Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 2 Issue 6

Article Number: 230109

Breast Cancer Detection using Machine Learning Algorithms

Bhoomi Jain*and Neetu Singla

Department of Computer Science and Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

Abstract

Machine learning employs classification methods on datasets. The Machine Learning repository provided the cancer datasets that were used in this study, which were used for categorization. Breast cancer databases come in two varieties. There are various numbers of characteristics dispersed among these datasets. Breast cancer observes around 14% of all female cancers. One in every 28 women will develop breast cancer. To analyse patterns in datasets, machine learning algorithms like SVM, KNN, and decision trees are used. Computers are able to "learn" from their past mistakes and come up with solutions that are difficult for humans to come up with. According to the study, there are many effective algorithms for analysing the properties of data sets. This study compares and implements several well known classification methods, including Decision Trees, K Nearest Neighbor, SVM, Bayesian Network, and Naive Bayes on the Wisconsin Diagnostic dataset by calculating its classification accuracy, and its sensitivity and specificity value.

Keywords: Breast Cancer Detection; Machine Learning Algorithms; Wisconsin Diagnostic Dataset; Algorithm Performance Comparison; SVM and Decision Trees

1 Introduction

As time progresses, there are increasingly more individuals who might develop cancer due to various causes. The Kaggle website hosts an extensive array of cancer data for research purposes. Data are available in multiple formats, such as text, image, micro-array, gene expression, and others. Cancer data can be categorized into two types: Malignant (M) and Benign (B) [1]. Benign tumors are considered less harmful as their cells do not proliferate, whereas malignant tumors are harmful and carcinogenic once they start growing inside a human body. The dataset used in this study is a large dataset with unstructured data sourced from the UCI ML repository, related to breast cancer. Initially, the data undergoes pre-processing using machine learning (ML) techniques, followed by data cleaning, data selection, determining variable dependencies, and removing independent variables. A breast cancer dataset is labeled and classified as malignant or benign using ML algorithms [2]. This paper comprises six sections. The introduction is addressed in the first section [3]. The second section reviews the literature from esteemed authors in this domain. The third section outlines the methodology and machine learning techniques applied to this dataset. The fourth section discusses the procedures for data acquisition and calculations to determine accuracy. The fifth section presents the experiment conducted for data analysis, using line chart graphs to illustrate the results [4]. The sixth and final section of the paper, followed by references, presents the conclusion and future research directions. A machine learning experiment is more likely to be successful if it is well-planned, executed, and the results are rigorously evaluated [5].

*Corresponding author: bhoomij1230@gmail.com
Received: 14 November 2023; **Revised:** 01 December 2023; **Accepted**: 04 December 2023; **Published:** 31 December 2023

DOI: 10.57159/gadl.jcmm.2.6.230109.

2 Related Work

This section discusses previous research on machine learning techniques utilized by researchers for diagnosing breast cancer. Arpita Joshi and Ashish Mehta compared the classification results obtained using Random Forest, KNN, SVM, and Decision Tree methods. The Wisconsin dataset from the UCI repository was employed in their study.

Table 1: Summary of Research Papers on Breast Cancer Diagnosis Using Machine Learning Techniques

Algorithms	Datasets	Results
Naïve Bayes, SVM, J48, GRNN	Breast Cancer	GRNN & J48: 91%, Naïve Bayes & SVM: 89%
Decision Tree	WBC, WDBC, Breast Cancer	Feature selection improves WBC: 97%, Breast Cancer: 71.45%
SVM, C4.5, Naïve Bayes, KNN	WBC	SVM outperforms others: 97.13%
Ensemble, Naïve Bayes, SVM	WDBC	Ensemble & NB: 97%, SVM: 98.5%
Naïve Bayes, J48	WDBC	Naïve Bayes: 98%
MLP, J48, Rough Set	Breast Cancer	J48: 80%, MLP: 76%, Rough Set: 72.3%
IBK, SMO, BF Tree	WBC	SMO: 96.2%, IBK: 95.9%, BF Tree: 95.5%
J48, SMO, MLP, Naïve Bayes, IBK	WBC, WDBC, WPBC	WBC: J48 & MLP: 97.5%, WDBC: SMO: 98%
Classification: KNN, SVM, Naïve Bayes, K-means	WPBC	SVM & C5.0: 82%
Naïve Bayes, C4.5, SVM	WPBM	Naïve Bayes: 68%, C4.5: 74%, SVM: 75.75%

