	2.152	0.031	1.954	1.0617	3.594
T4	1.43979	0.36581	3.936	<0.001	4.22	2.0602	8.643
N Stage:	N1 Ref						
N2	0.39872	0.14944	2.668	0.008	1.49	1.1116	1.997
N3	0.52133	0.2793	1.867	0.062	1.684	0.9742	2.912
Regional Node Examined	-0.03108	0.00804	-3.866	<0.001	0.969	0.9542	0.985
Reginal Node Positive	0.07811	0.01788	4.37	<0.001	1.081	1.044	1.12
Survival Months	-0.06141	0.00276	-22.272	<0.001	0.94	0.9354	0.946
Tumor Size	-0.00271	0.00469	-0.578	0.563	0.997	0.9882	1.007
Grade:	Ref						
1-anaplastic; Grade IV	-2.27296	0.80448	-2.825	0.005	0.103	0.0213	0.498
2-anaplastic; Grade IV	-1.68538	0.78332	-2.152	0.031	0.185	0.0399	0.861
3-anaplastic; Grade IV	-1.2529	0.78415	-1.598	0.11	0.286	0.0614	1.328
A Stage:	Ref						
Regional-Distant	0.17649	0.32326	0.546	0.585	1.193	0.6331	2.248
Estrogen Status:	Ref						
Positive-Negative	-0.38257	0.22699	-1.685	0.092	0.682	0.4372	1.064
Progesterone Status:	Ref						
Positive-Negative	-0.51744	0.15216	-3.401	<0.001	0.596	0.4423	0.803

Table 3 demonstrates the results of the prediction model from the SEER database after splitting the data set into a training set and testing set. Four classification models were developed and executed to the training data set, whereas the model's performance was checked using testing data. The performance of the various models was checked using various parameters like accuracy, sensitivity, specificity, and F1 score area under the curve (AUC).

$$\text{Accuracy} = \frac{TP+TN}{FP+FN+TP+TN}$$

Sensitivity is the proportion of accurately categorized high-risk samples to the total number of high-risk samples in the collection. The classifier's sensitivity is also known as the true-positive Rate (TPR), and it may be calculated using a formula as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

The true-negative rate (TNR) is the proportion of low-risk samples accurately identified as low-risk. The following formula may be used to calculate the specificity score:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

From the classification machine learning techniques, we evaluate the model performances using a test data set. The accuracy sensitivity specificity F1 score and AUC were calculated for each model. The accuracy of the decision tree was the best among all classification models.

Table 3. Performance matrix of classification models.

	Accuracy	Sensitivity	Specificity	F1 score	AUC
Random forest	0.898	0.892	0.856	0.877	0.866
KNN	0.891	0.883	0.769	0.876	0.766
Decision tree	0.914	0.908	0.916	0.907	0.858
Logistic regression	0.904	0.898	0.793	0.896	0.833

