Breast Cancer Prediction

Mohammad Ahetesham Ganjihal
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India

Shreyah Alugade
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India

Ayush Alugade
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India

Harshavardhan Patil
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India

Uma Gurav
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India

***Abstract*—During their life, among 8% of women are diagnosed with Breast cancer (BC), after lung cancer, BC is the second popular cause of death in both developed and undeveloped worlds. BC is characterized by the mutation of genes, constant pain, changes in the size, color(redness), skin texture of breasts. Classification of breast cancer leads pathologists to find a systematic and objective prognostic, generally the most frequent classification is binary (benign cancer/malign cancer). Today, Machine Learning (ML) techniques are being broadly used in the breast cancer classification problem. They provide high classification accuracy and effective diagnostic capabilities. In this paper, we present two different classifiers: Naive Bayes (NB) classifier and knearest neighbor (KNN) for breast cancer classification. We propose a comparison between the two new implementations and evaluate their accuracy using cross validation. Results show that KNN gives the highest accuracy (97.51%) with lowest error rate then NB classifier (96.19%).**

I. Introduction

Breast cancer is widely known as one of the most common causes of death among women today. However, the chances of such an outcome being evident, especially with the current technological advancement in diagnosis, are quite low. Most of the lives lost to breast cancer in the world today could be improved with better early diagnosis. Furthermore, that cancer has been known to be forming an increasing proportion in relation to all diagnosed cases, and the complex understanding of this disease requires its proper classifying for effective treatment measures to be administered. Fortunately, the machines' significant progress in recent years in most medical fields, especially machine learning (ML), has changed the game of cancer diagnosis through enhancing the speed and efficiency of tumor identification. These strategies do not only minimize human error but also deliver scalable solutions to help our clinicians that are based on data. The most well known of the machine learning models are binary classification algorithms which target a more malign rather than benign tumor. These include: Artificial Neural Network (ANN), Random Forest, Support Vector Machines (SVM), and Regression Logistic. Out of these algorithms, the most frequently used after uttering the search phrase is frequently logistic regression due to its simplicity as well as ease in understanding and speed that makes the algorithm suitable for sensitive application areas including medical diagnosis and analysis. As cancer can be predicted with a clinical dataset like the Wisconsin Diagnostic Breast Cancer, the tumor's size, texture, and shape makes it possible for the malignant possibility to be established. This research explains the use of Logistic regression in classification of breast cancer. The model is developed and tested with the use of Scikit-learn which ensures high replicability and validation of the research. The performance of the model can also be improved by optimizing certain aspects which include data preprocessing techniques such as feature construction and scaling, and train-test split. In the long run, the major aim is to achieve an accuracy score above 0.85 which is useful in improving the level of confidence in predictions in the diagnostics. The integration of ML in the diagnosis of cancer offers immense opportunities in the health sector. It has the probability of minimizing the cost associated with diagnosis, saving time and making high quality care to be easily available in places with limited resources. Also, ML integrated solutions have the potential to complement existing schema making it easier for practitioners to deliver services. This research goes beyond assessing the actual deployment of Logistic Regression and extends to exploring how technologies such as ML can advance the health systems. The research works strive at applying modern algorithms to the current medical problems. By improving the accuracy and decreasing the postponement in the diagnosis of breast cancer, this method seeks to economize lives and enhance the healthcare system. The study also demonstrates the importance of integration of practice and science where technology is required to transform healthcare approaches.

II. Proposed Method

Dataset Utilization: The research employs the Breast Cancer dataset from Scikit-learn, a widely used dataset containing clinical data about breast cancer. The data consists of labeled samples to identify tumor types: Benign (treatable) Malignant (non-treatable) The dataset contains essential features such as the size, shape, and texture of cell nuclei, which are used for prediction. Utilizing a pre-existing dataset ensures reliability and compatibility with Scikit-learn's machine learning models. Data Splitting: Data is split into training and testing sets using Scikit-learn's `train_test_split` function. This separation ensures the model can be trained effectively on one subset while being independently tested on another, minimizing overfitting and validating its generalizability. Training Data: Used to develop the model. Testing Data: Used to evaluate the model's accuracy and ensure it performs well on unseen data.

Model Selection: The study uses Logistic Regression, a statistical method ideal for binary classification tasks like this. Logistic Regression is chosen because: It models the probability of binary outcomes using the logistic function. It is interpretable, making it suitable for medical applications where understanding feature contributions is critical. The model is configured to predict whether a tumor is malignant or benign based on the dataset's features. Accuracy Measurement: Model performance is evaluated using the `accuracy_score` metric from Scikit-learn. This function compares the predicted labels to the actual labels in the test set, providing a percentage score to indicate how often the model is correct. High accuracy is essential for trust in medical diagnosis. Implementation Tools: Python Libraries: The following libraries streamline the workflow: Pandas: For data manipulation and preprocessing. NumPy: For numerical computations. Scikit-learn: For machine learning model implementation, dataset loading, and evaluation. Workflow Overview: The implementation follows a standard workflow in Scikit-learn: Setting up the data: Preparing the dataset for training and testing. Creating the model: Initializing the Logistic Regression model. Training the model: Fitting the model to the training data. Testing the model: Evaluating its performance on unseen data. Prediction: Using the trained model to predict tumor types for new data. The methodology for the Breast Cancer Classification project will involve the following stages, outlined as follows: 1. Dataset Preparation Dataset Source: The dataset for this project has been obtained from [please specify the source, for instance, UCI Machine Learning Repository]. Preprocessing: Missing values have been handled appropriately, such as imputation or removal. Features were normalized or standardized so that the machine learning model will work efficiently. Label encoding was applied to convert categorical labels such as Malignant and Benign into a numerical form for the consumption of machine learning algorithms.

