

Research Article

# Statistical Test-Based Feature Selection and Classification Techniques for Breast Cancer Data

Murfia Rahman Muna, Md. Alamgir Sarder\* 

Statistics Discipline, Khulna University, Khulna, Bangladesh

## Abstract

Breast cancer is a disease that affects the majority of women and it is the second most common cause of death among women globally. Medical scientists have proven that there are a vast number of genes that are responsible for breast cancer. Among them, all genes are not equally responsible. Therefore, the most relevant and informative genes are needed to find out to control the disease. The objectives of our study are: (i) To find the most informative and significant genes using different statistical test-based feature selection techniques (FST) as well as find the best classifier and (ii) To validate our experimental results using a simulated dataset. The breast cancer dataset is a benchmark dataset provided by Kent Ridge Biomedical Data Repository, USA. In our study, we have used different statistical test-based feature selection techniques such as the t-test and Wilcoxon signed rank sum (WCSRS) test. Naïve Bayes (NB), Adaboost (AB), linear discriminant analysis (LDA), artificial neural network (ANN), k-nearest neighbor (KNN), and random forest (RF) are treated as classification techniques. Our analysis included 24,188 genes and 97 patients. Among them, 46 patients were with cancer and 51 were in control. We considered 70% of the dataset as a training set and the rest is a test set and repeated this procedure about 1000 times. Among all the combinations of FST and classification techniques t-test-based Naive Bayes classifier gives us the highest classification accuracy. The analysis of our study indicates that the integration of t-test-based FST and Naïve Bayes classifier produces the maximum classification accuracy.

## Keywords

Breast Cancer, Feature Selection, Machine Learning, Classification

## 1. Introduction

Undoubtedly, cancer is regarded as one of the most feared diseases all over the world in the recent time. Cancer is a condition in which somebody's cells grow destructively and may spread to the other organs of the body [1]. The uncontrolled and abnormal divisions of cells cause this detrimental disease [2]. According to the latest global cancer statistics from GLOBOCAN 2020, there were projected to be 19.3 million new cases of cancer and almost 10.0 million cancer-related deaths worldwide in 2020 [3] and as per projec-

tions cancer stood as the second leading cause of mortality in the United States in 2022, resulting in approximately 609,360 deaths [4].

Breast cancer is one of the greatest malignant conditions that arises from the abnormal growth of cells in the breast tissue [5]. It can present itself as a lump, alteration in shape, or presence of fluid, among other symptoms. Breast cancer is a disease that affects the majority of women and it is the second leading factor of mortality among women globally [6].

\*Corresponding author: [mdalamgir592@gmail.com](mailto:mdalamgir592@gmail.com) (Md. Alamgir Sarder)

Received: 19 June 2025; Accepted: 7 July 2025; Published: 28 August 2025



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

According to the World Health Organization, the impact of breast cancer on women is 2.1 million per year. In 2020, there were about 2.3 million women were found to have breast cancer and 6,85,000 deaths worldwide [7]. This situation has motivated us to work with breast cancer.

The number of breast cancer cases are increasing day by day at an alarming rate because of different reasons. However, it is possible to be cured if the cancer is detected at an early stage and treated properly. To provide better treatment to the patients, it is important to predict precisely and identify the most informative genes of breast cancer. Many researchers suggested choosing the most relevant genes before performing classification to obtain promising results [8]. The accuracy of classification is increased by removing a vast number of irrelevant genes [9]. Much research has previously been done on feature selection techniques using data from microarray gene selection. Researchers believe that as microarray analysis matures, it will have a substantial impact on our capacity to investigate the genetic alterations linked to the genesis and progression of cancer. Microarray gene expression refers to the process of measuring the levels of gene expression for thousands of genes simultaneously across different samples, typically represented in a gene expression matrix where rows represent genes and columns represent samples [10]. This technology allows researchers to study how genes are turned on or off in response to various condi-

tions, toxins, or time points during biological processes, providing insights into up-regulation, down-regulation, and co-regulation of genes [11].

Our study aims to identify the most relevant genes for breast cancer by applying different statistical test-based feature selection techniques (FST) such as t-test and WCSRS test. Next, different classification techniques are applied to classify breast cancer and compare those classifiers to choose the best one based on accuracy (ACC) and area under the curve (AUC). Additionally, we want to validate our results with a simulated dataset.

