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

Prediction of Breast Cancer Using Machine Learning

Mr K.V.Acharyulu[1], Varanasi Vineela[2], Torati Lakshmi Ratnam[3], Pulagam Divya Veni Navya Swetha[4], Reddy Ganesh[5], Peyyala Subash[6]

[1]Associate Professor, CSE Department, Vsm College Of Engineering, Ramachandrapuram, Dr.B.R.Ambhedkar Konaseema Dt, Andhra Pradesh - 533255
[2,3,4,5,6]Student, CSE Department, Vsm College Of Engineering, Ramachandrapuram, Dr.B.R.Ambhedkar Konaseema Dt, Andhra Pradesh - 533255

ABSTRACT
One of the main causes of death for women globally is breast cancer. In order to increase survival rates and make effective treatment possible, early detection and diagnosis are essential. Despite their effectiveness, traditional diagnostic techniques like biopsies and mammograms are labor-intensive and prone to human error. The development of machine learning (ML) techniques in recent years has created new opportunities in the medical field, particularly in the identification of cancer. These methods are an effective tool for early breast cancer detection because they can analyze vast amounts of medical data with great accuracy, speed, and consistency.
The goal of this project is to create a prediction model for the identification of breast cancer by utilizing machine learning methods that are implemented in Python. Various classification techniques, such as Support Vector Machines (SVM), Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN), are applied to a publically available dataset—typically the Wisconsin Breast Cancer Dataset. A number of cell nucleus features, including texture, radius, perimeter, and compactness, are included in this collection. These attributes are crucial in identifying whether a tumor is benign or malignant.

Keywords: Long Short-Term Memory Network (LSTM), Convolutional Neural Network, Deep Learning, Image Captioning, and Natural Language Processing.

INTRODUCTION

One of the most prevalent and deadly illnesses impacting women worldwide is breast cancer. Global health statistics indicate that it
accounts for a significant percentage of cancer-related deaths among women. Early diagnosis and identification are crucial for increasing survival rates because they enable prompt treatment and improved illness management. However, conventional diagnostic techniques like physical exams, biopsies, and mammograms are frequently expensive, time-consuming, and prone to misunderstandings or human mistake. Particularly in areas with little access to sophisticated medical facilities or skilled radiologists, these restrictions may cause delays in diagnosis.
Machine learning has become a game-changing solution in the field of medical diagnostics as a result of

the advancement of computing technology. It makes it possible to analyze big datasets, find hidden patterns, and generate precise forecasts. The goal of this project is to use Python machine learning techniques to create a model for the identification of breast cancer. The model classifies tumors as benign or malignant by analyzing information from medical datasets, including the Wisconsin Breast Cancer Dataset, using classification techniques like Support Vector Machines, Decision Trees, and Random Forests. With the help of robust Python libraries like NumPy for numerical operations, Pandas for data manipulation, and Scikit-learn for modeling, the system seeks to assist medical professionals by offering quick, accurate, and consistent diagnostic results, which will ultimately improve patient care and results.

EXISTING SYSTEM

The K-means clustering algorithm, a well-liked approach based on centroid-based clustering approaches, is used in the current model for consumer segmentation. This algorithm is operated by a choosing a suitable value for K, which stands for the number of pre-established clusters in the dataset. Using unlabeled and raw data as input, it divides the data into K different clusters based on feature similarity. In each iteration, it updates the centroids until the most effective cluster arrangement is reached. This model's efficiency comes from how easy it is to speed up, which makes it appropriate for big datasets. Of Its sensitivity to the centroids' initial placement, however, is a drawback that, if improperly initialized, may result in less than ideal outcomes.to Outliers can also have a big impact on the model, causing it to distort to the cluster centers and lose accuracy.

