Proceedings of the Sixth International Conference on Trends in Electronics and Informatics (ICOEI 2022)
IEEE Xplore Part Number: CFP22J32-ART; ISBN: 978-1-6654-8328-5

Prediction of Breast Cancer based on Various Medical Symptoms Using Machine Learning Algorithms

Mayank Agrawal
Assistant Professor, Department of Computer Engineering and Applications,
GLA University, Mathura, India
mayank.agrawal@gla.ac.in

Vinod Jain,
Assistant Professor, Department of Computer Engineering and Applications,
GLA University, Mathura, India
vinod.jain@gla.ac.in
(ORCHID-0000-0003-0260-7319)

2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-8328-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOEI53556.2022.9777114

***Abstract*— Nowadays, breast cancer is a relatively common type of cancer. It's been seen in a lot of women lately, and it's been the cause of a lot of deaths. There is a need to predict the probability of cancer at early stages so that necessary actions can be taken timely to avoid this deadly disease. Machine Learning a a new Artificial Intelligence technique and its power is yet to be discovered for cancer prediction. The power of machine learning classifiers is used to predict breast cancer in this study. The prediction accuracy of four machine learning classifiers, Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier and Ada Boost Classifier is calculated on a standard data set. According to the results of the experiment, Support Vector Classifier trumps the other three in terms of prediction accuracy.**

Keywords*— *Artificial Intelligence, Machine Learning, Breast Cancer Prediction

I. INTRODUCTION

Breast Cancer (BC) is a serious type of problem in women now a days. Breast cancer recurrence can occur somewhere between 1 and 20 years after treatment of the original tumour. Breast cancer can be detected using MRI, mammography, biopsy, ultrasound, and physical examination. Breast cancer is divided into two types, benign and malignant, depending on the tumour. Next, it can be divided into two categories: regular and non-regular. The number of lymph nodes affected by the primary tumour determines the likelihood that the cancer will come back. Breast cancer can recur in the same place where it first spread to other parts of the body. The size of the original tumour, the number of lymph nodes involved, the area of the tumour, and other equivalent characteristics can all be used to predict the recurrence of breast cancer. The main advantage of using machine learning models to predict breast cancer recurrence is improved accuracy and reduced error. By setting a time frame for the last occurrence and knowing the size, shape, texture, and other characteristics of past tumours, it is possible to predict whether or not a recurrence will occur. Machine learning helps save time compared to traditional approaches such as biopsy, as the system can predict results in seconds.

A. Mangal and V. Jain [1] calculate breast cancer authors have use AI methods. For breast cancer analysing, ML methods like Decision Tree, Naïve Bayes, KNN and SVM can also be used. Training database has been used to train all these Machin learning methods then test database has been used to estimate their accuracy. Authors have implemented proposed model in the Python programming language. In this paper [1], authors have examined that SVM method give the better results as compare to other methods to this dataset. Future work of this paper [1] was that authors was examined SVM model on different database. The next section discussed the related work in this area.

II. LITERATURE SURVEY

V. A. Telsang and K. Hegde [2] given a novel method to examine the breast cancer and they have used various method of ML. In this paper [2], authors have used various evaluation matrix such as AUC and accuracy to match their examination. Here they have used database for examine the BC called Wisconsin Database. Here in the examination, authors have got that SVM method has more accuracy 96.25% with AUC of 99.4. Future work of this paper was that authors can be changed with their mathematical models to upsurge the calculation of breast cancer. L. Nithya et al. [3] used Find-S method and Candidate elimination method for cataloguing Breast Cancer. For Breast Cancer discovery, authors have used these algorithms. For cataloguing of breast cancer, they have used Navie Bayes Classifier. Limitation of this paper [3] was that Find-s method and candidate elimination method were failed if data was noisy data. P. Mekha and N. Teeyasuksaet [4] used grouping methods for the evaluation for Breast Cancer which is based on tumour cell. Here for kinds of breast cancer, authors have emphasis on using deep learning algorithms with various activation function such as Exprectifier, Tanh, Maxout and Rectifier. Here they have evaluated with various ML method such as SVM, RF, DT, AdaBoost, Vote(DT+NB+SVM) and NB. This paper [4], evaluation have done have on database called Wisconsin database. K. Uyar et al. [5] used RFNN and ANFIS which are based on Genetic algorithm. Here authors have done

experiments on database which is given by UCI Machine Learning Repository.

