Ahliya Journal of Allied Medico-Technology Science
[Published by Palestine Ahliya University]

Vol. 01 Issue 02 (2024)
ISSN: 3007-9489

The Analysis of Breast Cancer Classification Involves Utilizing Machine Learning (Ml) Techniques and Hyperparameter Adjustment - A Comparative Study

Mutaz Rasmi Abu Sara[1], Khaled Sabarna[2], Jawad H. Alkhateeb[3]

[1] IT Department, Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
✉ moutaz.a@paluniv.edu.ps

[2] Nursing Department, Faculty of Allied Medical Sciences, Palestine Ahliya University (Palestine)
✉ k.sabarna@paluniv.edu.ps

[3] Computer Engineering Department, College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Khobar, (Saudi Arabia)
✉ jalkhateeb@pmu.edu.sa

Received:13/09/2024 Accepted:03/11/2024 Published:15/12/2024

Abstract: ***This study aims to analyze and classify breast cancer (BC) cases using machine learning (ML) techniques and hyperparameter tuning. The BC dataset from the University of California (UCI) was utilized, which comprises 569 cases classified as malignant (M) and benign (B), with 32 features. The algorithms employed in the study included Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Decision Tree (DT), and Gaussian Naive Bayes (NB). The results indicated that the SVC algorithm performed the best, achieving an accuracy of 98% on the test set, along with a precision of 100%. Furthermore, all algorithms demonstrated high performance, reflecting the effectiveness of machine learning techniques in classifying breast cancer cases.***

Keywords: Breast Cancer (BC); Breast Cancer Diagnosis; Cancer Dataset; Machine Learning; Support Vector Classification (SVC); K-Nearest Neighbor Regression (KNN); And Logistic Regression (LR).

1. Introduction

Breast cancer (BC) is a malignant roughage tumor that develops in or around the breast tissue, primarily in the milk ducts and glands; however, benign breast lumps ty pically have smooth borders. When you push against them, they will move slightly. A malignant (BC) is one of the most common cancers among women worldwide, accounting for 23% of all female cancer cases [1]. Additionally, it is the primary cause of cancer-related deaths in low-resource countries. Women of all ages are more likely to get breast cancer in older age groups [2]. The fatality rate from breast cancer remains high despite the development of advanced screening tools. Breast cancer is the leading cause of cancer-related deaths among women aged 40–44. [3]. The risk is continually increasing even while high-income countries have seen considerable advances in survival [3]. Survival rates in low- and middle-income nations remain relatively low. Breast cancer is the most prevalent cancer in women, according to the National Cancer Registry (NCR).

Breast self-examination (BSE) was not a common practice; doctors comprised 31.3% of BSE practitioners, while midwives comprised 21.8% [4]. This implies that while health professionals perform BSE, only a few do it regularly. One kind of breast X-ray that can identify tissue anomalies, including cancerous growths, is a mammogram. It can identify breast cancer up to two years before a noticeable lump appears [5]. Mammograms may be used to check for breast cancer even in women who show no signs or symptoms of the disease. A breast tumor is an abnormal growth of tissues within the breast that can manifest itself in several ways, including changes in breast morphology, skin dimples, lump mass topology along the breast tissues, or the formation of red, scaly areas on the epidermal layer [6]. Breast cancer typically begins in the mammary glands and tissues responsible for nursing and aging in women. The fact that there are more than 18 different types of breast cancer is intriguing [7]. A biopsy of the suspected tumor will be performed to confirm the diagnosis of breast cancer. After a diagnosis, more medical tests are performed to determine the best course of treatment and evaluate the cancer's capacity to metastasize and spread outside (invasive or non-invasive) of the breast cells. Many different signs and symptoms can indicate breast cancer. A few of them include rosacea surrounding the breasts, flaking, peeling, or scaling of the pigmented area of the breast, a change in the size, appearance, or form of a breast, and a recent technique for inverted nipples called breast dimpling [8]. Machine learning for the analysis, classification, and prediction of breast cancer.

Early disease identification has become a more crucial topic in medical research in recent years due to the rapid population expansion. The rapid population growth is causing a significant increase in the risk of dying from breast cancer. In addition to helping medical professionals diagnose illnesses and provide a reliable, effective, and prompt response, an automated disease detection system reduces the risk of death [9]. To

diagnose illness, which reduces the risk of death and offers a reliable, effective, and speedy response, this study analyzes three supervised machine learning techniques: support vector classification (SVC), K-nearest neighbor regression (KNN), and logistic regression (LR).

