Volume: 15 Issue: 06 Pages: 6229– 6238 (2024) ISSN: 0975-0290

Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms

Satish Choudhary
Department of CSE, Lakshmi Narain College of Technology Excellence (LNCTE), Bhopal
Email : satish_choudhary25@yahoo.com
Priyanka Singh
Department of CSE, Lakshmi Narain College of Technology Excellence (LNCTE), Bhopal
Email : priyankas@lnct.ac.in
Mann Mittal
Department of CSE-AIML, Lakshmi Narain College of Technology Excellence (LNCTE), Bhopal
Email : Mannmittal622@gmail.com
Gaurav Singh
Department of CSE-AIML, Lakshmi Narain College of Technology Excellence (LNCTE), Bhopal
Email : gaurav07singh10@gmail.com

---ABSTRACT--

Breast cancer is the main reason for mortality in women. It was very difficult to predict breast cancer in the early stages by doctors and pathologists. They need some automated tools to make an early prediction of cancer and diagnosis as soon as possible. Some research found that Machine learning (ML) algorithm helps them to take decisions and perform diagnosis based on data collected by the medical field. In this paper, we use various ML algorithms and classifiers like K-NearestNeighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF) to find the accurate result of cancer in less intervals of time. It has been found that Support Vector Machine has highest accuracy 98.83% and Random Forest has second highest accuracy with 98.24% among all other models.

Keywords: Machine Learning**,** K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF)

Date of Submission: April 06, 2024 Date of Acceptance: April 29, 2024

1. Introduction

Breast cancer is a prevalent cause of death, and it is the only type of cancer that is widespread among women worldwide [1]. It is the second most infectious disease in women. In 2023, there were 2.3 million women diagnosed with breast cancer and 685000 deaths globally according to World Health Organisation (WHO) report [1,2]. But we can reduce this numbers by increasing awareness, early prediction and diagnosis [3].

Over the last two decades, the use of machine learning algorithms has spread to several fields including medicine. With the help of advanced processing units, now it’s possible to analyze medical data, which is very hard to analyze manually, via machine learning algorithms. There has been an increase in such studies over the last decade and more and more effective means to analyze medical data is introduced to the academic literature every day [4]. Many imaging techniques have been developed for early detection and treatment of breast cancer and to reduce the number of deaths, and many aided breast cancer diagnosis methods have been used to increase the diagnostic accuracy. In the last few decades, several data mining and machine learning techniques such as like SVM, Decision Tree, and many more models have been developed for breast cancer detection and classification [5-7], which can be divided into four main stages: preprocessing, feature extraction, and classification and evaluation.

To facilitate interpretation and analysis, the preprocessing of mammography films helps improve the visibility of peripheral areas and intensity distribution, and several methods have been reported to assist in this process [8,9]. Feature extraction is an important step in breast cancer detection because it helps discriminate between benign and malignant tumors. After extraction, image properties such as smoothness, coarseness, depth, and regularity are extracted by segmentation. Various transform-based texture analysis techniques are applied to convert the image into a new form using the spatial frequency properties of the pixel intensity variations. The aim is to predict and diagnosis breast cancer, using machine-learning algorithms such as K-NearestNeighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF) [10] find out the most efficient algorithm based on the performance and accuracy of each classifier in terms of confusion matrix, accuracy, F1 score, precision and sensitivity [11].

2. Literature Review

Varsha Nemade et al. used various ML Classification techniques: Naïve Bayes (NB) Logistic regression (LR), Support vector machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and ensemble techniques: Random-forest (RF), Adaboost, XGBoost ferent, both decision tree and XGBoost classifier has highest accuracy 97% among all model and highest AUC 0.999 obtained for XGBoost classifier [12].

Habib Dhahri et al. aim of this study was to optimize the learning algorithm by the help of genetic programming. The performance of the following machine learning algorithms: KNN, SVM classification, DT, RF, AB, GB, GNB, LDA, quadratic discriminant analysis, LR, and extras classifier. The AdaBoost classifier seemed to exhibit the best accuracy of 98.24% [13].

Obaid Oi et al. three machine-learning algorithms (Support Vector Machine, K-nearest neighbors, and Decision tree) on the dataset of Wisconsin Breast Cancer (Diagnostic). The outcomes of this study have revealed that quadratic support vector machine grants the largest accuracy of (98.1%) with lowest false discovery rates [14].Mohammed Amine Naji et al. applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic. It is observed that Support vector Machine achieved the highest accuracy (97.2%) [15].

