International Journal of Scientific Research in Computer Science and Engineering
Vol.6, Issue.5, pp.55-59, October (2018)

Research Paper

E-ISSN: 2320-7639

Prediction of Breast Cancer Using Futuristic Approach

Kamal Bunkar[1*], Chhaya Arya[2]

[1]Institute of Computer Science, Vikram University Ujjain, India
[2]Research Scholar Pt. JNIBM, Vikram University, Ujjain, India

Available online at: www.isroset.org

Received: 17/Oct/ 2018, Accepted: 24/Oct/ 2018, Online: 31/Oct/2018

Abstract—Breast Cancer is one of the diseases that causes a higher number of deaths in a year. Breast cancer is the second most common cause of mortality among women, and majority of death occur due to its unavailable fact in Canada. Breast cancer is the most treatable type of cancer among all others, and early detection and thorough screening for the disease guarantee a higher patient survival rate. In order to find a reliable approach of predicting breast cancer, this paper offers a study about breast cancer prediction based on machine learning techniques. This study compares numerous patient clinical records to find an accurate model that can predict the likelihood of developing breast cancer. In this paper, a few machine learning models— kNN (k Nearest Neighbour), SVM (Support Vector Machine), ANN (Artificial Neural Network), and Naive Bayes classifier—are used. Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995) are two commonly used test data sets that are used to assess the performance of these models. The 10-fold cross-validation approach is utilised to calculate each model's test error. We downloaded the dataset from Kaggle.com in order to conduct this study. With a class distribution of 357 benign and 212 malignant cells, it has a total of 32 attributes, including the I.D. number, diagnosis (M= malignant and B= benign), and additional 30 real-valued input properties. The analysis's findings show a thorough trade-off between these tactics and also give a thorough assessment of the models. In practical use, it is anticipated that feature identification results will help doctors and patients prevent breast cancer.

Keywords—Breast Cancer, Support Vector Machine, Random Forest, k-Nearest Neighbor, Artificial Neural Network

I. INTRODUCTION

Breast cancer is the most prevalent cancer in women globally [1], and over the past few years, there has been a marked increase in the number of occurrences. In 2012, there were 1.67 million new instances of breast cancer in women worldwide, which is roughly 25% of all cancer cases [2]. If found early, breast cancer can be one of the most treatable cancers. The most popular diagnostic methods—mammography and fine-needle aspiration cytology—don't have a strong diagnostic capacity [3]. In 2014, there were 585,720 cancer fatalities and 1,665,540 new cancer cases in the US. Breast cancer accounted for 30% of cancer diagnoses in women and 15% of cancer-related fatalities in 2014 [4]. Therefore, there is a critical need to create better cancer diagnosis methods that are practical, affordable, and compatible with current imaging technologies. Machine learning methods and expert systems have been used to develop a range of predictive tools to aid informing physician judgements, improving the diagnostic capability. The structure of this paper includes five sections. A general introduction about the background of breast cancer. Related works about data mining methodology and supervised learning in cancer prediction is presented in section 2. Section 3 presents a detailed description about four data mining models, support vector machine (SVM), artificial neural network (ANN), and Naïve Bayes classifier. Section 4 focuses on the experiment process, which includes data sets description, feature space reduction, test results and comparison. Section 5 concludes the study findings and suggests possible improvements of this paper.

II. RELATED WORK

Many studies on breast cancer have been published in the past, and the majority of them had accurate classification results. Tumour size classification of breast thermal imaging using fuzzy C-Means algorithm is one such paper [5]. Fuzzy C-Means technique is used to analyse thermal pictures . Here, they applied the C-Means technique to the idea of grouping or clustering. The methodology groups data by classifying according to a colour component. In order to determine the stage of the cancer, they recommended utilising the Fuzzy C-Means clustering algorithm to group the data received from the breast cancer thermography images according to the image's properties.

The most common forms of breast cancer, according to the American Cancer Society, are invasive ductal carcinoma, invasive lobular carcinoma, and ductal carcinoma in situ [6]. The exact breast cells that are damaged determine the type of breast cancer. 'Carcinomas' is one of the worst types of breast cancer. Sarcomas, phyllodes, Paget disease, and angio-sarcomas, which affect the connective tissues or cells of the muscles, are other less prevalent forms of breast cancer. Machine learning approaches can be useful for prognosis, classification, and prediction of breast cancer [7][8].

Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm [9] is another study. utilising logistic algorithms and neural networks, they diagnose breast cancer utilising the ensemble method approach. Multiple models are used in ensemble approaches to obtain better results. They typically yield results that are more precise than those of a single model. The "Voting model" and "Averaging model" are the easiest ensemble models to use because they are both simple to comprehend and put into practise. Regression is performed using the Average model, while classification is performed using the Voting model.

