FORECASTING OF BREAST CANCER USING VOTING CLASSIFIER

N.Dhanalakshmi[1], R.Karthikeyan[2], A.Thomas Paul Roy[3]
[1] Professor, PSNA College of Engineering and Technology, Dindigul, India
[2] Professor, PSNA College of Engineering and Technology, Dindigul, India
[3] Professor, PSNA College of Engineering and Technology, Dindigul, India
ndmugi@gmail.com

Abstract

Breast cancer is the most diagnosed and the leading cause of death in women. among women is breast cancer. Between 1 in 8 and 1 in 12 women in the developed world will experience breast cancer throughout their lives. The risk of breast cancer is of two primary types. The first type is that a person is likely to develop breast cancer over a specified time period. The second type reflects the likelihood of a high-risk gene mutation. Earlier work has shown that it has been better to predict the risk of breast cancer by adding input into the wide-spread Gail model. 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 a kind of artificial intelligence (AI) that ability to learn without being explicitly programmed. ML 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 ML 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 [1-2].

2 MOTIVATION

Breast cancers is most cancers that types in the mammary gland. At current there are many ladies struggling from this lethal sickness and the loss of life charge of humans struggling from this sickness are growing day with the aid of day[3-4]. In order to control, the solely answer is detecting it until now and present process terrific analysis and remedy based totally on the stage of most cancers which might also help to gradually decline the dying charge of patients. Implementation of desktop getting to know algorithm strategies such as Supervised ML algorithms which consists of Logistic regression, Decision Tree and Support vector computing device can helps to construct a classification mannequin to predict the breast most cancers from its two features. Some of the signs and symptoms and signs of breast most cancers that are used to classify the levels of most cancers are:

- A breast lump or thickening that feels one of a kind from the surrounding tissue.
- Change in size, form or look of a breast.
- Change in pores and skin over the breast, such as dimpling.
- A newly inverted nipple.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

• Peeling, scaling, crusting or flaking of the pigmented place of pores and skin surrounding the nipple (areola) or breast skin.
• Redness or pitting of the pores and skin over your breast, like the pores and skin of an orange.

3 Related Work

Comparison of ML Methods for Breast Cancer mentioned that two famous computer getting to know methods for Wisconsin Breast Cancer classification [3]. Artificial Neural Network and Support Vector Machine are used as ML methods for the classification of WBC (Original) dataset in WEKA tool. The effectiveness of utilized ML strategies is in contrast in time period of key overall performance metrics such as accuracy, precision, recall and ROC area. Based on the overall performance metrics of the utilized ML techniques, SVM (Sequential Minimal Optimization Algorithm) has confirmed the quality overall performance in the accuracy of 96.9957 percent for the prognosis and prediction from WBC dataset.

Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features explored a breast CAD technique supported function fusion with CNN deep features[5-7]. First, the mass detection approach is recommended, supported by profound aspects of CNN and unmonitored extreme machine learning (UEML). Secondly, we create a function set that combines profound characteristics, morphical characteristics, texture and density. Thirdly, the use of a fused function to classify benign and malignant breast masses is developed for an ELM classifier. Early lump detection can effectively reduce the carcinoma burden of death. This problem can be addressed through the computer-aided analysis for carcinoma (CAD). While the historical analytical method is widely used, it still has to be improved in its accuracy. The known crafted function configuration affects diagnostic precision immediately and thus an experienced doctor takes a position that is terribly necessary in the guiding characteristic extraction process. A mainly CNN deep points and US-ELM clustering approach is developed within the mass detection stage. At the time of mass diagnosis, the use of afused functions, fusing deeper properties with morphologic characteristical features, textures and density characteristics is classified using an ELM classifier in the case of benign and malignant breasts. The desire for aspects is the key to the diagnostic accuracy in the breast CAD approach [8-10].

