Learning Techniques for Prediction of Breast Cancer Disease: A Comparative Analysis

Chapter · November 2022

DOI: 10.1007/978-981-19-3148-2_42

5 authors, including:

Abhaya Kumar Sahoo
KIIT University
48 PUBLICATIONS **602** CITATIONS
SEE PROFILE

Amrendra Singh Yadav
Atal Bihari Vajpayee-Indian Institute of Information Technology and Management …
23 PUBLICATIONS **159** CITATIONS
SEE PROFILE

J. R. Mohanty
KIIT University
17 PUBLICATIONS **179** CITATIONS
SEE PROFILE

Learning Techniques for Prediction of Breast Cancer Disease: A Comparative Analysis

Chandramouli Das, Abhaya Kumar Sahoo, Amrendra Singh Yadav, Jnyana Ranjan Mohanty, and Rabindra Kumar Barik

Abstract Breast cancer and its detection is always one of the trending topics of all time. Many research activities have been done in this field using different technologies to make it effective. In this research, it has taken machine learning and deep learning algorithms to develop an effective breast cancer classifier. It has implemented ten best machines learning and deep learning classification algorithms like logistic regression, decision tree, random forest, support vector machine, K-nearest neighbor, Naïve Bayes, multi-layered perceptron, stochastic gradient descent, AdaBoost, and artificial neural network classifier for prediction of breast cancer disease. These classifiers are successfully able to classify whether the given data are coming under benign cell or malignant cell. It has also got different accuracy for every classifier. At the end, it has successfully implemented all the approaches and has got the good accuracy.

Keywords Breast cancer detection · Machine learning · Deep learning · Prediction

1 Introduction

Breast cancer is a very frequent malignancy nowadays. Breast cancer affects one in every eight women born today. Statistically, nearly 266,000 US women will be diagnosed with invasive breast cancer in 2018. Breast cancer strikes a woman every two minutes [1–3]. Breast cancer develops from breast cells. Statistically, one woman dies of breast cancer every thirteen minutes. On the plus side, breast cancer is treatable if diagnosed early. Survivorship of breast cancer has tripled in 60 years. Breast cancer

C. Das · A. K. Sahoo (✉)
School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
e-mail: abhayakumarsahoo2012@gmail.com

A. S. Yadav
School of Computing Science and Engineering, VIT Bhopal University, Sehore, Madhya Pradesh, India

J. R. Mohanty · R. K. Barik
School of Computer Applications, KIIT Deemed to be University, Bhubaneswar, India
e-mail: jmohantyfca@kiit.ac.in

A. Khanna et al. (eds.), *Proceedings of Third Doctoral Symposium on Computational Intelligence*, Lecture Notes in Networks and Systems 479,
https://doi.org/10.1007/978-981-19-3148-2_42

death rates have fallen by nearly 40% in 16 years [4–9]. 1.3 million breast cancer survivors live in the USA alone. Early detection is the only way to enhance outcomes. For this reason, researchers from all around the world are developing unique models that can detect breast cancer [4–6, 10].

Nowadays, machine learning and deep learning-based models are used to predict breast cancer disease, by which patients can be recovered from early stage of breast cancer disease [11]. In this paper, we develop different models based on deep learning and machine learning approaches to identify breast cancer where AdaBoost classifier-based model provides 99.4% accuracy among the models.

In this paper, Sect. 2 explains about related work. Section 3 depicts the model. Sections 4 and 5 describe the model descriptions with experiment analysis. Section 6 presents the result and discussion. Finally, Sect. 7 wraps off with concluding remarks.

2 Related Work

Breast cancer is the most common cancer in women, especially in cities. The newest figures reveal a 33% rise in cancer, predominantly breast [12–17]. The public must be aware of it and know how to safely identify and treat it. Breast cancer, like all cancers, has four stages. The first two stages are known as early breast cancer, whereas the third stage is known as locally advanced breast cancer. Stage 4 cancer has spread throughout the body and is incurable. Early-stage cancer has a survival rate of 80–93%. Locally advanced breast cancer is treatable in 55–70% of instances. In advanced breast cancer, chemotherapy can merely prolong life. Before reviewing breast cancer symptoms now, simple breast cancer symptoms must be known. Simple breast cancer symptoms must be known. Breast cancer typically affects women over 30. For any other substantial size variation in the breasts, a lady should contact her doctor, preferable an oncologist. Breast cancer has two types of causes: non-modifiable and modifiable risk factors (changeable risk factors). These are non-modifiable risk factors such as obesity, sedentism, hormone therapy, alcohol, and tobacco usage, all can be modified. Obesity, sedentism, hormone therapy, alcohol, and tobacco usage can all be modified. Table 1 highlights the research on learning strategies for breast cancer prediction.