Md. Milon Islam et al. compare five supervised machine learning techniques named support vector machine (SVM), K-nearest neighbors, random forests, artificial neural networks (ANNs) and logistic regression. The Wisconsin Breast Cancer dataset. The results reveal that the ANNs obtained the highest accuracy, precision, and F1 score of 98.57%, 97.82%, and 0.9890, respectively, whereas 97.14%, 95.65%, and 0.9777 accuracy, precision, and F1 score are obtained by SVM, respectively [16].

Ara Set al. used Wisconsin Breast Cancer Dataset and implemented Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes and Random Forest classifiers. Random Forest and Support Vector Machine outperform other classifiers with accuracy of 96.5% [17].

Mahesh TR et al. applied five machine learning techniques, namely support vector machine, K-nearest neighbors, decision tree Classifier, random forests, and logistic regression, in blended ensemble model. There is a 98.14 percent noticeable improvement with the Ensemble Learning model compared to the basic learners [18].

3. Material and methods

Breast cancer dataset is public and free available on Kaggle [19], which is widely used dataset for machine learning studies. In breast cancer dataset we applied all the which are necessary for the model prediction and its accuracy. The initial step was the preprocessing of data, after the feature scaling and feature extraction. With the help of python libraries such as pandas, Matplotlib, Sklearn, seaborn we were able to work on these features.
The dataset contains 31 features and 569 instances and a class attribute which classifies instances as either malignant or benign cancer which indicate if a tumor is cancerous or non-cancerous respectively. All the classifications were done using Goggle collab.

3.1 Preprocessing

Preprocessing of data is essential step before fitting the data into the respective machine learning model as it affects the performance and accuracy of model. Dataset may contain missing values, duplicates or outliers, if any of these are available in the dataset the accuracy of prediction of the model decline.
To overcome this, we perform data preprocessing. In this study, we check for the missing values, duplicate values which are present in this dataset but we found outliers. For finding outliers we go through some technique like z-score method, histogram method and boxplot plot method. The boxplot method gives the best results so used it to find the outliers for each attribute and replace it with the mean value of that respective attribute.

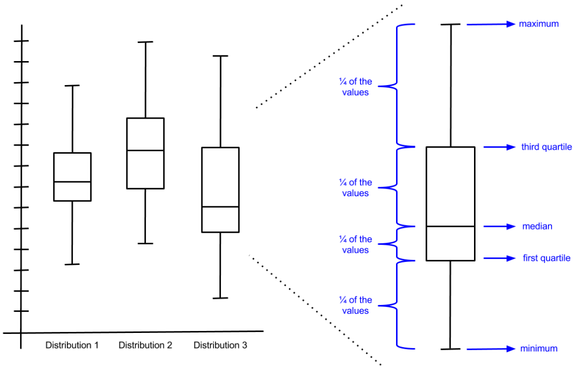

Fig-1 Method to find outliers

A box plot gives a five-number summary of set of data which is –

- Minimum – it is the minimum value in the dataset excluding the outliers.
- First Quartile (Q1) – 25% of the data lies below the first (lower) Quartile.
- Median (Q2) – It is mid-point of the dataset. Half of the values lie below it and half above.
- Third Quartile (Q3) – 75% of the data lies below third (upper) Quartile.
- Maximum – it is the maximum value in the dataset excluding the outliers.

3.2 Feature engineering

Feature engineering is the process of transforming raw data into features that can be used for creating a predictive model using Machine learning or statistical modelling.
Exploratory analysis or Exploratory data analysis (EDA) is an important step of features engineering, this step involves analysis, investing dataset, and summarization of the main characteristics of data. Different data visualization techniques are used to better understand the manipulation of data sources, to find the most appropriate statistical technique for data analysis, and to select the best features for the data.

Missing values within the dataset highly affect the performance of the algorithm, and to deal with them "Imputation" technique is used. Imputation is responsible for handling irregularities within the dataset.

For example, removing the missing values from the complete row or complete column by a huge percentage of missing values. But at the same time, to maintain the data size, it is required to impute the missing data, which can be done as:

- For numerical data imputation, a default value can be imputed in a column, and missing values can be filled with means or medians of the columns.
- For categorical data imputation, missing values can be interchanged with the maximum occurred value in a column.

3.3 Classification

Classification is a type of supervised machine learning algorithm which is used to predict the class of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data. For instance, an algorithm can learn to predict whether a cancer is malignant and benign.
There are various classification algorithm available which are used for prediction of output of data into the corresponding classes. Here we applied some of these classification algorithms which are K-NearestNeighbors (KNN) [19] , Support Vector Machine (SVM) [20] and Random Forest (RF) [21].

