International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Breast Cancer Classification using Machine Learning : A Review

Gurmanjit Kaur[1] , Manveen Kaur[2]
[1,2]DAV Institute of Engineering and Technology
[1]gurmanbhangu2002@gmail.com
[2]manveendhanju7@gmail.com

Abstract— Breast cancer represents a significant global health issue that profoundly affects women, emphasizing the critical importance of early detection for successful treatment. Machine learning algorithms have demonstrated remarkable potential in accurately categorizing different subtypes of breast cancer. This review paper focuses on providing a comprehensive summary of breast cancer classification using machine learning techniques. We delve into an examination of diverse machine learning algorithms, including Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB), and K Nearest Neighbors (k-NN), for the purpose of breast cancer classification. Their performance is thoroughly compared and evaluated. Furthermore, we discuss the challenges and future directions associated with machine learning-based breast cancer classification.

Overall, this review paper serves as an invaluable resource for researchers in the field by presenting a comprehensive overview of the current state-of-the-art in breast cancer classification using machine learning. Our objective is to explore different methodologies to effectively and efficiently detect breast cancer early using machine learning. Recognizing the critical significance of early cancer discovery, we employ various machine learning algorithms to predict whether a tumor is benign (noncancerous) or malignant (cancerous) based on the provided data features.

Keywords-Breast cancer classification, benign, malignant, Naïve Bayes, KNN, Support Vector Machine, Decision tree.

I.INTRODUCTION

Breast cancer ranks among the most prevalent cancers in women, characterized by abnormal growth of breast cells. Globally, there are 2.3 million new cases of female breast cancer reported annually, surpassing lung cancer (11.7%) [1]. Approximately 10% of breast cancer cases are attributed to hereditary factors, while the remaining 90% are influenced by various lifestyle factors. A significant rise in breast cancer prevalence has been observed across 15 population-based cancer registries focusing on women, with the majority of cases (97.7%) classified as epithelial tumors and undergoing multi-modality treatment.

In terms of breast cancer incidence in Asia, Israel exhibits the highest rate (84.6), while among Indian districts, Hyderabad has the highest incidence rate (48.0) [2]. According to a report by the National Cancer Registry Programme (NCRP), cancer cases are projected to increase from 13.9 lakh in 2020 to 15.7 lakh by 2025, assuming a 20% overall increase [3]. Timely treatment is crucial in preventing common cancers from becoming fatal, as early detection leads to effective management.

The objective of the review paper is to employ classification techniques to categorize tumor cases

as Malignant or Benign, thus achieving greater accuracy. The dataset used in this study is obtained from the Kaggle website. We have utilized supervised learning, a machine learning concept where we provide the machine with dependent and independent variables for training. After the learning process is completed, the machine is capable of predicting the value of the dependent variable for a given input in the form of an independent variable.

The classification techniques used for detecting the tumour are Decision tree, K Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) classification in google colab along with data visualization.

I. LITERATURE SURVEY

A. Nithya [4] conducted a study where three categorization methods, namely Decision Tree, k-Nearest Neighbour, and Naïve Bayes, were applied to different datasets. The authors also evaluated the error rate using specific attribute types in the datasets.

B. Shilpa M and C. Nandini [5] implemented an algorithm using Python and tested it on a single dataset. They achieved an accuracy of 94.74% and significantly reduced the required processing time.

C. A. Hafizah [2] compared the performance of Support Vector Machine (SVM) and Artificial Neural Network (ANN) using four different breast cancer datasets. The results demonstrated that SVM outperformed ANN in terms of both performance and outcomes.