From the foregoing argument, we can deduce that early identification of breast cancer is critical. Breast cancer can be readily cured if caught early. This is the main reason we started this investigation. The breast cancer dataset we used has many measurement columns and a goal column that indicates whether the malignancy is benign or aggressive. If the tumor is benign, there is no risk of cancer, but if it is malignant, there is a significant risk of cancer. We create a model that can categorize tumors whether it is benign or malignant.

Table 1 Different literature reviews on learning techniques for prediction of breast cancer

Author Name	Research Work	Work done	Method Name	Challenge
G Viale	Breast cancer classifier (2020) [18]	Histochemical characterization of breast cancer for the assessment of hormone receptor status	Breast cancer classifier	Scope of improvement
B Weigelt and HM Horlings et al	Clarification of breast cancer classification by molecular characterization (2018) [19]	To purify breast cancer classification systems by analyzing a series of 11 histological special using immune histochemistry and genome-wide gene expression profiling	Molecular classification	Many scopes of improvement
Gouda I. Sala et al	Breast cancer diagnosis on three different datasets using multi-classifiers (2012) [20]	Use of different classification and compare the result	MLP, KNN, SVM	More classification techniques to explore
Y.Ireaneus Anna Rejani et al	Early detection of breast cancer using SVM (2009) [21]	Detect tumor using SVM	SVM	More classification techniques to explore
D. Lavanya and Dr. K. Usha	Ensemble decision tree classifier for breast cancer data (2009) [22]	CART classifier with feature selection and bagging technique, which used to evaluate the performance in terms of accuracy using different breast cancer datasets	Decision tree classifier	More classification techniques to explore

(continued)

Table 1 (continued)

Author Name	Research Work	Work done	Method Name	Challenge
Murat Karabatak	A new classifier for breast cancer detection based on Naïve Bayesian (2015) [23]	Use of customized Naïve Bayes method	Naïve Bayes classifier	More classification techniques to explore
Wei Keat Lim et al	Master regulators used as breast cancer metastasis classifier (2009) [24]	Applying computational systems biology approach, it can infer robust prognostic markers by identifying upstream master regulators	Breast cancer metastasis classifier	Needed to enhance
Cuong Nguyen et al	Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic (2013) [25]	Use of random forest classification method	Random forest	More classification techniques to explore
Essam Amin et al	A breast cancer classifier based on a combination of case-based reasoning and ontology approach (2010) [26]	Detect benign or malignant cells case-based reasoning and ontology approach	Case-based reasoning and ontology approach	It is in its improvement state
Mohammad M. Ghiasi and SohrabZendehboudi	Use of decision tree-based ensemble learning (2021) [27, 28]	Authors contribute their work based on random forest and extremely randomized trees or extra trees algorithms to classify breast cancer	Random forest, extreme randomization of tree	Only works on a specific WBCD dataset
MoloudAbdar and VladimirMakarenkov	CWV-BANN-SVM ensemble learning classifier for better diagnosis of breast cancer (2019) [10, 29]	Authors use custom method CWV, which is a hybrid version of SVM and ANN	CWV-BANNSVM, SVM, ANN	Optimization techniques can be applied for better accuracy

(continued)

Table 1 (continued)

Author Name	Research Work	Work done	Method Name	Challenge
Yiqiu Shena et al	An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization (2021) [30]	This paper ensembles, ResNet 34 and R-CNN	ResNet 34, R-CNN	Very complex to use this approach, difficult to train this model
Moloud Abdara, et al	Application of modified nested ensemble technique for classification of breast cancer (2020) [31]	Voting and stacking techniques are used in this work	Voting, stacking, ensemble method	Need to use more ensemble methods for different types of cancer disease

3 Proposed Methodology

Now we will talk about our model's methodology [7, 32]. Ten machine learning and deep learning classifiers were used. Now, we will discuss each classifier in detail with examples [8]. The most prevalent machine learning categorization model is logistic regression. This classifier is popular for its ease of use in nature.

This method can tackle regression and classification problems (Fig. 1).