The most effective classifier, according to the simulation results, was KNN, followed by Random Forest, SVM, and Decision Tree. By integrating these techniques with feature selection/extraction methods, David A. Omondiagbe, Shanmugam Veeramani, and Amandeep S. Sidhu evaluated the performance of SVM, ANN, and Naive Bayes using the WDBC Dataset [2]. SVM-LDA was chosen over other methods due to its longer computation time, as indicated by the simulation results. Furthermore, data mining is frequently utilized in the medical field to predict and classify rare events, thereby aiding in the understanding of incurable diseases like cancer. The classification outcomes of data mining offer hope for early detection of breast cancer, which is why it is applied in this study. A summary of various research papers on breast cancer diagnosis using machine learning techniques is presented in Table 1.

3 Methodology

The methodology section outlines the application of various machine learning classifiers to the dataset. Specifically, Support Vector Machine (SVM), Decision Tree (C4.5), Logistic Regression, K-Nearest Neighbors (KNN), and Random Forests were utilized to analyze the data. The primary objective was to identify the most efficient and reliable algorithm for breast cancer detection [4].

3.1 Dataset

The dataset for the experiments as shown in Table 2 was sourced from Kaggle, focusing on the Wisconsin dataset which details characteristics of affected cell structures in breast cancer. This includes parameters like cell thickness, uniformity in cell size and shape, bare nuclei, single epithelial cell size, bland chromatin, normal nucleoli, and mitosis [1]. The dataset comprises 7,858 cases, organized into four expanded folders. Each folder corresponds to two types of tumors: benign and malignant.

Table 2: Description of the Wisconsin Diagnosis Breast Cancer (WDBC) Dataset

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Diagnosis Breast Cancer (WDBC)	32	569	2

3.2 Machine Learning (ML) Techniques for Classification

Several machine learning techniques are employed for data classification. These include Multilayer Perceptron, Bayesian Network, Naive Bayes, SVM, Decision Tree, Random Forest, and KNN. Key features of each method are as follows:

1. Support Vector Machine (SVM): SVM is is recognized as one of the effective approaches in the realm of machine learning, particularly when implementing kernel functions.

Its applications are diverse, encompassing facial recognition, database marketing, recommendation systems, text categorization, and cancer prediction, among other domains.

2. Random Forest (RF): Random Forest (RF) is an ensemble classifier that is based on the Decision Tree algorithm. RF is known to process large datasets, but it operates at a slower pace compared to other classifiers. RF generates a multitude of classification trees without the need for pruning. The pruning strategy, commonly associated with Classification and Regression Trees (CART), reduces the size of the tree by splitting the data into two subsets to find the best predictor in subsequent iterations. RF is capable of processing datasets with missing values and can estimate those missing values [6].

3. K-Nearest Neighbours (KNN): A classifier that uses the distance measure is called k-Nearest Neighbour. It is known as lazy learning or instance-based learning. The closest instance is used to complete the task locally [5]. The Manhattan distance method or the Euclidean distance are used to measure the distance. The classification in this technique is done using the smallest distance that was measured. The cost of learning the model is quite low, but it depends on the number of examples; as the number of instances rises, the cost climbs as well.

4. Bayesian Network: The probabilistic link between the relevant variables is represented by a directed acyclic graph called a Bayesian Network [2]. Each node represents a random (stochastic) variable with two or more potential states. It uses a set of variables on other variables to numerically deduce the probabilistic outcomes. It also has a reputation as a Belief Network (or Causal Probabilistic Network).

5. Decision Tree: A binary tree is constructed using the features present in datasets using the decision tree, a potent classification algorithm. ID3, C4.5, C5, J48, and CART are examples of popular algorithms. The process for selecting the root node is quite important. To find the variable and build a decision-making tree, this decision tree makes use of mathematical techniques like the Gini index, entropy, information gain, the chi-square test, etc. By dividing the variable into subgroups, homogeneity order must be preserved [7].

6. Naive Bayes: Conditional probability is the foundation of this classifier. The attributes included in the dataset are thought to be independent and reliable. In order to make it more effective, fewer parameters are used. This classifier can be used for applications such as sentiment analysis, language detection, and spam detection [8].

The comparative analysis of KNN, Naïve Bayes, and Random Forest in terms of various parameters is summarized in Table 3.