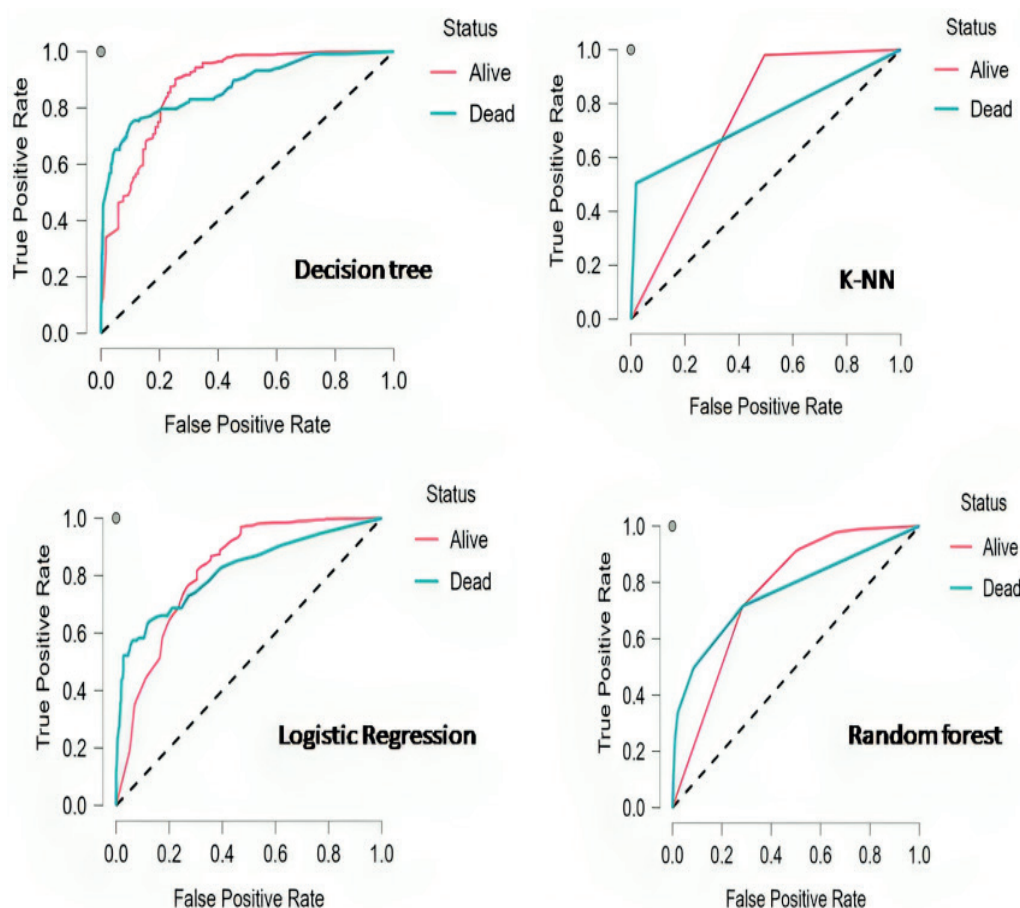


Figure 4. Roc curve of four classification models for breast cancer mortality prediction.

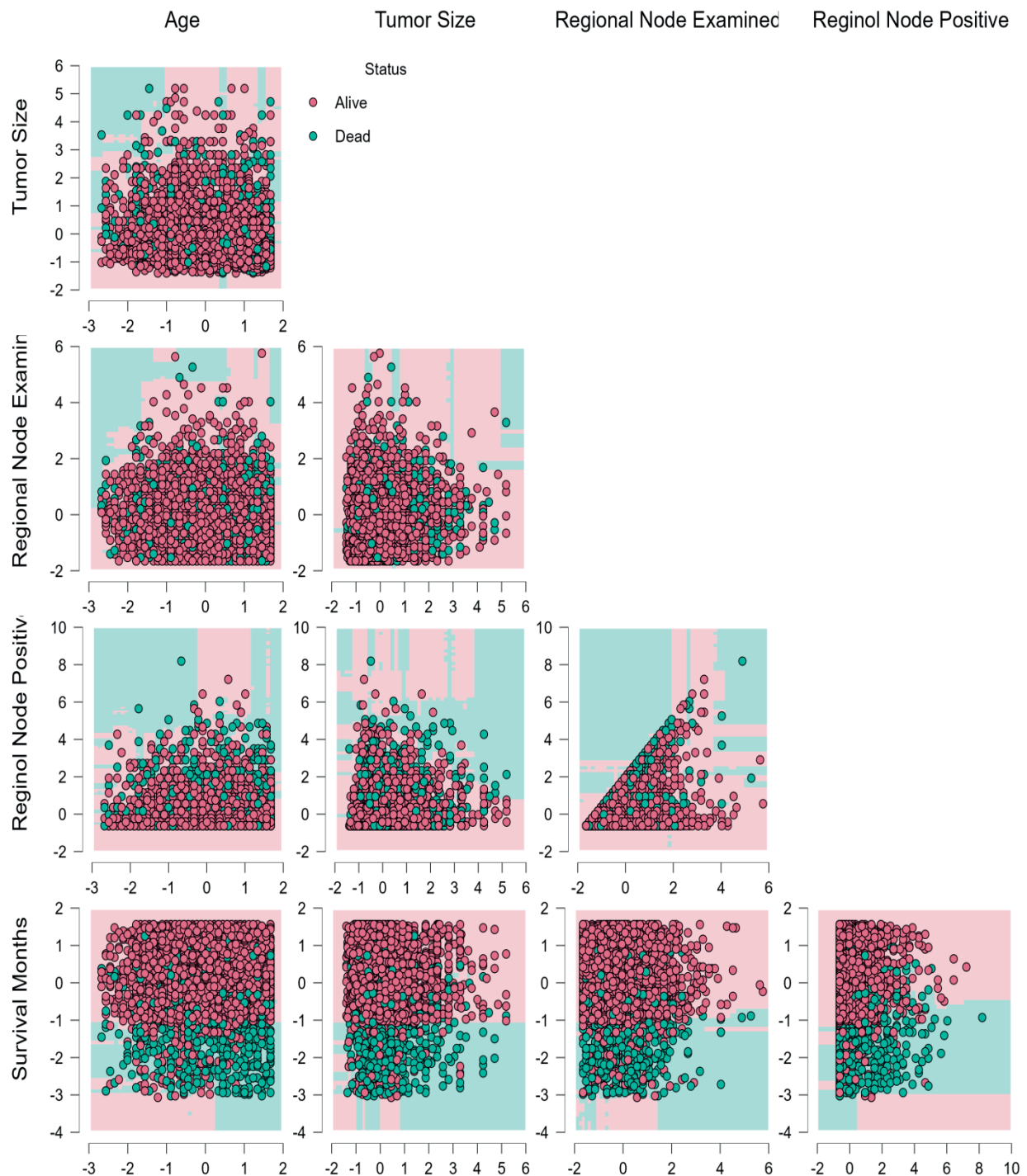


Figure 5. Decision boundary matrix of a classification model.

