BREAST CANCER CLASSFICATION AND PREDICTION USING ML ALGORITHMS

Krina Vipul Shah

Dwarkadas.J.Sanghvi College of Engineering, Mumbai.
Third Year EXTC.

Abstract*— Breast cancer is the most common identified cancer in women in developing countries and poses serious threat to the lives of people. Breast cancer cells usually forms a tumor that can be seen on an x-ray usually or felt as a lump. The development of breast cancer is a multi-step process involving multiple cell types, and its prevention remains challenging in the world. Early diagnosis of breast cancer is one of the best approaches to prevent this disease. It is fatal in under half of all cases and is the leading cause of death from cancer in women, accounting for 16% of all cancer deaths worldwide. The objective of this paper is to present a report on prediction of breast cancer where we took advantage of those available technological advancements by building the accurate models as the diagnosis of this disease manually takes long hours and has lesser availability of systems. Data mining techniques contribute a lot for such development. For the classification of cancer, we have used classification techniques of machine learning to predict the category of new input. Implementation of models is done using Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree Classification and Random Forest Classification. With respect to the results of accuracy and precision each algorithm is measured and compared. From the experiments, it showed that Random Forest Classification is the best for predictive analysis with an accuracy rate of 94.15%.

***Keywords*— Breast Cancer, Classification, Data mining techniques.**

I. INTRODUCTION

Breast cancer is one of the most common cancer along with lung and bronchus cancer, prostate cancer, colon cancer, and pancreatic cancer among others. According to WHO, it is estimated that total of 627,000 women lost their lives due to breast cancer. Doctors can easily identify breast cancer using diagnostic mammogram, Magnetic resonance imaging (MRI) biopsy, breast ultrasound. Based on these results, doctors may recommend further tests or therapy. But this procedure might take a lot of time. So, for an early detection, Machine learning algorithms play a vital role in predicting the cancer. If a patient gets to know about the cancer at an early stage then the chances of survivability of the patient may increase. Thus, fatality rate might get lower. ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction. The accuracy of cancer prediction outcome has significantly improved by 15%–20% over the last few years, with the help of ML techniques. ML algorithms are much faster, advances the system, reduces human errors, and lowers manual mistakes. Machine learning techniques are used for the classification of benign and malignant tumors. The prior diagnosis of Breast Cancer can enhance the prediction and survival rate. The utilization of data science and machine learning approaches in medical fields proves to be prolific as such approaches may be considered of great assistance in the decision-making process of medical practitioners. There are two broad categories of machine learning.

Supervised learning: In supervised learning, the model train itself using labeled data and is used to estimate or map the input data to the desired output]. It makes the predictions on future instances.

Unsupervised learning: In unsupervised learning, data sets are not labeled. Thus, the model has to itself create a structure that fits a data by finding patterns or discover the groups of the input data.

This paper presents a model which accurately predicts if the person has a cancer or not.

II. LITERATURE SURVEY

Breast cancer is the second most dangerous cancer worldwide, where advanced stages at diagnosis, and rising incidence and mortality rates, make it essential to detect the cancer at an early stage to reduce the mortality rates. For this, prediction as well as classification of disease with the help of data mining techniques play an important role. Lots of breast cancer research has been reported in the literature of medical data analysis, and most of them turn up with good classification accuracies.

In "*Breast Cancer Prediction Using Data Mining Method"* by *Haifeng Wang and Sang Won Yoon* published in May 2015, they have used two datasets, Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995) with implementation of 10-fold cross-validation method to estimate the test error of each model. The objective of this paper was to discover an effective way to predict breast cancer by comparing and identifying an accurate model to predict the incidence of breast cancer based on various patient's clinical records. In this paper, four data mining models are applied i.e., support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree. To reduce the computational time and memory, feature space reduction technique was implemented. For this, principal component analysis (PCA) method was used to reduce the feature space which gave the better results. PCs-SVM turned out to be the best predictor model for WBC dataset and PCi-ANN for the WBDC dataset with accuracy of 97.47% and 99.63% respectively. PCA, a dimension reduction technique, manifested some advantages in terms of prediction accuracy and efficiency. This paper concluded that by using a raw dataset or a SEER dataset prediction results can turn out really better and much beneficial in all clinics and hospitals.