## 2. Materials and Methods

### 2.1. Data Sources

In our study, we have used the breast cancer dataset which was provided by Kent Ridge Biomedical Data Repository, USA [12]. It is a benchmark dataset. Affymetrix Oligonucleotide complementary array used to analyze gene expression samples where more than 24,400 human genes. The data are collected at the gene expression level. A data matrix is used for contracting the gene expression matrix. The data matrix is given in Table 1.

**Table 1.** Data matrix for breast cancer dataset.

| Genes     | Gene 1 | Gene 2 | ..... | Gene 24,481 |
|-----------|--------|--------|-------|-------------|
| Sample 1  |        |        |       |             |
| Sample 2  |        |        |       |             |
| .....     |        |        |       |             |
| Sample 97 |        |        |       |             |

### 2.2. Data Management

The breast cancer data we are working with consists of 97 patients as our sample and each sample provides 24,481 genes responsible for breast cancer and a classification variable is the dependent variable expressing whether the patient is controlled or the patient is with cancer. However, in the dataset, some variables provide constant values for all the samples which hinders the calculation and analysis process. To solve the problem, we had to remove those variables with constant values. After removing the variables, we have the dataset consisting of 24,188 variables with individual respective values.

### 2.3. Overview of the Proposed Computational Method

The most informative features are used to train classifiers, which improves performance. Choosing pertinent features increases learning effectiveness, which frequently results in improved generalization and classification accuracy. Firstly, we have to normalize the breast cancer data to prevent the biasness. Next, we have to use two statistical test-based feature selection methods t-test and WCSRS-test to extract the most informative and significant genes. Then, we need to split the cancer dataset into two groups: the training set (70%) and the test set (30%). We have used six most classifiers (AB, NB, ANN, LDA, KNN, and RF) to classify the patients as

cancer vs. control. Classifiers are chosen on the basis of their effectiveness on this micro-array gene expression dataset i.e breast cancer. Naïve Bayesian is very effective on features are mostly independent of dataset which provides fast and moderate accuracy [13, 14]. An ensemble technique like AdaBoost that improves classification by combining weak learners, enhancing sensitivity and AUC [15]. LDA performs well in low-dimensional spaces and equal class covariance [16, 17]. ANN can capture complex non-linear relationships whereas KNN is a non-parametric technique and sensitive to irrelevant features and choice of cluster [18, 19]. Non-linearity and their interaction handle accurately by RF through ensemble of decision trees which tends to produce stable accuracy and high AUC for biomedical datasets [20]. After that, the six most important classifiers are adopted to classify the patients as cancer vs. control. We have used these different classifiers to estimate the training parameters. Later these parameters are used in a test set to predict breast cancer. The performance of the classifier is calculated through the given formula-

$$Accuracy(ACC) = \frac{TP+TN}{TP+TN+FP+FN} \quad [21]$$

Whereas, TP=True positive, TN=True negative, FP=False positive, FN=False negative.

And we use predicted probabilities to make ROC curve by calculation of TPR and FPR at various threshold, and then compute the area under curve [22]. The iteration method was used to obtain better and more reliable results for each classification technique. And at last, we use the mean values of the iteration results.

## 2.4. Gene Expression Data Normalization

We standardized our micro-array gene expression of breast cancer dataset and simulated data set. To accomplish this task, we used a standardized equation, given below:

$$Z = \frac{X - \mu}{\sigma}$$

Where  $X$  is said to be the variable that is normalized,  $\mu$  is the arithmetic mean,  $\sigma$  is the standard deviation of  $X$  and  $Z$  is the standardized variable whose values lie in 0 and 1.

## 2.5. Validation of Our Results

We consider another simulated dataset for checking the validation of the experimental result. Here the simulated dataset is generated by keeping the same number of genes (column) and patients (row). Each genes have the same mean and variance as breast genes and the dataset is created from normal distribution.

Mathematically, let the original gene expression matrix be,  $X_{real} \in \mathbb{R}^{97 \times 24,188}$  where 97 is the number of patients, and 24,188 is the number of genes. Let,  $x_{ij}$  denote the expres-

sion level of gene  $j$  for patient  $i$ .