PROBLEM STATEMENT

One of the biggest causes of death for women globally is still breast cancer, and early detection is essential for successful treatment and higher survival rates. Despite their effectiveness, traditional diagnostic techniques like mammography, ultrasound, and biopsy are frequently costly, time-consuming, and prone to human error, particularly when the disease is still in its early stages. Generally rule-based, current computer-aided diagnostic methods are not flexible enough to generalize across a variety of patient data. An intelligent, precise, and automated system that can consistently differentiate between benign and malignant tumors and evaluate medical data is desperately needed. By utilizing Python to create a machine learning-based breast cancer detection model, this project seeks to address this difficulty and give medical practitioners quick, reliable, and accurate diagnostic assistance.

PROPOSED SYSTEM

By identifying tumors as benign or malignant, the suggested system seeks to create an intelligent, machine learning-based model in Python that can identify breast cancer. Publicly accessible medical datasets, such the Wisconsin Breast Cancer Dataset, which includes attributes extracted from digital photos of breast mass cell nuclei, are used by this approach. The project entails preprocessing the dataset to separate it into training and testing sets, handle missing values, and normalize features. To find the most accurate and dependable model, a variety of supervised machine learning methods will be trained and assessed, such as support vector machines (SVM), random forests (RF), decision trees (DT), logistic regression, and K-nearest neighbors (KNN). Python tools such as Scikit-learn, Pandas, NumPy, and Matplotlib will be used for managing data, training models, and visualizing the results. It is anticipated that the finished model would offer a highly accurate, automated, and effective diagnostic tool that helps healthcare providers identify breast cancer early and accurately. This could save more lives by lowering the time and effort

needed for manual diagnosis

IMPLEMENTED ALGORITHMS

Logistic Regression

A well-liked supervised machine learning technique, logistic regression is mainly applied to binary classification problems in which the output variable might have one of two potential values, such as "yes" or "no," "malignant" or "benign," "spam" or "not spam." Logistic regression predicts the likelihood that a given input point belongs to a specific class, as opposed to linear regression, which predicts continuous values. The sigmoid (logistic) function is applied to the linear combination of input features in order to accomplish this. The output is mapped by the sigmoid function to a range of 0 and 1, which indicates the probability that the instance belongs to the positive class. The model identifies it as one class if the likelihood is more than or equal to a threshold, usually 0.5; if not, it classifies it as the other.

Decision Tree

A popular supervised machine learning technique for classification and regression applications is the decision tree. By dividing the dataset into subsets according to the input feature values, it creates a structure resembling a tree, with each internal node denoting a feature-based decision, each branch denoting an outcome, and each leaf node denoting a final prediction.

Using criteria like Gini impurity or Information Gain (entropy), the algorithm chooses the appropriate features to split the data with the goal of producing the most homogenous branches. Decision trees are particularly helpful for decision-making in domains like healthcare since they are simple to comprehend, interpret, and visualize. Based on input criteria including cell size, shape, and texture, they may successfully categorize tumors as benign or malignant in the detection of breast cancer. They may, however, be prone to overfitting, which is frequently prevented by employing ensemble approaches like Random Forest or pruning procedures.

Random Forest

An ensemble-based machine learning technique called Random Forest (RF) aggregates the predictions of several Decision Trees (DTs) to get a more reliable and accurate outcome. A Random Forest is made up of numerous DTs, each of which contributes to the final prediction, much like a forest is made up of numerous trees. Individual decision trees frequently have the tendency to overfit the training data, particularly when they are developed too deeply. Because of the large variation caused by this overfitting, even slight modifications to the input data might provide wildly disparate predictions. These trees are unreliable for generalization because they become extremely sensitive to the training set and may perform poorly on unseen data. In order to get around this, Random Forests use a technique known as bootstrap sampling to train each decision tree on a distinct subset of the training data. To further enhance variation among trees, each tree only splits nodes based on a random selection of features. A fresh input sample is run through every tree in the forest before being categorized. The final conclusion is reached by either choosing the average prediction (for regression tasks) or by majority voting (for classification tasks), with each tree providing its own classification based on the input vector. Random Forest greatly lowers variance and boosts robustness by combining the output of several different trees, which makes it more accurate and dependable than utilizing a single decision tree, particularly in complex tasks like detection.