"A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis" by *Chen, H.L., Yang, B., Liu, J., Liu, D.Y (2011)* suggested to use Rough Set based SVM (RSSVM) to classify the cancer diagnosis with 5 features after the selection of 5 best suited features of WBCD dataset. The accuracy indicated 99.41%

for 50-50% training test partition, 100% for 70-30% of training-testing partition with selected features including Clump thickness, Uniformity of cell shape, Marginal adhesion, Bare nucleoli, and Mitosis. Experimental results demonstrate the proposed RSSVM can not only achieve very high classification accuracy but also detect a combination of five informative features, which can give an important clue to the physicians for breast diagnosis. Their new model was tested on Wisconsin Breast cancer data set (WBCD).

"Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods" by *Dursun Delen , Glenn Walker, Amit Kadam* proposed a system to compare the performance among Bayesian Network ,ANN and Hybrid Network for the accurate prediction of Breast Cancer Survivability. Delen used the SEER dataset of breast cancer to predict the survivability of a patient using 10-fold cross validation method. To help with interpreting the classification results, a hybrid network which combined both ANN and Bayesian Network was studied in this paper. The result indicated that the decision tree is the best predictor with 93.6% accuracy on the dataset as compared to ANN and logistic regression model. In this paper by using sensitivity analysis on neural network models provided us with the prioritized importance of the prognostic factors used in their study.

"Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach" written by *Pragya Chauhan and Amit Swami* proposed a system where they found that Breast cancer prediction is an open area of research. In this paper different machine learning algorithms are used for detection of Breast Cancer Prediction. The different machine learning algorithms used in this paper for prediction were Decision tree, random forest, support vector machine, neural network, linear model and naive Bayes. An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which over- came the limitations of the classical weighted average method. Genetic algorithm based weighted average method is used for the prediction of multiple models. The comparison between Particle swarm optimization (PSO), Differential evolution (DE) and Genetic algorithm (GA) was studied in detail and concluded that the genetic algorithm performs better for weighted average methods. One more comparison between classical ensemble method and GA based weighted average method was shown and concluded that GA based weighted average method performs better.

"Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction" by *Yixuan Li and Zixuan Chen* published in October 2018 have explored the relationship between breast cancer and some attributes so that the death probability of breast cancer can be reduced. This study used five different classification models for prediction: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Logistics Regression (LR) for the classification of breast cancer prediction. Here, two different datasets related to breast cancer: Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Breast Cancer Database (WBCD) were used for the classification. After comparing the accuracy, F-measure metric and ROC curve of five classification models, the result has shown that RF is selected as the primary classification model in this study. The model of this study is approved to possess clinical and referential values in practical applications.

In paper "*An Expert System for Detection of Breast Cancer Based on Association Rules and Neural Network"* by *Karabatak, M., and Ince, M. C (2009)* they discussed about a breast cancer prediction model that hybrids association rule and Neural Network (NN). In the model, association rule was proposed to reduce the feature space of breast cancer database and NN was used for classification. The model was tested using Wisconsin breast cancer database and found that the automatic diagnostic systems outperformed NN in terms of effectiveness and efficiency.

"*Application of artificial neural network-based survival analysis on two breast cancer datasets"* by *C.L. Chi, W. N. Street, W. H. Wolberg (Nov 2007)* used the ANN model for Breast Cancer Prognosis on the two data set. Both the dataset used nuclear morphometric features. They predicted recurrence and non-recurrence based on probability of breast cancer and grouped patients with bad (<5 years) and good (>5 years) prognoses. But the results were not as clear when the separation was done within subgroups such as lymph node positive or negative.