3.3.1 K-NearestNeighbors (KNN)

K-NearestNeighbors (KNN) classification, machine learning algorithm is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing the value of K, we can select the number of nearby observations to use in the algorithm.
Suppose we have a new data point and need to put it in the required category. First, we will choose the number of neighbors as k=5. Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. By calculating the Euclidean distance, we got the nearest neighbors, consider the diagram below –

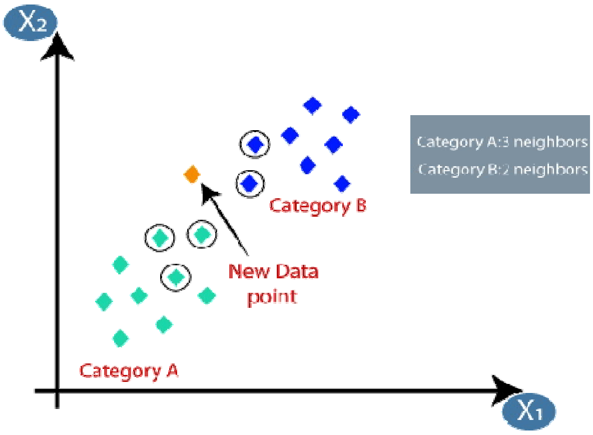

Fig-2 K-Nearest Neighbors

The three nearest neighbors in category A and two nearest neighbors in category B. Hence the new data point must belongs to the category A.

3.3.2 Random Forest Classification

Random Forest is one of the most popular and commonly used algorithms. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.
A random forest algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm provide the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

Key Benefits

- Reduced risk of overfitting: Decision trees run the risk of overfitting as they tend to tightly fit all the samples within training data.
- Provides flexibility: Since random forest can handle both regression and classification tasks with a high degree of accuracy, it is a popular method among data scientists.

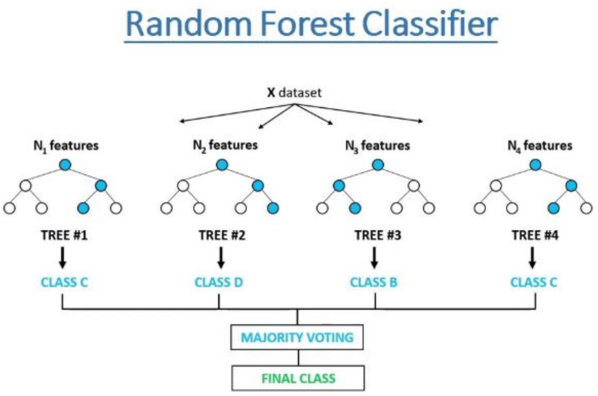

Fig-3 Random Forest

3.3.3 Support Vector Machine (SVM)

It is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. support vector machine is based on statistical approaches. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

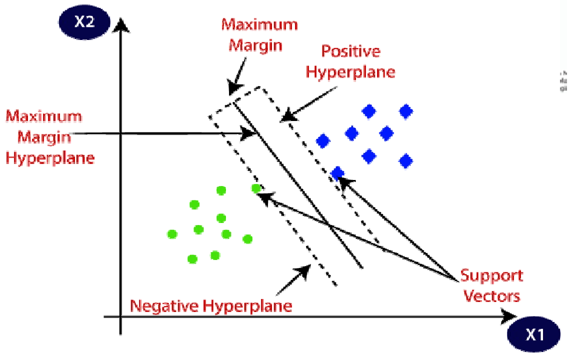

Fig-4 Support Vector Machine

4. Results

In this paper, we applied Wisconsin Breast Cancer dataset [22] dimension is 31 Features and 569 instances to validate and the designed models. As part of the research, basically three experiments were set up for the training of the input data.

The first experiment is to check whether there are any null values present in our dataset, if the result is True then it means there are some empty columns. To overcome this and make our dataset complete we have to fill the particular columns with a mean value of that column.

