Breast Cancer Classification using Machine Learning

1st R G Rohit
AIE
Amrita Vishwa Vidyapeetham
Bangalore, India

2nd K Rahul Rao
AIE
Amrita Vishwa Vidyapeetham
Bangalore, India

3rd Bharath Teja
AIE
Amrita Vishwa Vidyapeetham
Bangalore, India

***Abstract*—Breast cancer remains a very important health concern globally and many methods are being researched for effective treatment planning . Machine learning (ML) techniques have emerged as powerful tools in medical research, offering the potential to enhance the accuracy and efficiency of cancer diagnosis. In our project, we are training our data to figure our whether the tumor is Benign tumor or Malignant tumor. Malignant tumors are the ones that are cancerous and harmful meanwhile Benign are non cancerous tumors. In this study, we explore the application of ML algorithms for the classification of breast cancer, through comprehensive data preprocessing, feature selection, and model training, we evaluate the performance of various ML algorithms, including logistic regression in distinguishing between Malignant and Benign breast tumors.**

I. Introduction

Machine learning techniques, which encompass algorithms capable of learning from data and making predictions or decisions, have garnered significant attention in the field of medical and healthcare. As it wont be easy to process large datasets easily manually, ML algorithms can extract meaningful patterns and relationships in the datasets no matter the size. In the context of breast cancer, ML-based approaches can definitely help by improving the accuracy and reliability of classification tasks, such as distinguishing between malignant and benign tumors. As this classification type is a binary classification, it is well suited for breast cancer classification as we need to classify the tumor into either of the two types, whether it is Benign or Malignant.

II. Literature Survey

Machine Learning for Breast Cancer Prediction: Machine learning techniques gives us the most accurate predictions for early breast cancer detection , it helps the surgeons to make up the process fastly. Dataset Used: The Wisconsin Breast Cancer dataset from the University of California, Irvine, comprising records of 699 patients with features including clump thickness, cell size and shape uniformity, adhesion, epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Performance : It increases the performance in terms of accuracy, precision . It provides 97.07 percent accuracy.[1]

The paper titled "Breast Cancer Detection: Comparative Analysis of Machine Learning Classification Techniques" authored by Harsh Sharma, Pooja Singh, and Ayush Bhardwaj from various educational institutions in Noida, India, provides an in-depth investigation into the application of machine learning algorithms for the early detection of breast cancer using the Wisconsin Breast Cancer dataset. The introduction fives us the information about the importance of early detection of breast cancer due to its increasing incidence .It enhances the significance of machine learning techniques especially in cases where symptoms are not easily noticeable. The writer describes the significance of breast cancer globally,especially in india The algorithms that are discussed in the provided paper are such as Random Forest, and Decision Trees, achieving high accuracies and demonstrating the effectiveness of machine learning in this category. This paper helps in understanding and the step-by-step segregation of stages,and the side effects caused by Breast Cancer by using methodologies such as Random Forest and Descision Trees etc.... The implementation and result analysis section present the experimental setup, performance metrics, and accuracy percentages of the algorithms. Finally, the conclusion summarizes the importance of early breast cancer detection and the effectiveness of machine learning algorithms.[2]

this paper examines the use of supervised learning algorithms for breast cancer prediction focusing on factors like regression,decesion tree,svm,naive bayes and knn .evaluates the performance of these on wisconsindataset.we can compare using accuracy,precision,recall,f-1 score with other models.it is very Import to predict breast cancer and machine learning plays a vital role in it ,it even looks forward about optimization techniques.this literature survery talks about potential of machine learning[3]

The literature review provides an overview of the utilization of various machine learning (ML) approaches for early breast cancer prediction using the Wisconsin Breast Cancer Dataset. Numerous studies have employed a range of algorithms, including Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Neural Networks, and others, achieving accuracies ranging from 95percent to 99.42 percent. These investigations primarily focus on comparing the effectiveness of these algorithms, with SVM, Random Forest, Artificial Neural Networks (ANN), and KNN frequently demonstrating significant accuracy rates. Furthermore, certain studies explore the impact of different feature selection methods on classification accuracy. Collectively, the literature highlights the promising potential of ML techniques in ad-

vancing early breast cancer detection, with potential benefits for healthcare outcomes and cost reduction.[4]