WORKING OF MACHINE LEARNING ALGORITHMS

Linear regression is used in finding the best-fit line (or hyperplane) that divides data points into two classes

is how logistic regression operates. The sigmoid function is used to translate the weighted sum of the input information into a probability between 0 and 1. It then categorizes the input into one or more classes based on a selected threshold, typically 0.5. For linearly separable data, this approach is straightforward, efficient, and gives information about the influence of each feature on the prediction**.**

Decision Tree uses metrics like Gini impurity or entropy to determine the optimal splits, decision trees divide the dataset into smaller subgroups according to the most important properties. Until the data is fully classified or a halting condition is satisfied, the procedure keeps going backwards.

By building a collection (or forest) of decision trees, each trained on a random subset of characteristics and data.

Random Forest goes beyond this. In contrast to a single decision tree, the final output is based on the average (for regression) or majority vote (for classification) of all the trees, increasing accuracy and decreasing overfitting.

SYSTEM ARCHITECTURE

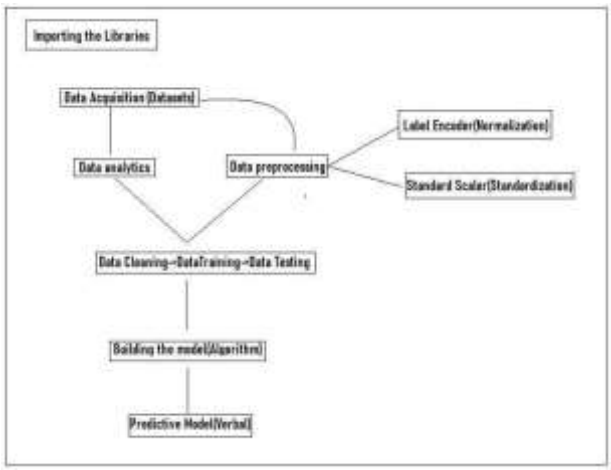


Fig:1 System Architecture Diagram

1.Importing the Libraries

To manage data, create models, and display results, necessary libraries like Pandas, NumPy, Scikit-learn, and Matplotlib are imported.These libraries include built-in features that make preprocessing, model training, and performance evaluation easier.

2. Data Acquisition (Datasets)

The dataset was gathered from trustworthy sources including Kaggle and the UCI Machine has characteristics that are important for detecting breast cancer, such as cell size, texture, and form.

3. Data Analytics

To comprehend the data distribution, spot trends, and find abnormalities, preliminary analysis is carried out. Finding associations between features is aided by visualization tools such as correlation heatmaps and histograms.

4. Data Preprocessing

To prepare raw data for model training, it is cleaned and converted.Methods such as feature scaling, categorical variable encoding, and handling missing values are used.

5. Label Encoder (Normalization)

Transforms text input with categories into numerical labels that are comprehensible to machine learning

algorithms.helpful for working with non-numeric features, including diagnosis labels (e.g., 'Benign', 'Malignant').

6. Standard Scaler (Standardization)

Eliminates the mean and scales to unit variance to standardize features.aids in enhancing the efficiency and rate of convergence of certain algorithms, such as SVM and logistic regression.

7. Data Cleaning → Data Training → Data Testing

To ensure an objective assessment, the data is divided into training and testing sets.Cleaning entails managing outliers, getting rid of duplicates, and making sure feature formats are consistent.

8. Building the Model (Algorithm)

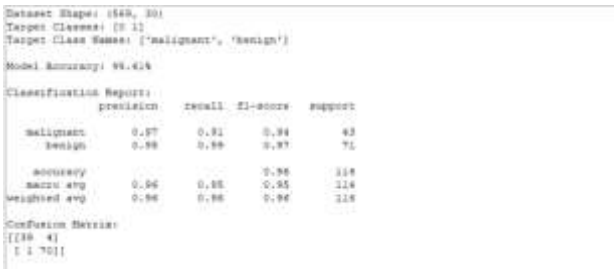
The training dataset is used to train machine learning models such as Random Forest, Decision Tree, and Logistic Regression.Accuracy, precision, recall, and other performance criteria are used to choose the optimal model.

