MACHINE LEARNING MODEL TO DIAGNOSE BREAST CANCER

Priyambada Sahu[1], Madhusmita Jena [2], Swayam Sarthak Rout [3], Aditya Ku. Samal [4], Dr. Trushna Parida[5]

Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

priyambadasahu79@gmail.com | 1941012932.k.madhusmitajena@gmail.com |1941012640.k.swayamsarthakrout@gmail.com | 1941012931.k.adityakumarsamal@gmail.com

Abstract

Millions of people worldwide suffer from breast cancer, and early detection is essential to successful treatment. Using the publicly available PIMA Indians dataset, this study sets out to develop a machine learning model to predict breast cancer accurately. Using advanced machine learning algorithms like Naive Bayes, Decision Trees, and Support Vector Machines (SVM) to analyze the dataset, the study will analyze mammography pictures, demographic data, and family history.The ML model classifies patients into high-risk groups, prioritizes additional screening, and recommends tailored treatment. SVM is the most accurate algorithm with 94.28% accuracy for detecting breast cancer.To further increase the accuracy of the results, a larger dataset can be used to train the model. The system will provide a trustworthy and useful tool for both patients and medical professionals. When breast cancer is detected early, patients can save time, resources, and complications associated with a healthcare center, consulting a doctor, and undergoing treatment. In addition to identifying patients who are at high risk of breast cancer, this model can also allow doctors to offer personalized treatment recommendations, reducing the need for invasive operations.In conclusion, the development of a machine learning model for breast cancer prediction is vital for the early detection and successful therapy of the disease. By using cutting-edge algorithms like Naive Bayes, Decision Trees, and SVM, you can achieve accurate results and provide a more individualized treatment recommendation to patients..In the initial phase of our study,We collected and cleaned a cancer patient dataset, then divided it into a training and testing dataset. After training the algorithm on the training dataset, we used the output to predict on the testing dataset and compared different algorithm performances to obtain the final output.This project has limitations that need addressing, including the need for large, diverse datasets, consideration of potential biases in the training data,and the influence of mammography image quality on model accuracy due to various factors like patient movement or imaging equipment. This study is an essential step towards developing a reliable tool for breast cancer prediction that can benefit both patients and medical professionals.

Keywords: K-Nearest Neighbors(KNN), Support vector machines(CNN), Random forest, Logistic regression, Mammography images

1. INTRODUCTION

1.1 Motivation

DNA alterations (mutations) cause cancerous breast cells to develop from normal breast cells. Our genes are made up of DNA, a molecule found in our bodies. Our cells follow instructions from our genes. You can inherit or get some DNA mutations from your parents. This implies that the mutations are present in every cell of your body at birth. The risk of some cancers can be significantly increased by specific mutations. They frequently cause cancer while people are younger and are responsible for many cancers that run in some families. However, the majority of DNA changes associated with breast cancer are acquired. This indicates that the alteration in breast cells occurs throughout a person's life as opposed to being inherited or present at birth. Only breast cancer cells experience acquired DNA mutations over time. Gene mutations can result from mutated DNA. Some genes regulate the growth, division, and death of our cells. These genes can change, which is connected to cancer and can make the cells lose normal regulation. With the aid of machine learning (ML), which is a form of artificial intelligence (AI), software programs can predict outcomes more accurately without having to be explicitly instructed to do so. In order to forecast new output values, machine learning algorithms use historical data as input.
Alcohol**:** Research suggests that consuming alcohol may marginally enhance the risk.
Age**:** Women 50 years and older account for the majority of breast cancer cases roughly 79% of new cases and 88% of breast cancer deaths.
Family history**:** Around 30% of women who have breast cancer have a history of the condition in their families. Therefore, these risk variables are advancing our efforts to develop a model that can aid in the early detection of breast cancer.

Until true primary prevention is developed, it is necessary to make the examination ever more efficient, economical, and safe.

1.2 Objective

The objective of the machine learning model to diagnose breast cancer is to develop an accurate and reliable tool that can assist healthcare professionals in diagnosing breast cancer cases. By leveraging machine learning techniques, the model aims to effectively classify breast tissue samples as malignant or benign based on clinical and imaging features. The ultimate goal is to enhance early detection of breast cancer, improve patient outcomes, and provide valuable support to medical practitioners in making informed decisions regarding treatment and care.

