Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

A Breast Cancer Detection: Malignant or benign- Using Machine Learning Classification Algorithms

Kajal Kumari[1], Mohan Rao Mamdikar[1], Advait Khare[2], Vishal Kumar Sahu[2]

[1]*Vishwavidyalaya Engineering College, Ambikapur, India*
[2]*Shri Shankaracharya Institute of Professional Management & Technology Raipur, India*
Email: Kajal30kumari2000@gmail.com

Breast cancer is common in humans (Women), is the most dangerous disease, and abnormally grows. Therefore, it is necessary to diagnose and detect deadly diseases accurately. Machine learning (ML) algorithms play a vital role in the detection of breast cancer cells in the human body. We proposed an ML techniques-based framework for the classification of breast cancer as malignant or benign using three important classification algorithms of ML- Support vector machine (SVM), K-nearest neighbors (KNN), and random Forest. We have evaluated accuracy and performance based on the accuracy, precision, recall, F1 score, and MCC parameters. In the proposed method, SVM gives the highest accuracy of 98% and MCC of 0.9623 compared to the other two ML algorithms KNN and Random Forest.

Keywords: autonomous breast cancer, classification, Random Forest, support vector machine, KNN.

1. Introduction

According to the World Health Organization (WHO), breast cancer is a fatal disease in which abnormal breast cells develop controllable and produce tumors rapidly. Such tumors are left untreated and then can spread into the whole body. Research says that, in the year 2022 only there were 670000 deaths worldwide with this disease, and 2.3 billion cases were diagnosed in women of breast cancer. As per the report, 99% of cases of breast cancer are found in women. However, there is the possibility of breast cancer in men of 0.5%-1%. If symptoms and breast cancer are diagnosed early in the stage accurately, it can be cured. Unfortunately, most of the persons will experience any symptoms in the early stage. Therefore, researchers and scientists are rigorously working on detecting such deadly diseases with the use of technology growing day by day.

Recent studies have shown that using ML algorithms has given significant results for the detection of breast cancer. Various researcher scholars and scientists have used different classification algorithms on the dataset Wisconsin Breast cancer dataset, these algorithms are

Support Vector Machine(SVM), Artificial Neutral Network (ANN), Naïve Bayes, K-Nearest Neighbour's (KNN), Decision Tree, Random Forest, and XGboost(Malik Adeiza Rufai, Ahmad Shehu Muhammad, Garba Suleiman, 2020)(Kumar et al., 2022)(Manav Mangukiya, Anuj Vaghani, 2022).

These models can classify the breast as malignant based on extracted features from images. The SVM has shown better accuracy which is between 94.3% to 98.8%(Malik Adeiza Rufai, Ahmad Shehu Muhammad, Garba Suleiman, 2020). Other algorithms have also shown good results such as Decision Tree 95.90%(Kumar et al., 2022), and XGBoost gives an accuracy of 98.24% (Manav Mangukiya, Anuj Vaghani, 2022). Machine learning models have the ability to increase the diagnosis or detection of breast cancer in the early stage present women. The Survival rates, if ML algorithms are used for breast cancer detection. Such facilities can be used in healthcare.

Classification of malignant or benign breast cancer correctly and accurately is a challenging issue in the field of healthcare. Therefore, it is essential to develop an ML classification technique for the classification of breast cancer cells that are malignant or benign accurately so that experts can necessary medical actions to cure the deadly disease.

In this research, we develop a machine learning algorithm for classification using K-Nearest Neighbors (KNN), SVM, and Random Forest. The organization of the rest of the papers is as follows- section 2 describes a mathematical background, and section 3 addresses the related work. In section 4, we discuss the proposed methodology. Section 5 addresses the result and discussion, and in section 6 conclusion and future work.

2. Mathematical Background

As the breast cancer detection process is a classification problem, some of the essential and mathematical terminologies are frequently used in classification and machine learning(Jalloul et al., 2023) . These terminologies are as follows:

Formulation of Classification Problem

The process of "classification of things" into smaller classes is referred to as classification. Labeled data is used for training in supervised machine learning, which includes classification(Parundekar, 2018). A classification process is illustrated in Figure 1.