The decision tree [16] method is one of the most effective methods for both regression and classification. It is a decision-making tree. The random forest classifier is also one of the most useful classifiers. Random forest approach can be used for both regression and classification problems [9]. To understand the concept of random forest, we need to learn about the characteristics of ensemble learning. In most simple word ensemble learning is, we take multiple machine learning algorithms and put them together to make a modified version of machine learning algorithm. Support vector classifier is a part of support vector machine. The characteristics of support vector machine is a little different from rest of machine learning algorithms. Support vector machine also can be used for both regression and classification problems [5, 33].

It is simple and straightforward to implement KNN algorithm. It is classified as supervised [27] machine learning and is mostly used for categorization. The Naive Bayes classifier is used here. This classifier uses the Bayes theorem as its basis. Stochastic gradient descent is a powerful but simple technique. Like logistic regression, it fits linear regressor or classifier into a convex loss function. Large datasets require SGD classifiers. We talked about ensemble learning in the random forest classifier. Bagging and boosting are two types of ensemble learning. Adaptive boosting classifier (abc) is a boosting method. Considering boosting algorithms instead of a

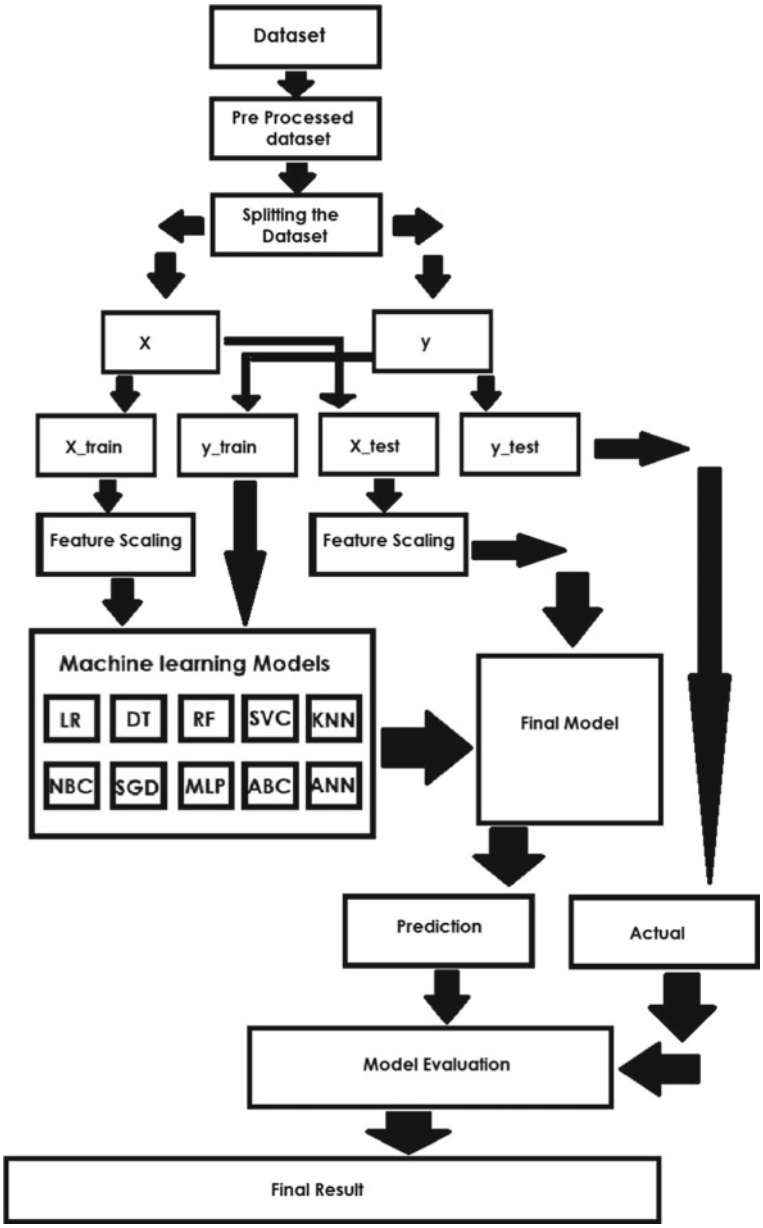

Fig. 1 Working flow

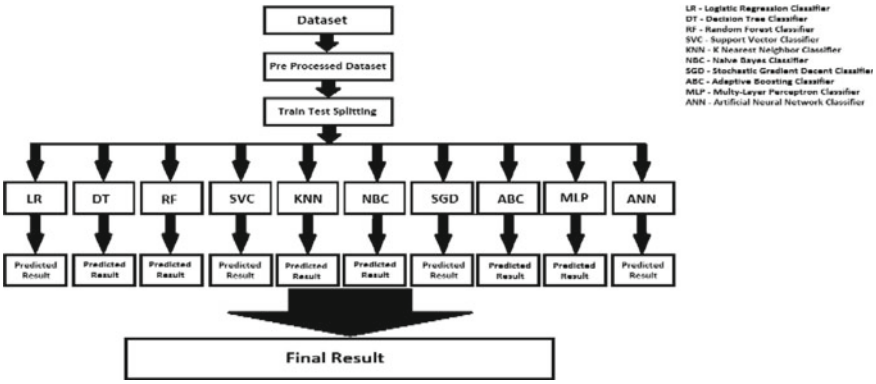

Fig. 2 Model specifications

strong classifier, each tree fed as an improved dataset. It is a type of artificial neural network. This is a forward-feed type. Sooner or later, we will talk about ANN. This model uses perceptron layers. It is supervised machine learning. A dataset is used to train an activation function. A deep learning method called artificial neural network [5] was employed here instead of machine learning algorithms previously used. It is an extremely powerful supervised deep learning algorithm.

4 Model Specifications

Till now, we have discussed all the ten classifiers in detail, that we have used. Now let us understand about this breast cancer classifier model.