For each gene  $j$ ,  $j \in \{1, 2, \dots, 24,188\}$ , we compute the sample mean and sample variance from the real datasets as

$$\mu_j = \frac{1}{97} \sum_{i=1}^{97} x_{ij} \quad \sigma_j^2 = \frac{1}{96} \sum_{i=1}^{97} (x_{ij} - \mu_j)^2$$

Now the simulated dataset  $X_{sim} \in \mathbb{R}^{97 \times 24,188}$  was generated by drawing each element  $x_{ij}^{sim}$  from an independent univariate normal distribution with the corresponding gene-specific mean and variance:

$$x_{ij}^{sim} \sim N(\mu_j, \sigma_j^2), \text{ for all } i=1, \dots, 97 \text{ and } j=1, \dots, 24,188.$$

## 3. Results

### 3.1. Figuring Out the Best Classification and Feature Selection Method

One of the main objectives of our experiment was to identify the relevant and significant genes responsible for breast cancer. Figure 1. shows the number of significant and relevant genes at different p-values using two feature selection techniques: t-test and WCSRS-test. The t-test provides 2265 ( $p=0.05$ ), 902 ( $p=0.01$ ), 587 ( $p=0.005$ ), and 218 ( $p=0.001$ ) genes, on the contrary, WCSRS provides 3829 ( $p=0.05$ ), 1530 ( $p=0.01$ ), 1030 ( $p=0.005$ ), and 381 ( $p=0.001$ ) genes. It is seen that the number of significant genes dropped concurrently for the both test as the p values dropped.



**Figure 1.** Number of important genes using t-test and WCSRS-test of microarray data of breast cancer dataset.

Table 2 displays the results of the t-test and WCSRS-test, as well as the changes in the mean accuracy of six classifiers, together with the corresponding p-values for breast cancer. The accuracy of 87.96% (91.18% AUC) is achieved by combining a Naïve Bayesian classifier with a t-test, even though the t-test alone selects just 2265 genes. However, with a combined WCSRS based AB classifier score of accuracy drops to 71.98% (79.89% AUC).

### 3.2. Verification of the Suggested Approach

By utilizing the mean and variance of the associated 24,188 genes in the breast cancer dataset, we generate 97 observations for 24,188 genes, following a normal distribution, in order to verify the proposed method. The recommended computational approach's validation is discussed in Table 3. According to the

results, 351 genes were selected using the t-test ( $p=0.05$ ), and NB achieved the best classification accuracy (84.21%) and area under the curve (94.64%). Whereas, AB ( $p=0.05$ ) yields the worst classification accuracy (68.78%) and AUC (76.88%) as compared to NB. Consequently, the best classification accuracy is achieved when t-test and NB-based classifier are used together.

**Table 2.** Changes in six classifiers' mean accuracy (ACC) and area under curve (AUC) versus t-test and WCSRS-test p-values of breast cancer microarray data.

| Tests      | P-value | Genes | Measure | AB    | ANN   | KNN   | LDA   | RF    | NB*   |
|------------|---------|-------|---------|-------|-------|-------|-------|-------|-------|
| t-test     | 0.05    | 2265  | ACC     | 81.00 | 81.84 | 82.57 | 86.21 | 86.79 | 87.96 |
|            |         |       | AUC     | 86.00 | 86.05 | 87.35 | 88.26 | 88.66 | 91.18 |
|            | 0.01    | 902   | ACC     | 79.98 | 80.00 | 82.76 | 84.83 | 85.00 | 85.26 |
|            |         |       | AUC     | 82.56 | 83.59 | 85.22 | 86.27 | 88.68 | 89.42 |
|            | 0.005   | 587   | ACC     | 78.76 | 79.62 | 81.00 | 82.76 | 84.62 | 85.97 |
|            |         |       | AUC     | 82.21 | 82.80 | 86.14 | 86.38 | 87.24 | 88.87 |
|            | 0.001   | 218   | ACC     | 45.65 | 78.03 | 80.09 | 82.07 | 82.00 | 83.59 |
|            |         |       | AUC     | 81.00 | 81.57 | 85.32 | 84.84 | 85.94 | 86.88 |
|            | 0.05    | 3829  | ACC     | 81.02 | 81.03 | 82.41 | 84.66 | 84.75 | 85.31 |
|            |         |       | AUC     | 80.45 | 80.76 | 83.78 | 80.65 | 82.19 | 86.33 |
|            | 0.01    | 1530  | ACC     | 80.11 | 80.31 | 81.93 | 81.97 | 82.09 | 84.45 |
|            |         |       | AUC     | 81.87 | 82.59 | 86.00 | 86.08 | 86.66 | 86.72 |
| WCSRS test | 0.005   | 1030  | ACC     | 79.56 | 79.38 | 80.59 | 82.93 | 82.97 | 83.55 |
|            |         |       | AUC     | 80.74 | 81.31 | 82.30 | 84.17 | 84.51 | 85.65 |
|            | 0.001   | 381   | ACC     | 71.98 | 72.07 | 77.59 | 78.65 | 78.72 | 80.91 |
|            |         |       | AUC     | 79.89 | 80.93 | 81.92 | 82.97 | 83.00 | 84.31 |