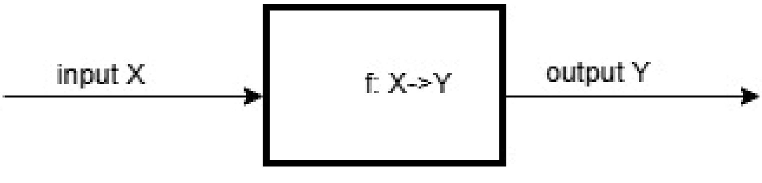

Figure 1 Classification Process

Input features: Let R^d is the d – dimensional feature space for medical dataset (Breast cancer dataset), then $S=\{s_1, s_2,s_3,\ldots s_n,\}$, $s_i \in R^d$ representing d- dimensional feature vector.

Output (label): Let y is target (output variable), then for classification of breast cancer malignant or benign, $y=\{0,1\}$ where 0 indicates benign and 1 is for malignant.

Aim: Select the machine learning model by mapping f:X→Y such that the accuracy must be satisfactory.

Formulation Mathematical Foundation of ML Algorithm

Logistic Regression: To compute the correlation in between two variables, the machine learning analysis algorithm method called as logistic regression employs mathematics. Using Logistic regression, classification problem can be solved with one possible solution such either yes or no that is binary classification. The logistic regression uses logistic function and it is mathematically defined as-

$$\mathrm{lrP}(Y = 1|x) = \frac{1}{1+e^{-(\alpha_0+\alpha_1 x_1+\cdots+\alpha_n x_n)}} \qquad (1)$$

Where $\mathrm{lrP}(Y = 1|x)$ is the probability of the output Y.

Support Vector Machine: To address regression and classification issues, the most prominent supervised learning method is termed a Support Vector Machine. However, SVM is generally used for classification applications based on classification. To find the hyperplane in the dataset with multi-dimensional which classifies the data points as the target. This helps resolve an optimization issue and it is mathematically represented as-

$$\text{minimize } \frac{1}{2}||w|^2 \quad \text{subject to } y_i(w\ \ x_i + b) \geq 1, \forall i \qquad (2)$$

subject to $y_i(w\ \ x_i + b) \geq 1, \forall i$

Decision Tree: The decision trees are supervised learning based on the non-parametric machine learning method significantly used for regression and classification. With the help of a Decision tree predict the value of the output variable. The tree can be used for constant approximation in a piecewise manner. To optimize the information gain, split the dataset into training and testing based on features using metrics such as those given mathematically-

$$\mathrm{Infgain} = E(Y) - E(X|Y) \qquad \text{here E is entropy} \qquad (3)$$

K-Nearest Neigbhors (KNN): It is the simplest, widely used classification and regression classification method used in Machine Learning in practice. The KNN is based on supervised learning and non-parametric which utilizes proximity to predict the grouping of data points of classification. The KNN depends on the Euclidian distance that metrics that represented as-

$$d(s,t) = \sqrt{\sum_{i=1}^{n}(s_i - t_i)^2} \qquad (4)$$

3. Related Work

(Naji et al., 2021)Devised a framework for the prediction and detection of breast cancer using different ML classification techniques. Authors have used SVM, Random Forest, Logistic Regression, Decision Tree, and KNN for the prediction of disease. In this article, classification accuracy was found 97%.

(Safdar et al., 2022) proposed a framework for the detection of breast cancer using classification algorithms of ML such as SVM, Random Forest, Naïve Bayes, DT, and KNN. In this research, XGboost is specially used to get a high accuracy of 98%.

(Safdar et al., 2022)presented an article for segmentation, and detection of cancerous/non-cinereous from the BCWD Dataset with the help of classification algorithms such as SVM, LR, and KNN. Authors have classified breast cancer as malignant or benign. However, in the proposed work there are some "false positive (FP)" and "false negative (FN)" predictions that shows further improvement required in the computation of accuracy.

(Allugunti, 2022)proposed a computer-aided diagnosis framework using Deep Learning and Machine Learning algorithms for thermographic images to classify the image dataset into three categories namely cancer, no cancer, and non-cancerous. In this research, authors have used convolution Neural Networks (CNNs), SVM, and Random Forest for the accurate classification of images.

