International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Breast Cancer Detection Using Machine Learning

Rishabh Kumar[1], Rohit Kumar Yadav[2], Md. Shahbaz Hassan[3], Dr. Bhoopendra Dwivedy[4]

[1,2,3]Master of Computer Application, Galgotias University
[4]Professor, Master of Computer Application, Galgotias University

Abstract
Cancer mortality remains a significant challenge in developing nations, despite various preventive measures. Some cancer types still lack effective treatments. Breast cancer, one of the most prevalent forms, relies heavily on early detection for successful treatment. Accurate diagnosis plays a crucial role in managing breast cancer. Numerous studies have explored methods for predicting breast tumor types. This study utilized breast cancer tumor data from kaggle datasets forecast tumor types. The research employed data visualization and machine learning techniques, including logistic regression and Python were used to implement these techniques and visualizations. The research aimed to conduct a comparative analysis of data visualization and machine learning applications for breast cancer detection and diagnosis. The diagnostic performance of these applications was found to be comparable in identifying breast cancers. The study demonstrated that data visualization and machine learning techniques could significantly benefit the decision-making process in cancer detection. Various machine learning and data mining techniques were proposed for breast cancer detection. The logistic regression model, incorporating all features, yielded the highest classification accuracy at 98.1%. This approach showed improved accuracy performance, suggesting potential new avenues for breast cancer detection.

Keywords: Breast Cancer, Dataset, CNN, KNN, Naïve Bayes, Random Forest, SVM, Logistic Regression

1. Introduction

The Kaggle datasets identifies breast cancer as the most prevalent cancer among women. The survival rates for breast cancer vary significantly, influenced primarily by the cancer type and stage at diagnosis. Breast cancer originates in breast cells, typically in the lobules or ducts, but can also develop in the breast's fatty or fibrous connective tissue. The uncontrolled cancer cells often infiltrate healthy breast tissue and may spread to axillary lymph nodes. Medical professionals attribute breast cancer to abnormal cell growth in the breast, which can expand like Meta Size from the breast to lymph nodes or other body parts.

Early detection and halting of these unwanted cells' growth is crucial to prevent further complications. Upon tumor diagnosis, doctors first determine whether it is benign or malignant, as treatment approaches differ. Benign tumors are non-cancerous and do not spread, while malignant tumors are cancerous and can metastasize. A significant challenge in breast cancer management is the lack of effective early-stage diagnostic tools, which hinders timely treatment initiation to impede unwanted cell or tumor growth. Early diagnosis of diseases often leads to better outcomes with minimal intervention. Many individuals fail to

detect their illness before it becomes chronic, contributing to increased global mortality rates. Breast cancer is potentially curable when identified in its early stages, before it spreads throughout the body. The scarcity of prognosis models makes it difficult for doctors to devise treatment plans that may extend patient survival. Consequently, there is a need for developing techniques with minimal error to enhance accuracy. Traditional breast cancer detection methods like mammograms, ultrasounds, and biopsies are time-consuming, necessitating a computerized diagnostic system employing Machine Learning methodology. This approach incorporates algorithms that aid in tumor classification and more precise cell detection while reducing time requirements.

2. Literature Review

This segment examines prior research on breast cancer diagnosis using various machine learning techniques.

A study by Arpita Joshi and Dr. Ashish Mehta [1] evaluated the performance of different classification methods, including KNN, SVM, Random Forest, and Decision Tree (both Recursive Partitioning and Conditional Inference Tree). They utilized the Wisconsin Breast Cancer dataset from the UCI repository for their analysis. The results of their simulations indicated that KNN was the most effective classifier, followed by SVM, Random Forest, and Decision Tree, in that order**.**

In a study by David A. Omondiagbe, Shanmugam Veeramani, and Amandeep S. Sidhu [2], the researchers evaluated the effectiveness of three machine learning algorithms: Support Vector Machine, Artificial Neural Network, and Naïve Bayes. They utilized the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and combined these techniques with feature selection and extraction methods to determine the most appropriate approach. The results of their simulations indicated that SVM-LDA was preferred over the other methods due to the extended computational time required by the alternatives.

A comparative analysis of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) was conducted by Kalyani Wadkar, Prashant Pathak and Nikhil Wagh [3]. They also incorporated various other classifiers such as Convolutional Neural Networks (CNN), K-Nearest Neighbors (KNN), and Inception V3 to enhance dataset processing. The study's findings and performance evaluation revealed that ANN outperformed SVM as a classifier, demonstrating a higher rate of efficiency.

By utilizing machine learning techniques such as logistic regression, random forest, K-nearest neighbor, decision tree, support vector machine, and naïve bayes classifier, as well as deep learning techniques such as artificial neural networks, convolutional neural networks, and recurrent neural networks.

Monica Tiwari, Rashi Bharuka, Praditi Shah, and Reena Lokare [4] introduced a novel approach to breast cancer detection. According to the results of the comparative study between machine learning and deep learning approaches, the accuracy achieved by the CNN model (97.3%) and ANN model (99.3%) was superior to that of the machine learning models.

Using machine learning techniques like the Convolutional Neural Network (CNN) method for breast image classification, conventional Neural Network (NN), Random Forest (RF) algorithm, Support Vector Machines (SVM), and Bayesian methods, Abdullah-Al Nahid and Yinan Kong [5] presented a novel approach to detect breast cancer by image classification. Given that Convolutional Neural Network (CNN) techniques typically use kernels to extract features globally, which are then used for image classification, the CNN approach turned out to be the most effective for detecting breast cancer.