Table 3: Comparison Among KNN, Naïve Bayes, and Random Forest

Parameter	KNN	Naïve Bayes	Random Forest
Time Complexity (Training Phase)	O(1)	O(Nd)	O(MKlogN)
Problem Type	Classification and Regression	Classification	Classification and Regression
Accuracy	High	Requires a large number of records for high accuracy	High
Model Parameter	Non-Parametric	Parametric/Non-Parametric	Non-Parametric

4 Experimental Environment

4.1 Dataset Acquisition

The Breast Cancer Wisconsin Diagnostic dataset, obtained from the University of Wisconsin Hospitals Database, was utilized for the analysis[1]. This dataset provides comprehensive insights into the characteristics of breast cancer cases. It encompasses a total of 569 instances of Breast Cancer Wisconsin, with a distribution of 212 malignant (37.26%) and 357 benign (62.74%) cases, classified into two categories: malignant and benign.

4.2 Preprocessing

The initial data samples are acquired with a variety of attributes and values, often containing a wide range of issues such as outliers, noisy data, duplicates, missing values, and skewed data. To address these issues, preprocessing of the data is necessary. The data cleaning process involves eliminating or reducing missing data and noisy information. This can be achieved by deleting tuples, inputting missing values, and replacing numerical values with the mean attribute or the attribute mean of the corresponding class.

Additionally, data preprocessing techniques like feature selection, dimension reduction, and feature extraction are employed to modify data collection, making it compatible with machine learning algorithms [8].

4.3 Feature Extraction

After preprocessing, feature extraction is the subsequent step, where relevant features significant for breast cancer detection are identified and extracted from the pre-processed images. Techniques for feature extraction may include edge detection, texture analysis, or shape analysis. Following feature extraction, feature selection methods are employed to choose the most pertinent features that could enhance the machine learning model's accuracy. Some common feature selection methods are mutual information, principal component analysis (PCA), and recursive feature elimination [9].

4.4 Model Training and Validation

Post data preprocessing, machine learning methodologies such as classification, prediction, and estimation are applied to develop the model. To prevent overfitting, the model is trained and validated on a dataset separate from the one used for training. Test datasets estimate model error, while training sets are used for model construction. Techniques like Artificial Neural Networks (ANN), Decision Trees, SVM, and Bayesian Networks are utilized for predicting breast cancer[10]. The model's efficacy is tested by feeding it with new, labeled data, typically divided into training and testing sets through the train-test split method. About 75% of the data is used for constructing the model—known as the training set—while the remaining 25% serves as the test set to assess the model's performance. Post evaluation, the outcomes are analyzed to identify the algorithm that provides the highest accuracy and predictability for the presence of breast cancer. In the provided work, a comparative analysis of machine learning algorithms was conducted, focusing on key performance metrics like Accuracy, Precision, Recall, and F1 score. Figure 1 illustrates these comparisons in a comprehensive manner.

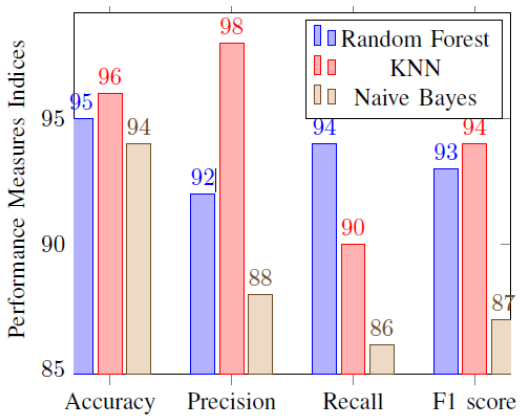

Figure 1: Comparative analysis of machine learning algorithms using performance metrics such as Accuracy, Precision, Recall, and F1 score.

Furthermore, a summary of different Machine Learning Techniques was compiled, as shown in Table 4. This table provides a concise overview of each algorithm, including their descriptions and a comparative analysis of their advantages and disadvantages.

Table 4: Summary of Machine Learning Techniques

Name of Algorithms	Descriptions	Advantages/Disadvantages
ANN	The output is generated through the combination of input and hidden layers.	Laborious operations and potentially subpar performance due to generic layered structure.
Decision Tree	Classification tree formed by nodes (variables) and leaves (decision outcomes).	Easy to interpret and fast learning process.
SVM	Identifies multiple hyperplanes in a high dimensional feature space and selects the best hyperplane for classifying input data into two classes.	Difficulty in handling large datasets.
Bayesian Networks	Makes estimates of probabilities rather than predictions [11].	Computationally expensive.