The proposed LR [10] model for classification can be applied to mammography pictures by utilising feature extraction methods in image processing. The suggested model includes a cost function that, after a certain number of repetitions, returns the ideal value. The proposed model's prediction accuracy for LR and BPNN is 93.7%. In this method of constructing a system, binary logistic regression (BLR) [11] can distinguish between malignant and benign tumours. With an accuracy of 98.9%, the model will quantify a smaller set of attributes using those that have the best segregation between the two classes.

The classification problem utilising the WBCD was solved in 2015 using genetic programming (GP) based on ANN [18], i.e., a novel genetically optimised ANN (GOANN) algorithm. The best characteristics of GP were used to optimise the weights and architecture of ANN. Additionally, it put out redesigned crossover and mutation operators that widened the search window and improved efficiency. The destructive aspect of the crossover and mutation operators was abolished by this method. Using 10-fold cross validation, an accuracy of about 99.26% was attained.

FNAC is also widely adopted in the diagnosis of breast cancer, but the average correct identification rate is only 90% [17]. However, many methods have been generated and proposed for detecting a process that is more efficient than X-ray procedures such as, artificial intelligence and data mining.

2.1 The Architecture of Machine Learning Model

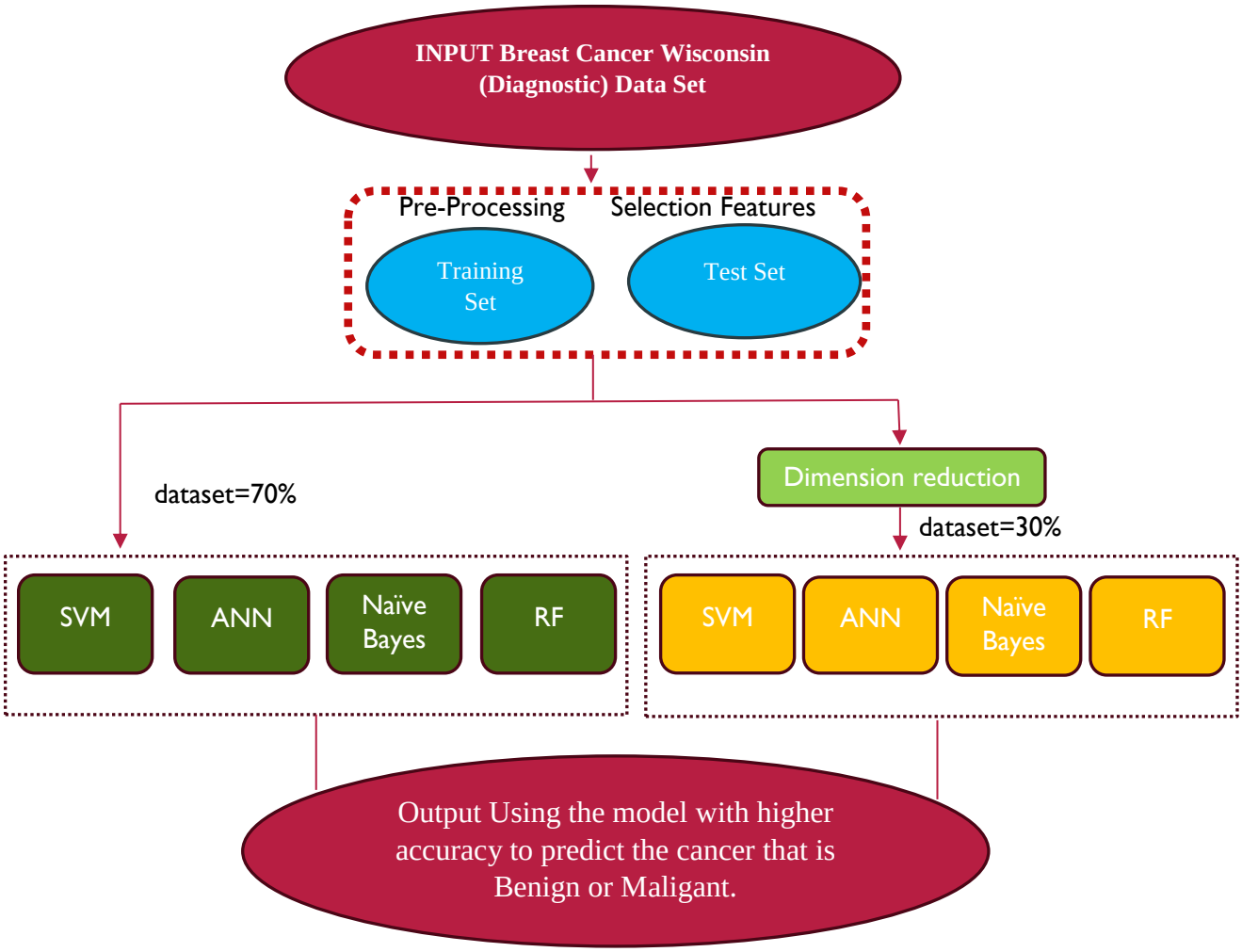

Figure 1 Architecture of Machine Learning Model

III. MACHINE LEARNING TECHNIQUE

3.1 Support Vector Machine (SVM): SVM is a powerful machine learning algorithm commonly used for classification tasks, including breast cancer detection. SVM works by finding a hyperplane that best separates data into different classes[12]. In the context of breast cancer detection, it can be used to classify breast cancer tumors as either malignant (cancerous) or benign (non-cancerous) based on various features extracted from medical images or other relevant data.

Algorithm 1 Support Vector Machine (SVM) Algorithm:

Begin
Step 1: Input a labeled data set (u1, v1), ...,(un, vn), vi ∈ Rd, and vi ∈ (1, +1), ui is a vector for feature and vi is a class label.
Step 2: The optimal hyperplane is defined as M * u + c = 0 to achieve vector for feature selection. And the binary classification can then be expressed as a function F(x) = sign (M * u + c). M is the weight vector, u is input feature and c is the bias
Step 3: All the elements with dissimilarity from the training data set must satisfy if $M * ui + c \geq +1$ if vi = +1 and $M * ui + c \leq -1$ if vi = −1. Repeat Step 3 until all the elements.
Step 4: Find M and c for the hyperplane to divide the data as Malignant or Benign.
End

