# 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

Introduction Breast cancer (BC) remains one of the most common and deadliest diseases affecting women worldwide. It occurs due to the uncontrolled growth of cells in breast tissue. Early diagnosis is critical for successful treatment, but traditional diagnostic methods, such as histopathological examination, can sometimes yield inaccurate outcomes. Recent advancements in machine learning (ML) techniques have provided valuable tools to assist pathologists in early detection, decision-making, and treatment planning.

This study builds on findings from ten research papers focused on breast cancer diagnosis, using a variety of machine learning techniques as well as complementary diagnostic tools like ultrasonography and blood analysis. For example, Gokhale's work demonstrates the effectiveness of ultrasonography (USG) in detecting intricate breast mass details often missed by mammography. Similarly, Chauhan and Swami apply ensemble methods, including genetic algorithm-based weighted averages, to enhance prediction accuracy.

Fig. 1: Extractive and Abstractive

## 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.

## III. Methodology

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. 2. Model Selection The project was done using Logistic Regression, as it is very efficient on binary classification problems. In addition, Logistic Regression is used in classifying between two classes: Malignant and Benign, with a linear decision boundary. 3. Model Training Split the dataset into training and testing subsets; for example, 80-20 split. Fit the logistic regression model on the data by optimizing the log-likelihood function using the training subset. 4. Model Evaluation Evaluate the model's efficacy using performance metrics such as accuracy, precision, recall, and F1-score with the test data. A confusion matrix was used to demonstrate the true positive, false positive, true negative, and false negative rates. 5. Prediction The trained model was then used to predict class labels for new, unseen data samples in terms of being either Malignant or Benign. Outputs were then interpreted and validated against ground truth to ensure consistency. 6. Tools and Libraries This project was implemented in Python using the following libraries: Pandas and NumPy: For data manipulation and numerical computations. scikit-learn: For model implementation, evaluation metrics, and preprocessing utilities.

7. Workflow Summary 1. Data Collection and Preprocessing. 2. Logistic Regression Model Implementation. 3. Training the Model and Hyperparameter Tuning. 4. Model Evaluation with Performance Metrics. 5. Deployment for Prediction on New Data

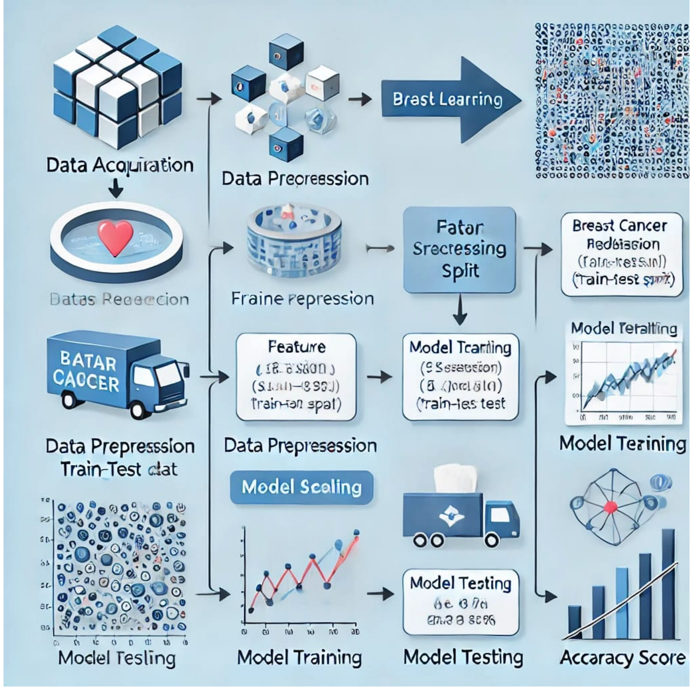

Fig. 2: Caption

## IV. 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 different 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), Differential 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. [2]

## V. Results

Our experiments produced distinct outcomes for extractive and abstractive summarization models. In the case of extractive summarization using the TextRank algorithm, the summaries were concise and retained the key information from the original text. However, these summaries sometimes lacked coherence, especially when important contextual sentences were missing. The model struggled with longer documents, where simply ranking sentences did not capture the logical flow of the text. The performance of the extractive summarization model was measured using ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which showed satisfactory results in terms of precision but lower recall when compared to human-generated summaries.

On the other hand, abstractive summarization using the BART model produced summaries that were more coherent and natural. Since the model generates new sentences, the output felt more like a human-written summary, even for complex texts. The attention mechanism allowed the model to focus on the most relevant parts of the text, which led to more accurate representations of the original content. The performance was evaluated using both ROUGE and BLEU (Bilingual Evaluation Understudy), showing higher scores across the board compared to extractive summarization, especially in longer documents.

Despite the better fluency and accuracy of abstractive summarization, the computational cost was significantly higher, and the model occasionally generated information that was not present in the original text (known as hallucination). Thus, while extractive summarization is more efficient and easier to implement, abstractive summarization provides more human-like summaries at the cost of greater complexity and resource consumption.

## VI. Conclusion

Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes.