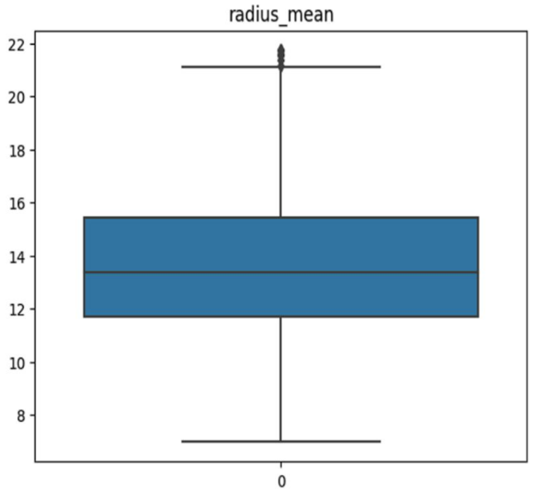

Fig-5 Boxplot Diagram before Removing Outliers

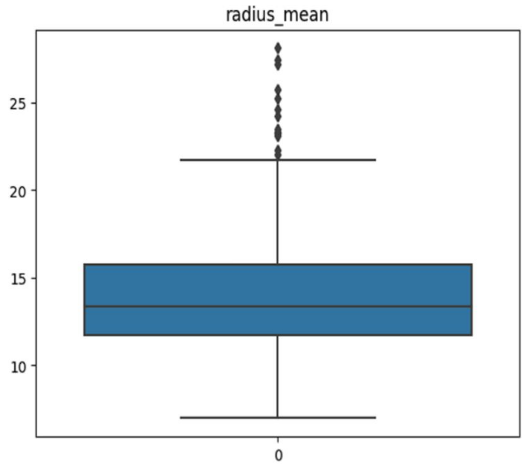

Fig-6 Boxplot Diagram after Removing Outliers

After successfully checking the null values, we have to find if is there any duplicates present in our dataset because it will affectthe result and accuracy so we check the duplicates using the duplicate () function and if it is True then we remove that particular duplicate row from the dataset.

After that, we check outliers using Boxplot or scatter plot as in Figure 5. We remove this outlier by finding the outlier value in a particular column and replacing it with the mean of that column. After replacing it with a mean we again check it using the Boxplot graph in Figure 6.

After completing the first part, the second experiment compared the popular supervised learning algorithms applied for classification of the problem.

In machine learning algorithms, various metrics are used to evaluate the proposed model.In this study, the metrics used were accuracy, AUC, confussion matrix, and precision-recall.

Accuracy (ACC) is the measure of correct prediction of the classifier, and it provides general information about how many samples are misclassified. It is defined as

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} \quad (1)$$

WhereTP is True posivite, TN is True negative, FP is False postive and FN is False negative.

Confussion Matrix: The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known.

The other metrics derived from a confussion matrix are defined as follows:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Recall:It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Precision: It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true.

$$F\text{-}1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \qquad (4)$$

F-1 measure: If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time.

As already mentioned, the ROC [23] space is defined with false positives and true positives as the x and y coordinates, respectively.

The ROC curve summarizes the performance across all possible thresholds. A perfect classifier would fall into the top-left corner of the graph with a true positive rate of 1 and a false positive rate of 0.

Classification Techniques	**Parameters**	**Confusion Metrics**	**Class**	**Recall**	**F-1 Score**	**Accuracy**
KNearest Neighbor	n_neighbors = 9	[107 1] [4 59]	Benign Malignant	0.96 0.98	0.98 0.96	0.9707
Support Vector Machine	C = 100 Kernel = rbf Gamma = scale	[108 0] [2 61]	Benign Malignant	0.98 1.00	0.99 0.98	0.9883
Random Forest	n_estimator = 21	[106 2] [1 62]	Benign malignant	0.99 0.97	0.99 0.98	0.9824

Table-1 Classifier Performance

Table 1 show the performance of different classifiers applied on the dataset. And SVM shows the best accuracy 98.83% among all other models.

In addition to this, we employed ROC curve to represent the relation between recall (sensitivity) and specificity metrics.

The models applied in this experiment were K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM).