## 4. Discussion

A novel approach to patient categorization in breast cancer was introduced in this research, in which ninety-six systems were developed by combining six classifiers (NB, RF, LDA, KNN, ANN, and AB) with two feature selection methods (t-test and WCSRS test) for each combination.

Every possible combination of FST and classifier was evaluated for classification accuracy and AUC after a series of stages. We chose the most important genes with FST at different levels of p-values (0.05, 0.01, 0.005, and 0.001). Using the findings of various FSTs and classification algorithms, we were able to arrive at the unique conclusion that the FST and classifier based on t-test ( $p=0.05$ ) performed the best. In comparison to other efforts, our suggested strategy significantly improves upon statistical test-based feature selection.

Prior research by the majority of the authors established that logistic regression and support vector machines were the most effective classifiers [23, 24]. None of them, however, had employed a statistical test to identify a breast cancer trait that was particularly significant. Instead of this they used LASSO, Sequential Forward Selection [23], Relief and Pearson algorithms [25]. A paper introduces three metaheuristic feature selection algorithms: Gravitational Search Optimization Algorithm (GSOA), Emperor Penguin Optimization (EPO), and an integrated (hGSEPO) algorithm for breast cancer classification [26]. Maniruzzaman et al (2019) extracted useful features from colon microarray gene expression data using a number of statistical tests-based feature selection methods, including the t-test, WCSRS test, Kruskal-Wallis (KW), and F-test [27]. For instance, Sharma et al. (2017) used multiple machine learning models on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset and discovered that their predictive

classifier had an accuracy of around 80-90% without any FSTs [28].

Islam et al. (2024) found an average higher accuracy using XGBoost for primary breast cancer dataset, indicating consistent performance across different folds, while Ashika (2025) emphasized the importance of incorporating feature selection approaches such as t-tests to increase classification accuracy and reduce model overfitting. The AUC value greater than 0.90 demonstrates the model's robustness in differentiating across classes, making it a reliable candidate for clinical decision support [29, 30].

One of our aims is to use statistical test-based feature se-

lection as it is making them suitable for initial screening in high-dimensional datasets (e.g., microarray data with thousands of genes) and also they measure real, quantifiable relationships (e.g., correlation or difference in means) between features and the target variable [31, 32]. Our proposed method achieved an accuracy of 87.96% (AUC 91.18%), being the highest accuracy to date using statistical test-based FST in breast cancer dataset. However, we can assert that the accuracy of our Naïve Bayes classifier much surpasses that of classifiers based on statistical test-based feature selection techniques. This prediction is corroborated by comparing it to the identical outcome of a simulated dataset.

