Implementation of Prediction Model for Medical Diagnosis of Breast Cancer Using Machine Learning Algorithms and Classification Techniques

G. VINODA REDDY[1], M. SREENU NAIK[2], G PARVATHI DEVI[3], ARFA MAHVISH[4]
[1] *Professor, Department of CSE (AI & ML), CMR Technical Campus, Hyderabad.*
[2, 3] *Assistant Professor, Department of CSE (AI & ML), CMR Technical Campus, Hyderabad.*
[4] *Assistant Professor, Department of IT, CMR Technical campus, Hyderabad.*

Abstract— Cancer is the one of the uncured disease in the world, its variants too, among which breast cancer is one spreading over the world in female community. The death rate of this disease was setting a great immutable example day by day. Even best treatment facilities and excellent diagnosis medical equipements doctor are given a challenge in curing of the disease. In this paper we attempted to develop a prediction model to predict breast cancer at early stages by using machine learning algorithms and classification techniques using several machine-learning algorithms that are Random Forest, Naïve Bayes, Support Vector Machines SVM, and K-Nearest Neighbours K-NN, and chose the most effective. The experimental results show that SVM gives the highest accuracy 98.5%. The finding will help to select the best classification machine-learning algorithm for breast cancer prediction. We studied the results with breast cancer dataset. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis.

Indexed Terms-- Machine Learning, classification, Breast cancer, SVM, K-NN, Naïve Bayse, Random Forest, Efficiency.

I. INTRODUCTION

The second major cause of women's death is breast cancer (after lung cancer) 1. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women. There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparison between the performances of four classifiers: SVM, NB, C4.5 and k-NN which are among the most influential data mining algorithms in the research community and among the top 10 data mining algorithms. Our aim is to evaluate efficiency and effectiveness of those algorithms in terms of accuracy, sensitivity, specificity and precision.

The main objective of this paper is to analyze data from a breast cancer dataset using a classification technique in the field of medical bioinformatics to accurately predict the class in each case, using the weka data-mining tool and its use for classification. It first classifies the data set and then determines the best algorithm for the diagnosis and prediction of breast cancer disease. Prediction begins with identifying symptoms in patients, then identifying sick patients from a large number of sick and healthy patients. The main contributions of this work are Select the best classifier for breast cancer prediction, Comparison of different data mining algorithms on the breast cancer dataset and Identification of the best performance-based algorithm for disease prediction.

Total number of women dying in 2021 is approximately 963,000, according to the World Health Organization (WHO), Still, the organization predicts that the number could reach 2.9 million globally. Breast cancer can occur in women and rarely in men. The ICMR (Indian Council of Medical Research) recently published a report which stated that in 2020 the total number of new cancer cases is expected to be about 17.3 lakhs. An Indian woman is diagnosed with breast cancer in every four minutes. Breast cancer is a

disease that occurs but when a woman or a man is aware of this symptom, it immediately goes beyond its original stage. Breast cancer is a common and dangerous disease in women, cancer is the creation of abnormal cells that come into these cells genetically and mutated. Spreads throughout the body, leading to death in diagnosis and treatment. There are two types of breast cancer, Malignant and Benign. The first is classified as harmful has the ability to infect other organs and is cancerous, Benign is classified as non-cancerous. This disease infects the women's chest and specifically glands and milk ducts; the spread of breast cancer to other organs is frequent and could be through the bloodstream. Different techniques are used to capture breast cancer such as Ultrasound Sonography, Computerized Thermography, Biopsy (Histological images).

II. RELATED WORK

In recent work, Latchoumiet al. [13] proposed a weighted particle swarm optimization (WPSO) with smooth support vector machine (SSVM) for classification reached 98.42% .

C. Junaid Ahmed achieved 84.21% accuracy by using Adaptive Reasoning Theory, the Wisconsin data set was used, that contains 569 rows of data, and also contains 32 attributes.

Asri et al. [14]s howed that SVM can predict breast cancer better than Naive Bayes.

F. Hafizah [2] compared SVM and ANN using four different datasets of breast cancer. The researchers have demonstrated that SVM was better than ANN in performance and result.