III. METHODOLOGIES AND IMPLEMENTATION

Machine learning has become a key technique in solving problems for Image processing and computer vision, for face recognition, motion detection, and object detection, and in many health-related problems, including the development of new medical procedures, the handling of patient data and records and the treatment of chronic diseases and predicting the disease at an early stage. Over the few years, machine learning algorithms have proved to be smarter, faster and much effective than regular process. Machine learning algorithms help to generate the data from insight by finding natural patterns which makes better decisions and predictions. Machine learning includes both supervised learning and unsupervised learning. Supervised learning develop predictive model based on both input and output data whereas unsupervised learning group and interpret the data based on only input data. Supervised learning uses classification and regression techniques to develop predictive models. Clustering is the most common technique used in unsupervised learning. In this paper, the Breast Cancer dataset used is obtained from Kaggle.com that provides many attributes for the prediction of breast cancer. In this dataset, there are 569 instances with 6 features viz. and there are no missing values in this dataset. Jupyter notebook is used for the implementation. For this dataset, classification techniques have been tried and tested such as Logistic Regression, Support Vector Machine (SVM), K- Nearest Neighbor, Decision Tree classification and Random forest Classification.

The detail description of the components and the activities performed against each component is mentioned below.

1) Data preprocessing:

Data processing is the first step in building the model. It is the most important step as it helps to clean, format, and organize the raw data, thereby making it suitable for the next process.

1.1. Importing the libraries:

Firstly, all the libraries are imported (numpy, pandas, matplotlib, etc.) in python. Then the dataset is to be imported.

1.2. Encoding the Categorical data:

Categorical data refers to the information that has specific categories within the dataset i.e. they contain label values instead of numeric values. But since in our dataset there are no label values so we have skipped this part.

1.3. Splitting the dataset:

Dataset is split into training set and set test. Training set denotes the subset of a dataset that is used for training the machine learning model. A test set is the subset of the dataset that is used for testing the machine learning model. The ML model uses the test set to predict outcomes. In this project, the ratio of training set to test set is 70:30.

1.4 Feature Scaling:

It is the last step in data preprocessing. Generally, datasets contains features which highly varies in range and unit. So these unequal range of features should be bought down to the same level of magnitudes for a better output.

2) ML algorithms

2.1 Logistic Regression:

Logistic regression in machine learning is used to predict the probability of dependent data variable by analyzing the relationship between one or more existing independent variables. In our dataset we have dependent variable in the binary form as '0' and '1'. It is one of the simplest ML algorithm and provides a constant output.

The estimated accuracy of logistic Regression achieved in this model is 92.397%

Classification report:

	precision	recall	f1-score	support
0	0.90	0.89	0.90	63
1	0.94	0.94	0.94	108
accuracy			0.92	171
macro avg	0.92	0.92	0.92	171
weighted avg	0.92	0.92	0.92	171

2.2 Support vector Machine:

Support vector machine is another type of supervised learning used in both regression and classification problems. It creates a hyper plane while classifying the objects. An SVM algorithm builds a model that assigns new examples to at least one category or the opposite, making it a non-probabilistic binary linear classifier. It is an powerful method to build a classifier. SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature space.

The accuracy achieved using SVM algorithm is 91.228%

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.90	0.88	63
1	0.94	0.92	0.93	108
accuracy			0.91	171
macro avg	0.90	0.91	0.91	171
weighted avg	0.91	0.91	0.91	171

2.3 K-Nearest Neighbor (k-NN)

KNN is a classifier algorithm where the training is predicted how similar may be data from other. It works by finding the distances between a point and all examples within the data. It is a lazy algorithm. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

The accuracy obtained from KNN algorithm is 86.842 %

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.89	0.85	47
1	0.93	0.94	0.88	67
accuracy			0.87	114
macro avg	0.86	0.88	0.87	114
weighted avg	0.88	0.87	0.87	114

2.4 Decision Tree:

A decision tree is a tree-like structure in which internal node represents a "test" on an attribute each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. They use a layered splitting process, where at each layer they struggle to separate the info into two or more groups. It is a top down approach.

The accuracy achieved from Decision Tree is 91.812 %.

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.86	0.89	63
1	0.92	0.95	0.94	108
accuracy			0.92	171
macro avg	0.92	0.91	0.91	171
weighted avg	0.92	0.92	0.91	171

2.5 Random Forest Classification:

Random forest algorithm is a supervised learning. In this classifier, higher the number of trees it gives high accuracy result. It can be used for both classification and regression problems and it also handles the missing values. The best thing about this algorithm is that it does not overfit the model, even if there are more number of trees.

Random forest classification has achieved the highest accuracy of about 94.152%.