**Table 3.** Changes in six classifiers' mean accuracy (ACC) and area under curve (AUC) versus t-test and WCSRS-test p-values of breast cancer simulated data.

| Tests      | P-value | Genes | Measure | AB    | ANN   | KNN   | LDA   | RF    | NB    |
|------------|---------|-------|---------|-------|-------|-------|-------|-------|-------|
| t-test     | 0.05    | 351   | ACC     | 80.21 | 80.47 | 82.15 | 82.58 | 83.52 | 84.21 |
|            |         |       | AUC     | 86.78 | 88.37 | 89.14 | 89.62 | 90.41 | 94.64 |
|            | 0.01    | 65    | ACC     | 79.08 | 80.89 | 81.28 | 82.05 | 82.55 | 83.94 |
|            |         |       | AUC     | 85.45 | 86.15 | 88.23 | 88.75 | 90.91 | 92.27 |
|            | 0.005   | 30    | ACC     | 71.98 | 72.89 | 76.11 | 78.22 | 79.39 | 80.17 |
|            |         |       | AUC     | 81.97 | 86.39 | 88.07 | 88.77 | 89.39 | 89.90 |
|            | 0.05    | 337   | ACC     | 76.56 | 78.84 | 80.00 | 81.78 | 82.22 | 82.78 |
|            |         |       | AUC     | 80.87 | 81.95 | 82.34 | 83.87 | 83.25 | 89.63 |
| WCSRS test | 0.01    | 64    | ACC     | 74.11 | 74.56 | 76.45 | 77.56 | 78.22 | 80.56 |
|            |         |       | AUC     | 78.67 | 80.16 | 81.86 | 84.37 | 84.80 | 86.03 |
|            | 0.005   | 28    | ACC     | 68.78 | 69.95 | 70.50 | 75.17 | 76.72 | 77.11 |
|            |         |       | AUC     | 76.88 | 77.37 | 82.31 | 80.32 | 83.22 | 84.57 |

## 5. Strengths and Limitations of the Study

Our study serves as an accurate predictor of breast cancer diseases using a high-risk stratification technique. There are 97 patients in the breast cancer dataset, divided into two categories: cancer and control. Our study demonstrated that statistical test based, t-test, the feature selection technique with NB-based classifier has the highest statistical performance and the best classification accuracy. As cancer is very expensive to medication, this study will provide highest prediction accuracy with limited significant genes by a ML-based classifier that will help policy maker (pharmacist) to control those genes via different medication. Also researchers, authorities, the government, and other decision-makers who are interested in this research topic will be benefited. After finding the most significant genes responsible for breast cancer by the feature selection techniques and the ML-based classifiers, we will hopefully try to suggest the

remedies to focus on the factors and the most robust classifier for prediction. To get a better performance, other feature selection methods such as F-test, KW test etc can be used. This is also true for classification techniques like Support vector machine (SVM), Classification and Regression Tree (CART), Logistic Regression etc.

## 6. Conclusion

This work demonstrated a comprehensive assessment of breast cancer gene expression using two main criteria. At first, significant genes have been selected by statistical test-based feature selection techniques (t-test & WCSRS test). And then a variety of classifiers were employed in order to determine which one best predicted breast cancer. Six classification method such as NB, RF, LDA, KNN, ANN & AB, the study has provided a comparative accuracy. t-test based Naïve

Bayesian classifier provides highest accuracy (ACC) and AUC. Simulated results have validated our results.

## Abbreviations

|        |                                                              |
|--------|--------------------------------------------------------------|
| FST    | Feature Selection Techniques                                 |
| USA    | United State of America                                      |
| WCSRS  | Wilcoxon Signed Rank Sum                                     |
| NB     | Naïve Bayes                                                  |
| AB     | Adaboost                                                     |
| LDA    | Linear Discriminant Analysis                                 |
| ANN    | Artificial Neural Network                                    |
| KNN    | K-nearest Neighbor                                           |
| RF     | Random Forest                                                |
| ACC    | Accuracy                                                     |
| AUC    | Area Under Curve                                             |
| LASSO  | Least Absolute Shrinkage and Selection Operator              |
| GSOA   | Gravitational Search Optimization Algorithm                  |
| EPO    | Emperor Penguin Optimization                                 |
| hGSEPO | Hybrid Gravitational Search and Emperor Penguin Optimization |
| ML     | Machine Learning                                             |
| SVM    | Support Vector Machine                                       |
| CART   | Classification and Regression Tree                           |

## Ethical Approval

No ethical approval is required for this dataset.