2. Related Work

İlkuçar et al. (2014), introduced the UCI BC Dataset. Mainly, two types of algorithms for Artificial Neural Networks were used. Both the Back Propagation and the Harmony Search algorithms were used in training the Artificial Neural Networks feed-forward (ANN). The classification performance was tested in terms of precision, SSE, and regression parameters. The values of the back-propagation performance were obtained as 94.1/0.007/0.92 and Harmony Search 97.57/0.005/0.96 respectively [10].

Douangnoulack & Boonjing (2018), introduced the Principal Component Analysis (PCA) with the Wisconsin BC (WBC) dataset for a lossless data reduction technique with good classification performance. The goal is to find the best performance classier by giving minimal classification rules by employing PCA. The best accuracy among the three classifiers is the J48 decision tree classifier. The J48 decision tree is 97.36%, Minimized Error Pruning Tree 96.77%, and Random Tree 94.72% [11].

Bayrak and Ansari (2019), introduced the most popular techniques in ML techniques, both the SVM and the ANN in the Wisconsin BC Dataset. The comparison was done on the classification performance of these techniques to each other. It was concluded that the SVM classifier had the best percentage split accuracy of about 95% and the ANN of about 88% While using precision, accuracy, recall, and the ROC area. [12].

Yedjou et al., (2021), introduced a novel prediction diagnosis by using the computer-aided diagnosis system for classification of the BC by applying ML. Particularly, they discussed the concepts of ML and outlined its application in the classification of BC. By using various ML approaches, their findings revealed that among the 569 patients involved in this study, 63% were diagnosed with benign tumors and 37% were diagnosed with malignant tumors. Various feature characteristics were used such as radius, perimeter, area, texture, concavity, compactness, and concave points of the cell [13].

Alkhateeb et al., (2020), proposed various techniques in machine learning and deep learning for recognizing the Arabic handwritten text. These techniques can used in classifying the BC to obtain an excellent performance [14].

3. Methodology

3.1 Dataset:

Two forms of breast cancer were classified using the UCI Machine Learning Repository's breast cancer dataset [15]: benign (B) and malignant (M). 32 parameters that represent attributes associated with breast cancer, including size, shape, and degree of cellular differentiation, are included in this dataset, which comprises 569 categorized cases. Because of its thoroughness and dependability, this dataset was selected for use in breast cancer research. The 'id' column was eliminated as part of the data pretreatment process to get accurate findings because it had no useful information for prediction. By eliminating it, the accuracy of the model is improved, leaving 30 columns as inputs and one column as output. In terms of the result (diagnostic), 37.3% of the classifications were for benign tumors and 62.7% were for malignant tumors. The ratio between them is illustrated in Figure 1.

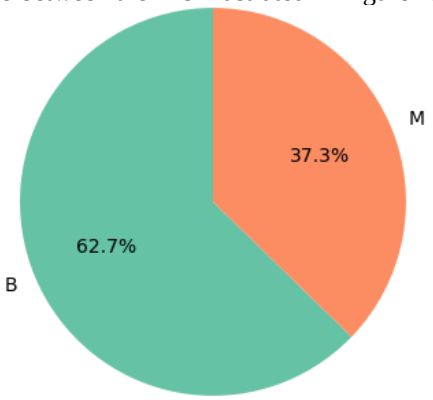

Figure 1. Distribution of Target Feature

3.2 Data Preprocessing

Target Encoding: The "diagnosis" column's categorical values were transformed into numerical values, where benign tumors were assigned a value of (0) and malignant tumors a value of (1). Since most algorithms need numerical values to detect patterns, this encoding makes machine learning easier.

Feature Scaling: The dataset was standardized using StandardScaler, which guarantees that every feature falls within the same range. Because scaling the data speeds up training and lessens bias brought on by variations in feature scales, this procedure helps distance-based algorithms like KNN and SVC perform better.

Data Splitting: T In order to preserve the relative distribution of benign and malignant tumors in both groups, the dataset was split into two groups: a training set (80%) and a testing set (20%) using stratified sampling. This method guarantees correct evaluation and improves the model's dependability.