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.89	0.92	63
1	0.94	0.97	0.95	108
accuracy			0.94	171
macro avg	0.94	0.93	0.94	171
weighted avg	0.94	0.94	0.94	171

Implementation:

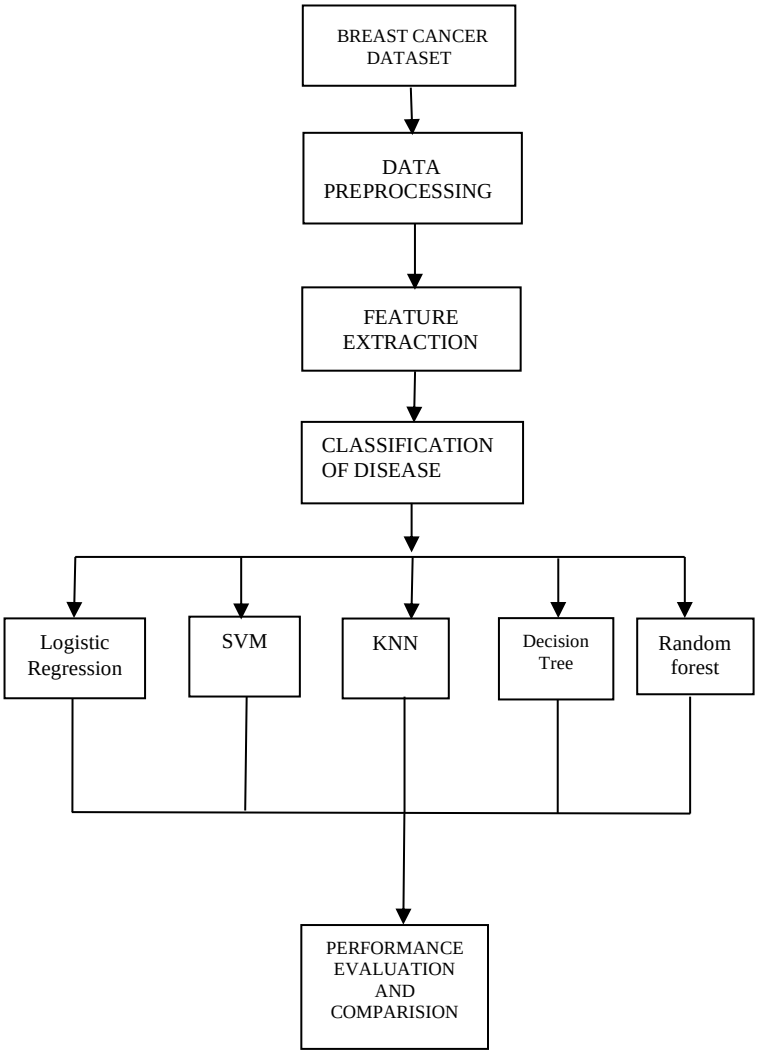

Fig 1.: System Architecture

IV. RESULTS AND DISCUSSION

In this paper, I have verified 5 different types of machine learning algorithms for the accurate prediction of Breast Cancer. Their accuracies are compared and computed with each other, so that the best prediction model can be used by doctors in real life to identify the Breast Cancer relatively faster than previous methods. Out of these models, Random Forest Classification turned out to be the best predictor with 94.152% accuracy. Logistic regression came out to be second with accuracy of 92.397%.

SR. NO	MODEL	SCORE
1.	Random Forest Classification	94.152
2	Logistic Regression	92.397
3	Decision Tree	91.812
4.	Support Vector Machine	91.228
5.	KNN	86.842

V. CONCLUSION

In this paper, I have taken the breast cancer dataset for creating an efficient Classifier since it is highly essential in any clinical investigation to determine the nature of a disease, especially a life-threatening ailment like cancer. The results of all classification techniques are clearly mentioned with necessary results. The comparative study of multiple prediction models for breast cancer survivability using a dataset provided us with an insight into the relative prediction ability of different data mining methods. This type of projects will really help doctors to identify the cancer much easier than previous methods. It will also help in the development of medical research. Further accurate classification would enable clinicians to propose drugs for a new patient based on whether his/her features correspond to a good or bad prognosis. So this algorithms can be deployed in clinics and hospital for early detection and hence cure and increase the survival rate of the patients.