(Naseem et al., 2022) proposed a framework for the automatic identification of cancer (breast) diagnosis and prognosis using an ensemble classifier. The authors have firstly, provided a systematic review on several machine learning algorithms and an ensemble of machine algorithms. Authors have shown ensemble of ML classifiers gives outperformance of the state-of-art with an accuracy of 98%.

(Ara et al., 2021) proposed a methodology for the analysis of the WBC Dataset and estimated the performance of several ML algorithms for accurately predicting breast cancer. In this research, authors have used different models such as SVM, KNN, Logistic Regression, Decision Tree, Naïve Bayes, and Random Forest etc. SVM and Random forest machine learning algorithms give an accuracy of 96.5% as compared to other ML classifiers.

(Jasti et al., 2022) proposed a model that combines feature extraction, feature selection, image preprocessing, and ML algorithms for the detection of skin cancer and classification. In this article, a geometric mean filter is used for image enhancement and AlexNet is used for extracting features respectively. However, the authors have not shown accuracy in the form of percentages.

(Khalid et al., 2023) proposed a framework using a Deep Learning model and ML algorithms that has the ability to recognize breast cancer. The framework depends on the elimination of low-variance features and univariate feature selection. The medical-lateral views and craniocaudally of mammograms are used and tested with a large dataset. The six ML algorithms are used for the classification and categorization of breast cancer.

(Ak, 2020) proposed a framework using a Deep Learning model and ML algorithms that has the ability to recognize breast cancer. The framework depends on the elimination of low-variance features and univariate feature selection. The medical-lateral views and craniocaudally of mammograms are used and tested with a large dataset. The six ML algorithms are used for the classification and categorization of breast cancer.

(Ahmad et al., 2024) proposed a framework for enhancing breast cancer diagnosis by independently categorizing and identifying breast lesions, and segmenting mass lesions based on the pathology using deep learning.

4. Proposed Framework

The proposed framework has six steps as illustrated in Figure 2. In the first step, loading of the dataset that is taken from the data repository is available for the public at https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

In this breast cancer dataset, all the features are calculated from a fine needle aspiration(FNA) that describes the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus. Figure 3 shows 357 benign, and 212 malignant.

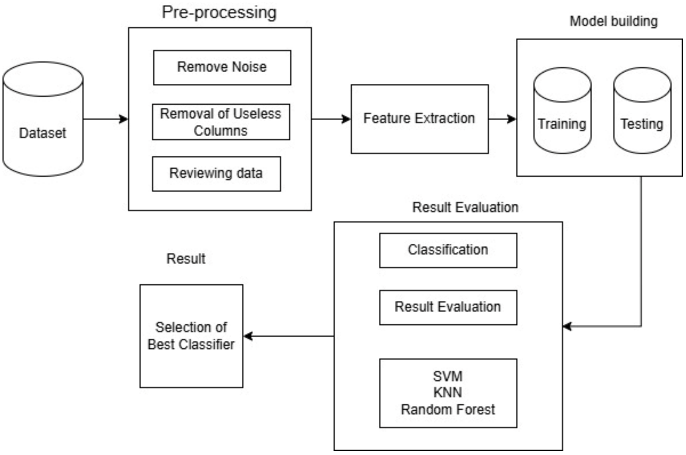

Figure 2 Proposed Framework

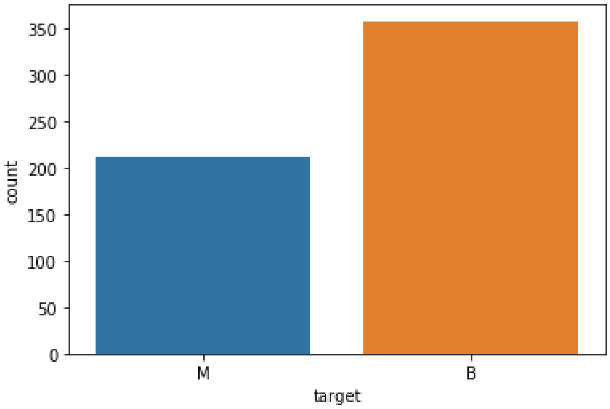

Figure 3 Dataset as Malignant and benign

In the 2nd step, the collected dataset need to be refined by applying pre-processing techniques in which raw data is improved for better classification result. In the pre-processing, data noise is removed, the redundancy data is removed, useless columns are removed, removed missing values. By applying data transformation, data files are transformed into human readable.