3.3 Study Algorithms:

This study compares five different machine-learning classification algorithms:

1. **Logistic Regression (LR)**: This algorithm is commonly used for binary classification, creating a linear model to estimate probabilities, which helps determine the potential class for each case. For this

algorithm, the solver was set to 'liblinear' for faster training, providing high efficiency when dealing with small datasets, the equation for the Logistic Regression (LR) algorithm as follows [16]:

$$\hat{y} = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}} \quad (1)$$

2. **Support Vector Classifier (SVC)**: This algorithm is used to enhance classification effectiveness in high-dimensional spaces and is effective in non-linear cases due to the Radial Basis Function (RBF) kernel. For this algorithm, C was set to 1.0 to control the prediction power, where the value of C represents the confidence in the classification model. Gamma was adjusted to 0.1 to regulate the impact of individual points in space, ensuring model accuracy, the equation for the Support Vector Classifier (SVC) algorithm is as follows [16].

$$B_0 + (B_1.X_1) + (B_2.B_2) = 0 \quad (2)$$

3. **K-Nearest Neighbors (KNN)**: This algorithm relies on the neighborhood principle, making it simple and effective for classification based on nearby data points. For this study, K was set to 3, meaning that the model considers the three nearest neighbors to determine the class. This number was chosen after evaluating performance through hyperparameter tuning techniques, the class probabilities for binary classification can be calculated by computing the normalized frequency of samples belonging to each class among the *K* nearest neighbors of a new data instance as follows [17]:

$$p(class = 0) = \frac{count(class=0)}{count(class=0)+count(class=1)} \quad (3)$$

4. **Decision Tree (DT)**: This algorithm is used to provide a clear and easily understandable interpretation of predictions, facilitating decision-making. For this algorithm, max_depth was set to 5 to limit complexity, reducing the risk of overfitting and enhancing the model's generalization capability, figure 1 illustrates a basic decision tree model with a binary target variable *Y* (0 or 1) and two continuous predictors (X1, X2), both ranging from 0 to 1. As illustrated, a decision tree consists of nodes and branches, with its construction involving three key steps: splitting, stopping, and pruning [18].

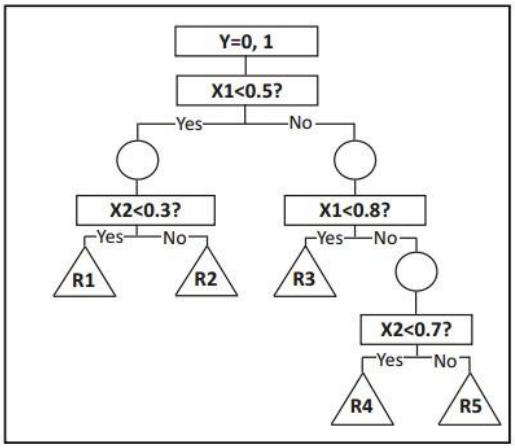

Figure 2. Basic decision tree based on a binary target label Y [13].

5. **Gaussian Naive Bayes (NB)**: This algorithm is effective for classifying large datasets and performs well with independent features, based on the assumption that each feature contributes independently to the class probability. Notably, this algorithm does not require hyperparameter tuning, relying instead on estimating the normal distribution of the features, Gaussian Naive Bayes assumes with probability (P(xi|y)) follows the Gaussian Distribution for each x_i within y_k, the equation for Gaussian Naive Bayes (NB) algorithm as follows [19].

$$P(x_i|y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

3.4 Model Evaluation:

To ensure the effectiveness of each model, several metrics were used to assess performance [20]:

- **Accuracy**: Calculated as the ratio of correctly predicted cases to the total number of cases. This metric is used to determine the model's success in making predictions, the formula for calculating accuracy as follows:

$$Accuaracy = \frac{(TP)+(TN)}{(TP)+(TN)+(FP)+(FN)} \quad (5)$$

- **Precision**: Represents the ratio of true positive results to all predicted positive results. It reflects the model's accuracy in correctly identifying malignant tumors, the formula for calculating precision as follows:

$$Precision = \frac{(TP)}{(TP)+(FP)} \quad (6)$$

- **Recall**: Shows the proportion of actual positive cases to true positive outcomes. It makes it easier to comprehend how well the model captures all positive cases, the formula for calculating precision as follows:

$$Recall = \frac{(TP)}{(TP)+(FN)} \quad (7)$$

- **F1 Score**: Shows the precision and recall harmonic mean. When the accuracy and efficacy of the model must be balanced, this score can be helpful, the formula for calculating precision as follows:

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (8)$$

Hyperparameters in the SVC and KNN algorithms were tuned using GridSearchCV, ensuring optimal model performance by evaluating a range of parameter combinations through cross-validation.