Figure 2 presents the model descriptions. We start with the raw data. Then, we pre-processed the model to understand the column-feature relationships. Our dataset is now usable. Then, we split into the train and test model. Then, we use one classifier at a time. First, we trained the models and then tested them on test data. This provides us classifier accuracy. Then, we updated all ten. It sorts the new data. It produces output. We aggregate ten classifier outputs to get our final output. Every classifier we tested is 94% accurate or better. Finally, we have high confidence in its output forecast.

5 Experimental Data Analysis

These are the dataset, accuracy calculation, and confusion matrix. We employed eleven different classification [33–35] approaches. Each categorization strategy has a F1-score. So now let us go over each one. We used Kaggle data. These are UCI

Machine Learning Repository data. It has 569 columns and 32 rows. This dataset has no nulls. Another two unneeded columns complete the list. These are later removed. Each column is associated. This dataset has no null values; thus, no pre-processing is required. We normalized the entire dataset before splitting it into training and testing. We split the data 70/30 between training and testing. This column has no data. So we went binary to use our dataset in our model. The confusion matrix is used to classify data. It is a table that shows an algorithm's performance. Rows reflect anticipated values, while columns indicate actual values. Recall, precision, F1-score, and accuracy are calculated using Eqs. (1), (2), (3), and (4), respectively.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{1}$$

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False Positive}} \tag{2}$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{4}$$

In this paper, we have implemented confusion matrix and also calculated above metric parameters for every classifier. We have used ten different classifiers and achieved ten different accuracy, recall, precision, and F1-score. Then, we have compared all the values in a single table. Table 2 is depicted in the result section. We also have plotted these graphically.

Table 2 Results of ten classifiers along with their accuracy

Method	Accuracy	Recall Value	Precision Value	F1-Score
Logistic regression classifier	0.98831	0.99048	0.99048	0.99048
Decision tree classifier	0.95321	0.94286	0.9802	0.96117
Random forest classifier	0.96491	0.98095	0.96262	0.9717
Support vector classifier	0.97076	0.98095	0.9717	0.9763
K-nearest neighbor classifier	0.98245	1	0.97222	0.98592
Naïve Bayes classifier	0.94736	0.97143	0.94444	0.95775
Stochastic gradient decent classifier	0.97661	0.98095	0.98095	0.98095
AdaBoost classifier	0.99415	0.99048	1	0.99522
Multi-layer neuron classifier	0.97661	0.98095	0.98095	0.98095
Artificial neural network classifier	0.98245	1	0.97222	0.98592