Furthermore, we apply feature extraction to make features in the dataset as much as simple. These simplified features can capture the original characteristics of the dataset which is efficient for analysis.

Now, the dataset is broken into categories as training and testing respectively for building ML algorithms. The training data is 70% and testing data is 30% respectively.

Now, we apply classification algorithms of machine learning such as SVM, KNN, and Random Forest for breast cancer of malignant or benign. Three classification algorithms are applied for the detection of breast cancer and to make a comparison of the results of each of the algorithms.

5. Evaluation Result

The literature review or related work, has shown that accuracy as mentioned in section 3. Here we use accuracy, F1_score, and Matthews Correlation Coefficient (MCC) as parameters for accuracy evaluation. Also, precision and recall are taken into account for the result evaluation. The accuracy is expressed as given in the equation (5).

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TN+FP+FN+TP}} \qquad (5)$$

In equation (5), TP =True Positive, TN = True Negative, FN = False Negative, and FP = False Positive.

Recall is an important parametric metric measure that evaluates the percentages of data sets that an algorithm identifies correctly. It is also called sensitivity and is calculated by dividing the number of TPs by the number of positive instances as given in equation (6)-

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{FN+TP}} \qquad (6)$$

The precision (Lotter et al., 2021) is used to evaluation of percentage in all positive class, predicting all positive patterns accurately and it is given as equation (7)-

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{FP+TP}} \qquad (7)$$

F1 score is the important metrics in the ML that is used for evaluating performance of the classification algorithm. It is a combination of recall and precision and merged into single one. It is represented as equation (8)-

$$\mathrm{F1\ score} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision+recall}} \qquad (8)$$

We set the limit 0.75 for better correlation between different features as shown in the Figure 4.

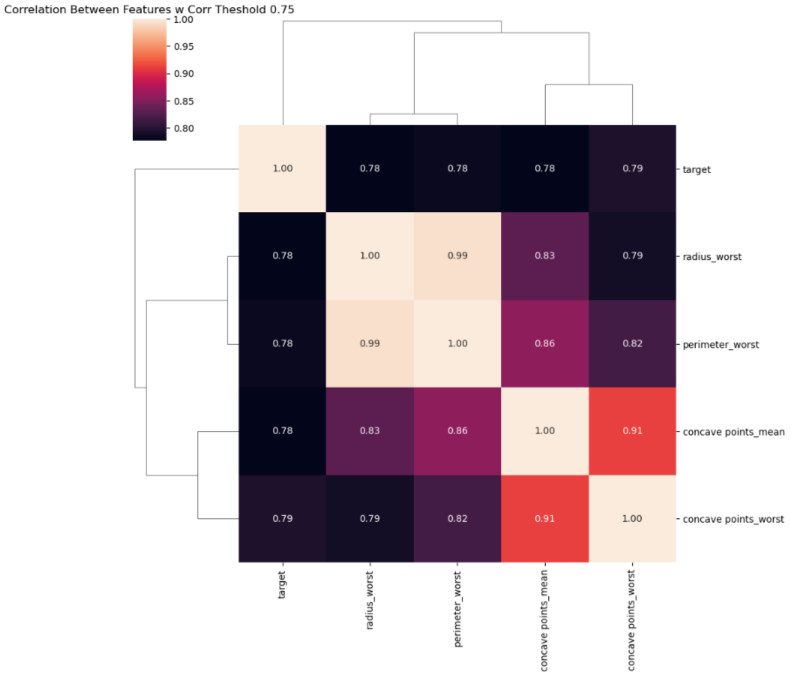

Figure 4 Correlation between features

In the Figure 5, visualize a breast cancer diagnosis that indicated values high in some features for patients and low values in some features for few patients.