5 Results and Discussion

The performance of machine learning algorithms was evaluated using the Wisconsin Dataset. The models' performance was compared using metrics such as sensitivity, F1 score, confusion matrix, precision, and accuracy. Confusion matrices are particularly useful in assessing classification problems with two or more class types. They provide counts of True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP). Accuracy, the most common metric, is defined by the percentage of correctly predicted predictions from a specific sample size[4]. The accuracy rates for the Wisconsin dataset are displayed in Table 5 and Table 6

It was observed that while all classifiers demonstrated varying levels of accuracy, the Support Vector Machine (SVM) consistently outperformed others in the testing phase with an accuracy of 97.2%. Table V presents the confusion matrix, illustrating the performance of the classifiers in actual class conditions. According to the confusion matrix, SVM correctly predicted 556 out of 569 cases, including 201 actual cases of malignancy and 356 actual cases of benignity [11]. However, SVM also misclassified 11 benign cases as malignant and 1 malignant case as benign. This leads to SVM having higher accuracy compared to other classification methods [11]. The results indicate that SVM surpasses other classifiers with respect to sensitivity, precision, and F-Measure, all at 0.97%. In diagnosing malignant and benign classes in the Breast Cancer Wisconsin data, SVM is consistently superior to other classifiers [5].

Table 5: Confusion Matrix for Classifier Performance

Algorithm	Malignant	Benign
KNN	TP: 201, FP: 7	FN: 11, TN: 350
Logistic Regression	TP: 201, FP: 5	FN: 11, TN: 352
Random Forest	TP: 196, FP: 7	FN: 16, TN: 350
SVM	TP: 201, FP: 1	FN: 11, TN: 356

Table 6: Performance of Models During Testing Phase

Model	Recall	Precision	F1 Score	Accuracy
RF	94	92.2	93	93.74
K-nearest neighbour	91	98.3	94.3	96
Naive Bayes	86	89	86.4	94.5

6 Conclusion

Breast cancer remains the most prevalent cancer among women worldwide. Enhancements in diagnosis and prognosis are critical for health preservation. This study investigated two popular ML approaches for the categorization of Wisconsin Breast Cancer: Support Vector Machine (SVM) and Artificial Neural Networks (ANN). The findings indicate that SVM, with a precision of 97.5% and an Area Under the Curve (AUC) of 96.6%, outperforms all other algorithms, consistently delivering superior results in terms of diagnostic precision and accuracy for breast cancer. It is important to acknowledge the limitations of this study, which are primarily related to the confinement of the results to the WBCD database. Future work should consider applying the same methodologies to other databases to validate and compare the findings across different datasets. Furthermore, there is an intention to apply our and other ML algorithms to larger datasets with more disease classifications and additional parameters, aiming to achieve even more accurate results.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

Bhoomi Jain: Methodology, Software, Validation, Investigation, Data curation, Writing-original draft, visualisation.
Neetu Singla: Conceptualisation, Formal analysis, Resources, Writing-review & editing, supervision.

References

[1] B. Gayathri and C. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–5, IEEE, 2016.

[2] S. Kharya and S. Soni, "Weighted naive bayes classifier: a predictive model for breast cancer detection," *International Journal of Computer Applications*, vol. 133, no. 9, pp. 32–37, 2016.

[3] G. M. Hadi, "Benign and malignant breast cancer features based on region characteristics."

[4] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 114–118, IEEE, 2018.

[5] K. Sivakami and N. Saraswathi, "Mining big data: breast cancer prediction using dt-svm hybrid model," *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, vol. 1, no. 5, pp. 418–429, 2015.

[6] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[7] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 114–118, IEEE, 2018.

[8] M. F. bin Othman and T. M. S. Yau, "Comparison of different classification techniques using weka for breast cancer," in *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006: Biomed 2006, 11–14 December 2006 Kuala Lumpur, Malaysia*, pp. 520–523, Springer, 2007.

[9] B. Gayathri and C. Sumathi, "Mamdani fuzzy inference system for breast cancer risk detection," in *2015 IEEE international conference on computational intelligence and computing research (ICCIC)*, pp. 1–6, IEEE, 2015.

[10] Y. Buttan, A. Chaudhary, and K. Saxena, "An improved model for breast cancer classification using random forest with grid search method," in *Proceedings of Second International Conference on Smart Energy and Communication: ICSEC 2020*, pp. 407–415, Springer, 2021.

[11] S. Goyal, M. Sinha, S. Nath, S. Mitra, and C. Arora, "Breast cancer detection using machine learning," in *Communication, Software and Networks: Proceedings of INDIA 2022*, pp. 613–620, Springer, 2022.