M. S. Yarabarla et al. [6] advantage of recent developments in CAD systems and related methodologies. The project's major goal is to determine whether or not the person has breast cancer. Machine learning is the process of teaching computers to learn and perform on their own without the need for an explicit code or direction. As a result, using training data, it is possible to predict if a person has breast cancer or not. S. V. J. Jaikrishnan et al. [7] used 6 machine learning techniques to present a novel strategy for breast cancer detection that improves performance using machine learning. The methods' unbiased estimations are determined by using k-fold cross-validation technique. Conventional machine learning techniques are shown to be more efficient when using this developed model. Limitation of this paper was that the suggested model, when supplemented by feature selection approaches, can be used to different datasets to uncover hidden insights and improve performance.

LR, KNN, DT, SVM, NB, and RF are six machine learning classification algorithms used by M. R. Ahmed et al. [8] to assess the prediction capacity. The major objective of this work is to calculate the performance of ML models on Wisconsin Breast Cancer (new) database.

The Malwa reason has shown an increase in total death rates due to breast cancer, according to N. Hooda et al. [9]. The paper looks into the disease's mortality and its links to a variety of risk variables, including demographics, pesticide residue levels in the water and soil are measured in the blood, tumour, and surrounding tissue. The extent to which individuals have been exposed to contaminants such as heavy metals is also studied. Bagoost, an effective ensemble machine learning-based system for estimating the risk of cancer in women using experimental data, is given.

Naveen et al. [10] given new method and with the help of this method authors can calculate Breast Cancer with the help of cancer features. In this paper [10], authors have experiments on database called breast cancer Coimbra database. The authors employ the following steps: feature scaling, cross validation, and multiple ensemble machine learning models utilising bagging technique. As a consequence of their research, the authors may now give this model to the medical community, allowing doctors or diagnostic professionals to enter patients' characteristics into a diagnosis report and predict whether or not they would get breast cancer.

Many researchers also contributed in prediction of this disease [12]-[20]. But there is a further need to do more research in exploring the power of ML for cancer prediction so that this disease can be controlled in future. The next section discussed the proposed work in this area.

III. PROPOSED WORK

This work illustrates the capacity of ML models to predict the chances of breast cancer. Four ML algorithms such as 'SVM', 'Random Forest', 'Gradient Boosting' and 'Ada Boost' classifier are implemented in python and their results are investigated.

A. Classification Task

One of the classification problems considered by the suggested method was breast cancer prediction. The issue will determine whether the patient is Benin or Malignant. The remaining attributes are considered inputs, whereas the single attribute is considered output. The proposed methodology for breast cancer prediction utilising a machine learning system is shown in Figure 1.

The data set is first downloaded. After that, the data is pre-processed. The data set is then split into two parts: training and testing. Then the four machine learning algorithms are applied and their prediction accuracy is calculated and compared.

Here the training and testing data set is selected on 60-40 rule i.e. 60% data is taken for training and 40% data is taken for testing set.

B. Data Set and Attributes

The data set for breast cancer prediction in this study comes from the Wisconsin Hospitals Madison Breast Cancer Database which is available at Kaggle.com [11]. The data collection contains 32 patient features that are used to predict cancer. A total of 569 instances are stored in the data collection. As a result, each event has one of two possible outcomes: Benin or Malignant.

In this paper this data set is analyzed by machine learning algorithms. The machine learning algorithms are first trained on the training data set and then their prediction accuracy is tested on testing data set.

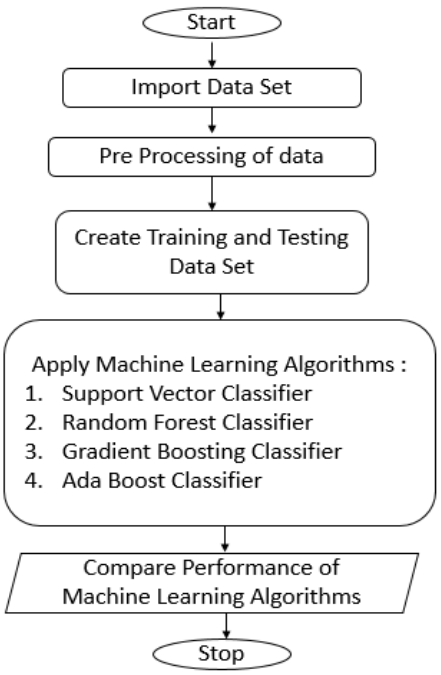

Fig. 1. The proposed methodology of breast cancer prediction.