D. S. Gc [1] focused on feature extraction, specifically variance, range, and compactness. They employed SVM classification to analyze the performance, resulting in a highest variance of 95% and compactness of 86%. Based on their findings, SVM was considered a suitable method for breast cancer prediction.The following is a summary of work done in the following domain:

Author	Dataset Used	Technique Used	Tools Used	Result
Muktevi et al. [4]	Wisconsin breast cancer-Kaggle	SVM, Random Forest, KNN, LR, NB	Python	The Random Forest model had the highest accuracy with 98.24%
Ramik Rawal [5]	Wisconsin-Kaggle	SVM, Logistic Regression, Random Forest, K-NN	Jupyter Notebook-Python	SVM-97.13% highest efficiency and accuracy
Gaurav Singh [6]	UCI repository	K-NN, SVM, LR, NB	Python	K-NN- 99% SVM- 96% LR- 97% NB-95%
Min-Wei et al. [7]	UCI Repository, ACM SIGKDD Cup 2008	SVM classifier, SVM ensemble	Weka	SVM ensembles perform slightly better than single SVM classifiers
Deepika et al. [8]	UCI repository	Naïve Bayes, MLP	Weka	Naïve Bayes had better accuracy
Ch. Shravya et al. [9]	UCI repository	SVM, K-NN, LR	Spyder Platform	SVM predicted the best accuracy of 92.78% followed by KNN-92.23%
Wang et al. [10]	Wisconsin Breast Cancer Database (1991) Wisconsin Diagnostic Breast Cancer (1995)	SVM, ANN, Adaboost, PCA	WEKA	8 PCs - 92.6% correlation, 10 PCS- 95%

III. METHODOLOGY

A. *Dataset Description:*

We obtained the Breast Cancer Wisconsin (Diagnostic) Dataset from Kaggle, consisting of 569 patient datasets. Each instance in the dataset contains 32 attributes related to the diagnosis and characteristics of breast cancer. The aim is to predict the presence of cancer solely based on the provided features. The feature values are in numeric format. The "target" variable represents the diagnosis, with "benign" mapped to 0 indicating non-cancerous cells and "malignant" mapped to 1 indicating cancerous cells.

B. *Section Headings*

As our coding platform, we utilized Google Colab. Our approach involved employing supervised learning algorithms and classification techniques such as Support Vector Classifier (SVM), Naïve Bayes, Decision Tree, and K-Nearest Neighbors (KNN). Given the varying units and magnitudes of the features in the dataset, it was necessary to normalize them to the same order of magnitude, which we accomplished using the default scaling in Sklearn. Model selection plays a crucial role in machine learning

C. *METHOD*

The primary objective of our experiment was to identify an effective and predictive algorithm for breast cancer detection. To achieve this, we applied several machine learning classifiers, including Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN), to the Wisconsin breast cancer diagnosis dataset. We evaluated the results obtained from each model to determine which one exhibited higher accuracy.

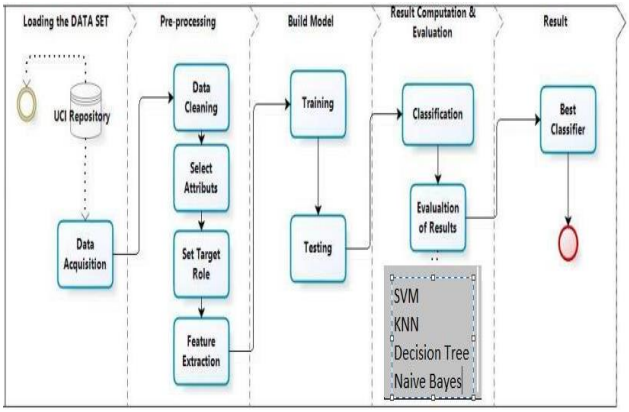


Fig. 1 Process of breast cancer Diagnostic using Machine Learning Algorithm

D. *BACKGROUND STUDY: MACHINE LEARNING ALGORITHMS*

• *Decision Tree Classifier:*

The decision tree classifier, developed by J. Ross Quinlan in 1980, exemplifies supervised machine learning. It operates by creating a tree-like structure where each node represents a decision based on specific conditions. The decision tree classifier assigns classifications to the conditions encountered at each node, ultimately leading to a solution.