Osman et al. [15] proposed a two-step SVM algorithm was presented by combining a two-step clustering algorithm with an efficient probabilistic vector support machine to analyze the
Wisconsin Breast Cancer Diagnosis WBCD with a classification accuracy of 99.10%.

G. S. Gc [1] worked on extracting features including variance, range, and compactness. They used SVM classification to analyse the performance. Their findings showed the highest variance of 95% and compactness 86%. According to their results, SVM can be considered as an appropriate method for Breast Cancer Prediction.

Turgut Machine learning procedure compared with SVM, KNN, DT, Logistic Regression, Random Forest, ADA Boost. In this various method checked and conclude that highest efficiency is 89% of random forest.

B. Narasingarao.M presents a survey of the work conducted to detect breast cancer using with different algorithm and conclude the efficiency of algorithm.

D. Nithya [13] applied the three categorizing methods such as Decision Tree, k-Nearest Neighbour, and Naïve Bayes for the different datasets. The authors also inspect the evaluation metrics of error rate. The implementation was focused on a type of attribute of a dataset.

E. Shilpa M and C. Nandini [14] implemented the algorithm using python and tested the same using dataset and achieved an accuracy of 94.74 and also reduces the time taken.

In this paper, we focused on the use of classification techniques in medical science and bioinformatics. Classification is the most commonly used data mining technique and uses a set of pre-classified examples to develop a model to classify the population of records The main objective of the classification technique is to accurately predict the target class for each case in the data.

III. METHODOLOGY

Our research uses a publicly available data set from the University of Wisconsin Hospitals Madison Breast Cancer Database . There are 11 attributes for each sample. Attributes 2 to 10 were used to represent instances respectively. The number of cases is 699. However, some instances are deleted due to missing attributes. There is one class attribute in addition to 9 other attributes. Each instance has one of the possibilities: Benin or malignant. One of the other numeric value columns is the instance ID column. Our data set includes two classes, as mentioned earlier. They are benign (B) and malignant (M). We further

analyzed the data and arrived at 30 attributes with 569 attributes.

The complete of 569 cases with 32 attributes was amassed for the Breast Cancer data set from kaggle. The attribute “diagnosis” described as the measurable and zero indicates patients are not having Breast cancer(B=Benign) and one indicates patients are having Breast Cancer (M = Malignant).Table I suggests the attributes values of Breast Cancer data set .The data set having consists of 569 tuples out of which 357 no breast cancer (B=Benign) cases and 212 breast cancer yes cases .

Table 1: Breast Cancer Data Set

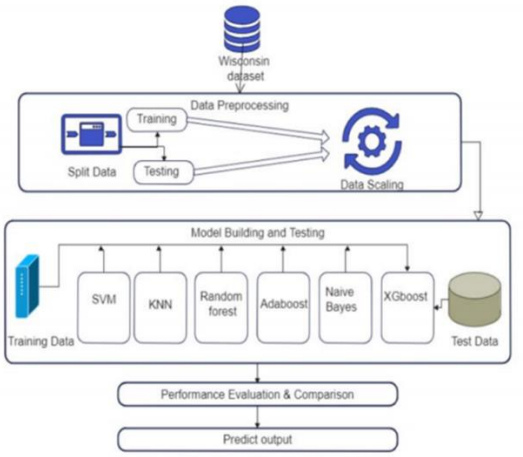
Attribute	Description
ID	Identification Number
Diagnosis	The diagnosis of breast tissues(M = malignant,B = benign)
Radius_ mean	Mean of distances from center to points on the perimeter
Texture_ mean	Standard deviation of gray-scale values
Perimeter_ mean	Mean size of the core tumor
Area_mean	Area Mean
Smoothness_mean	Mean of local variation in radius lengths
Compactness_ mean	Mean of perimeter^2 / area - 1.0
Concavity_ mean	Mean of severity of concave portions of the contour
concave points_ mean	Mean for number of concave portions of the contour
Symmetry_ mean	symmetry mean
Fractal_dimension_mean	mean for "coastline approximation" - 1
Radius_se	standard error for the mean of distances from center to points on the perimeter
Texture_se	standard error for standard deviation of gray-scale values
Perimeter_se	perimeter_se
Area_se	Area_se
Smoothness_se	standard error for local variation in radius lengths
Compactness_se	standard error for perimeter^2 / area - 1.0
Concavity_se	standard error for severity of concave portions of the contour
concave points_se	standard error for number of concave portions of the contour
symmetry_se	symmetry_se
Fractal_dimension_se	standard error for "coastline approximation" - 1

IV. PREDICTION MODEL

The principle destinations of this examination are to propose a technique that can create best Machine Learning algorithm for prediction of Breast Cancer disease.