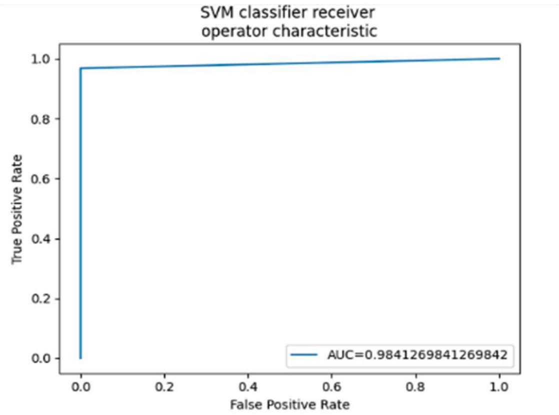

Fig-7 AUC-ROC curve of SVM Algorithm

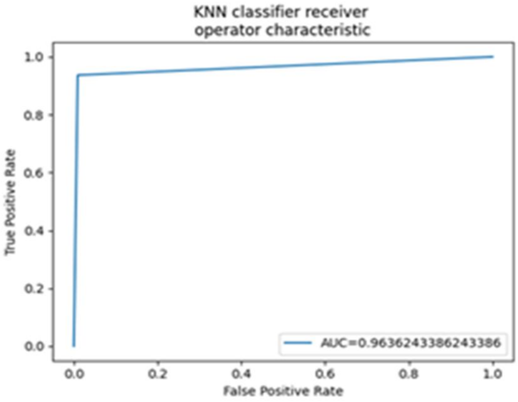

Fig-8 AUC-ROC curve of KNN Algorithm

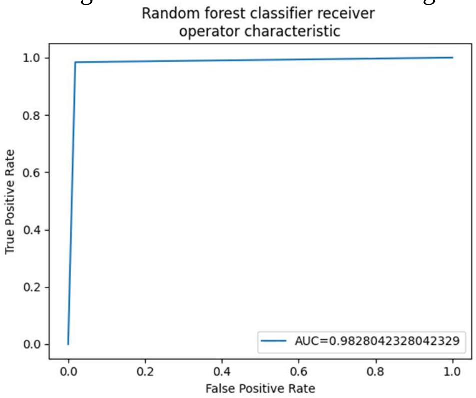

Fig-9 AUC-ROC curve of Random Forest Algorithm

Thus, it is shown that applied models can predict more accurately. Figures 7-9 compare the performances of the three computational models. In our experiment, we found that SVM obtained a higher AUC of 98.41%.

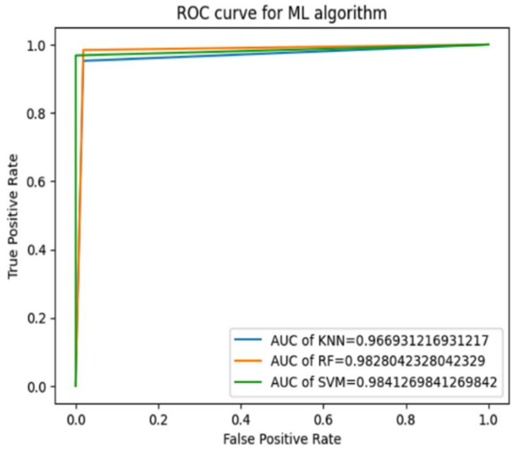

Fig-10 AUC-ROC curve of ML Algorithms

Figure 10 shows the compares the performance of different ML algorithm in one AUC- ROC curve.

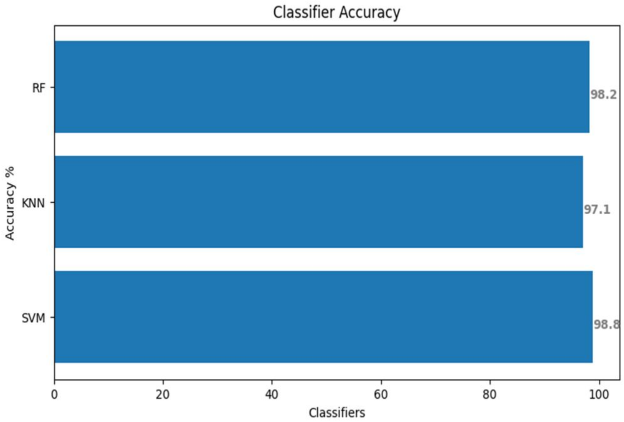

Fig-11 Accuracy of different ML Algorithm

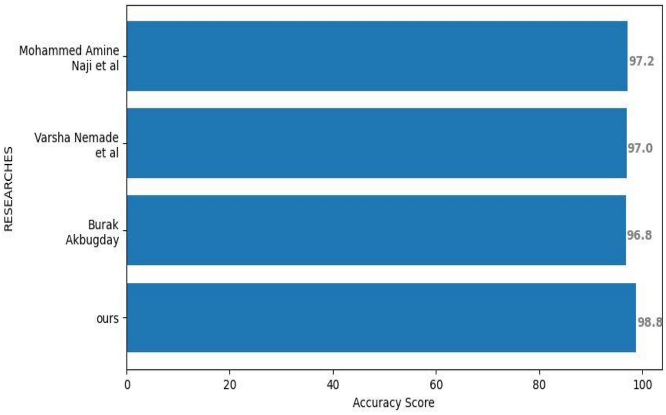

Fig-12 Accuracy Comparison

Fig-11 shows the accuracy of three ML algorithm using bar graph and Fig-12 compares the accuracy of different researches.