4. Results and Discussion:

The performance of five different classification algorithms—KNN, SVC, LR, DT, and NB—was evaluated using the breast cancer dataset to estimate the effectiveness of each model. The results for each algorithm were summarized according to various performance metrics, namely Accuracy, Precision, Recall, and F1-score, for both the training and testing sets, as illustrated in table 1:

Table 1. Comparison of Machine Learning Classification Algorithms on the Dataset

		KNN	SVC	LR	DT	NB
Accuracy	Train	0.98	0.98	0.98	0.99	0.94
	Test	0.94	0.98	0.96	0.95	0.92
Precision	Train	1.0	1.0	1.0	1.0	0.94
	Test	0.97	1.0	0.97	1.0	0.92
Recall	Train	0.96	0.96	0.96	0.98	0.90
	Test	0.86	0.95	0.93	0.88	0.86
	Train	0.98	0.98	0.98	0.98	0.92
	Test	0.91	0.97	0.95	0.94	0.89

The results presented in the previous table indicate that all algorithms achieved high accuracy on the training set, with values ranging between (94%) and (99%). Specifically, the Decision Tree achieved the highest accuracy rate at (99%), followed by the KNN, SVC, and LR algorithms, each recording an accuracy of (98%). In the testing set, SVC continued to excel, achieving an accuracy of (98%), followed by LR and KNN, with rates of (96%) and (94%) respectively, indicating the robustness of these models in making accurate predictions.

Regarding **Precision**, which reflects the ratio of true positive predictions to all positive predictions, the training set results were identical for KNN, SVC, and LR, each recording a precision of (100%). Meanwhile, the Decision Tree and Naive Bayes recorded precision values of (100%) and (94%) respectively. In the testing set, SVC achieved the highest precision of (100%), followed by KNN and LR at (97%), demonstrating a high efficiency in avoiding false positive errors.

For **Recall**, which reflects the model's ability to capture all actual positive cases, the training set results showed that KNN, SVC, and LR achieved high recall values ranging between (96%) and (98%), indicating their strong capability in identifying malignant tumors. However, the testing set results were lower, with KNN achieving (86%), SVC at (95%), and LR at (93%), suggesting some challenges faced by the models in detecting all positive cases.

As for the **F1 Score**, which is a comprehensive metric that considers both precision and recall, all models recorded high F1 values in the training set, with KNN, SVC, and LR all averaging (98%). In the testing set, SVC achieved the highest F1 Score of (97%), followed by LR at (95%) and DT at (94%). These results indicate that the studied models are not only accurate but also balanced in their output.

The findings suggest that the studied algorithms, particularly SVC, KNN, and LR, demonstrated excellent performance in classifying breast cancer cases, achieving high levels of accuracy, precision, recall, and F1 Score. However, it is noteworthy that Naive Bayes, despite its reasonable performance, remains less efficient in some metrics, figure 2 illustrates the confusion matrices for each algorithm.

Among the five algorithms used in this study, the SVC algorithm stands out as the best model based on performance measured by the four metrics. This is due to its strength as an algorithm that relies on the principle of separating classes in multi-dimensional space. The SVC algorithm employs hyperplane techniques to determine the boundaries that separate the classes. Thanks to its ability to operate in non-linear spaces using techniques such as the kernel trick, SVC can differentiate classes more accurately, even in complex data sets.

SVC demonstrated excellent performance in terms of accuracy and precision on the testing set, making it one of the most precise models. When it comes to precision, SVC achieved a score of (100%) in the testing set, indicating its ability to completely avoid false positive errors. These figures suggest that the model is not only accurate but also reliable in estimating positive cases.

There is also a notable balance between precision and recall, indicating that SVC proves to be a balanced model in terms of recall as well, having recorded a value of (95%) in the testing set. This means it was effective in identifying all true positive cases while maintaining a low level of false positives. This balance highlights the importance of SVC in medical applications, where reducing diagnostic errors is crucial.