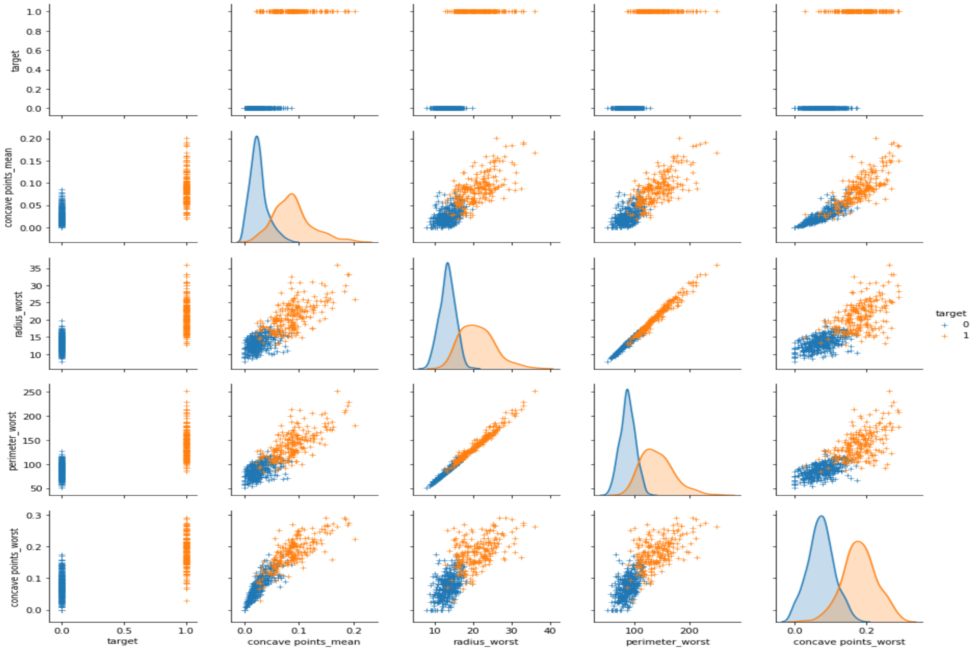

Figure 4 Breast cancer diagnosis visualization

We run three ML algorithms namely, KNN, SVM, and Random Forest algorithm to capture the accuracy and performance of the breast cancer present in the patients. To achieve this, we use different parameters such as accuracy, recall, precision, F1 score, MCC, ROC, and AUC. Table 1, Table 2, and Table 3 show the computations of accuracy, recall, F1 score, precision, and MCC for KNN, SVM, and Random Forest.

Table 1 Accuracy, Recall, F1 score, Precision, and MCC using KNN

KNN	Precision	Recall	F1 Score	MCC
0	0.96	0.97	0.97	0.911203
1	0.95	0.94	0.94	
Accuracy	0.96			

Table 2 Accuracy, Recall, F1 score, Precision, and MCC using SVM

SVM	Precision	Recall	F1 Score	MCC
0	0.97	1	0.99	0.962353
1	1.00	0.95	0.98	
Accuracy	0.98			

Table 3 Accuracy, Recall, F1 score, Precision, and MCC using Random Forest

Random Forest	Precision	Recall	F1 Score	MCC
0	0.96	0.99	0.98	0.936764
1	0.98	0.94	0.96	
Accuracy	0.97			

The ROC curve for SVM in which AUC=1.0, ROC curve for KNN in which AUC=0.99 and the ROC curve for KNN in which AUC=1.00 as shown in the Figure 6.