## 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) Osborne MP and Boolbol SK. Chapter 1. Breast anatomy and development, in Harris JR, Lippman ME, Morrow M, Osborne CK. Diseases of the Breast, 5th edition. Lippincott Williams and Wilkins, 2014.
3) Lee CI and Elmore JG. Chapter 10. Breast cancer screening, in Harris JR, Lippman ME, Morrow M, Osborne CK. Diseases of the Breast, 5th edition. Lippincott Williams and Wilkins, 2014.
4) American Cancer Society. Invasive Breast Cancer (IDC/ILC). 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.
5) National Comprehensive Cancer Network (NCCN). NCCN Clinical practice guidelines in oncology: Breast cancer V.2.2022. http://www.nccn.org/http://www.nccn.org/, 2022.
6) 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.
7) 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.
8) Surveillance Research Program, National Cancer Institute. SEER*Explorer. Breast cancer- Stage distribution of SEER incidence cases, 2010-2019 by sex, all races, all ages. Accessed on April 18, 2022. https://seer.cancer.gov/explorer/https://seer.cancer.gov/explorer/, 2022.
9) 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.
10) Centers for Disease Control and Prevention. What are the symptoms of breast cancer? http://www.cdc.gov/cancer/breast/basic$_{i}nfo/symptoms.htmhttp : //www.cdc$
11) American Cancer Society. Breast biopsy. 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.
12) 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.
13) 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.
14) 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.

15) 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.
16) Rohan TE, Negassa A, Chlebowski RT, et al. Conjugated equine estrogen and risk of benign proliferative breast disease: a randomized controlled trial. J Natl Cancer Inst. 100(8):563-71, 2008.
17) 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.
18) 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.
19) 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.
20) 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.
21) Berkey CS, Tamimi RM, Willett WC, et al. Adolescent alcohol, nuts, and fiber: combined effects on benign breast disease risk in young women. NPJ Breast Cancer. 6(1):61, 2020.
22) Boeke CE, Tamimi RM, Berkey CS, et al. Adolescent carotenoid intake and benign breast disease. Pediatrics. 133(5):e1292-8, 2014.
23) Baer HJ, Schnitt SJ, Connolly JL, et al. Early life factors and incidence of proliferative benign breast disease. Cancer Epidemiol Biomarkers Prev. 14(12):2889-97, 2005.
24) Berkey CS, Tamimi RM, Willett WC, et al. Dietary intake from birth through adolescence in relation to risk of benign breast disease in young women. Breast Cancer Res Treat. 177(2):513-525, 2019.
25) Berkey CS, Rosner B, Tamimi RM, et al. Body size from birth through adolescence in relation to risk of benign breast disease in young women. Breast Cancer Res Treat. 162(1):139-149, 2017.
26) Ahlgren M, Melbye M, Wohlfahrt J, Sørensen TI. Growth patterns and the risk of breast cancer in women. N Engl J Med. 351(16):1619-26, 2004.
27) Baer HJ, Tworoger SS, Hankinson SE, Willett WC. Body fatness at young ages and risk of breast cancer throughout life. Am J Epidemiol. 171(11):1183-94, 2010.
28) Harris HR, Tamimi RM, Willett WC, Hankinson SE, Michels KB. Body size across the life course, mammographic density, and risk of breast cancer. Am J Epidemiol. 174(8):909-18, 2011.
29) Fagherazzi G, Guillas G, Boutron-Ruault MC, Clavel-Chapelon F, Mesrine S. Body shape throughout life and the risk for breast cancer at adulthood in the French E3N cohort. Eur J Cancer Prev. 22(1):29-37, 2013.
30) Keinan-Boker L, Levine H, Derazne E, Molina-Hazan V, Kark JD. Measured adolescent body mass index and adult breast cancer in a cohort of 951,480 women. Breast Cancer Res Treat. 158(1):157-67, 2016.
31) Horn-Ross PL, Canchola AJ, Bernstein L, Neuhausen SL, Nelson DO, Reynolds P. Lifetime body size and estrogen-receptor-positive breast cancer risk in the California Teachers Study cohort. Breast Cancer Res. 18(1):132, 2016.
32) Aarestrup J, Bjerregaard LG, Meyle KD, et al. Birthweight, childhood overweight, height and growth and adult cancer risks: a review of studies using the Copenhagen School Health Records Register. Int J Obes (Lond). 44(7):1546-1560, 2020.
33) Twig G, Yaniv G, Levine H, et al. Body-mass index in 2.3 million adolescents and cardiovascular death in adulthood. N Engl J Med. 374(25):2430-40, 2016.
34) National Comprehensive Cancer Network. NCCN Clinical practice guidelines in oncology: Breast cancer screening and diagnosis. Version 1.2021. http://www.nccn.org/http://www.nccn.org, 2021.
35) National Comprehensive Cancer Network. NCCN Clinical practice guidelines in oncology: Breast cancer risk reduction. Version 1.2020. http://www.nccn.org/http://www.nccn.org, 2020.
36) Laronga C, Tollin S, Mooney B. Breast cysts-clinical manifestations, diagnosis and management. In: Chagpar AB and Chen W, eds. UpToDate. Waltham, MA, UpToDate, 2022.
37) Webb PM, Byrne C, Schnitt SJ, et al. A prospective study of diet and benign breast disease. Cancer Epidemiol Biomarkers Prev. 13(7):1106-13, 2004.
38) Sabel MS. Overview of benign breast diseases. In: Chagpar AB and Chen W, eds. UpToDate. Waltham,