## Funding

No fund received for this project.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

- [1] Jones PA, Baylin SB. The Epigenomics of Cancer. *Cell* [Internet]. 2007 Feb 23 [cited 2025 Apr 13]; 128(4): 683-92. Available from: <https://www.cell.com/action/showFullText?pii=S0092867407001274>
- [2] Dasari S, Wudayagiri R, Valluru L. Cervical cancer: Biomarkers for diagnosis and treatment. *Clin Chim Acta*. 2015 May 20; 445: 7-11.
- [3] Sung H, Ferlay J, ... RSC a cancer journal, 2021 undefined. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Wiley Online Libr Sung, J Ferlay, RL Siegel, M Laversanne, I Soerjomataram, A Jemal, F BrayCA a cancer J Clin 2021•Wiley Online Libr [Internet]. 2021 May [cited 2024 Sep 8]; 71(3): 209-49. Available from: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660>
- [4] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin* [Internet]. 2022 Jan 1 [cited 2025 Apr 13]; 72(1): 7-33. Available from: <https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21708>
- [5] Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk Factors and Preventions of Breast Cancer. *Int J Biol Sci* [Internet]. 2017 [cited 2025 Apr 13]; 13(11): 1387-97. Available from: <http://www.ijbs.com>
- [6] Altaf MM. A hybrid deep learning model for breast cancer diagnosis based on transfer learning and pulse-coupled neural networks. *Math Biosci Eng* 2021 55029 [Internet]. 2021 [cited 2025 Apr 13]; 18(5): 5029-46. Available from: <http://www.aimspress.com/article/doi/10.3934/mbe.2021256>
- [7] WHO. WHO EMRO | Breast Cancer Awareness Month 2022 | Campaigns | NCDs [Internet]. 2022 [cited 2025 Apr 13]. Available from: <https://www.emro.who.int/noncommunicable-diseases/campaigns/breast-cancer-awareness-month-2022.html?format=html>
- [8] D áz-Uriarte R, bioinformatics SA de AB, 2006 undefined. Gene selection and classification of microarray data using random forest. SpringerR D áz-Uriarte, S Alvarez Andr éBMC bioinformatics, 2006•Springer [Internet]. 2006 Jan 6 [cited 2024 Sep 8]; 7. Available from: <https://link.springer.com/article/10.1186/1471-2105-7-3>
- [9] Ruiz R, Riquelme J, Recognition JARP, 2006 U. Incremental wrapper-based gene selection from microarray data for cancer classification. ElsevierR Ruiz, JC Riquelme, JS Aguilar-RuizPattern Recognition, 2006•Elsevier [Internet]. 2006 [cited 2024 Sep 8]; Available from: <https://www.sciencedirect.com/science/article/pii/S0031320305004140>
- [10] Chen JJ, Chen CH. Microarray Gene Expression. 2003.
- [11] Gonz ález Calabozo JM, Pel áez-Moreno C, Valverde-Albacete FJ. Gene Expression Array Exploration Using  $\mathcal{K}$ -Formal Concept Analysis. *Lect Notes Comput Sci* (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. 2011 [cited 2024 Sep 10]; 6628 LNAI: 119-34. Available from: <https://www.infona.pl/resource/bwmata1.element.springer-e61fdfa4-8690-3607-ba62-9b62807c030f>
- [12] Zhu Z, Ong Y, Recognition MDP, 2007 undefined. Markov blanket-embedded genetic algorithm for gene selection. ElsevierZ Zhu, YS Ong, M DashPattern Recognition, 2007 Elsevier [Internet]. 2007 [cited 2024 Sep 8]; Available from: <https://www.sciencedirect.com/science/article/pii/S0031320307000945>
- [13] Swaminathan M, Bhatti OW, Guo Y, Huang E, Akinwande O. Bayesian Learning for Uncertainty Quantification, Optimization, and Inverse Design. *IEEE Trans Microw Theory Tech*. 2022 Nov 1; 70(11): 4620-34.