IV. RESULTS AND ANALYSIS

For cancer prediction, classification machine learning methods are used. Four ML algorithms are compared. The prediction accuracy of the machine learning techniques Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Ada Boost Classifier is shown in Table 1.

Table 1 shows that the Support Vector Classifier outperforms the Random Forest Classifier, Gradient Boosting Classifier, and Ada Boost Classifier in terms of prediction accuracy.

The Support Vector Classifier has the highest accuracy of the four, at 97.36 percent. Figure 2 shows a bar graph comparing the accuracy of three machine learning classifiers in terms of prediction.

TABLE I : Comparison of Prediction Accuracy of Machine Learning Algorithms

Sr. No.	Machine Learning Model	Prediction Accuracy
1	Support Vector Classifier	0.9736
2	Random Forest Classifier	0.9561
3	Gradient Boosting Classifier	0.9649
4	Ada Boost Classifier	0.9605

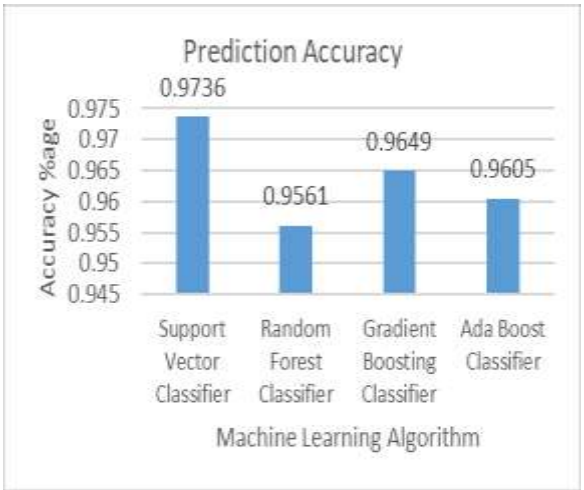

Fig. 2. Bar graph showing prediction accuracy

The goal behind the SVM is in finding a hyperplane that separated the two patterns and SVM tries to find to achieve the maximum separation with the closest point which is known as support vectors [21]. The objective is to maximize the minimum distance.

If the point lies on the positive group then we get a positive value for the equation.

$$w^T(\Phi(x)) + b > 0 \quad (1)$$

If the point lies on the negative group, then we get a negative value for the equation.00

$$w^T(\Phi(x)) + b < 0 \quad (2)$$

The support vector machine shows a better accuracy among all the four machine learning algorithm. The mathematical model discussed above is the key behind the high accuracy of the SVM classifier [21].

V. CONCLUSION AND FUTURE SCOPE

Different ML algorithms for breast cancer prediction are discussed in this research. On the well-known WBCD breast cancer database, three machine learning classifiers are used. It has been discovered that the 'Support Vector Classifier' outperforms the other three classifiers, 'Random Forest Classifier', 'Gradient Boosting Classifier', and 'Ada Boost Classifier', with a score of 97.36 percent. As a result, Support Vector Classifier was discovered to be the most effective method for predicting breast cancer.

Other ML techniques could be used to predict breast cancer in the future. Other soft computing approaches, such as neural networks, could be used in the future to predict breast cancer.

Further a research can be done in future to analyze SVM is achieving best result for cancer prediction and its accuracy can also be checked to predict other types of cancers.