According to S. Vasundhara, B.V. Kiranmayee, and Chalumuru Suresh [6], mammogram images can be automatically classified as benign, malignant, or normal. utilizing different machine learning algorithms.

A comparison of Random Forest, Convolutional Neural Networks, and Support Vector Machines is conducted. According to the simulation results, CNN is the best classifier because it uses morphological and filtering operations to automatically classify digital mammograms.

Dr. William H. Walberg's dataset from the University of Wisconsin Hospital was utilized by Muhammad Fatih Ak [7]. This dataset was subjected to a variety of data visualization and machine learning techniques, such as logistic regression, k-nearest neighbors, support vector machines, naïve Bayes, decision trees, random forests, and rotation forests and Python were selected to be used with these visualization and machine learning methods. A comparative analysis of all the methods was carried out. The recommended method demonstrated an improvement in accuracy performances, and the results obtained with the logistic regression model with all features included showed the highest classification accuracy (98.1%).

SVM, Logistic Regression, Naïve Bayes, and Random Forest have all been compared by Sivapriya J, Aravind Kumar V, Siddarth Sai S, and Sriram S [8]. The comparison is carried out using the Wisconsin Breast Cancer dataset. The Random Forest algorithm demonstrated the highest accuracy (99.76%) with the lowest error rate, according to the results of the conducted experiments. All of the experiments were carried out in a simulated setting using the ANACONDA Data Science Platform.

3. Methodology

1. Dataset

Fine Needle Aspiration-it is a type of biopsy procedure. In fine needle aspiration, a thin needle is inserted into an area of abnormal appearing tissue or body fluid. As with other type of biopsies, the sample collected during fine needle aspiration can help make a diagnosis or rule out conditions such as cancer.

2. Workflow

The first step is data collection from kaggle.com so we will try to figure out the data which we are going to work on and once we have this data set the data collection process is completed.

The second step is process this data because we cannot feed this raw data to our machine learning model so there are some data pre-processing steps which we will cover here.

The third step is to split the data into training data and testing data, the purpose of this is we will train our machine learning model using the training data and then we will evaluate our model using the test data.

Once we do that we will train our machine learning model so in this case we are going to use a logistic regression model because logistic regression is one of the best model when it comes to binary classification. classification in the sense we are going to classify the data points into two types one is benign and other is malignant.

Once you train logistic regression model with this particular training data set we will have a trained logistic regression model, and we will also do evaluation on this model using the test data so once all these things are completed we can have a new data and once we give this new data the model can tell you whether that particular tumor is benign or malignant.

3. Import Dependencies

Numpy- Numpy is a widely helpful cluster handling package that provides tools for working with exhibitions and an elite multidimensional exhibit item. The essential Python package for logical processing is called Numpy.

Pandas- Panda is an open-source Python library that uses its incredible information structures to provide excellent information control and analysis tools. Python was heavily used for information robbing and

preparedness. It made hardly no effort to examine the information. Pandas addressed this problem. This allows five standard ventures to handle and investigate information while ignoring the reason for the information load, preparation, control, modeling, and examination. Python and Pandas are used in a variety of academic and professional domains, such as finance, statistics, research, and so on.

Sklearn- usecase from we can get dataset.

Model_selection- here we split our datasets into Training data and Testing data

Logistic_regression- The dependent variable in linear regression cannot be predicted using the independent variable and the resulting linear regression hyperplane. Thus, logistic regression is employed when categorical data is present. Rather of making a continuous prediction, logistic regression makes a true-false prediction. It serves as a classification tool. The independent variable is transformed into a probability expression pertaining to the dependent variable that spans from 0 to 1 using the sigmoid function. It is a well-liked machine learning algorithm since it can categorize new samples using both continuous and discrete measurements and return probabilities. The assumption of linearity between the dependent and independent variables is a disadvantage of logistic regression.

Accuracy_score- it is used to determine how many correct predictions our model is making.

4. Result

1 --> Benign

0 --> Malignant

```
[ ] data_frame.groupby('label').mean()
```

label	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	radius error	texture error	perimeter error	area error
0	17.462830	21.604906	115.365377	978.376415	0.102898	0.145188	0.160775	0.087990	0.192909	0.062680	0.609083	1.210915	4.323929	72.672406
1	12.146524	17.914762	78.075406	462.790196	0.092478	0.080085	0.046058	0.025717	0.174186	0.062867	0.284082	1.220380	2.000321	21.135148

Conclusion

In this study, we looked at machine learning methods for detecting breast cancer. We use Logistic Regression and found that it performs better in terms of accuracy, precision, and dataset size.

References

1. Arpita Joshi and Dr. Ashish Mehta “Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer” (2017).
2. David A. Omondiagbe, Shanmugam Veer Amani and Amandeep S. Sidhu “Machine Learning Classification Techniques for Breast Cancer Diagnosis” (2019).
3. Kalyani Wadkar, Prashant Pathak and Nikhil Wagh “Breast Cancer Detection Using ANN Network and Performance Analysis with SVM” (2019).
4. Monika Tiwari, Rashi Bharuka, Praditi Shah and Reena Lokare “Breast Cancer Prediction using Deep learning and Machine Learning Techniques”.
5. Abdullah-Al Nahid and Yinan Kong “Involvement of Machine Learning for Breast Cancer Image Classification: A survey” (2017).
6. S.Vasundhara , B.V. Kiranmayee and Chalumeau Suresh "Machine Learning Approach for Breast Cancer Prediction" (2019).

7. Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications" (2020).
8. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning" (2019).