Machine Learning for Breast Cancer Classification

Subjects: Computer Science, Artificial Intelligence

Contributor: Onyinyechi Jessica Egwom , Mohammed Hassan , Jesse Jeremiah Tanimu , Mohammed Hamada , Oko Michael Ogar

Breast cancer is a prevalent disease that affects mostly women, and early diagnosis will expedite the treatment of this ailment. Recently, machine learning (ML) techniques have been employed in biomedical and informatics to help fight breast cancer. Extracting information from data to support the clinical diagnosis of breast cancer is a tedious and time-consuming task. The use of machine learning and feature extraction techniques has significantly changed the whole process of a breast cancer diagnosis.

breast cancer Feature extraction Linear Discriminant Analysis Machine Learning

1. Introduction

According to the World Health Organization (WHO) [1], cancer is a large group of diseases that occurs in any part or tissue of the body when abnormal cells grows uncontrollably beyond their usual boundaries, invading adjoining parts of the body and destroying body tissues.

The WHO reports that cancer, such as breast, cervical, ovarian, lung and prostate cancer, has accounted for over 10 million deaths in 2022. Breast cancer is the most prevalent cancer at 2.26 million cases and is the leading cause of premature mortality among women globally, with 685,000 deaths [2]. Breast cancer (BC) is one of the most prevailing cancers among women worldwide, with fewer cases in men [3].

Breast cancer is a medical abnormality in which the cells lining the breast ducts form clumps with malignant characteristics. It is the most common cancer in women, found mostly in middle- and-low-income countries of sub-Saharan Africa, most especially Nigeria [4][5].

The primary concern of breast cancer treatment begins with accurate predictions of the cancer presence and classifying the cancer type to determine how to treat the cancer [6]. However, predicting breast cancer type is among the classic problems in health-related research [3]. The accurate classification of breast cancer would translate to its early detection, diagnosis, treatment, and, where possible, full eradication. Furthermore, the accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments [7].

Over the last few decades, several organizations have acquired vast repositories of data collected from diverse sources in distinct formats [8][9]. These collected data could be used in different application domains such as medicine, agriculture and weather forecasting [10]. These increasingly large amounts of data surpass the ability of the traditional methods used in analyzing, searching for patterns and information hidden in them for decision making [11][12]. Data obtained from medical data repositories could be analyzed using machine learning algorithms such as classification, clustering, and regression algorithms. Machine learning algorithms and their usefulness in knowledge detection from medical data repositories have been valuable tools for the success of disease prediction [13][14]. A good number of research works have reported the use of machine learning algorithms for breast cancer predictions [15]. Machine learning algorithms have been prevalent in the development of predictive models to support effective decision-making for breast cancer predictions [16].

Machine learning algorithms as tools have been used to create predictive models for BC to support physicians' decisions with acceptable accuracy [17]. However, these models show some limitations, such as the use of appropriate methods to fit the model depending on the dataset without considering feature extraction techniques [18]; proper feature extraction techniques effectively reduce dimensionality for the better prediction of the disease [19]. There is also an increasing concern regarding the methods of handling missing values in the dataset [20]. Hence, researchers developed an improved machine learning model to give accurate breast cancer predictions and increase survivability rates in women.

2. Breast Cancer Classification

Prediction is one of the most important and essential tasks in machine learning [21]. Extensive research has been conducted using machine learning algorithms on different medical datasets, especially in BC prediction. Most ML techniques used showed good prediction accuracy.

In 2015, [19] used SVM, an artificial neural network, a naïve Bayes classifier and Adaboost for breast cancer prediction using machine learning techniques, where principal component analysis was used for feature space reduction.

In 2020, [22] used an artificial neural network (ANN) and SVM for the prognosis of breast cancer recurrence as well as patient's death within 32 months of undergoing surgery. SVM had the best performance, with an accuracy of 96.86%

Khourdifi [23] applied four machine learning techniques, namely SVM, RF, Naïve Bayes, and K-NN, on the Wisconsin breast cancer dataset from the UCI machine learning repository. The authors used Waikato Environment for Knowledge Analysis (Weka) software for the simulation of the algorithm. In their results, SVM had the best overall performance in terms of effectiveness and efficiency.

Chaurasia et al. [24] used naïve Bayes, RBF network, and J48 for the prediction of benign and malignant breast cancer in the Wisconsin breast cancer database (WBCD) to improve the accuracy of the BC prediction model; the results showed that naïve Bayes was the best predictor. Kumar [25] used naïve Bayes, logistic regression, and decision tree for the performance analysis of data mining algorithms for breast cancer cell detection.

