EARLY PREDICTION OF BREAST CANCER BY VOTING CLASSIFIER

[1]Dr.N.Dhanalakshmi, [2]Dr.S.Satheesbabu, [3]Dr.P.Gokulakrishnan

Abstract

Among women, breast cancer is the most frequently diagnosed cancer and a leading cause of death. In the developed world, between 1 in 8 and 1 in 12 women will have breast cancer during her lifetime. There are two primary types of breast cancer risk. The first type represents the probability that an individual will contract breast cancer during a specified period of time. The second type reflects the probability that a mutation will occur in a high-risk gene. Previous works found that adding inputs to the widely-used Gail model improved its ability to predict breast cancer risk. The main objective is to predict analytics model to diagnose breast cancer stages of patients. The main objective of this work is to detect and analyze breast cancer. It predicts the stages of the cancer and gives as the accurate result. **In** this work, to investigate a dataset of medical patient records for hospital sector using machine learning technique and to identify patients having breast cancer stages from given dataset attributes. Then the accurate result is found by naive Bayesian algorithm with precision, recall, F1score.

1 INTRODUCTION

Machine learning (ML) is used to predict the future from past data and it is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data. It has three subtypes supervised, unsupervised and reinforcement. In our analysis we have used Supervised Machine Learning and its algorithms to build a classification model from the data collected. The dataset used for analysis may contain inconsistencies like missing values, outliers and it has to be handled before being used to build the model. After the implementation of all the algorithms with the information provided the result is determined based on the accuracy of the used algorithms.

[1] *Professor, Department of CSE., PSNA College of Engineering and Technology, ndmugi@gmail.com*
[2] *Associate Professor, Department of CSE., PSNA College of Engineering and Technology, sbsdgl@gmail.com*

[3]*Professor, Department of CSE., PSNA College of Engineering and Technology, gokulakrishnan@psnacet.edu.in*

2 MOTIVATION

Breast cancer is cancer that forms in the mammary gland. At present there are many women suffering from this deadly disease and the death rate of people suffering from this disease are increasing day by day. In order to control, the only solution is detecting it earlier and undergoing appropriate diagnosis and treatment based on the stage of cancer which may help to steadily decline the death rate of patients. Implementation of machine learning algorithm techniques such as Supervised Machine Learning algorithms which includes Logistic regression, Decision Tree and Support vector machine can helps to build a classification model to predict the breast cancer from its 2 features. Some of the signs and symptoms of breast cancer that are used to classify the stages of cancer are:

- A breast lump or thickening that feels different from the surrounding tissue.
- Change in size, shape or appearance of a breast.
- Change in skin over the breast, such as dimpling.
- A newly inverted nipple.
- Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin.
- Redness or pitting of the skin over your breast, like the skin of an orange.

3 Related Work

Comparison of Machine Learning Methods for Breast Cancer Diagnosis (Ebru Aydındag Bayrak, Pınar Kırcı,Tolga Ensari 2019) discussed that two popular machine learning techniques for Wisconsin Breast Cancer classification. Artificial Neural Network and Support Vector Machine are used as ML techniques for the classification of WBC (Original) dataset in WEKA tool. The effectiveness of applied ML techniques is compared in term of key performance metrics such as accuracy, precision, recall and ROC area. Based on the performance metrics of the applied ML techniques, SVM (Sequential Minimal Optimization Algorithm) has showed the best performance in the accuracy of 96.9957 % for the diagnosis and prediction from WBC dataset.Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features (Zhiqiong Wang, Mo Li, Huaxia Wang 2019) explored a breast CAD method supported feature fusion with Convolution Neural Network (CNN) deep features. First, we propose a mass detection method supported CNN deep features and Unsupervised Extreme Learning Machine (UELM) clustering. Second, we build a feature set fusing deep features, morphological features, texture features, and density features. Third, an ELM classifier is developed using the fused feature set to classify benign and malignant breast masses. Early detection of lumps can effectively reduce the death rate of carcinoma.The computer-aided diagnosis (CAD) for carcinoma can help address this issue. Although the old diagnosis method has been widely used, its accuracy still has to be improved. The standard of the handcrafted feature set directly affects the diagnostic accuracy, and hence an experienced doctor plays a awfully important role within the process of manual feature extraction.Within the stage of mass detection, a method based on sub-domain CNN deep features and US-ELM clustering is developed. In the stage of mass diagnosis, an ELM classifier is utilized to classify the benign and malignant breast masses using afused feature set, fusing deep features,morphological features, texture features, and density features. In the process of breast CAD, the choice of features is the key in determining the accuracy of diagnosis.

4 Proposed System

Fig.4.1 shows the overall system architecture diagram of the proposed system.In this, one input is taken for processing from user and the other from past dataset.