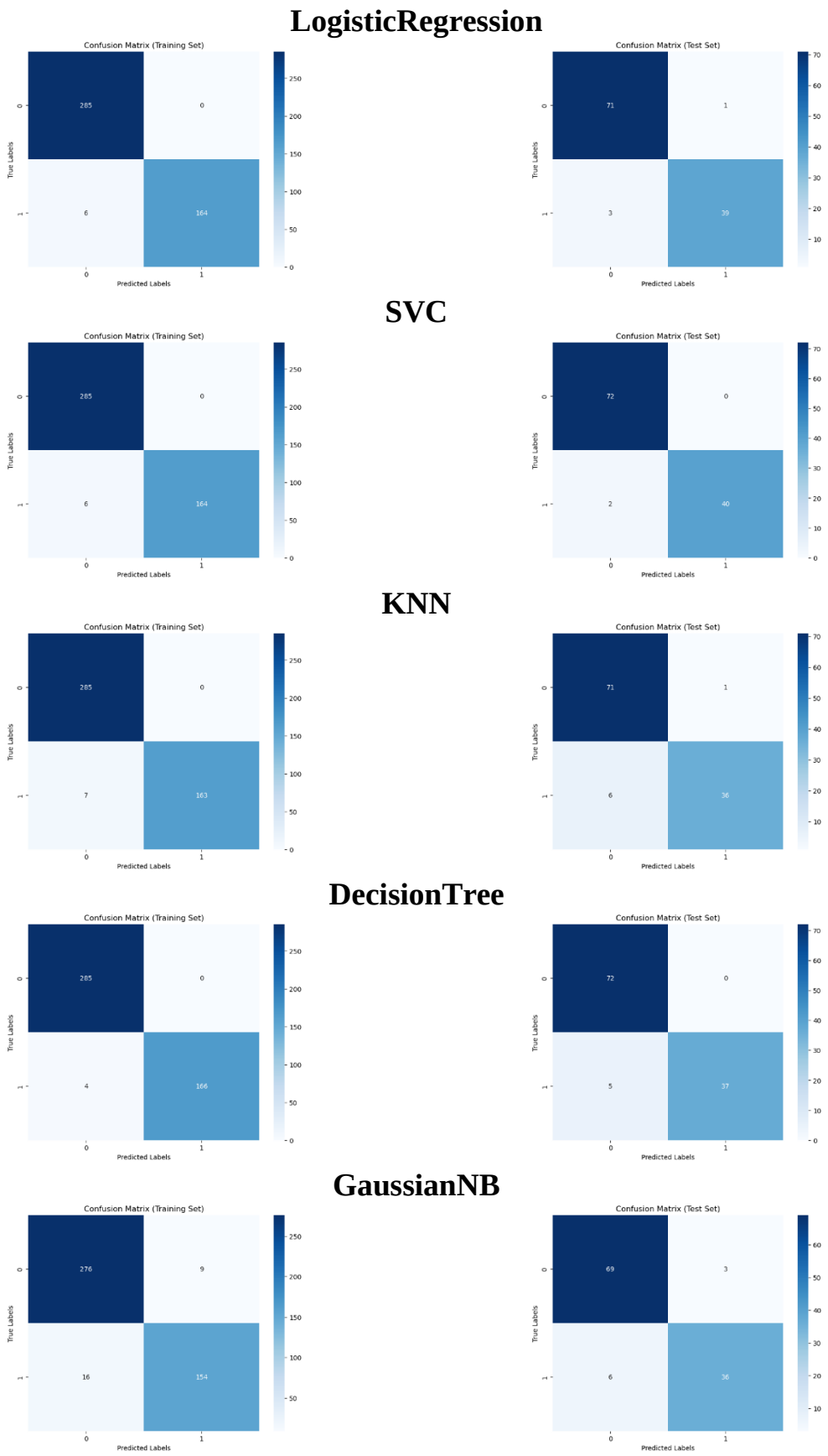

Figure 2. Confusion metrics for each algorithm

SVC also allows users to customize its parameters, such as the C parameter and the type of kernel transformation, enabling it to adapt to specific data characteristics. Using the appropriate kernel transformation can enhance model performance based on the nature of the data. In this study, SVC was effectively utilized to tune these parameters, contributing to performance improvement.