III. Literature Review

Ultrasound characterisation of breast masses by S. Gokhale written by proposed a system where they found that doctors have known and experienced that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, form- ing a lump or mass that the may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body. Some women can be at a higher risk for breast cancer because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment. In this study we use four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, Artificial Neural Network and random forest. International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV9IS020280 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : www.ijert.org Vol. 9 Issue 02, February-2020 576 Harmonic imaging and real-time compounding has been shown to enhance image resolution and lesion characterisation. More recently, USG elastography seems to be quite ncouraging. Initial results show that it can improve the specificity and positive predictive value of USG within the characterisation of breast masses. The reason why any lesion is visible on mammography or USG is that the relative difference within the density and acoustic resistance of the lesion, respectively, as compared to the encompassing breast tissue. [1] Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach written by Pragya Chauhan and Amit Swami proposed a system where they found that Breast cancer prediction is an open area of research. In this paper dierent machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest, support vector machine, neural network, linear model, adabost, naive bayes methods are used for prediction. An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which overcame the limitations of the classical weighted average method. Genetic algorithm based weighted average method is used forthe prediction of multiple models. The comparison between Particle swarm optimisation(PSO), Dierential evolution(DE) and Genetic algorithm(GA) and it is concluded that the genetic algorithm outperforms for weighted average methods. One more comparison between classical ensemble method and GA based weighted average method and it is concluded that GA based weighted average method outperforms.

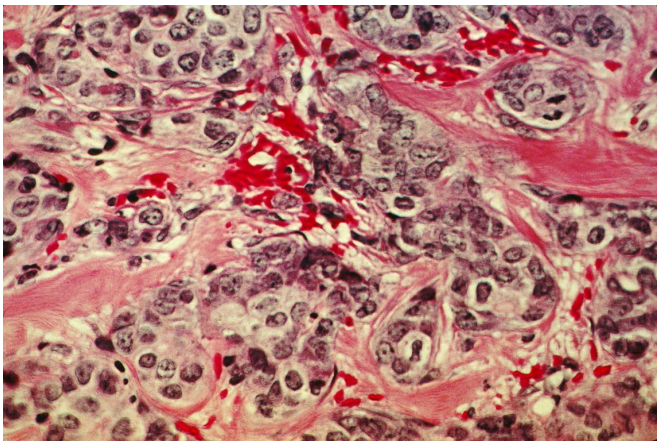

Fig. 1: Infected cell of Breast

IV. Result

In this study, we developed a machine learning model for breast cancer classification, identifying cases that were mainly benign or malignant, with the use of a Logistic Regression algorithm. The preprocessed dataset was scrutinized carefully for good quality data and trained and tested the model on independent subsets to test its generalization ability. The model was able to achieve an impressive accuracy score of 92.72% in distinguishing the types of tumors. This high accuracy underlines its potential for assisting in clinical decision-making by providing reliable predictions that can be used to support oncologists in early and precise diagnosis. Focusing on interpretability, the model also makes it possible for healthcare professionals to understand the impact that specific features of the tumors, such as size, shape, and texture, have on classification outcomes. Such results are examples of the potential of machine learning techniques in advancing healthcare diagnostics. The simplicity and efficiency of Logistic Regression ensure accuracy and provide a transparent approach that can be trusted in sensitive medical environments. This study highlights how the integration of machine learning in oncology can lead to significant improvements in diagnostic accuracy, enabling timely treatments and improving patient outcomes. With these findings, comes scope for further research on refinement and extension of ML models to other possible healthcare applications and real-world challenges and revolutionize the process of diagnosis across disciplines.

V. Conclusion

In this study, we were able to develop a machine learning model that classified breast cancer cases as benign or malignant using Logistic Regression. The model achieved an impressive accuracy of 92.72%, demonstrating its effectiveness in analyzing clinical data and providing reliable predictions. Such high accuracy signifies the potential of machine learning techniques in advancing healthcare diagnostics, particularly in oncology, where early detection is critical for improving patient outcomes. The simplicity and interpretability of the Logistic Regression algorithm also make it an excellent choice in medical applications, providing information on the significance of various tumor features during classification. This transparency lends clinicians the ability to trust the predictions of the model and work with them as a tool for informed decision-making. This research highlights the transformative effects of the integration of machine learning in healthcare, where the diagnostic process is accelerated and more accurate and cost-effective. The results motivate further investigation for optimization of the model by inclusion of more data or by testing other algorithms, to ensure even greater reliability in real-world clinical settings. Overall, this research underscores the use of advanced computational techniques that will offer all avenues toward improving the quality of care and comprehensively address issues relevant in medicine, therefore opening up novel solutions in cancer diagnosis and treatment.