3.2 Random Forest: Random Forest It is a supervised learning algorithm. An ensemble of decision trees is created, the bagging method is used to train the system. The ground methodology on which this technique is based is recursion. A random sample of size N is picked from the data set in each instance of an iteration. The dataset has been divided into training and testing set. It is obvious that diagnosis, radius_mean, texture_mean, perimeter_mean are influential variables, the other variables are of moderate influence but none of them can be neglected to increase the model accuracy. The confusion matrix of random forest is quite promising. There are only five observations that are misclassified as Benign and four observations are misclassified as Malignant and the accuracy equals 94.74%.

Algorithm 2 Random Forest (RF) Algorithm:

Begin
Step 1: Input dataset with S number of attributes, and the attributes s are chosen arbitrarily from S to form the nodes for decision tree.
Step 2: Choose a training set m for the above decision tree, which is error free. Step 3: Split tree based on chosen m and carry out preparation on each decision tree.
Step 4: Poll to find the optimal solution.
End

3.3 K-Nearest-Neighbor (kNN): K may be seen as the representation of the data points for training in close proximity to the test data point which we are going to use to find the class. A k-nearest-neighbor may be defined as the algorithm used to determine where a data set belongs to on the basis of the other data sets present around it. K-Nearest Neighbors (kNN)[14][15] is a simple and effective machine learning algorithm that can be used for breast cancer detection. It's a type of instance-based or lazy learning algorithm that classifies a data point based on the majority class of its k-nearest neighbors in the feature space. The accuracy of kNN is found to be 95.90% , there is only one observation that is misclassified as Benign and four observations are misclassified as Malignant.

Algorithm 3 K- Nearest Neighbor (kNN) Algorithm:

Begin
Step 1: Input Data set, kNN acquires all neighborhood data points. Data points that have a large amount of variation are significant elements in distance determination.
Step 2: Applying Euclidean distances formula to find distances between attributes (P1, P2, ..., Pn), and place them in 2-dimensional planes as Eq.(1)

$$D(P1, P2)^n = \sum_{i=0}^{n} D(P1 - P2)^2 \text{------1}$$

Step 3: Assume k as a positive integer and the first k distances are taken from the above arrangement. Using such distances k to measure the points k. For k greater than 0, ki is the number of points corresponding to the i^{th} category.
Step 4: The condition of ki > kj , only when i6=j, then keep x(data point) in the category i.
End

3.4 Naïve Bayes: Naive Bayes (NB) is one of the simplest, most effective and commonly used, machine learning techniques. It is a probabilistic classifier that classifies using the hypothesis of conditional independence with the pretrained datasets [16] it can also be applied to other types of classification tasks like medical diagnosis, including breast cancer detection.