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.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

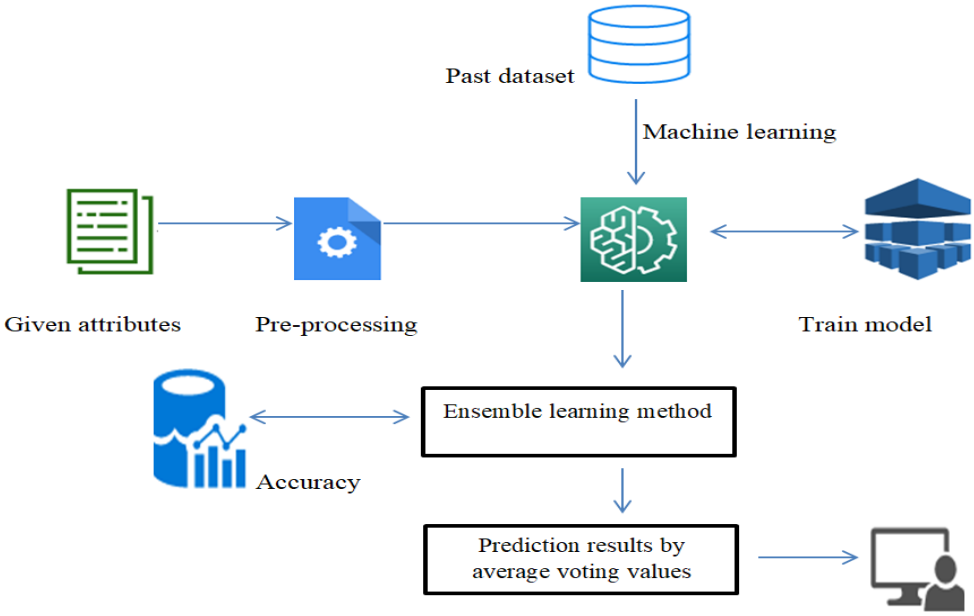

Fig.4.1 Architecture Diagram

Then these are preprocessed in order to keep away from duplication and error values, which then undergoes implementation of more than a few supervised computing device getting to know algorithms. The end result of every algorithm is once more processed by using ensemble approach and it offers the fantastic accuracy end result amongst the effects of algorithms with the aid of balloting classifier approach and the remaining effects 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 easy Ensemble getting to know strategies in which predictions from more than one fashions are combined. The approach begins with developing two or extra separate fashions with the equal dataset. Then a Voting based totally Ensemble mannequin can be used to wrap the preceding fashions and mixture the predictions of these models. After the Voting primarily based Ensemble mannequin is constructed, it can be used to make a prediction on new data. The predictions made via the sub-models can be assign weights. Stacked aggregation is a method which can be used to research how to weight these predictions in the quality viable way.

Comparing Algorithm with prediction in the shape of excellent accuracy. It is essential to evaluate the overall performance of more than one one-of-a-kind computing device gaining knowledge of algorithms persistently and it will find out to create a check harness to evaluate a couple of one-of-a-kind laptop getting to know algorithms in Python with sk-learn. It can use this take a look at harness as a template on your personal computer mastering issues and add extra and special algorithms to compare. Each mannequin will have distinctive overall performance characteristics. Using resampling techniques like move validation, you can get an estimate for how correct every mannequin can also be on unseen data. It desires to be capable to use these estimates to pick out one or two satisfactory fashions from the suite of fashions that you have created.

The equal thought applies to mannequin selection. A way to do this is to use exceptional visualization strategies to exhibit the common accuracy, variance and different houses of the distribution of mannequin accuracies.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

4.2 FINDING DIFFERENT STAGES OF CANCER

The stage of a cancer is a dimension of the extent of the most cancers and its spread. Opted therapy can be supplied primarily based on the stage. The trendy staging gadget for breast most cancers makes use of a gadget recognized as TNM, where: T stands for the primary (primary) tumor ,N stands for unfold to close by lymph nodes which is placed underneath palms and ,M stands for metastasis (spread to far-off components of the body). stage zero is the earliest stage in breast cancers which is ranged from stage I (1) via IV (4). As a rule, the decrease the wide variety suggests that the unfold of most cancers low and suggests decrease stage and greater wide variety like stage IV indicates cancer has unfold extra and shows vital stage. The grade of a tumor suggests what the cells appear like and offers an notion of how rapidly the most cancers may additionally develop 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.