Rajbharath and Sankari [7] used a hybrid of random forest (RF) and logistic regression (LR) algorithms for building a breast cancer survivability prediction model. RF was used to perform a preliminary screening of the variables for ranking. The new data set was extracted from the initial WDBC dataset and input into the logistic regression procedure, which is responsible for building interpretable models for predicting breast cancer survivability.

In 2016, Asri et al. [26] performed a comparison between different machine learning algorithms, support vector machine (SVM), decision tree (C4.5), naïve Bayes (NB) and k nearest neighbors (k-NN), in the Wisconsin Breast Cancer (original) datasets for breast cancer risk prediction and diagnosis. The experimental SVM gave the highest accuracy with low error rate.

Ricciardi et al. [27] used a combination of linear discriminant analysis (LDA) and principal component analysis (PCA) for the classification of coronary artery disease with principal component analysis used to create new features and linear discriminant analysis for the classification, which improved the diagnosis of patients.

Kumar et al. [3] predicted malignant and benign breast cancer using 12 algorithms: Ada Boost M1, decision table, J-Rip, J48, Lazy IBK, Lazy K-star, logistic regression, multiclass classifier, multilayer perceptron, naïve Bayes, random forest and random tree. The primary data were drawn from the Wisconsin breast cancer database, and Lazy K and the random tree had the highest accuracy.

Furthermore, Gupta and Gupta [28] performed a comparative analysis of four widely used machine learning techniques, namely, multilayer perceptron (MLP), decision tree (C4.5), support vector machine (SVM), and K-nearest neighbor (KNN) performed on the Wisconsin Breast Cancer dataset to predict the breast cancer recurrence. The main objective of their work was to obtain the best classifier of the four in terms of accuracy, precision and recall In their work, they concluded that MLP performed better than the other techniques, including when 10-fold cross-validation.

Zheng et al. [29] studied K-means and support vector machine (K-SVM) algorithms based on 10-fold cross-validation, and the proposed methodology improved the accuracy of breast cancer prediction to 97.38% when tested on the Wisconsin Diagnostic Breast Cancer (WDBC). The authors proposed a new combination of machine learning algorithms, specifically using K-means for the separate recognition of the hidden patterns in the malignant and benign tumors and then SVM to generate the new classifier within the 10-fold cross-validation. Their new approach obtained an accuracy of 97.38%, which was higher than the scores for the other six algorithms.

In another study, Sivakami and Saraswathi [20] worked on breast cancer prediction using a DT–SVM hybrid model of decision tree and support vector machine. The decision tree was used for feature selection, and the proposed methodology improved the accuracy of breast cancer prediction to 91%. In another recent study on breast cancer, Wu and Hicks [30] investigated four ML algorithms: support vector machine, K-nearest neighbor, naïve Bayes and decision trees to classify triple-negative breast cancer and non-triple-negative breast cancer for patients using gene expression data. SVM gave better classifications than the other three algorithms.

References

1. WHO. 2022 Cancer. Available online: https:www.who.int/news-rooms/factsheet/details/cancer (accessed on 2 May 2022).
2. Labrèche, F.; Goldberg, M.S.; Hashim, D.; Weiderpass, E. Breast cancer. In Occupational Cancers; Springer: Berlin/Heidelberg, Germany, 2020; pp. 417–438.
3. Kumar, V.; Misha, B.K.; Mazzara, M.; Thanh, D.N.; Verma, A. Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In Advances in Data Science and Management; Springer: Berlin/Heidelberg, Germany, 2019; pp. 435–442.
4. Meera, C.; Nalini, D. Breast cancer prediction system using data mining methods. Int. J. Pure Appl. Math. 2018, 119, 10901–10911.
5. Rathi, M.; Pareek, V. Hybrid approach to predict breast cancer using machine learning techniques. Int. J. Comput. Sci. Eng. 2016, 5, 125–136.
6. Way, G.P.; Sanchez-Vega, F.; La, K.; Armenia, J.; Chatila, W.K.; Luna, A.; Greene, C.S. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep. 2018, 23, 172–180.e3.
7. Rajbharath, R.; Sankari, I.; Scholar, P. Predicting breast cancer using random forest and logistic regression. Int. J. Eng. Sci. Comput. 2017, 7, 10708–10813.
8. Luque, C.; Luna, J.M.; Luque, M.; Ventura, S. An advance review on text mining in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019, 9, 1302.
9. Hassan, M.; Hamada, M. Genetic algorithm for improving prediction accuracy of multi-criteria recommender systems. Int. J. Comput. Intell. Syst. 2018, 11, 146–162.
10. Hassan, M.; Hamada, M. Enhancing Learning Objects Recommendation Using Multi-Criteria Recommender Systems. In Proceedings of the 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering , Bangkok, Thailand, 7–9 December 2016; pp. 62–64.
11. Tanimu, J.J.; Hamada, M.; Hassan, M.; Yusuf, S.I. A contemporary machine learning method for accurate prediction of cervical cancer. In Proceedings of the 3rd ETLT 2021. ACM International Conference on Information and Communication Technology, Aizu, Japan, 27–30 January 2021; p. 04004.
12. Abba, A.H.; Hassan, M. Design and Implementation of a CSV Validation System. In Proceedings of the 3rd international Conference on Applications in information Technology, Wakamatsu, Japan, 1–3 November 2018; pp. 111–116.
13. Osianwo, F.Y.; Akinsola, J.E.T.; Awodele, O.; Hinimikaiye, J.O.; Olakanmi, O.; Akiniobi, J. Supervised machine learning algorithm: Classification and comparisiom. Int. J. Comput. Trends