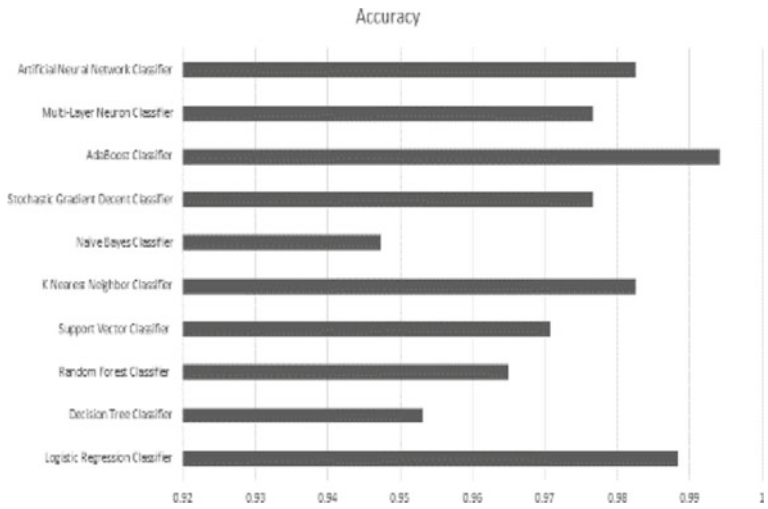

Fig. 3 Accuracy graph

6 Result and Discussion

Different classifiers provide the accuarcy which is depicted in Fig. 3. As we can see, the AdaBoost classification method works the best among all the classifiers. To understand the performance of the classifiers better, we have plotted another accuracy graph. It displays the model results in Fig. 4.

So far, we have shown our model's performance. We can anticipate if a cell is benign or cancerous based on data. So we tried to add a new datapoint to our model, and it successfully categorized it. Table 2 shows the results and accuracy of ten classifiers. The classifiers' accuracy varies. All classifiers can correctly predict 94% of the time. AdaBoost or adaptive boosting classifiers are the most accurate at 99.4%. Then it follows the logistic regression classifier (98.8%). The remaining classifiers then performed accurately. The Naive Bayes classifier performs poorly here. The other nine function better. Our model is 97% accurate. To summarize, the chance of adding a benign cell to our model is 97%. So, here is how our model accurately diagnoses breast cancer.

7 Concluding Remarks

As we have shown, our research model can successfully classify between malignant and benign cell. But there is some scope of improvement too. Like we have deal with text data. In future, we will work with image data. Here, we have used only one deep learning algorithm, so in future, we will use more deep learning model so that we can

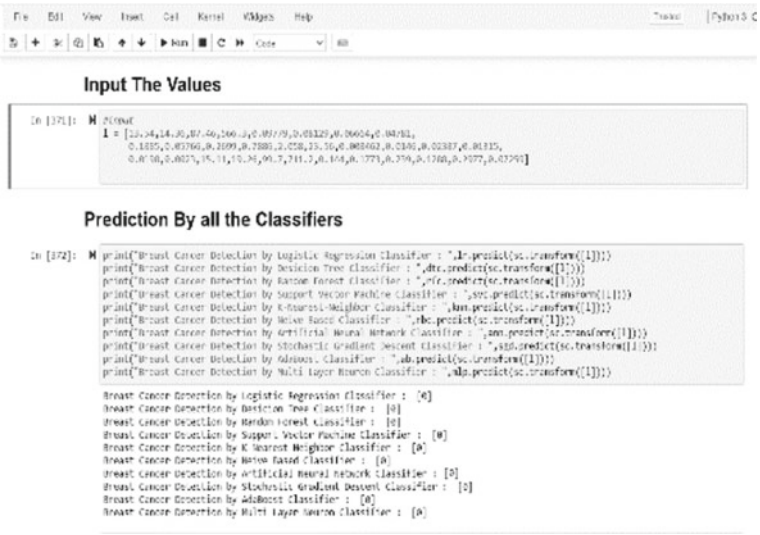

Fig. 4 Model result

get more improved results. These are some steps that can be applied in our model. The main challenge in breast cancer is the detection of the cancer at its early stage. Our model can fulfill this challenge to some extent. We have started this research activity with the aim of detecting breast cancer. Now, we can say that our aim is successfully completed. In future, we will take forward our aim by implementing new machine learning and deep learning algorithms.

References

1. U.S. Cancer Statistics Working Group (2012) 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
2. Lyon IAfRoC: World Cancer Report (2003) International agency for research on cancer press. pp 188–193
3. Elattar I (2005) Breast cancer: magnitude of the problem. Egyptian Society of Surgical Oncology Conference, Taba, Sinai, Egypt, 30 Mar–1 Apr 2005
4. American Cancer Society (2013) Breast cancer: facts and figures. ACS, Atlanta
5. Pena Reyes CA, Sipper M (1999) A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 17(2):131–155
6. Goodman D, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Proceedings of the artificial neural networks in engineering, pp 179–183
7. Kopans D (1998) Breast imaging. Lippincott-Raven, Philadelphia
8. Bator M, Nieniewski M (2012) Detection of cancerous masses in mammograms by template matching: optimization of template brightness distribution by means of evolutionary algorithm. J Digit Imaging 25(1):162–172