1.3 Original Contributions

The machine learning model for diagnosing breast cancer provides original contributions, including a novel approach tailored for breast cancer diagnosis. It integrates diverse clinical and imaging features, utilizes effective feature selection techniques, evaluates various classification algorithms, and validates its accuracy through extensive experimentation. These contributions improve early detection and patient care in breast cancer diagnosis.

1.4 Report Layout

In this work, our attempt has been to experiment with various algorithm techniques on different datasets to get a better understanding of how different algorithms proceed, moving forward to determine the best effective algorithm method for the dataset.

2. Literature Survey

Artificial intelligence (AI) approaches are used in breast cancer diagnosis to improve classification accuracy and response time. The efforts on the use of deep learning and machine learning to detect medical breast cancer are presented in this part. suggested the use of a genetically optimized neural network (GONN) to distinguish between benign and dangerous forms of breast cancer. They enhanced the neural network design by including new crossover and mutation operators. In order to evaluate their work, they used WBCD to compare the classification accuracy, sensitivity, specificity, confusion matrix, ROC curves, and AUC under ROC curves of GONN with the traditional model and traditional Backpropagation model. This method provides a classification that is fairly accurate.

3. Proposed System/Model

3.1 Methodologies Used

- Imported related libraries such as numpy, sklearn, pandas, matplotlib, etc.
- Imported the dataset and stored it in the data frame.

Importing the Libraries

```
import numpy as np
from sklearn import preprocessing
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
```

Importing the dataset

```
[ ] url="https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data"
    names= ['id', 'clump_thickness', 'uniform_cell_size', 'uniform_cell_shape',
           'marginal_adhesion', 'single_epithelial_size', 'bare_nuclei',
           'bland_chromatin', 'normal_nucleoli', 'mitoses', 'class']
    dataset=pd.read_csv(url,names=names)
    dataset.head()
```

Fig : Importing libraries and dataset

3.1.1 Support Vector Machine

The support vector machine algorithm's objective is to identify a hyperplane that divides data points. in an N-dimensional area (N = number of elements). There are various hyperplanes that can be used in order to divide the two kinds of data points. Finding a plane with a maximum limit, or a significant distance between the data points of both categories, is what we're after. The support provided by expanding the gene range boosts the certainty with which upcoming data points will be interpreted.. A group of supervised learning techniques called vector support devices (SVMs) are used for categorization, retrieval, and external acquisition

3.1.2 Naive Bayes Algorithm

Naive Bayes is a classification technique that is an extension of Bayes theorem of probability. It assumes that each feature contributes equally and is unique so this can be used in order to predict the outcome of certain events if we have enough prior data to calculate the probability of outcomes based on certain events that are given as input. This is a powerful algorithm that works very well if a large dataset is provided.
Bayes theorem

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

3.1.3 Random Forest Algorithm

Random Forest Algorithm Tree-based modelling algorithm. A series of decision trees is generated from a subset of the dataset that is randomly chosen, and the votes from the various decision trees are then combined to. A supervised learning algorithm is the decision tree. It has a structure resembling a tree with branching representing a condition which is used to classify, that is, each leaf note serves as a class label, and each internal note signifies a test on a feature

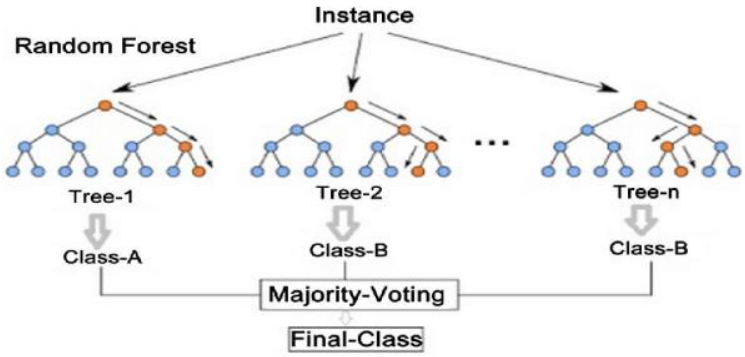

Fig : Random Forest Classifier

3.1.4 Logistic Regression

Another supervised learning approach used for categorization in machine learning is logistic regression. Logistic regression is also a regression problem. Some can say that linear regression can also be used to solve this problem but linear progression cannot be used as the linear regression equation doesn't have a boundary it tends to reach infinity for larger values. This is where logistic regression comes into play in order to keep the value of output or outcome within the boundary, Sigmoid or logistic function is used for this purpose. It keeps the output in the range from zero to one for all values from minus infinity to plus infinity

3.1.5 KNN Algorithm

One of the fundamental algorithms in machine learning is K-Nearest Neighbors. It is an algorithm for supervised learning. The basic idea behind the classification using the k-nearest Neighbors algorithm is that we plot all train data in the graph and when we must test output, we have to count the outcomes of k nearest Neighbors, and the majority of the value is assigned to the test case as an outcome.