Discussion

To identify relevant characteristics from Wisconsin breast cancer data, methods for selecting features such as Recursive Feature Elimination, Forward Feature Selection, f-test, and correlation are utilized.²⁵ The Random Forest, KNN (k-Nearest-Neighbor), and Naive Bayes models are also employed to classify the Wisconsin dataset.²⁶ An evaluation of recent literature for categorizing breast cancer datasets was also conducted. To categorize breast cancer datasets, artificial intelligence approaches such as machine learning techniques and Deep Learning

techniques were used.²⁷ Fuzzy-based techniques are used for forecasting micro-RNA-regulated protein interaction networks and for evaluating *Arabidopsis Thaliana*.²⁸

Huang *et al.* created their prognostic models using Taiwan Cancer Registry (TCR) data and a multivariate Cox proportional hazard model.²⁹ For internal validation, a data-splitting mechanism was applied, and for prognostic continuous variables, a multivariable fractional polynomial approach was used. External validation was performed on subjects who were Asian, black, or white in the Surveillance, Epidemiology, and End Results (SEER)

database in the United States. Both internal and external datasets were used to assess model discrimination and calibration.

Ganggayah *et al.* and co-authors retrieved a large breast cancer data set. They applied machine learning algorithms to predict breast cancer survival.³⁰ Prediction algorithms have been developed utilizing decision trees, random forests, neural networks, extreme boost, logistic regression, and support vector machines to determine the main prognostic features of breast cancer survival rate.

Conclusion

From the current research and investigation, we may conclude that prediction models play a vital role in predicting the survival status of breast cancer patients. Prognostic models can be developed and validated using different feature selection methods. We developed and validated four machine learning algorithms and compared each other with different performance matrices. ML approaches have demonstrated an extraordinary capacity to enhance prediction and classification accuracy.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

NA

References

- [1] DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. *CA Cancer J Clin.* 2011; 61(6): 408-18. doi: 10.3322/caac.20134.
- [2] Ataollahi MR, Sharifi J, Paknahad MR, Paknahad A. Breast cancer and associated factors: a review. *J Med Life.* 2015; 8(Spec Iss 4): 6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319297>
- [3] Venturi S. Is there a role for iodine in breast diseases?. *The Breast.* 2001; 10(5): 379-82. doi.org/10.1054/brst.2000.0267
- [4] Heravi Karimovi M, Pourdehqan M, Jadid Milani M, Foroutan SK, Aieen F. Study of the effects of group counseling on quality of sexual life of patients with breast cancer under chemotherapy at Imam Khomeini Hospital. *J Maz Univ Med.* 2006; 16(54): 43-51. <http://jmums.mazums.ac.ir/article-1-133-en.html>
- [5] Wu R, Luo J, Wan H, Zhang H, Yuan Y, Hu H, Feng J, Wen J, Wang Y, Li J, Liang Q. Evaluation of machine learning algorithms for the prognosis of breast cancer from the Surveillance, Epidemiology, and End Results database. *PlosOne.* 2023; 18(1): e0280340. doi.org/10.1371/journal.pone.0280340
- [6] Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. *Eurasip J Adv.* 2016; 2016: 1-6.
- [7] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Comput Sci.* 2021; 2(3): 160. doi.org/10.1007/s42979-021-00592-x
- [8] Gardezi SJ, Elazab A, Lei B, Wang T. Breast cancer detection and diagnosis using mammographic data: Systematic review. *J Med Internet Res.* 2019; 21(7): e14464. doi.org/10.2196/14464
- [9] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* 2015; 13: 8-17. doi.org/10.1016/j.csbj.2014.11.005
- [10] Toth R, Schiffmann H, Hube-Magg C, Büscheck F, Höflmayer D, Weidemann S, Lebok P, Fraune C, Minner S, Schlomm T, Sauter G. Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clin Epigenetics.* 2019; 11: 1-5. doi.org/10.1186/s13148-019-0736-8
- [11] Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics.* 2008; 9(1): 1-0. doi.org/10.1186/1471-2105-9-319
- [12] Panda NR. A review on logistic regression in medical research. *Natl J community Med.* 2022; 13(04): 265-70. doi.org/10.55489/njcm.134202222
- [13] Zhou X, Liu KY, Wong ST. Cancer classification and prediction using logistic regression with Bayesian gene selection. *J Biomed Inform.* 2004; 37(4): 249-59. doi.org/10.1016/j.jbi.2004.07.009
- [14] Liu L. Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In: 2018 International Conference on Robots & Intelligent System (ICRIS) 2018 May 26 (pp. 157-160). doi.org/10.1109/ICRIS.2018.00049
- [15] Sharma M, Singh SK, Agrawal P, Madaan V. Classification of clinical dataset of cervical cancer using KNN. *Indian J Sci Technol.* 2016; 9(28): 1-5. doi: 10.17485/ijst/2016/v9i28/98380
- [16] Odajima K, Pawlovsky AP. A detailed description of the use of the kNN method for breast cancer diagnosis. In: 2014 7th International Conference on Biomedical Engineering and Informatics 2014 Oct 14 (pp. 688-692). doi.org/10.1109/BMEI.2014.7002861
- [17] Pawlovsky AP, Nagahashi M. A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. In: IEEE-EMBS International conference on biomedical and health informatics (BHI) 2014 Jun 1 (pp. 189-192). doi.org/10.1109/BHI.2014.6864336
- [18] Chaudhari P, Agarwal H, Bhateja V. Data augmentation for cancer classification in oncogenomics: an improved KNN based approach. *Evol Intell.* 2021; 14: 489-98. doi.org/10.1007/s12065-019-00283-w
- [19] Panda NR, Mahanta KL, Pati JK, Varanasi PR, Bhuyan R. Comparison of Some Prediction Models and their Relevance in the Clinical Research. *Int. J. Stat. Med.* 2023; 12: 12-19. doi.org/10.6000/1929-6029.2023.12.02
- [20] Venkatesan EV, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. *Indian J Sci Technol.* 2015; 8(29): 1-8.
- [21] Hamsagayathri P, Sampath P. Performance analysis