9. Predictive Model (Verbal)

The finished model forecasts results for fresh, unobserved data (tumor classification, for example).A verbal or visual explanation of the forecast is frequently included in the results' interpretation and user-friendly presentation.

RESULT ANALYSIS

Fig:2 To run the prediction, press the button.



Dataset Shape: (569, 31)
Target Classes: [0 1]
Target Class Names: ['malignant' 'benign']

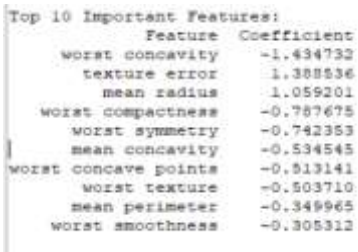
Model Accuracy: 99.41%

Classification Report:

	precision	recall	f1-score	support
malignant	0.97	0.91	0.94	43
benign	0.96	0.99	0.97	71
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

Confusion Matrix:
[[39 4]
[1 70]]

Fig:2 Classification Report



Top 10 Important Features:

Feature	Coefficient
worst concavity	-1.434732
texture error	1.388536
mean radius	1.059201
worst compactness	-0.787675
worst symmetry	-0.742353
mean concavity	-0.534545
worst concave points	-0.513141
worst texture	-0.503710
mean perimeter	-0.349965
worst smoothness	-0.305312

Fig:3 Important Features

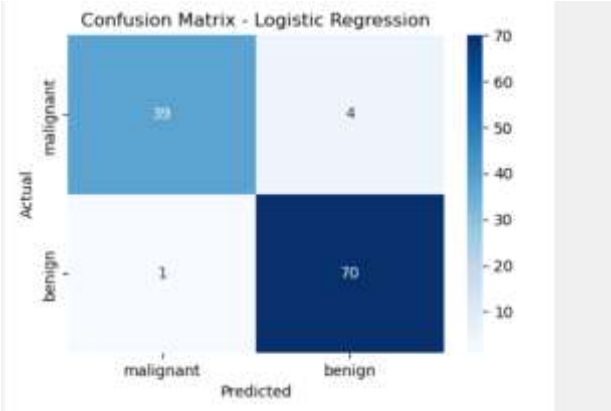


Fig:4 Confusion Matrix

[14]:

	Feature	Coefficient
26	worst concavity	-1.434732
11	texture error	1.388536
0	mean radius	1.059201
25	worst compactness	-0.787675
28	worst symmetry	-0.742353
6	mean concavity	-0.534545
27	worst concave points	-0.513141
21	worst texture	-0.503710
2	mean perimeter	-0.349965
24	worst smoothness	-0.305312

Fig:5 Data Set

CONCLUSION

To sum up, this experiment shows how machine learning can be applied to the identification of breast cancer. The system is capable of reliably classifying tumors as either benign or malignant by utilizing algorithms such as Random Forest, Decision Tree, and Logistic Regression. The system offers a dependable tool that can help medical practitioners with early diagnosis, which can improve treatment outcomes, provided that the data is properly preprocessed and the model is evaluated. All things considered, this experiment demonstrates how machine learning may enhance medical diagnosis and save lives.

The Random Forest classifier, which handles non-linear interactions well and minimizes overfitting because it is an ensemble model, often had the highest accuracy and robustness among the models studied. Although the Decision Tree model was simple to understand, it had a propensity to overfit, particularly when used to smaller datasets. Despite being straightforward and easy to understand, logistic regression proved less accurate than random forest in this situation even if it worked well with linear correlations.

With Random Forest emerging as the most successful algorithm among the three, the study shows that

machine learning models can be helpful tools in supporting early identification of breast cancer. These models may help medical professionals make prompt and accurate diagnoses with additional optimization and integration of real-world data.

REFERENCES

1. W3 School – (https://www.w3schools.com/python/)
2. Geek for Geeks – (https://www.geeksforgeeks.org/python-programming-language/learn- python-tutorial/
3. Scikit-learn: Machine learning in Python. (https://scikit-learn.org/stable/)
4. Breast Cancer Wisconsin (Diagnostic) Data Set. (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)