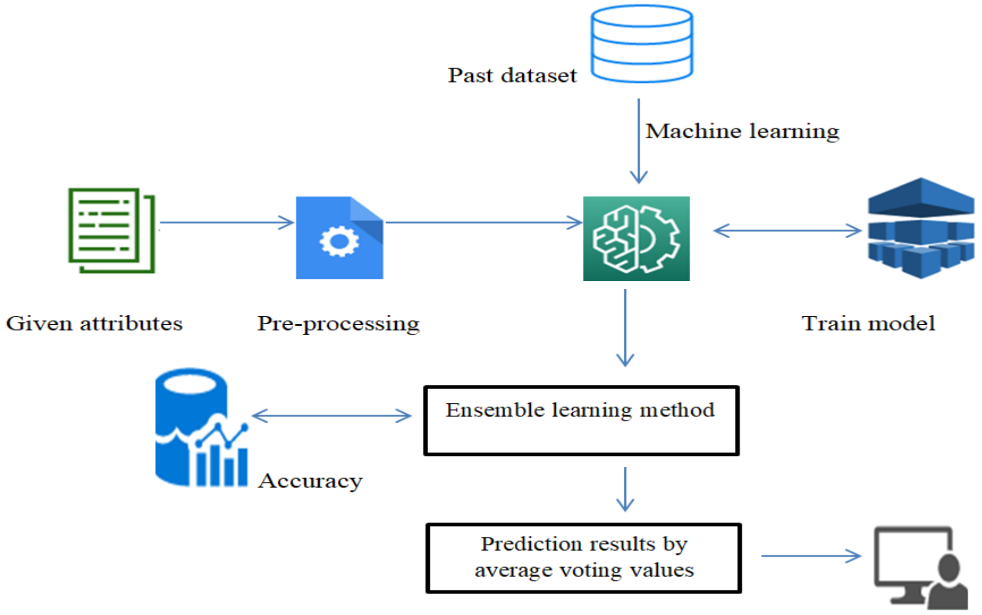


Fig.4.1 Architecture Diagram

Then these are preprocessed in order to avoid duplication and error values, which then undergoes implementation of various supervised machine learning algorithms. The result of each algorithm is again processed by ensemble method and it gives the best accuracy result among the results of algorithms by voting classifier method and the final results are displayed in the GUI screen.

4.1 Improvisation of Machine Learning by ensemble learning method using voting Classifier

Voting is one of the most straightforward Ensemble learning techniques in which predictions from multiple models are combined. The method starts with creating two or more separate models with the same dataset. Then a Voting based Ensemble model can be used to wrap the previous models and aggregate the predictions of those models. After the Voting based Ensemble model is constructed, it can be used to make a prediction on new data. The predictions made by the sub-models can be assign weights. Stacked aggregation is a technique which can be used to learn how to weight these predictions in the best possible way.

Comparing Algorithm with prediction in the form of best accuracy It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with sk-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created.

The same idea applies to model selection. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

4.2 FINDING DIFFERENT STAGES OF CANCER

The stage of a cancer is a measurement of the extent of the cancer and its spread. Opted treatment can be provided based on the stage. The standard staging system for breast cancer uses a system known as TNM, where: T stands for the main (primary) tumor ,N stands for spread to nearby lymph nodes which is located under arms and ,M stands for metastasis (spread to distant parts of the body). stage 0 is the earliest stage in breast cancers which is ranged from stage I (1) through IV (4). As a rule, the lower the number shows that the spread of cancer low and indicates lower stage and higher number like stage IV shows cancer has spread more and indicates critical stage. The grade of a tumor indicates what the cells look like and gives an idea of how quickly the cancer may grow and spread. Tumors are graded between 1 and 3.

Understanding the stage of the cancer in Fig.4.2 helps doctors to predict the likely outcome and design a treatment plan for individual patients.

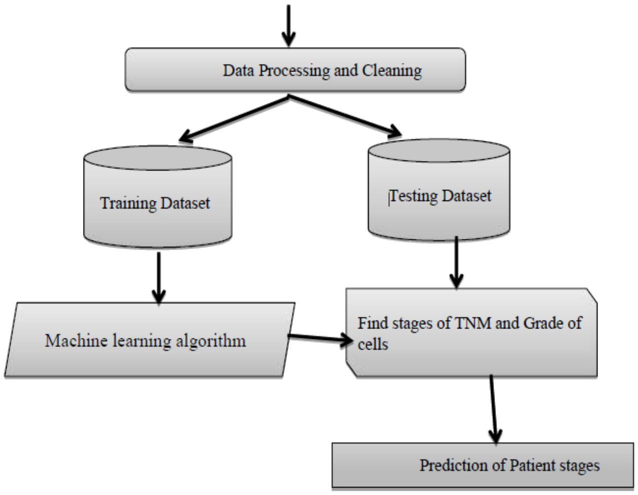

Data Processing and Cleaning
Training Dataset
|Testing Dataset
Machine learning algorithm
Find stages of TNM and Grade of cells
Prediction of Patient stages