The high-level pseudocode for the Decision Tree Classifier algorithm involves various attribute selection measures such as Entropy (Information Gain), Gain Ratio, and Gini Index. The entropy of a split can be calculated using the formula:

$$H(s) = -P_{(+)} \log_2 P_{(+)} - P_{(-)} \log_2 P_{(-)}$$

$Here \frac{P_{(+)}}{P_{(-)}}$ =% of +ve class 1% of -ve class

The Gini impurity of features after partitioning can be calculated using the following formula:

GI = 1-$\sum_{i=1}^{n}(p)^2$

GI=1-$[(P_{(+)})^2 + (P_{(-)})^2]$

Algorithm:

1. Begin at the root node.
2. Compare the value of the root node with the corresponding attribute value from the actual dataset.
3. Move to the next node based on the outcome of the comparison.
4. Compare the attribute value with the value of the subnode and proceed to the next node accordingly.
5. Repeat this process until reaching the leftmost node of the tree.

• ***Naïve Bayes:***

The Naïve Bayes classification technique utilizes Bayes' theorem and is referred to as "naïve" because it assumes independence among input variables. This fast and simple machine learning algorithm is commonly used for predicting large classes. When applying the Naïve Bayes classifier to a class variable, it assumes that the presence or absence of one feature is unrelated to the presence or absence of other features. This algorithm is particularly useful and effective for handling complex problems.

In Gaussian Naive Bayes, it is assumed that the continuous values associated with each feature used for prediction follow a normal distribution (Gaussian Distribution). This results in a bell-shaped curve when plotted, which is symmetric around the mean of the feature, as shown in the

$$P(X \mid Y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

equation.

Algorithm:

1) Separate the training data based on class.
2) Calculate the mean and standard deviation for each attribute.
3) Summarize and organize the dataset by class.
4) Calculate the Gaussian Probability Density function.
5) Finally, calculate the class probabilities.

• ***Support Vector Machine:***

Support Vector Machine (SVM) is a machine learning algorithm used for both classification and regression problems. The linear SVM classifier operates on the principle of margin maximization. It creates a decision boundary, known as the hyperplane, to divide the n-dimensional space into different classes, allowing for easy addition of new data points. The linear classifier aims to maximize the space between the decision hyperplane and the nearest data points by identifying the most suitable hyperplane.

The decision function in SVM is defined as:
f(x) = sign(w^T x + b)
where x is the input instance, w is the weight vector, b is the bias term, and sign() is the sign function that returns -1 or +1 depending on the sign of its argument.

Algorithm:

1) Accurately classify the training dataset based on lines/boundaries.
2) Select the line/boundary that has the greatest distance to the nearest data point.

• ***K-Nearest Neighbour:***

K-Nearest Neighbor (KNN) is a supervised learning algorithm used for both classification and regression tasks. It is a non-parametric method that does not make assumptions about the underlying data. Often referred to as a "lazy learner," the KNN algorithm does not immediately learn from the training set but instead stores the dataset and performs actions during the classification phase.

In the KNN algorithm, similarity between a new instance and existing data is assumed, and the new instance is assigned to the category with the

highest similarity to the existing instances. All available data is stored, and classification of new data points is based on their similarity to existing data. This allows for easy classification of new data points when they appear. The KNN algorithm stores records during the training phase and assigns new records to a category similar to the new data.

In our implementation, we have utilized the Euclidean distance metric, which is widely popular and recommended by experts. Let the points P and Q be represented by feature vectors P = (x1, x2, ..., xm) and Q = (y1, y2, ..., ym), where m is the dimensionality of the feature space. The Euclidean distance between P and Q is calculated using the following formula:

$$\text{dist}(P,Q)=\sqrt{\frac{\sum_{i-1}^{m}(x_i-y_i)^2}{m}}$$

Algorithm:

1) Retrieve the training and testing data from the dataset.
2) Select the desired value for the number of nearest data points (K).
3) For each data point in the testing set, compute the Euclidean distance between the data point and each row of the training data, and arrange them in ascending order.
4) Choose the top K rows with the smallest distances and determine the most frequent class among these K data points.
5) Assign the test point to the class that appears most frequently among the K nearest neighbors.

IV. CONCLUSION

This paper analyses a Breast Cancer (Diagnostic) Data Set with 32 attributes and makes a prediction using different classifiers whether the tumour is benign or malignant.

It can be seen that the highest accuracy is obtained using the SVM model with an accuracy for the linear kernel. We can summarize by saying that the SVM model showed the most efficient and effective accuracy out of all the 4 models used for the classification of benign and malignant tumours. We can see the number of Malignant (M) (harmful) or Benign (B) cells (not harmful) cells and plot it in a graph.