- of breast cancer classification using decision tree classifiers. *Int J Curr Pharm Res.* 2017; 9(2): 19-25. doi.org/10.22159/ijcpr.2017v9i2.17383
- [22] Hazra R, Banerjee M, Badia L. Machine learning for breast cancer classification with Ann and decision tree. In: 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2020 Nov 4 (pp. 0522-0527). doi.org/10.1109/IEMCON51383.2020.9284936
- [23] Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. *Comput Biol Med.* 2021; 128: 104089. doi.org/10.1016/j.compbimed.2020.104089
- [24] Tarawneh O, Otair M, Husni M, Abuaddous HY, Tarawneh M, Almomani MA. Breast cancer classification using decision tree algorithms. *Int J Adv Comput Sci Appl.* 2022; 13(4): 676-80
- [25] Dhanya R, Paul IR, Akula SS, Sivakumar M, Nair JJ. A comparative study for breast cancer prediction using machine learning and feature selection. In: 2019 International conference on intelligent computing and control systems (ICCS) 2019 May 15 (pp.1049-1055). doi.org/10.1109/ICCS45141.2019.9065563
- [26] Sharma S, Aggarwal A, Choudhury T. Breast cancer detection using machine learning algorithms. In: 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS) 2018 Dec 21 (pp. 114-118). doi.org/10.1109/CTEMS.2018.8769187
- [27] Abdulla SH, Sagheer AM, Veisi H. Breast cancer classification using machine learning techniques: A review. *TURCOMAT.* 2021; 12(14): 1970-9.
- [28] Manikandan P, Ramyachitra D, Nandhini R. Fuzzy based algorithms to predict MicroRNA regulated protein interaction pathways and ranking estimation in *Arabidopsis thaliana*. *Gene.* 2019; 692: 170-5. doi.org/10.1016/j.gene.2018.12.066
- [29] Huang CC, Chan SY, Lee WC, Chiang CJ, Lu TP, Cheng SH. Development of a prediction model for breast cancer based on the national cancer registry in Taiwan. *Breast Cancer Res.* 2019; 21(1): 1-9. doi.org/10.1186/s13058-019-1172-6
- [30] 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: 1-7. doi.org/10.1186/s12911-019-0801-4
- [31] Chen MT, Sun HF, Zhao Y, Fu WY, Yang LP, Gao SP, Li LD, Jiang HL, Jin W. Comparison of patterns and prognosis among distant metastatic breast cancer patients by age groups: a SEER population-based analysis. *Sci Rep.* 2017; 7(1): 9254. doi.org/10.1038/s41598-017-10166-8