From the perspective of automatic learning, breast cancer detection can be seen as a classification or clustering problem. On the other hand, we formed a model on the vast set of malicious and benign file data, we can reduce this problem to classification. For known families, this problem can be reduced to one classification only - having a limited set of classes, certainly including the breast cancer sample, it is easier to identify the right class, and the result would be more accurate than with clustering algorithms. In this section, the theoretical context is given on all the methods used in this research. After the features were extracted and selected, we can apply the machine learning methods to the data that we obtained. The machine learning methods to be applied, as discussed previously, are K-Nearest Neighbors, Support Vector Machines, Naive Bayes, Random Forest.

Fig 1. Predictipon Model



As Shown in diagram, we first Uploaded dataset From Wisconsin Breast Cancer Dataset. After that We did Preprocessing to the data and applied Machine Learning Models, which is used in this project to predict Breast cancer.

V. EXPERIMENTAL DISCUSSION

This Section describes the parameters and discusses the results of the assessment of the implemented machine learning methods.

Accuracy: The accuracy of detection is measured as the percentage of correctly identified instances. This is the number of correct predictions divided by the total number of instances in the dataset. It should be noted that the accuracy is highly dependent on the threshold was chosen by the classifier and may, therefore, vary between different sets of tests. Therefore, this is not the optimal method to compare different classifiers, but it can give an overview of the class. Therefore, the accuracy can be calculated using the following equation.

$$Accuracy = \left(\frac{(TP + TN)}{TP + FP + TN + FN}\right)$$

Where: TP = True positive; FN= False negative; FP= False positive; TN = True negative.

Recall: Recall, also commonly known as sensitivity, is the rate of the positive observations that are correctly predicted as positive. This measure is desirable, especially in the medical field because how many of the observations are correctly diagnosed the sensitivity or the true positive rate (TPR) is defined by: TP/(TP+FN)

Precision: Percentage of correctly classified elements for a given class: Precision =TP(TP+TN)

Comparison Between Techniques

Techniques	Accuracy Without Standard scale	Accuracy With Standard scale
SVM	57.89%	96.49%
KNN	93.85%	57.89%
Random Forest	97.36%	75.43%
Decision Tree	94.73%	75.43%
Naïve Bayes	94.73%	93.85%
Adaboost	94.73%	94.73%
XGboost	98.24%	98.24%

Table 2. Comparison of accuracy measures for C4.5, SVM, NB and k-NN.

	TP	FP	Precision	Recall	F-Measure	Class
C4.5	0.95	0.05	0.96	0.95	0.96	Benign
	0.94	0.04	0.91	0.94	0.93	Malignant
SVM	0.97	0.03	0.98	0.97	0.97	Benign
	0.96	0.02	0.95	0.96	0.95	Malignant
NB	0.95	0.02	0.98	0.95	0.96	Benign
	0.97	0.04	0.91	0.97	0.94	Malignant
k-NN	0.97	0.08	0.95	0.97	0.96	Benign
	0.91	0.02	0.94	0.91	0.93	Malignant

Table 3. Confusion matrix

	Benign	Malignant	class
C4.5	438	20	Benign
	14	227	Malignant
SVM	446	12	Benign
	9	232	Malignant
NB	436	22	Benign
	6	235	Malignant
k-NN	445	13	Benign
	20	221	Malignant

Table 4. Classifiers Performance

Evaluation criteria	Classifiers			
	K-NN	SVM	RF	NB
Time to build model (s)	0	0.08	0.28	0.01
Correctly classified instances	547	557	546	527
Incorrectly classified instance	22	12	23	42
Accuracy (%)	96.1	97.9	96	92.6
TP Rate	0,961	0,979	0,960	0,926
FP Rate	0,046	0,034	0,055	0,086
Recall	0,961	0,979	0,960	0,926
Precision	0,961	0,979	0,960	0,926

This is followed by SVM with an accuracy of 98.59% and the difference is 0.18% with AdaBoost. Based on the sensitivity results, AdaBoost and SVM have the same highest sensitivity at 99.44%, which the number

of patients who is correctly identified as identified with breast cancer is high. While AdaBoost has a huge percentage of specificity at 97.66% respectively, which has the highest number of patients who are not identified with breast cancer. Based on the performance evaluation, it is concluded that AdaBoost has the highest classification accuracy for the breast cancer data at 98.77% respectively.