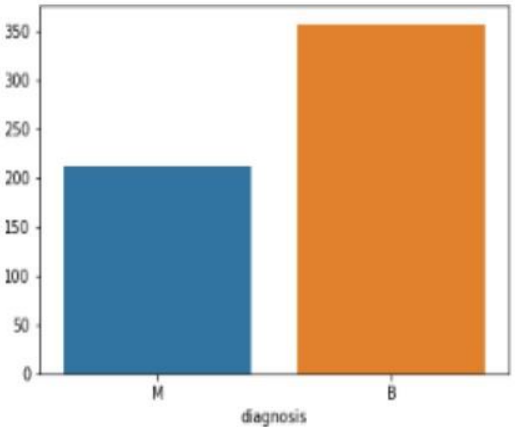
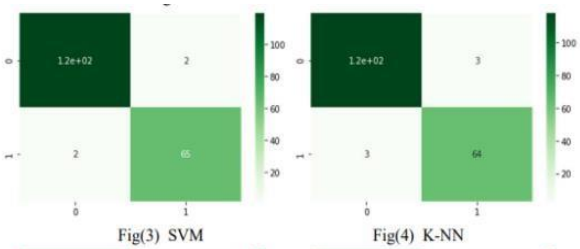


Fig. 2

A confusion matrix was plotted to calculate the miscalculations in each model. A confusion matrix is used to evaluate and recount the performance of a classifier.

	Predicted Class		
Actual Class		Yes	No
	Yes	True Positive(TP)	False Negative(FN)
	No	False Positive(FP)	True Negative(TN)

- True Positive (TP): Correctly predicts positive class.
- False Positive (FP): Incorrectly predicts positive class.
- False Negative (FN): Incorrectly predicts false class.
- True Negative (TN): Correctly predicts false class.



Fig(3) SVM Fig(4) K-NN

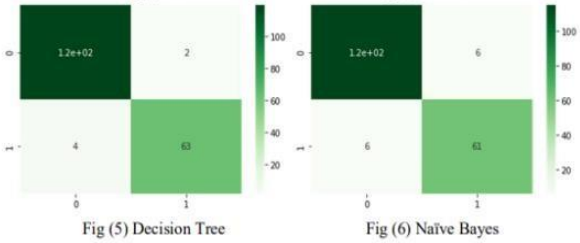
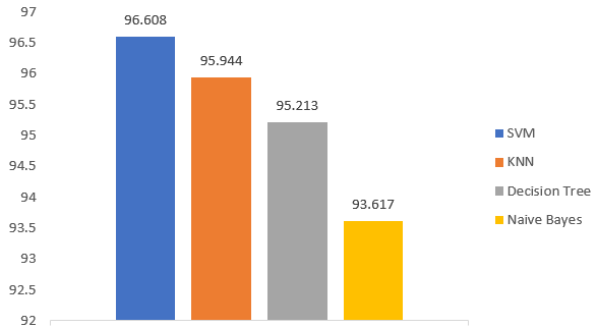


Fig (5) Decision Tree Fig (6) Naïve Bayes

The data was then standard scaled for analysis.
It had the following predictions:

Techniques	Accuracy
SVM	96.808%
KNN	95.744%
Decision tree	95.213%
Naïve Bayes	93.617%



V.REFERENCES

[1] S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, "Variability Measurement for Breast Cancer Classification Mammographic adaptive and convergent systems (RACS), Prague, Czech Republic, 2015, pp. 177–182.

[2] S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, "Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study," J. Teknol, vol. 65, pp. 73–81, 2013.

[3] A. T. Azar, and S. A. El-Said, "Performance analysis of support vector Neural Compute. Appl., vol. 24, no. 5, pp. 1163–1177, 2014.

[4] Tüba Kiyanand Tülay Yildirim (2004), Breast cancer diagnosis using statistical neural networks, Journal of electrical & electronics engineering, vol.4, pp.1149- 1153.

[5] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, Breast Cancer Prediction using Machine Learning, ISSN: 2277-3878, Volume-8 Issue-4, November 2019.