9. Abonyi J, Szeifert F (2003) Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recogn Lett 24(14):2195–2207
10. Chen HL, Yang B, Wang G, Wang SJ, Liu J, Liu DY (2012) Support vector machine based diagnostic system for breast cancer using swarm intelligence. J Med Syst 36(4):2505–2519
11. Sahoo AK, Pradhan C, Das H (2020) Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. Nature inspired computing for data science. Springer, Cham, pp 201–212
12. Fear EC, Meaney PM, Stuchly MA (2003) Microwaves for breast cancer detection. IEEE Potentials 22:12–18, February-March 2003
13. Homer MJ (1997) Mammographic interpretation: a practical approach. 2nd edn. McGraw hill, Boston, MA
14. American College of Radiology, Reston VA (1998) Illustrated breast imaging reporting and data system (BI-RADSTM), 3rd edn.
15. Astley SM (2004) Computer-based detection and prompting of mammographic abnormalities. Br J Radiol 77:S194–S200
16. Burhenne LJW (2000) potential contribution of computer aided detection to the sensitivity of screening mammography. Radiology 215:554–562
17. Aruna S, Rajagopalan SP, Nandakishore LV (2011) An empirical comparison of supervised learning algorithms in disease detection. Int J Inf Technol Convergence Serv (IJITCS) 1(4)
18. Viale G (2020) The current state of breast cancer classification. Elsevier. Science Direct, Jan 7
19. Weigelt B, Horlings HM, Kreike B, Hayes MM, Hauptmann M, Wessels LF, De Jong D, Van de Vijver MJ, Veer LV, Peterse JL (2018) Refinement of breast cancer classification by molecular characterization of histological special types. Wiley Online Library, 14 Jul 2018
20. Salama GI, Abdelhalim M, Zeid MA (2012) Breast cancer diagnosis on three different dataset using multi-classifiers. Int J Comput Inf Technol (2277–0764) 01(01)
21. Rejani Y, Selvi ST (2009) Early detection of breast cancer using SVM classifier technique. Int J Comput Inf Technol 1(3):127–130
22. Lavanya D, Rani KU (2009) Ensemble decision tree classifier for breast cancer data. Int J Comput Inf Technol Convergence Serv (IJITCS) 2(1)
23. Karabatak M (2015) A new classifier for breast cancer detection based on Naïve Bayesian. Elsevier, 6 May 2015
24. Lim WK, Lyashenko E, Califano A (2009) Master regulators used as breast cancer metastasis classifier
25. Nguyen C, Wang Y, Nguyen HN (2013) Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic
26. Abdrabou EAML, Salem ABM (2010) A breast cancer classifier based on a combination of case-based reasoning and ontology approach. In: Proceedings of the international multiconference on computer science and information technology. IEEE, Oct 2010, pp 3–10
27. Vlahou A, Schorge JO, Gregory BW, Coleman RL (2003) Diagnosis of ovarian cancer using decision tree classification of mass spectral data. J Biomed Biotechnol 2003(5):308–314
28. Ghiasi MM, Zendehboudi S (2021) Application of decision tree-based ensemble learning in the classification of breast cancer. Comput Biol Med 128:104089
29. Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, Gururajan R (2020) A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recogn Lett 132:123–131
30. Shen Y, Wu N, Phang J, Park J, Liu K, Tyagi S, Geras KJ (2021) An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal 68:101908
31. Abdar M, Makarenkov V (2019) CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement 146:557–570
32. Aruna S, Rajagopalan SP, Nandakishore LV (2011) Knowledge based analysis of various statistical tools in detecting breast cancer. Comput Sci Inf Technol 2(2011):37–45
33. Vidyarthi A (2020) Multi-scale dyadic filter modulation based enhancement and classification of medical images. Multimedia Tools Appl 79(37):28105–28129

34. Vidyarthi A, Mittal N (2014) Comparative study for brain tumor classification on MR/CT images. In: Proceedings of the third international conference on soft computing for problem solving. Springer, New Delhi, pp 889–897
35. Vidyarthi A, Mittal N (2015). Brain tumor segmentation approaches: review, analysis and anticipated solutions in machine learning. In: 2015 39th national systems conference (NSC). IEEE, Dec 2015, pp 1–6
36. Vidyarthi A, Nagpal J (2021) Malignancy grade identification and classification of brain MR images with new 2D co-occurrence matrix and wavelet transformation. In: 2021 Thirteenth international conference on contemporary computing (IC3–2021), Aug 2021, pp 43–50