VI. FUTURE SCOPE

The world is turning towards ML for every sort of decision making, start-ups, medical platforms and so many. Effective machine learning implementation will help healthcare professionals in better decision-making, identifying trends and innovations, and improving the efficiency of research and clinical trials. In the recent past, cancer research has been performed continuously. The majority of scientists have embraced various methods that seek to establish types of cancers. Also, the research practices have led to the development of new strategies through which cancer

treatment outcomes could be predicted early. The proposed system greatly reduces the cost of treatment and improves the quality of life by predicting breast cancer at early stage of development which will help save lives of thousands of women. In this study, there are still some limitations that should be solved in further work. For example, though there also exist some indices people have not found yet, this study only collects the data of 6 attributes in this experiment. There is also scope for analysis using other classifiers and dimensionality reduction techniques which may help in better understanding of larger datasets with many mor e features in near future. The limited raw data has an effect on the accuracy of results. In addition, the Random Forest Algorithm can also be combined with other data mining technologies to obtain more accurate and efficient results in the future work. The future scope will focus on exploring more of the dataset values and yielding more interesting outcomes. This study can help in making more effective and reliable disease prediction and diagnostic system which will contribute towards developing better healthcare system by reducing overall cost, time and mortality rate.

REFERENCES:

[1] C.L. Chi, W. N. Street, W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets", American Medical Informatics Association Annual Symposium, pp. 130-134, Nov. 2007.

[2] "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction", by Yixuan Li, Zixuan Chen October 18, 2018

[3] "Breast Cancer Prediction Using Data Mining Method" by Haifeng Wang and Sang Won Yoon, Department of Systems Science and Industrial Engineering State University of New York at Binghamton Binghamton, May 2015.

[4] "Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach" by Pragya Chauhan and Amit Swami, 18 October 2018

[5] D. Delen, G. Walker, A. Kadam, " Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113127, 2004.

[6] "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis" by Wenbin Yue, Zidong Wang, 9 May 2018.

[7] "Ultrasound characterisation of breast masses", The Indian journal of radiology imaging by S. Gokhale., Vol. 19, pp. 242-249, 2009. K. Elissa, "Title of paper if known," unpublished.

[8] "BREAST CANCER PREDICTION USING MACHINE LEARNING TECHNIQUES" by K. Varshini, Ram Kishore Sethuramamoorthy, Vipin Kumar, S. Abitha Shree, S. Deivarani.

[9] D. Dubey, S. Kharya, S. Soni and – "Predictive Machine Learning techniques for Breast Cancer Detection", International Journal of Computer Science and Information Technologies, Vol.4(6),2013,1023-1028.

[10] Vikas Chaurasia, BB Tiwari and Saurabh Pal – "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms and Computational Technology.

[11] Karabatak, M., and Ince, M. C.,2009, "An Expert System for Detection of Breast Cancer Based on Association Rules and Neural Network," Expert Systems with Applications, 36(2), 3465-3469.

[12] Lavanya, D.,2012,"Ensemble Decision Tree Classifier for Breast Cancer Data," International Journal of Information Technology Convergence and Services, 2(1), 17-24.

[13] Zheng, B., Yoon, S. W., and Lam, S. S., 2014, "Breast Cancer Diagnosis Based on Feature Extraction Using a Hybrid of K-means and Support Vector Machine Algorithms," Expert Systems with Applications, 41(4), 1476-1482.

[14] Huang, M. L., Hung, Y. H., and Chen, W. Y., 2010, "Neural Network Classifier with Entropy Based Feature Selection on Breast Cancer Diagnosis," Journal of Medical Systems, 34(5), 865-873.

[15] Fallahi, A. and Jafari S., 2011, "An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network," International Journal of Advanced Science and Technology, 34, 65-70.

[16] "Breast Cancer Diagnosis by Different Machine Learning Algorithms Using Blood Analysis Data" by Muhammet Fatih Aslan*, Yunus Celik, Kadir Sabanci, Akif Durdu, Department of Electrical-Electronic Engineering, Karamanoglu Mehmetbey University, Karaman-70100, Turkey.

[17] "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model" by K.Sivakami , Department of Computer Application, Nadar Saraswathi College of Arts &Science , Theni, August-2015.