REFERENCES

[1] A. Mangal and V. Jain, "Prediction of Breast Cancer using Machine Learning Algorithms," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 464-466. doi: 10.1109/I-SMAC52330.2021.9640813

[2] V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2020, pp. 1-5. doi: 10.1109/C2I451079.2020.9368911

[3] L. Nithya, S. Dixit and B. I. Khodhanpur, "Prediction of breast cancer using Find-S and Candidate elimination algorithm," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-4. doi: 10.1109/CSITSS47250.2019.9031046

[4] P. Mekha and N. Teeyasuksaet, "Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells," 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), Nan, Thailand, 2019, pp. 343-346. doi: 10.1109/ECTI-NCON.2019.8692297

[5] K. Uyar, U. Ilhan, A. Ilhan and E. I. Iseri, "Breast Cancer Prediction Using Neuro-Fuzzy Systems," 2020 7th International Conference on Electrical and Electronics Engineering (ICEEE), Antalya, Turkey, 2020, pp. 328-332. doi: 10.1109/ICEEE49618.2020.9102476

[6] M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 121-124. doi: 10.1109/ICOEI.2019.8862533

[7] S. V. J. Jaikrishnan, O. Chantarakasemchit and P. Meesad, "A Breakup Machine Learning Approach for Breast Cancer Prediction," 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), Pattaya, Thailand, 2019, pp. 1-6. doi: 10.1109/ICITEED.2019.8929977

[8] R. K. Bhogal, P. D. Suchit and C. Naresh, "Review: Breast Cancer Detection Using Deep Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 847-854. doi: 10.1109/ICOEI51242.2021.9452835

[9] M. R. Ahmed, M. A. Ali, J. Roy, S. Ahmed and N. Ahmed, "Breast Cancer Risk Prediction based on Six Machine Learning Algorithms," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, pp. 1-5. doi: 10.1109/CSDE50874.2020.9411572

[10] N. Hooda, R. Gupta and N. R. Gupta, "Prediction of Malignant Breast Cancer Cases using Ensemble Machine Learning: A Case Study of Pesticides Prone Area," in IEEE/ACM Transactions on Computational Biology and Bioinformatics. doi: 10.1109/TCBB.2020.3033214

[11] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[12] Uttam, A.K., Mangal, A., "Application of extreme gradient boosting ensemble model for sleep quality prediction on personalized wearable device data", International Journal of Advanced Science and Technology, 2020, 29(5), pp. 3755–3762

[13] Mangal, A., Uttam, A.K., "Sleep prediction by various supervised machine learning model", International Journal of Advanced Science and Technology, 2020, 29(5), pp. 3786–3792

[14] Mishra, Satyasis, T. Gopi Krishna, Harish Kalla, V. Ellappan, Dereje Tekilu Aseffa, and Tadesse Hailu Ayane. "Breast Cancer Detection and Classification Using Improved FLICM Segmentation and Modified SCA Based LLWNN Model." In Computational Vision and Bio-Inspired Computing, pp. 401-413. Springer, Singapore, 2021.

[15] Batra, Kushal, Sachin Sekhar, and R. Radha. "Breast cancer detection using CNN on mammogram images." In International Conference On Computational Vision and Bio Inspired Computing, pp. 708-716. Springer, Cham, 2019.

[16] Devakumari, D., and V. Punithavathi. "Noise Removal in Breast Cancer Using Hybrid De-noising Filter for Mammogram Images." In International Conference On Computational Vision and Bio Inspired Computing, pp. 109-119. Springer, Cham, 2019.

[17] Inamdar, Vijaylaxmi, S. G. Shaila, and Manoj Kumar Singh. "FNAB-Based Prediction of Breast Cancer Category Using Evolutionary Programming Neural Ensemble." In Computational Vision and Bio-Inspired Computing, pp. 653-663. Springer, Singapore, 2021.

[18] Dawngliani, M. S., N. Chandrasekaran, R. Lalmawipuii, and H. Thangkhanhau. "Breast Cancer Recurrence Prediction Model Using Voting Technique." In International Conference on Mobile Computing and Sustainable Informatics, pp. 17-28. Springer, Cham, 2020.

[19] Balasubramaniam, Vivekanadam. "IoT based Biotelemetry for Smart Health Care Monitoring System." Journal of Information Technology and Digital World 2, no. 3 (2020): 183-190.

[20] Manoharan, Samuel. "Early diagnosis of Lung Cancer with Probability of Malignancy Calculation and Automatic Segmentation of Lung CT scan Images." Journal of Innovative Image Processing (JIIP) 2, no. 04 (2020): 175-186.

[21] https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/