5. Conclusion

On the Wisconsin Breast Cancer Diagnostic dataset (WBCD) we applied three main algorithms which are: SVM, Random Forests, K-NN algorithms. We calculate, compare and evaluate different results obtained based on confusion matrix, accuracy, sensitivity, precision, AUC to identify the best machine learning algorithm that are precise, reliable and find the higher accuracy. All algorithms have been programmed in Python using numpy, pandas, matplotlib and scikit-learn libraries in Google Colab. After an accurate comparison between our models, we found that Support Vector Machine achieved a higher accuracy of 98.83%, Precision of 98.5%, AUC of 98.41% and outperforms all other algorithms. In conclusion, Support Vector Machine has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieve the best performance in terms of accuracy and precision. It should be noted that all the results obtained are related just to the WBCD database, it can be considered as a limitation of our work, it is therefore necessary to reflect for future works to apply other algorithms and methods on same databases to find more accurate results. Which helps us to find more accuracy and precision in the same dataset. And we apply this on other datasets to find the results with different parameters.

References

[1]. Rebecca L Siegel , Kimberly D Miller , Nikita Sandeep Wagle , Ahmedin Jemal (2023). Cancer statistics 2023. CA: a cancer journal for clinicians, 73(1):17-48.

[2]. Rebecca L Siegel , Angela N Giaquinto, Ahmedin Jemal . Cancer statistics, 2024. CA Cancer J Clin, 74(1):12-49.

[3]. Patil S, Kirange D, Nemade V. Predictive modelling of brain tumor detection using deep learning. Journal of Critical Reviews. 2020;7.

[4]. Nemade, V., Pathak, S., Dubey, A. K., &Barhate, D. (2022).A Review and Computational Analysis of Breast Cancer Using Different Machine Learning Techniques,12(3)111-118.

[5]. A. J. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” Cancer Informatics, vol. 2, pp. 59–77, 2006.

[6]. G. Valvano, G. Santini, N. Martini et al., “Convolutional neural networks for the segmentation of microcalcification in mammography imaging,” Journal of Healthcare Engineering, vol. 2019, Article ID 9360941, 9 pages, 2019.

[7]. M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” Expert Systems with Applications, vol. 36, no. 2, pp. 3240–3247, 2009.

[8]. D. Narain Ponraj, M. Evangelin Jenifer, P. Poongodi, and J. Samuel Manoharan, “A survey of the preprocessing techniques of mammogram for the detection of breast cancer,” Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 12, pp. 656–664, 2011.

[9]. A. P. Charate and S. B. Jamge, “*e preprocessing methods of mammogram images for breast cancer detection,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 1, pp. 261–264, 2017.

[10]. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., &Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science, 191, 487-492.

[11]. Yellamma, P., Chowdary, C. S., Karunakar, G., Rao, B. S., & Ganesan, V. (2020). Breast Cancer Diagnosis Using MLP Back Propagation. International Journal, 8(9)

[12]. Varsha Nemade , Vishal Fegade (2023, January). Machine Learning Techniques for Breast Cancer Prediction, Volume 218, ISSN 1877-0509.

[13]. Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi. Automated

Breast Cancer Diagnosis Based on Machine Learning Algorithms, Volume 2019, Article ID 4253641.

[14]. Obaid OI, Mohammed MA, Ghani MK, Mostafa A, Taha F. (2018) Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Tecnology,7(4.36):160-6.

[15]. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science, 191, 487-492.

[16]. Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), 1-14.

[17]. Ara S, Das A, Dey A. Malignant and benign breast cancer classification using machine learning algorithms. In2021 International Conference on Artificial Intelligence (ICAI) 2021 Apr 5 (pp. 97-101). IEEE.

[18]. Jabbar MA. Breast cancer data classification using ensemble machine learning. Engineering and Applied Science Research. 2021 Jan 27;48(1):65-72.

[19]. C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 415–425, 2002.

[20]. J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, no. 4, pp. 580–585, 1985.

[21]. J. Ham, Y. Yangchi Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 492–501, 2005

[22]. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

[23]. T. E. Cohn, "Receiver operating characteristic analysis of photoreceptor sensitivity," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-13, no. 5, pp. 873–881, 1983