Algorithm 5 Naive Bayes (NB) Algorithm

Begin
Step 1: Input data set with S attributes.
Step 2: There are two classes malignant or benign. For data X, we predict which class it belongs to such that P(Already occurred/Occurring) is maximum. P(Occurring/Already occurred) = (P(Already occurred/Occurring)*P(Occurring))/P(Already occurred)
Step 3: Repeat the class predictor such that P(Already occurred/Occurring) is maximum for all classes and pick maximum.
Step 4: Get the maximum class it belongs to.
End

4 Experimental Setup and Results: The Wisconsin breast cancer dataset [19] was utilised and examined. They were gathered at the University of Wisconsin Madison Hospitals by Dr. William H. Wolberg. There are 699 instances in the database, and Table I shows the 10 features and their types with class level. experimental tasks are performed using PyCharm that is an integrated development environment (IDE) used for programming in Python.

Table 1: Description of features in dataset

No.	Features	Information Factor
1	ID	Numerical
2	Clump Thickness	Numerical
3	Uniformity of Cell size	Numerical
4	Uniformity of Cell shape	Numerical
5	Marginal Adhesion	Numerical
6	Single Epithelial Cell size	Numerical
7	Bare Nuclei	Numerical
8	Bland Chromatin	Numerical
9	Normal Nucleoli	Numerical
10	Mitoses	Numerical

In the work, we used some statistical analysis that measure the test performance of different metrics. The performance of the classification techniques was evaluated by different metrics such as accuracy, sensitivity, specificity, precision and f1 measure and AUC[20].

True Positive (TP): The result of prediction correctly identifies that a patient has breast cancer.

False Positive (FP): The result of prediction incorrectly identifies that a patient has breast cancer. **True Negative (TN):** The result of prediction correctly rejects that a patient has breast cancer.

False Negative (FN): The result of prediction incorrectly rejects that a patient has breast cancer.

AUC: The AUC is used in the medical diagnostics framework which offers a standard assessment methodology based on the average of each point on the ROC curve. The AUC score will always in the range '0' to '1' the variant with a higher AUC rating provides better results in the classifier.

Table 2: Comparison of Accuracy, AUC

Algorithm	Accuracy%	AUC
SVM	96.35	98.5
RF	94.25	95.6
KNN	92.25	90.2
Naïve Bayes	93.00	95.4

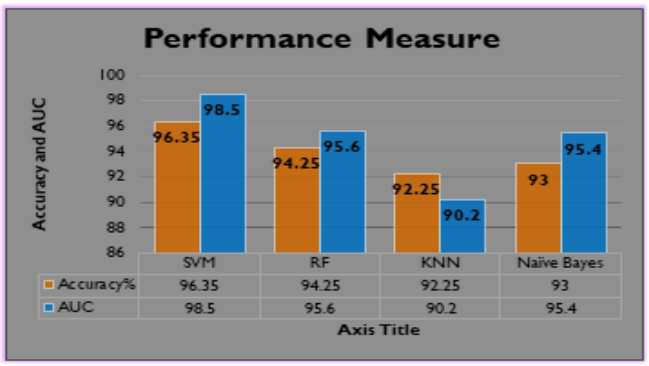

Figure 2: Graphical representation of Performance Measure Indices

Table 3: Comparison of Performance Parameters of Algorithms

Algorithm	Cancer Type	Performance Parmeter		
		Precision	Recall	F1-Score
SVM	Benign	0.96	.99	0.98
	Malignant	1.0	0.89	0.96
RF	Benign	0.98	0.96	0.95
	Malignant	0.81	0.86	0.84
KNN	Benign	0.92	.99	0.97
	Malignant	1.0	0.71	0.86
Naïve Bayes	Benign	0.97	0.96	0.97
	Malignant	0.90	0.95	0.92

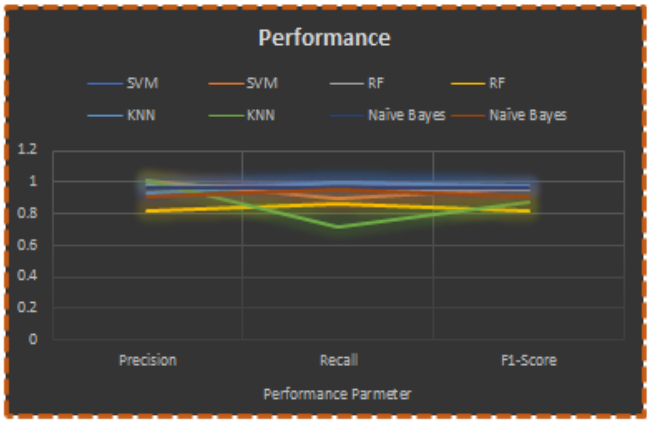

Figure 3: Graphical representation of Performance Measure Precision, Recall and F1 Score