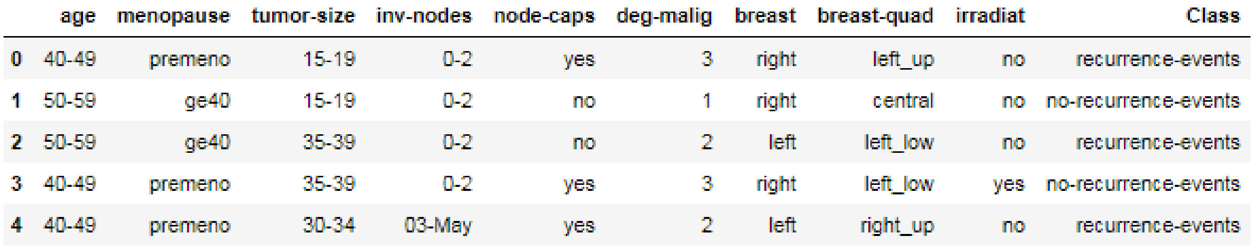
Fig.4.2 Identification of stages of breast cancer

5 PERFORMANCE ANALYSIS

5.1 PREPARING THE DATASET

The dataset is now supplied to machine learning model on the basis of this data set the model is trained. Every new patient's details filled at the time of appointments form acts as a test data set. After the operation of testing, model predicts whether the new patient is a fit case for affecting breast cancer or not based upon the inference it concludes on the basis of the training data.

sets.



	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
0	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurrence-events
1	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recurrence-events
2	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurrence-events
3	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recurrence-events
4	40-49	premeno	30-34	03-May	yes	2	left	right_up	no	recurrence-events

5.2 Results and Discussion

The analysis of dataset is done by supervised machine learning algorithm to capture information such as, variable identification, univariate analysis, bi-variate analysis etc. Additionally, we have to discuss the performance from the given hospital dataset with evaluation classification report and identify the confusion matrix. The data validation, preparing and visualization will be applied on the entire given dataset. Thus the result shows the effectiveness of the proposed machine learning algorithm technique which provides us with best accuracy, precision, Recall and F1 Score.Table 5.1 shows the performance of various algorithms measured by several parameters like precision,recall,F1-Score ,Sensitivity, Specificity and accuracy.

Parameters	Logistic Regression (LR)	Decision Tree (DT)	Random Forest (RF)	Support Vector Machines (SVM)	K-Nearest Neighbour (KNN)	Naive Bayes algorithm (NB)
Precision	1	1	1	0.89	0.82	1
Recall	1	1	1	1	0.98	1
F1-Score	1	1	1	0.94	0.89	1
Sensitivity	1	1	1	1	0.98	1
Specificity	1	1	1	0.72	0.48	1
Accuracy (%)	100	100	100	91.66	83.33	100

Table 5.1 shows the performance of various algorithms

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall.

Sensitivity = (True Positive) / (True Positive + False Negative)

Specificity: Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative).

Specificity = (**T**rue **N**egative) / (**T**rue **N**egative + **F**alse **P**ositive)

Accuracy calculation is made as Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision calculated as TP / (TP + FP)

Recall calculated as TP / (TP + FN)

F1-Score calculated by 2*(Recall * Precision) / (Recall + Precision)

Table 5.2 shows confusion matrix which describes the visualization of the performance of a classification model on a test data.It provides a summary of number of correct and incorrect predictions with count values and provide best accuracy among these algorithms .

Parameters	**LR**	**DT**	**RF**	**SVC**	**KNN**	**NB**
TP	25	25	25	18	12	25
TN	59	59	59	59	58	59
FP	0	0	0	0	1	0
FN	0	0	0	7	13	0
TPR	1	1	1	0.72	0.48	1
TNR	1	1	1	1	0.98	1
FPR	0	0	0	0	0.01	0
FNR	0	0	0	0.28	0.52	0
PPV	1	1	1	1	0.92	1
NPV	1	1	1	0.89	0.81	1

Table 5.2 Performance measurements confusion matrix

6 Conclusion and Future Work

The process of analysis started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finding the patient stages and grade with parameter like accuracy, classification report and confusion matrix on public test set of given attributes by ensemble learning method of voting classifier accuracy is 100%.

Hospital wants to automate the detection of the breast cancer from eligibility process (real time) based on the account detail. To automate this process we show the prediction result in web application or desktop application. To optimize the work to be implemented in Artificial Intelligence environment

References

1. Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-3). IEEE.

2. Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018, April). Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE.

3. Turgut, S., Dağtekin, M., & Ensari, T. (2018, April). Microarray breast cancer data classification using machine learning methods. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-3). IEEE.

4. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.

5. Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., & Xin, J. (2019). Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. *IEEE Access*, *7*, 105146-105158.

6. Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C. D., & Cha, K. H. (2018). Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, *38*(3), 686-696.

7. Rakhlin, A., Shvets, A., Iglovikov, V., & Kalinin, A. A. (2018, June). Deep convolutional neural networks for breast cancer histology image analysis. In *International Conference Image Analysis and Recognition* (pp. 737-744). Springer, Cham.

8. Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, *16*(3), 841-850.