Detection Of Breast Cancer Based on Fuzzy Logic

In this paper the researcher are making an advanced system for detectin of Breast Cancer : - They have introduces an innovative method which uses machine learning techniques like the Support Vector Machine (SVM) and Decision Tree (DT) algorithms - Their aims is the early detection and diagnosis of breast cancer. -They have utilized the Wisconsin data set. Their fuzzy based SVM and DT classification approach shows exceptional performance metrics, accuracy, precision, recall and specificity - They also have an outstanding accuracy rate of 98.2- They have also had analysis and comparison with alternative classification models such as K-Nearest Neighbor (KNN) and Naive Bayes (NB) - Their research shows efficacy of the fuzzy-based SVM and DT method in identifying breast cancer. - These findings shows the potential of machine learning in the field of breast cancer detection.[5]

The paper discuss the use of Naive Bayes(NB) classifier and k-nearest neighbor(KNN) in machine learning for breast cancer classification. It underscores the significance of accurate classification methods for breast cancer. The literature survey likely goes into existing research on machine learning-based choice of NB and KNN. The paper lastly concludes that with a comparative analysis, indicates that KNN model achieves higher accuracy than NB model on the Winconsin Breast Cancer dataset. This also concludes that by demonstrating that KNN performs better than NB in terms of accuracy on the Winsconsin Breast Cancer dataset.[6]

This paper discuss the importance of breast cancer as a leading cause of cancer-related deaths among women. The literature survey explores the evolution of large data in the health system and its impact on breast cancer detection. It discuss existing research on the application of machine learning algorithms to breast cancer datasets, highlighting the importance of early detection and prevention. The authors have reviewed studies employing Random Forest, Naive Bayes, Support Vector Machines(SVM), and K-Nearest Neighbors(KNN) for breast cancer prediction. The whole focus is on selecting the most efficient algorithm, and the paper concludes with experimental results indicating that SVM performs the highest accuracy at 97.9 percent. This finding conveys to thne identification of the optimal machine-learning algorithm for breast cancer prediction.[7]

The paper starts by highlighting breast cancer as prevalent tumor that affecting breast tissues, especially in women, and highlights its status as a main cause of female death rate worldwide. The focus of the article is a comparative analysis of machine learning, deep learning, and data mining techniques for breast cancer detection. The literature survey likely explores the efforts of different researchers in the field of breast cancer diagnosis and prognosis, approaching the varied accuracy rates across different techniques, situations, tools, and datasets. The article aims to provide a comprehensive overview of already existing machine learning algorithms used in the breast cancer detection, offering useful information for beginners entering the field. The conclusion summarizes the main findings of the review, highlighting the search for the most suitable algorithm for effective breast cancer prediction. Future work is suggested, including addressing issues such as limited dataset availability, the imbalance of positive and negative data, and the need for solving challenges related to correct diagnosis and prediction of breast cancer.[8]

This paper authorized by S. Turgut, M. Dağtekin, and T. Ensari present their groundbreaking research at the highly esteemed 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) in Istanbul, Turkey. This groundbreaking paper delves into the exciting possibility of using machine learning methods for the classification of microarray breast cancer data. Specifically, the study focuses on the crucial elements of feature extraction and support vector machines (SVM) within the larger context of machine learning algorithms. Relevant keywords such as feature extraction, machine learning algorithms, SVM, neurons, and microarray technology are woven seamlessly into the narrative, providing a comprehensive overview of the cutting-edge research presented. The authors likely conducted a thorough literature survey, investigating previous studies on microarray-based breast cancer classification methods to inform their own innovative approach.[9]