Fig. 2. Comparison of performance evaluation for classification methods.

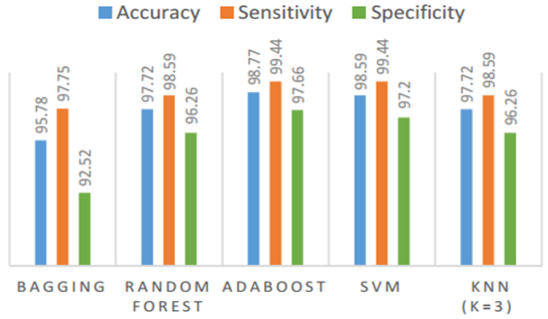
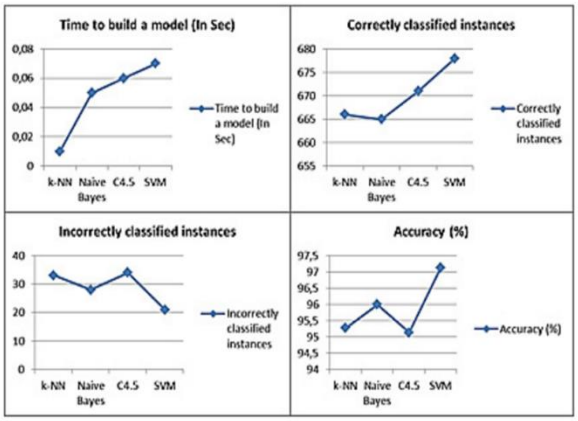


Fig. 3. Comparative graph of different classifiers.



The proposed method used in this study almost surpassed the other methods proposed by the author based on literature. Nonetheless, this is not including Rotation Forest with the highest accuracy rate at 99.48%, respectively. The author has implemented the feature selection namely genetic algorithm based to identify the best attributes, and thus the method has outperformed among others. While the classification accuracy of MLP is 99.04%, the second-highest accuracy respectively. The author has conducted the ratio of training and testing at 60:40 and 70:30, respectively. Based on the accuracy result, the training and testing of 70:30 have produced the best accuracy rate. AdaBoost has the same accuracy rate as the author in [11] that implemented C5.0, SVM, ANN methods. Moreover, it is the third-highest classification accuracy for the breast cancer dataset. Thus, AdaBoost model has improved the accuracy rate equitably and sensitivity rate significantly.

VI. SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is a supervised learning technique used for classification and regression. SVM is a well-known technique in machine learning and extensively implemented in cancer diagnosis. Principally, the function of SVM to classify the outcomes by mapping data between input vectors to a huge perspective space. Thus, the main objective of SVM is to determine the optimal hyperplane by dividing the dataset into classes. Linear classifier aims to fully utilize the distance between the decision hyperplane and the marginal distance, which is the nearest data point [3]. In this study, SVM is implemented to obtain the performance accuracy for malignant tumors and benign tumors. The complexity parameter C is used to control the flexibility of the process to draw lines to isolate the classes, and in this case, C-Support Vector Classification (C-SVC) is selected as the SVM type. The type of SVM is used Linear regression is selected as the key parameter in SVM classifier with normalized data.