Technol. 2017, 3, 128–138.

14. Hassan, M.; Hamada, M. A computational model for improving the accuracy of multi-criteria recommender systems. In Proceedings of the 2017 IEEE 11th International Symposium of Embedded Multicore/Many-core Systems-on-chip (MCSoc), Seoul, Korea, 18–20 September 2017; pp. 114–119.
15. Huang, M.W.; Chen, C.W.; Lin, W.C.; Ke, S.W.; Tsai, C.F. Svm and Svm ensembles in breast cancer prediction. PLoS ONE 2017, 12, e0161501.
16. Bazazeh, D.; Shubair, R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In Proceedings of the 2017 International Conference on Electronic Devices, Systems, and Applications, Kuching, Malaysia, 7–8 August 2017; pp. 2–5.
17. Agarap, A.F.M. On Breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, Phu Quoc Island, Vietnam, 2–4 February 2018; pp. 5–9.
18. Elgedawy, M. Prediction of breast cancer using random forest, support vector machines and naïve Bayes. Int. J. Eng. Comput. Sci. 2017, 6, 19884–199889.
19. Wang, H.; Yoon, S.W. Breast Cancer Prediction Using Data Mining Method. In Proceedings of the IIE Annual Conference Proceedings, Institute of Industrial and System Engineers (IISE), New Orleans, LA, USA, 30 May–2 June 2015; p. 818.
20. Sivakami, K.; Saraswathi, N. Mining big data: Breast cancer prediction using DT-SVM hybrid model. Int. J. Sci. Eng. Appl. Sci. 2015, 1, 418–429.
21. Jessica, E.O.; Hamada, M.; Yusuf, S.I.; Hassan, M. The Role of Linear Discriminant Analysis for Accurate Prediction of Breast Cancer. In Proceedings of the 2021 IEEE 14th International Symposium of Embedded Multicore/Many-core Systems-on-chip (MCSoc), Singapore, 20–23 December 2021; pp. 340–344.
22. Boeri, C.; Chiappa, C.; Galli, F.; de Berardinis, V.; Bardelli, L.; Carcano, G.; Rovera, F. Machine learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med. 2020, 9, 3234–3243.
23. Khourdifi, Y. Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. In Proceedings of the 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 5–6 December 2018; pp. 1–5.
24. Chaurasia, V.; Pal, S.; Tiwari, B.B. Prediction of benign and malignant breast cancer using data mining techniques. J. Algorithms Comput. Technol. 2018, 12, 119–126.

25. Kumar Mandal, S. Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. Int. J. Eng. Comput. Sci. 2017, 6, 2319–7242.
26. Asri, H.; Mousannnif, H.; al Moatassime, H.; Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. 2016, 83, 1064–1069.
27. Ricciardi, C.; Valente, S.A.; Edmund, K.; Cantoni, V.; Green, R.; Fiorillo, A.; Picone, I.; Santini, S.; Cesarelli, M. Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Inform. J. 2020, 26, 2181–2192.
28. Gupta, S.; Gupta, M.K. A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques. In Proceedings of the 2nd International Conference on Computing Methodologies and Communication (ICCMC 2018), Erode, India, 15–16 February 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 997–1002.
29. Zheng, B.; Yoon, S.W.; Lam, S.S. Breast cancer diagnosis based on feature extraction using a hybrid of k-mean and support vector machine algorithms. Experts Syst. Appl. 2014, 41, 1476–1482.
30. Wu, J.; Hicks, C. Breast Cancer Type Classification Using Machine Learning. J. Pers. Med. 2021, 11, 61.

Retrieved from https://encyclopedia.pub/entry/history/show/60012