5. Conclusion:

The results indicate that machine learning techniques, particularly the SVC algorithm, provide effective tools for classifying breast cancer cases. This algorithm has shown excellent performance across all metrics used, making it a reliable model in medical applications. It is important to note that performance enhancement through parameter tuning plays a critical role, underscoring the importance of employing advanced methods in medical data analysis. This study recommends further research to explore other algorithms and new approaches to improve model accuracy in the fields of medicine and diagnosis.

References:

1. Parkins R, Fernández G, Yip CH. The spectrum of breast cancer in Malaysian women: Overview. World J Surg. 2007; 6:921-923.
2. Omotara CM, Mok T. Knowledge, perceptions, and attitudes of Hong Kong Chinese women on screening mammography and early breast cancer management. Breast J. 2012; 11:52-56.
3. Alwan, N. A. S., et al. "Knowledge, attitude and practice regarding breast cancer and breast self-examination among a sample of the educated population in Iraq." EMHJ-Eastern Mediterranean Health Journal, 18 (4), 337-345, 2012 (2012).
4. Alwan, N. A. S., et al. "Knowledge, attitude and practice regarding breast cancer and breast self-examination among a sample of the educated population in Iraq." EMHJ-Eastern Mediterranean Health Journal, 18 (4), 337-345, 2012 (2012).
5. Akpinar-Nustus W, Mikhail B. Factors associated with breast self-examination among Jordanian women. Pub Health Nurs. 2011;19:263-271.
6. Smith, Robert A., Vilma Cokkinides, and Otis W. Brawley. "Cancer screening in the United States, 2009: a review of current American Cancer Society guidelines and issues in cancer screening." CA: a cancer journal for clinicians 59.1 (2009): 27-41.
7. Jabbar, Meerja Akhil. "Breast cancer data classification using ensemble machine learning." Engineering & Applied Science Research 48.1 (2021).
8. Kabel, Ahmed M., and Fahad H. Baali. "Breast cancer: insights into risk factors, pathogenesis, diagnosis and management." Journal of Cancer Research and Treatment 3.2 (2015): 28-33.
9. Harmon, David M., et al. "Integrating online community support into outpatient breast cancer care: Mayo Clinic Connect online platform." Digital Health 7 (2021): 20552076211048979.
10. Islam, Md Milon, et al. "Breast cancer prediction: a comparative study using machine learning techniques." SN Computer Science 1 (2020): 1-14.
11. İlkuçar, M., Işik, A. H., & Çifci, A. 2014. Classification of breast cancer data with harmony search and backpropagation-based artificial neural network. In 2014 22nd signal processing and communications applications conference (SIU) (pp. 762-765). IEEE.
12. Douangnoulack, P., & Boonjing, V. 2018. Building minimal classification rules for BC diagnosis. In 2018 10th International Conference on Knowledge and Smart Technology (KST) (pp. 278-281). IEEE.
13. Bayrak, E. A., Kırcı, P., & Ensari, T. 2019. Comparison of machine learning methods for BC diagnosis. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). IEEE.
14. Yedjou, C. G., Tchounwou, S. S., Aló, R. A., Elhag, R., Mochona, B., & Latinwo, L. 2021. Application of Machine Learning Algorithms in BC Diagnosis and Classification. International Journal of Science Academic Research, 2(1), 3081–3086.
15. J. H. Alkhateeb, A. A. Turani, and A. A. Alsewari, "Performance of Machine Learning and Deep Learning on Arabic Handwritten Text Recognition," *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, Bangladesh, 2020, pp. 1-7.
16. M. Zwitter and M. Soklic. "Breast Cancer," UCI Machine Learning Repository, 1988. [Online]. Available: https://doi.org/10.24432/C51P4M.
17. Zalloum, H. N., Al Zeer, S., Manassra, A., Abu Sara, M. R. & Alkhateeb, J. H. (2022). “Breast Cancer Grading using Machine Learning Approach Algorithms”, Journal of Computer Science, Vol. 18, No 12, 2022. PP 1213-1218
18. Jason Brownlee, Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End, 2021.
19. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015 Apr 25;27(2):130-5. doi: 10.11919/j.issn.1002-0829.215044. PMID: 26120265; PMCID: PMC4466856.
20. H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," *2019 International Engineering Conference (IEC)*, Erbil, Iraq, 2019, pp. 165-170, doi: 10.1109/IEC47844.2019.8950650.
21. Željko Đ. Vujovic, “Classification Model Evaluation Metrics” International Journal of Advanced Computer Science and Applications (IJACSA), 12(6), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120670