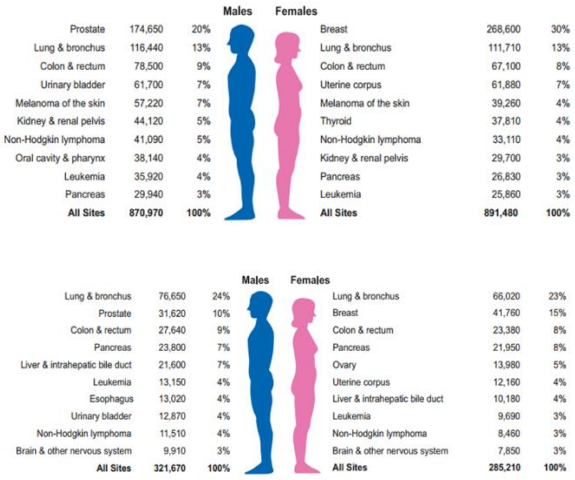


Fig 4. Ten Leading Cancer Types For The Estimated New Cancer Cases And Deaths by Sex

VII. CONCLUSION

To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we

employed four main algorithms: SVM, NB, k-NN and C4.5 on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. Of SVM reaches and accuracy of 97.13% and outperforms, therefore, all other algorithms. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate. Variation of machine learning techniques has been proposed to correctly classify the data namely as bagging, random forest, AdaBoost, SVM, and k-NN. According to the results, AdaBoost has achieved the highest accuracy at 98.77%. The accuracy results in this study are compared to the previous research works that used the breast cancer dataset. In addition to accuracy, specificity and sensitivity are also counted in this study to find the number of patients with breast cancer and without having breast cancer, approximately. The experiment with 2-fold, 3-fold, and 5-fold cross validation also has been implemented to the proposed methods. The result shows AdaBoost has produced best accuracy with 98.41% and 98.24%, respectively, for 2-fold and 3-fold cross validation. However, the accuracy rate is decreased with 5-fold cross validation, whereas SVM shows the highest accuracy at 98.60% with a 0.01 error rate. In future work, various ensemble techniques may be employed on the newly proposed methods to enhance the diagnostic accuracy of breast cancer. Furthermore, numerous feature selection techniques to manage complexity and a huge number of breast cancer data can be extended in the future.

REFERENCES

[1] Youness Khourdifi,"Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification",978-1-5386-4225-2/18/$31.00 ©2018 IEEE.

[2] Hiba Asri,"Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science 83 (2016) 1064 – 1069.

[3] Ravindra Changala, “Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods In Text Mining” in ARPN Journal of Engineering and Applied Sciences,Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.

[4] Nurul Amirah Mashudi,"Comparison on Some Machine Learning Techniques in Breast Cancer Classification",978-1-7281-4245-6/21/$31.00 ©2021 IEEE.

[5] Manav Mangukiya,"Breast Cancer Detection with Machine Learning",International Journal for Research in Applied Science & Engineering Technology,Volume 10 Issue II Feb 2022.

[6] Ravindra Changala, “Statistical Models in Data Mining: A Bayesian Classification” in International Journal of Recent Tredns in Engineering & Research (IJRTER), volume 3, issue 1, pp.290-293. in 2017.

[7] Rebecca L. Siegel,"Cancer Statistics, 2019",VOLUME 69 | NUMBER 1 | JANUARY/FEBRUARY 2019 .

[8] Mahesh Kotha, "A Survey on Predicting Uncertainty of Cloud Service Provider Towards Data Integrity and Economic" 2019 IJSRST | Volume 6 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X.

[9] Vinoda Reddy, “Recurrent Feature Grouping and Classification for action model prediction in CBMR”, International Journal of Data Management and Knowledge Process, Vol.7, No.5/6, November 2017, pp. 63 74. http://dx.doi.org/10.5121/ijdkp.2017.7605.

[10] Ravindra Changala, “Integrating Different Machine Learning Techniques for Assessment and Forecasting of Data” in Springer series, August-2015.

[11] Bharath Kumar Enesheti, Naresh Erukulla, Kotha Mahesh, “Edge Computing to Improve Resource Utilization and Security in the Cloud Computing System”, Journal of Engineering, Computing & Architecture, ISSN NO:1934-7197, Volume 11, Issue 12, DECEMBER - 2021.

[12] Ravindra Changala, “Development of Predictive Model For Medical Domains To Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms And Classification Techniques”, ARPN Journal of Engineering and Applied Sciences, VOL. 14, NO.6, MARCH 2019, ISSN 1819-6608.

[13] E. F. Hall, M., I. Witten, Data mining: Practical machine learning tools and techniques, Kaufmann,. 2011.

[14] M. Lichman, "UCI Machine Learning Repositry [Online]," Available:https://archive.ics.uci.edu/, 2013.

[15] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.