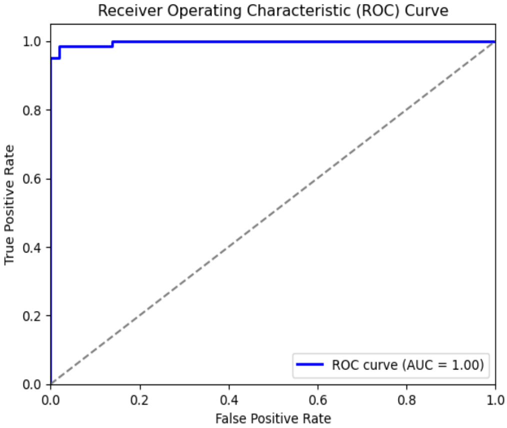

Figure 6(a) The ROC for SVM

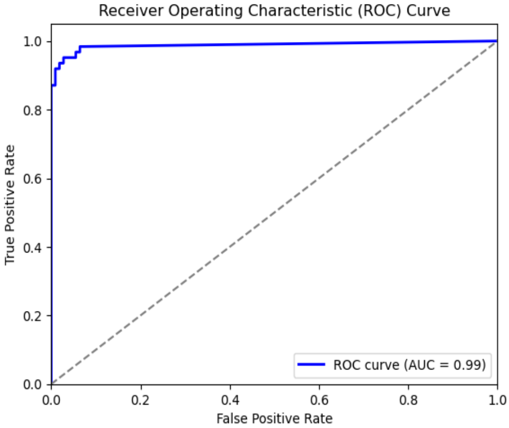

Figure 6(b) The ROC for KNN

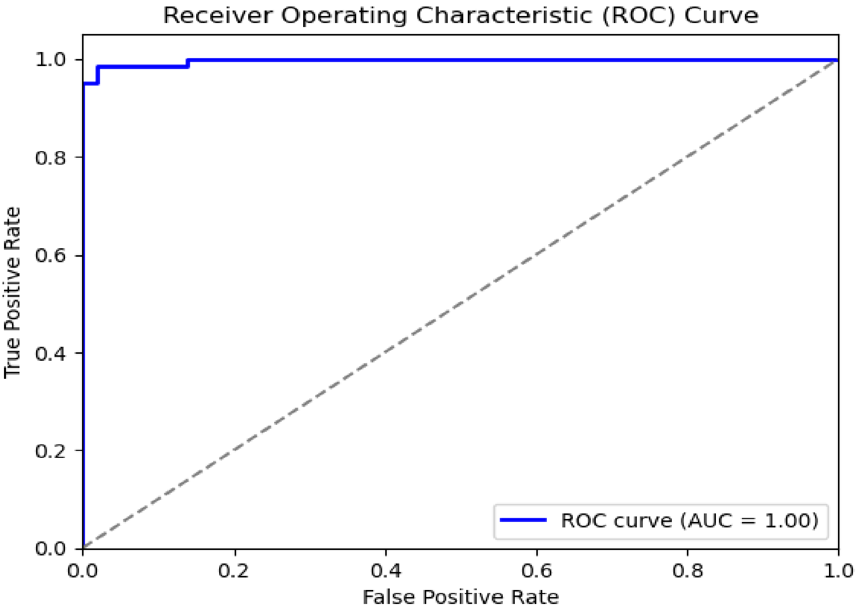

Figure 6(c) The ROC for Random Forest

Figure 6 The ROC for SVM, KNN and Random Forest

From the Table 1, Table 2 and Table 3 the accuracy of the KNN is 0.96, accuracy of SVM is 0.98 and accuracy of Random Forest algorithm is 0.97 respectively. The comparison of the accuracy is shown in the Figure 7.

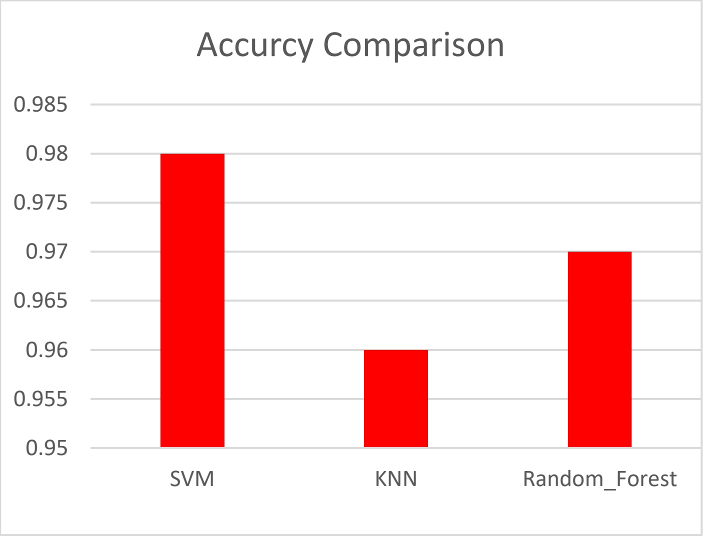

Figure 7 The Matthews correlation coefficient (MCC) value

The Matthews correlation coefficient (MCC) values for the SVM, KNN, and Random Forest are shown in Table 1 to Table 3. It is obtained with the help of the True Positives (TP), True Negatives (TN), False positives (FP), and False Negatives (FN). The value of MCC =0 means no correlation, and MCC=1 means prediction is perfect. The result shows that the accuracy of SVM is 98% and MCC is 0.9623, the accuracy of KNN is 96%, and MCC is 0.912, and the accuracy of Random Forest is 97%, and MCC is 0.9367

