COMP-7745

Machine Learning

PREDICTION OF BREAST CANCER

Project Report

Pooja Sree Gurrala - U00885267

Kavya Bojja - U00893842

ABSTRACT

Breast cancer is the most prevalent cancer in women, responsible for approximately 30% of all cancer diagnoses. Traditional diagnostic tests such as mammography, ultrasound, and MRI are not always definitive, and better means of identifying and diagnosing breast cancer are required. Machine learning algorithms can evaluate a significant amount of data, including patient demographics, clinical characteristics, and imaging findings, to predict the possibility of malignant growth with a high degree of accuracy. Researchers have developed various machine learning models and algorithms to predict the aggressiveness of breast cancer, which has improved the accuracy of breast cancer detection and reduced the need for unnecessary biopsies and procedures.

INTRODUCTION

Breast cancer is a common type of cancer that affects women worldwide. Early detection and diagnosis of breast cancer are essential for successful treatment and better patient outcomes. Mammography, ultrasound, and magnetic resonance imaging (MRI) are standard diagnostic tests that may help in identifying abnormalities, but they are not always definitive. Machine learning algorithms have become effective tools for the early identification and diagnosis of breast cancer in recent years.

The motivation behind developing a breast cancer prediction model using machine learning algorithms is to provide a more accurate and reliable means of detecting and diagnosing breast cancer. Breast cancer is the most common cancer among women worldwide, and early detection and treatment are crucial to improving patient outcomes and reducing medical expenses. While traditional diagnostic methods such as mammography and ultrasound have been useful, they are not always definitive, and there is a need for better and more dependable means of identifying and diagnosing breast cancer.

Machine learning algorithms offer the potential to improve breast cancer detection by evaluating large amounts of data, including patient demographics, clinical characteristics, and imaging findings. By identifying trends and forecasting the possibility of malignant growth with a high

degree of accuracy, these algorithms can assist medical professionals in making informed decisions and provide patients with better outcomes.

It can provide a more accurate and reliable means of detecting breast cancer, reducing the need for unnecessary biopsies and procedures, and improving patient outcomes. Ultimately, the development of this model can save lives and improve the quality of life for those who have been diagnosed with breast cancer.

METHODOLOGY:

DATA COLLECTION

For this step, we utilized data that is available on Kaggle.

https://www.kaggle.com/code/jssicapinto/cancer-machine-learning/input

It consists of 32 attributes.

DATA PREPROCESSING

- In this dataset some attributes contain continuous values with many decimal places, some we have rounded those values to two decimal places.
- We have transformed the target attribute “diagnosis” datatype from categorical to numeric using a label encoder.
- In this dataset some attributes have less correlation with the target variable so we have dropped those from the dataset.

  Dropped attributes:

  texture_se','smoothness_se','symmetry_se','fractal_dimension_mean','id','fractal_dimension_se'
- PCA has been used to extract the most relevant features from the data by finding the principal components of the data that capture the maximum amount of variation in the data.

MODEL IMPLEMENTATION

After pre-processing the data, out of all samples of data, randomly selected 70% of the samples are taken for training the model and the remaining 30% of the samples are taken for testing the model. Using the training data, supervised machine learning algorithms i.e., Logistic Regression, Random Forest, and Decision Tree are used to fit the model.

MODEL EVALUATION

Using 30% of testing data, diagnosis values are predicted by the models. Among the three implemented models Random Forest is having highest accuracy compared to other models.

MODEL DESCRIPTIONS

Supervised learning

Supervised learning is a category of machine learning algorithms that learn from labeled training data to make predictions about unseen or future data. The objective of supervised learning is to learn a mapping function from input variables to output variables based on a set of labeled examples.

- **Logistic regression :**
  Logistic regression is a supervised learning algorithm used for binary classification tasks. It is a linear model that makes predictions using a sigmoid function to transform the linear combination of input variables into a probability between 0 and 1.
- **Random Forest**
  Random Forest is a supervised learning algorithm used for classification and regression tasks. It is an ensemble method that combines multiple decision trees to make predictions. Each tree in the forest is trained on a subset of the data and a subset of the input variables. The algorithm randomly selects input variables and data samples to build a set of diverse decision trees. The final prediction is made by aggregating the predictions of all the trees in the forest.

- **Decision tree**

  A decision tree is a supervised learning algorithm used for classification and regression tasks. It is a tree-based model that recursively splits the data into subsets based on the value of a selected input variable. The splits are made based on a criterion such as Gini impurity or entropy to maximize the separation between classes. The algorithm continues splitting until a stopping criterion is met, such as a maximum depth or a minimum number of data points in a leaf node. The final prediction is made by traversing the decision tree based on the values of the input variables.

Neural Networks

The neural network model consists of layers of interconnected nodes or neurons. Each neuron takes inputs from the previous layer, performs a weighted sum, applies an activation function, and produces an output. The output of the last layer represents the predicted class of the input data.

In the context of breast cancer prediction, a neural network model can be trained on a labeled dataset to learn patterns and relationships between various clinical and demographic features and the presence or absence of breast cancer. The neural network can then be used to predict the probability of breast cancer for new, unseen data.

- **Multi-Layer Perceptron**

  The MLP is a type of artificial neural network that uses multiple layers of nodes (perceptrons) to classify input data. The MLP is a feedforward neural network that uses backpropagation to adjust the weights and biases of the nodes in each layer during the training process. We chose the MLP algorithm for breast cancer prediction due to its ability to handle complex, nonlinear relationships between input features and output labels.

EXPERIMENT AND RESULTS

- Database:

  The dataset consists of 32 attributes of which 30 are float, 1 is int, and the other is object.

- Training and testing logs:

  Loaded the dataset which has been split into training and testing data, preprocessed the data, and evaluated the confusion matrix to assess the performance of different algorithms.

- Discussion and comparison
    i. Logistic Regression
    ii. Random Forest
    iii. Decision tree
    iv. Multi-Layer Perceptron

CONCLUSION

Breast cancer is a prevalent cancer in women worldwide and early detection is critical for successful treatment and better patient outcomes. In recent years, machine learning models have been developed to increase the efficiency of breast cancer detection. In this project, we implemented three supervised learning algorithms - Logistic Regression, Random Forest, and Decision Tree - to predict breast cancer.

In conclusion, our study demonstrated the potential of machine learning models in improving the detection of breast cancer. The developed Logistic Regression model can be used as a useful tool for physicians to help in the early identification and diagnosis of breast cancer. However, further studies are needed to validate the performance of the model on larger and more diverse datasets.

REFERENCES

- Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis.
- SYSTEMATIC ANALYSIS ON BREAST CANCER PREDICTION
  G Shanmugasundaram;S Balaji;R Saravanan;V Malarselvam;S Yazhini
  2018 IEEE International Conference on System, Computation, Automation.
- A Comparative Study on Breast Cancer Prediction using Optimized Algorithms
  S. Nathiya;J. Sumitha

2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)

- Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification

  Youness Khourdifi;Mohamed Bahaj

  2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)