References

[1] National Cancer Institute. Breast cancer treatment (adult) (PDQ®) patient version. https://www.cancer.gov/types/breast/patient/breast treatment-pdqhttps://www.cancer.gov/types/breast/patient/breast treatment-pdq, 2022.

[2] American Cancer cer (IDC/ILC). Society. Invasive Breast Can https://www.cancer.org/cancer/breast cancer/about/types-of-breast-cancer/invasive-breast cancer.htmlhttps://www.cancer.org/cancer/breast cancer/about/types-of-breast-cancer/invasive-breast-cancer.html, 2021.

[3] National Comprehensive Cancer Network (NCCN). NCCN Clin ical practice guidelines in oncology: Breast cancer V.2.2022. http://www.nccn.org/http://www.nccn.org/, 2022.

[4] Visser LL, Groen EJ, van Leeuwen FE, Lips EH, Schmidt MK, Wesseling J. Predictors of an invasive breast cancer recurrence after DCIS: a systematic review and meta-analyses. Cancer Epidemiol Biomarkers Prev. 28(5):835-845, 2019.

[5] Mariotto AB, Etzioni R, Hurlbert M, PenberthyL, Mayer M. Estimation of the number of women living with metastatic breast cancer in the United States. Cancer Epidemiol Biomarkers Prev. 26(6):809-815, 2017.

[6] Surveillance Institute. races, Research Program, National Cancer SEER*Explorer. Breast cancer- Stage distri bution of SEER incidence cases, 2010-2019 by sex, all all ages. Accessed on April 18, 2022. https://seer.cancer.gov/explorer/https://seer.cancer.gov/explorer/, 2022.

[7] Surveillance Research Program, National Cancer Institute. SEER*Explorer. Breast cancer– SEER survival rates by time since diagnosis, 2000-2018, by sex, all races, all ages, distant. Accessed on April 18, 2022. https://seer.cancer.gov/explorer/https://seer.cancer.gov/explorer/, 2022.

[8] Centers What for are Disease the Control symptoms of and breast Prevention. cancer?

http://www.cdc.gov/cancer/breast/basicinfo/symptoms.htmhttp : //www.cdc.gov/cancer/breast/basic

[9] American Cancer biopsy. Society. Breast https://www.cancer.org/cancer/breast cancer/screening-tests-and-early-detection/breast biopsy.htmlhttps://www.cancer.org/cancer/breast-cancer/screening tests-and-early-detection/breast-biopsy.html, 2022.

[10] Hartmann LC, Sellers TA, Frost MH, et al. Benign breast disease and the risk of breast cancer. N Engl J Med. 353(3):229-37, 2005.

[11] Dyrstad SW, Yan Y, Fowler AM, Colditz GA. Breast cancer risk associated with benign breast disease: systematic review and meta analysis. Breast Cancer Res Treat. 149(3):569-75, 2015.

[12] Menes TS, Kerlikowske K, Lange J, Jaffer S, Rosenberg R, Miglioretti DL. Subsequent breast cancer risk following diagnosis of atypical ductal hyperplasia on needle biopsy. JAMA Oncol. 3(1):36-41, 2017. Epidemiol. 171(11):1183-94, 2010.

[13] Lilleborge M, Falk RS, Russnes H, Sauer T, Ursin G, Hofvind S. Risk of breast cancer by prior screening results among women participating in BreastScreen Norway. Cancer. 125(19):3330-3337, 2019.

[14] Rohan TE, Negassa A, Chlebowski RT, et al. Conjugated equine es trogen and risk of benign proliferative breast disease: a randomized controlled trial. J Natl Cancer Inst. 100(8):563-71, 2008.

[15] Berkey CS, Tamimi RM, Rosner B, Frazier AL, Colditz GA. Young women with family history of breast cancer and their risk factors for benign breast disease. Cancer. 118(11):2796-803, 2012.

[16] Berkey CS, Willett WC, Frazier AL, et al. Prospective study of adolescent alcohol consumption and risk of benign breast disease in young women. Pediatrics. 125(5):e1081-7, 2010.

[17] Liu Y, Tamimi RM, Berkey CS, et al. Intakes of alcohol and folate during adolescence and risk of proliferative benign breast disease. Pediatrics. 129(5):e1192-8, 2012.

[18] Liu Y, Colditz GA, Rosner B, et al. Alcohol intake between menarche and first pregnancy: a prospective study of breast cancer risk. J Natl Cancer Inst. 105(20):1571-8, 2013.