IV. CONCLUSIONS

This work is the proposed some futuristic method of machine learning for diagnosis breast cancer, in which we have applied SVM, RF, KNN, and Naïve Bayes approach on Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995) dataset and used only 10 features for diagnosis of cancer. In future we will try on all features of dataset. Based upon the result obtained it shows that the classification performance alters based on the method that is selected. It has been observed form table 2 that each of the algorithm had an accuracy of more than 92%, to determine benign tumor or malignant tumor. From Table 3, it is found that RF and Naïve Bayes are the most effective in detection of the breast cancer as it had the best precision, recall and F1 score over the other algorithms. Thus, supervised machine learning techniques will be very supportive in early diagnosis and prognosis of a cancer type in cancer research.

REFERENCES

[1] A. Toloie Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A. R. Razavi, and L. Ghasem Ahmad, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," **2013.**

[2] Y. Li, H. Chen, L. Cao, and J. Ma, "A Survey of Computer-aided Detection of Breast Cancer with Mammography," **2016.**

[3] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Syst. Appl., Vol.**38**, Issue.**7**, pp.**9014–9022, 2011.**

[4] Forouzanfar, M. H., Foreman, K. J., Delossantos, A. M., Lozano, R., Lopez, A. D., Murray, C. J., and Naghavi, M., "Breast and Cervical Cancer in 187 Countries between 1980 and 2010: A Systematic Analysis," The Lancet, 378(9801), pp.**1461-1484, 2011.**

[5] Siegel, R., Ma J., Zou Z., and Jemal A., "Cancer Statistics 2014," CA: A Cancer Journal for Clinicians, Vol.**64**, Issue.**1**, pp.**9-29, 2014.**

[6] Octa Heriana, Indah Soesanti. Tumor size classification of breast thermal image using fuzzy C-Means algorithm. 2015 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications (ICRAMET), IEEE, **2015**

[6] B.M.Gayathri and C.P.Sumathi,"Mamdani fuzzy inference system for breast cancer risk detection", **2015.**

[7] Mohd,F.,Thomas,M, "Comparison of different classification techniques using WEKA for Breast cancer" **2007.**

[8] T Choudhury, V Kumar, D Nigam ,An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter Way, International Journal of Computer Science and Mobile Communication (IJCSMC) **2014.**

[9] Naresh Khuriwal, Nidhi Mishra. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. 2018 IEEMA Engineer Infinite Conference (eTechNxT), IEEE, **2018.**

[10] Mohd Rasoul Al-hadidi, Abdulsalam Alarabeyyat, Mohannad Alhanahnah, "Breast cancer detection using k-nearest neighbor machine learning algorithm", 9th International Conference on Developments in eSystems Engineering, pp.**35–39, 2016.**

[11] Ahmed F. Seddik, Doaa M. Shawky, "Logistic regression model for breast cancer automatic diagnosis", SAI Intelligent Systems Conference, pp. 150-154, 2015.

[12] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), pp.**1–4, 2016.**

[13] C. M. Dayton, "LOGISTIC REGRESSION ANALYSIS," **1992.**

[14] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit., Vol.**40**, Issue.**7**, pp.**2038–2048, 2007.**

[15] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," Springer, Berlin, Heidelberg, pp.**986–996, 2003.**

[16] W. H. Wolberg and O. L. Mangasarian, "Multi-surface method of pattern separation for medical diagnosis applied to breast cytology.," Proc. Natl. Acad. Sci., Vol.**87**, Issue.**23**, pp.**9193–9196, 1990.**

[17] Koza, J.R.; Rice, J.P. Genetic generation of both the weights and architecture for a neural network. In Proceedings of the IJCNN-91- Seattle International Joint Conference on Neural Networks, Seattle, WA, USA, 8–12 July; Vol.**2**, pp.**397–404, 1991.**

[18] Bhardwaj, A.; Tiwari, A. Breast cancer diagnosis using genetically optimized neural network model. Expert Syst. Appl., 42, pp.**4611– 4620, 2015.**

[19] W. H. Wolberg and O. L. Mangasarian, "Multi-surface method of pattern separation for medical diagnosis applied to breast cytology.," Proc. Natl. Acad. Sci., Vol.**87**, Issue.**23**, pp.**9193–9196, 1990.**

[20] A. D.-N. C. and Applications and undefined, "Performance evaluation of different machine learning techniques for prediction of heart disease," Springer. **2016.**