- [14] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* [Internet]. 2005 Jun 1 [cited 2025 Jul 3]; 34(2): 113-27. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S093365704001010?via%3Dihub>
- [15] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J Comput Syst Sci* [Internet]. 1997 Aug 1 [cited 2025 Jul 3]; 55(1): 119-39. Available from: <https://www.sciencedirect.com/science/article/pii/S00220009791504X>
- [16] McLachlan GJ. Discriminant Analysis and Statistical Pattern Recognition. 2004; 552.
- [17] Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* [Internet]. 2015 Jan 1 [cited 2025 Jul 3]; 13: 8-17. Available from: <https://www.sciencedirect.com/science/article/pii/S2001037014000464>
- [18] Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks* [Internet]. 2006 May 1 [cited 2025 Jul 3]; 19(4): 408-15. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S0893608005002844>
- [19] Aha DW, Kibler D, Albert MK, Quinlan JR. Instance-based learning algorithms. *Mach Learn* 1991 61 [Internet]. 1991 Jan [cited 2025 Jul 3]; 6(1): 37-66. Available from: <https://link.springer.com/article/10.1007/BF00153759>
- [20] Chaurasia DV, Pal S. A Novel Approach for Breast Cancer Detection Using Data Mining Techniques. 2014 Jun 29 [cited 2025 Jul 3]; Available from: <https://papers.ssrn.com/abstract=2994932>
- [21] Powers DMW, Ailab. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020 Oct 11 [cited 2025 Jul 3]; Available from: <https://arxiv.org/pdf/2010.16061>
- [22] Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* [Internet]. 2006 Jun 1 [cited 2025 Jul 3]; 27(8): 861-74. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S01676550500303X>
- [23] Akkur E, TURK F, Eroglu O. Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization. *Comput Syst Sci Eng* [Internet]. 2022 Nov 3 [cited 2025 Jun 15]; 45(2): 1017-31. Available from: <https://www.techscience.com/csse/v45n2/50445/html>
- [24] Naji MA, Filali S El, Aarika K, Benlahmar EH, Abdelouahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Comput Sci* [Internet]. 2021 Jan 1 [cited 2025 Jun 15]; 191: 487-92. Available from: <https://www.sciencedirect.com/science/article/pii/S1877050921014629>
- [25] López NC, García-Ordás MT, Vitelli-Storelli F, Fernández-Navarro P, Palazuelos C, Alaiz-Rodríguez R. Evaluation of Feature Selection Techniques for Breast Cancer Risk Prediction. *Int J Environ Res Public Health* [Internet]. 2021 Oct 1 [cited 2025 Jun 15]; 18(20): 10670. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8535206/>
- [26] Singh LK, Khanna M, Singh R. Efficient feature selection for breast cancer classification using soft computing approach: A novel clinical decision support system. *Multimed Tools Appl* [Internet]. 2024 Apr 1 [cited 2025 Jun 15]; 83(14): 43223-76. Available from: <https://link.springer.com/article/10.1007/s11042-023-17044-8>
- [27] Maniruzzaman M, Jahanur Rahman M, Ahammed B, Abedin MM, Suri HS, Biswas M, et al. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Comput Methods Programs Biomed* [Internet]. 2019 Jul 1 [cited 2025 Jun 15]; 176: 173-93. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S0169260718317681>
- [28] Sharma A, Kulshrestha S, Daniel S. Machine learning approaches for breast cancer diagnosis and prognosis. 2017 Int Conf Soft Comput its Eng Appl Harnessing Soft Comput Tech Smart Better World, *icSoftComp* 2017. 2017 Jul 2; 2018-January: 1-5.
- [29] Islam T, Sheakh MA, Tahsin MS, Hena MH, Akash S, Bin Jordan YA, et al. Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. *Sci Rep* [Internet]. 2024 Dec 1 [cited 2025 Jul 3]; 14(1): 1-17. Available from: <https://www.nature.com/articles/s41598-024-57740-5>
- [30] Ashika T, Grace GH. Enhancing Classification Performance through Rough Set Theory Feature Selection: A Comparative Study across Multiple Datasets. *Eur J Pure Appl Math* [Internet]. 2025 May 1 [cited 2025 Jul 3]; 18(2): 5934-5934. Available from: <https://www.ejpm.com/index.php/ejpm/article/view/5934>
- [31] Guyon I, Elisseeff A. An Introduction of Variable and Feature Selection. *J Mach Learn Res*. 2003; 1.
- [32] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* [Internet]. 2007 Oct 1 [cited 2025 Jun 15]; 23(19): 2507-17. Available from: <https://dx.doi.org/10.1093/bioinformatics/btm344>