6. Conclusion

We applied three ML algorithms (SVM, KNN, and Random Forest) for detecting breast cancer on the breast cancer dataset. The obtained results are compared. The comparison is based on various parameters- confusion matrix, accuracy, precision, recall, F1 score, MCC, and AUC for correct identification of ML models for best accuracy and performance. In this research, SVM gives the highest accuracy of 98% and MCC of 0.9623 as compared to the other two ML algorithms KNN and Random Forest. This research paper will be helpful for researchers and scientists who are working healthcare field for cancer detection.

References

[1] A. L. Malik Adeiza Rufai, Ahmad Shehu Muhammad, Garba Suleiman, "MACHINE LEARNING MODEL… FUDMA Journal of Sciences (FJS) Malik, Ahmad, Garba and Audu ISSN online: 2616-1370 ISSN print: 2645 - 2944 MACHINE LEARNING MODEL FOR BREAST CANCER DETECTION," vol. 3, no. 1, pp. 210–224, 2020.

[2] S. M. N. Kumar, V. N. Ganesh, J. A. Mayan, and A. Jesudoss, "Prediction of Breast Cancer using Machine Learning Algorithm's," 2022 6th Int. Conf. Trends Electron. Informatics, ICOEI 2022 - Proc., pp. 1–6, 2022, doi: 10.1109/ICOEI53556.2022.9776803.

[3] M. S. Manav Mangukiya, Anuj Vaghani, "Breast Cancer Detection with Machine Learning," 2022, no. February, pp. 1–23.

[4] R. Jalloul, H. K. Chethan, and R. Alkhatib, "A Review of Machine Learning Techniques for the Classification and Detection of Breast Cancer from Medical Images," Diagnostics, vol. 13, no. 14, 2023, doi: 10.3390/diagnostics13142460.

[5] R. Parundekar, "Classification of Things in DBpedia Using Deep Neural Networks," Int. J. Web Semant. Technol., vol. 9, no. 1, pp. 01–18, 2018, doi: 10.5121/ijwest.2018.9101.

[6] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis," Procedia Comput. Sci., vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.

[7] S. Safdar et al., "Bio-Imaging-Based Machine Learning Algorithm for Breast Cancer Detection," Diagnostics, vol. 12, no. 5, pp. 1–18, 2022, doi: 10.3390/diagnostics12051134.

[8] V. R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," Int. J. Eng. Comput. Sci., vol. 4, no. 1, pp. 49–56, 2022, doi: 10.33545/26633582.2022.v4.i1a.68.

[9] U. Naseem et al., "An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers," IEEE Access, vol. 10, no. July, pp. 78242–78252, 2022, doi: 10.1109/ACCESS.2022.3174599.

[10] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 Int. Conf. Artif. Intell. ICAI 2021, no. April, pp. 97–101, 2021, doi: 10.1109/ICAI52203.2021.9445249.

[11] V. D. P. Jasti et al., "Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis," Secur. Commun.

Networks, vol. 2022, 2022, doi: 10.1155/2022/1918379.

[12] A. Khalid, A. Mehmood, A. Alabrah, B. F. Alkhamees, and F. Amin, “Breast Cancer Detection and Prevention Using Machine Learning,” pp. 1–21, 2023.

[13] M. F. Ak, “A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine,” 2020.

[14] J. Ahmad et al., “Deep learning empowered breast cancer diagnosis : Advancements in detection and classification,” pp. 1–24, 2024, doi: 10.1371/journal.pone.0304757.

[15] W. Lotter et al., “Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach,” Nat. Med., vol. 27, no. 2, pp. 244–249, 2021, doi: 10.1038/s41591-020-01174-9.