This paper "Breast Cancer Detection Using Machine Learning Algorithms," presented at the International Conference on Computational Techniques, Electronics, and Mechanical Systems, S. Sharma, A. Aggarwal, and T. Choudhury dive into the crucial role of machine learning in detecting breast cancer. Through thorough research, the authors likely explore the history and progression of using machine learning for breast cancer diagnosis. This includes in-depth discussions on popular algorithms such as random forest, k-Nearest-Neighbor, and naive bayes, with a focus on their success in training and classifying data. To provide a comprehensive examination on the effectiveness of machine learning in breast cancer detection, highlighting key methodologies and advancements. The paper's includes the importance on classification algorithms, machine learning techniques, and the use of random forest, K-Nearest Neighbor(KNN), and Naive Bayes for breast cancer prediction.[10]

III. Papers

Breast Cancer Risk Prediction based on Six Machine Learning Algorithms 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering

Breast Cancer Detection: Comparative Analysis of Machine Learning Classification Techniques 2022 International Conference on Emerging Smart Computing and Informatics (ESCI)

Model Selection for Predicting Breast Cancer using Supervised Machine Learning Algorithms 2020 IEEE International Conference for Convergence in Engineering

Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset 2022 3rd International Informatics and Software Engineering Conference (IISEC)

Detection Of Breast Cancer Based on Fuzzy Logic 2023 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)

Breast cancer classification using machine learning 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)

Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)

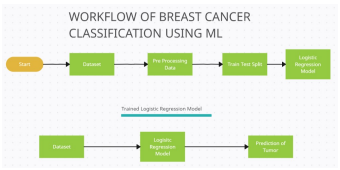
2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) NOREEN FATIMA, LI LIU SHA HONG AND HAROON AHMED

Microarray breast cancer data classification using machine learning methods 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)

Breast Cancer Detection Using Machine Learning Algorithms 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)

IV. Methodology

In order to build a breast cancer classification system using machine learning, the required steps are given below, The foremost is data collection and understanding, we need to collect datasets that contain data about breast cancer tumors that mainly consists of features such as radius, texture, perimeter, area and smoothness. These parameters are necessary as they are important in classifying the data. This dataset can be obtained from websites such as Kaggle. Tumors are of two types Benign which is nothing but non-cancerous and malignant which is cancerous. Malignant tumors are the tumors that have the ability to spread to other parts of the body. The second step is Data processing, this is the most important step in training data for machine learning because if the data is not pre processed, then the accuracy will be affected severely. We process the raw data to make it easier for the machine learning models. Then we split the data into training and testing set respectively in order to evaluate its performance The third step is Model training, we have chosen the most suitable binary classification machine learning model that is Logistic regression. We train the model in this which is learning the relationships between the labels and the target variable. The fourth step is Model evaluation, this is where we test the performance of the machine learning model on the testing data. We calculate the accuracy, precision and F1 score to look at how well the model is performing in classifying the tumors. The final step is Prediction, we use the trained model to predict the tumors whether they are Benign or Malignant.



V. Progress

A. Progress report 1.

We have found the dataset that is to be used in Kaggle website. It is a analysis of Breast Cancer. The dataset contains the following terms in it, id, diagnosis, radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, fractal dimension mean, radius, texture , perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension, radius worst, texture worst, perimeter worst, area worst, smoothness worst, compactness worst, concavity worst, concave points worst, symmetry worst, and fractal dimension worst. Currently we are in the stage of pre processing the dataset as it is a necessary step in training machine learning models to improve the accuracy of the result. First we are importing the necessary dependencies that is the packages that are necessary such as scikit-learning, numpy, pandas, and in scikit learning we are using it to train and test split, logistic regression and to calculate the accuracy score. We loaded the data from sklearn and then we loaded it later on to data frame. When the dataset loaded into the sklearn is printed, it will give values in 0's and 1's where 1's are Benign and 0's are Malignant. Then we calculated the shape of the dataset that it nothing but the amount of rows and columns present in it, which has turned out to be (569,31). The next step we took is to find out the missing values. This is the work that has been completed so far and we have partially completed the stage of pre processing the dataset.

B. Report assignment

As we increase the value of K, the decision boundaries increases which in turn helps in figuring out the best value to get the best accuracy. Smaller values of K might lead to the case of overfitting. Larger values of K leads to underfitting.Yes, the data can be said to have a regular fit as it has performed well in both training sets an testing sets. Overfitting can occur when the given data is small.