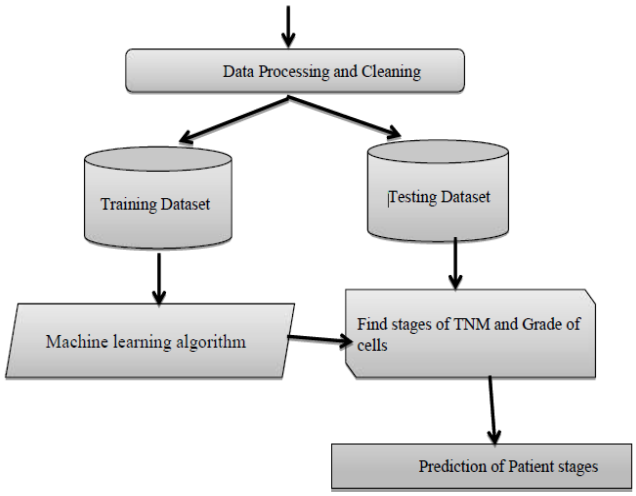

Fig.4.2 Identification of stages of breast cancer

5 PERFORMANCE ANALYSES

5.1 PREPARING THE DATASET

The dataset is now provided to laptop studying model on the foundation of this information set the model is trained. Every new patient's small print stuffed at the time of appointments structure acts as a take a look at records set. After the operation of testing, mannequin predicts whether or not the new affected person is a in shape case for affecting breast most cancers or no longer primarily based upon the inference it concludes on the foundation of the coaching 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

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

5.2 Results and Discussion

The evaluation of dataset is completed with the aid of supervised desktop studying algorithm to seize data such as, variable identification, univariate analysis, bi-variate evaluation etc. Additionally, we have to talk about the overall performance from the given sanatorium dataset with assessment classification file and become aware of the confusion matrix. The statistics validation, getting ready and visualization will be utilized on the whole given dataset. Thus the end result suggests the effectiveness of the proposed computing device mastering algorithm method which gives us with great accuracy, precision, Recall and F1 Score.Table 5.1 shows the overall performance of a range of algorithms measured with the aid of various parameters like precision,recall,F1-Score ,Sensitivity, Specificity and accuracy.

Table 5.1 shows the performance of various algorithms

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

Sensitivity is a calculation probability 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.

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

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

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 manner of evaluation commenced from records cleansing and processing, lacking value, exploratory analysis and sooner or later mannequin constructing and evaluation. Finding the affected person levels and grade with parameter like accuracy, classification record and confusion matrix on public check set of given attributes by using ensemble gaining knowledge of approach of vote casting classifier accuracy is 100%.Hospital wishes to automate the detection of the breast most cancers from eligibility system (real time) primarily based on the account detail. To automate this system we exhibit the prediction end result in net software or computing device application. To optimize the work to be carried out in AI surroundings

References

[1]. J.Jayanthi, T.Jayasankar, N.Krishnaraj, N.B.Prakash, A.Sagai Francis Britto, K.Vinoth Kumar, "An Intelligent Particle Swarm Optimization with Convolutional Neural Network for Diabetic Retinopathy Classification Model," Journal of Medical Imaging and Health Informatics (2020), Volume 11, Number 3, March 2021.

[2]. Priyanka Parvathy, D Kamalraj Subramaniam, G.K.D PrasannaVenkatesan, P. Karthikaikumar , Justin Varghese, T.Jayasankar, "Development of Hand Gesture Recognition System Using Machine Learning", Journal of Ambient Intelligence and Humanized Computing (2020),

[3]. 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.

[4]. S.JacilyJemila and T. Jayasankar, *"*An Automated Cancer Recognition System for MRI Images Using Neuro Fuzzy Logic*",* International Journal of Computer Information Systems, Vol. 2, No. 5, 2011, pp.18-22

[5]. M.Anuradha, T.Jayasankar, PrakashN.B[3], Mohamed Yacin Sikkandar, G.R.Hemalakshmi, C.Bharatiraja,A. Sagai Francis Britto, "IoT enabled Cancer Prediction System to Enhance the Authentication and Security using Cloud Computing," Microprocessor and Microsystems (Elsevier 2021), vol 80, February, (2021) https://doi.org/10.1016/j.micpro.2020.103301

[6]. 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.

[7]. 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.

[8]. 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

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, *38*(3), 686-696.

[9]. J.Jayanthi, T.Jayasankar, N.Krishnaraj, N.B.Prakash, A.Sagai Francis Britto, K.Vinoth Kumar, "An Intelligent Particle Swarm Optimization with Convolutional Neural Network for Diabetic Retinopathy Classification Model*," Journal of Medical Imaging and Health Informatics* (2020), Volume 11, Number 3, March 2021, pp. 803-809, https://doi.org/10.1166/jmihi.2021.3362

[10]. M. Buvana, K. Muthumayil, T. Jayasankar, "Content-Based Image Retrieval based on Hybrid Feature Extraction and Feature Selection Technique Pigeon Inspired based Optimization", Annals of the Romanian Society for Cell Biology (2021), Vol. 25, Issue 1, 2021, pp. 424 – 443,

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731