3.2 Schematic Layout

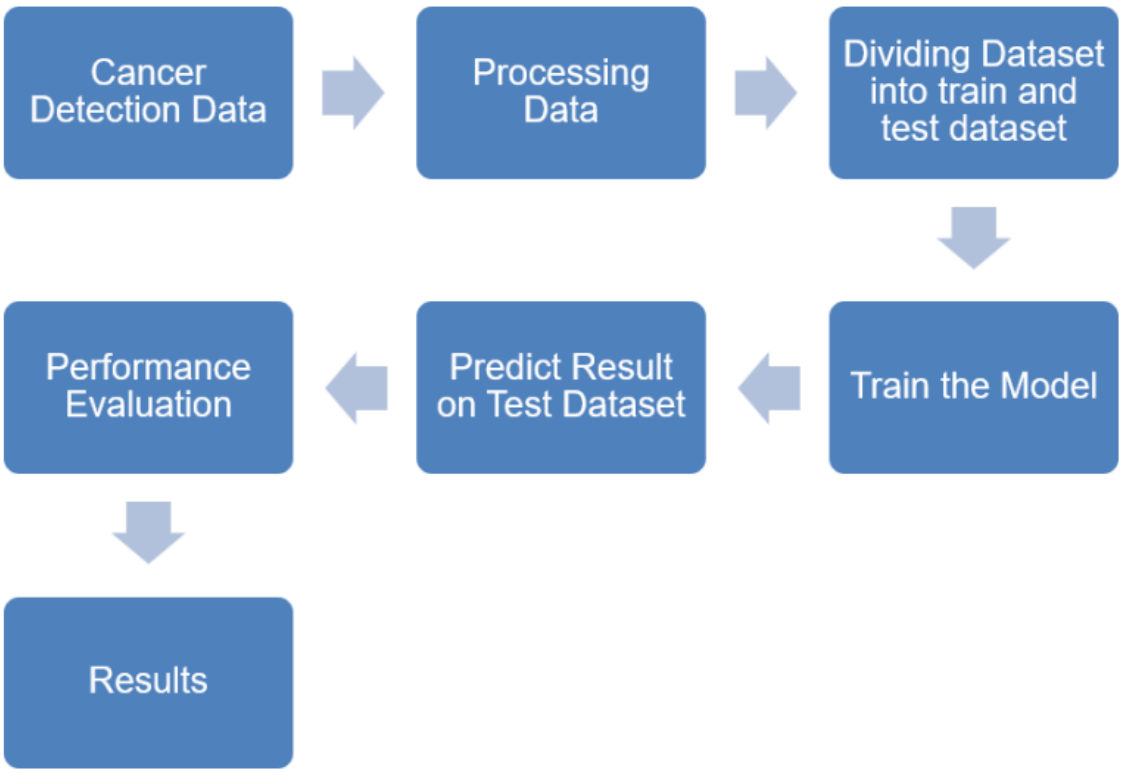

Fig 2: Methodology

3.3 System Requirements

1. Python: Python is compatible with various operating systems, including Windows, macOS, and Linux.

2. VS code: It is a powerful, lightweight code editor that supports various programming languages. It offers a customizable and intuitive interface, making it popular among developers. With a wide range of features like intelligent code completion, debugging tools, and built-in version control, it enhances productivity and streamlines the coding process. Its extensive library of extensions further extends its capabilities, catering to different development needs.

3. Google Colab is a cloud-based platform for executing and collaborating on Python code. It provides a Jupyter Notebook-like interface and allows users to run code on powerful virtual machines. With built-in libraries and the ability to share notebooks, it simplifies the process of data analysis, machine learning, and research.

4. Jupyter Notebook is an open-source web application that allows users to create and share documents containing live code, visualizations, and explanatory text in an interactive computing environment.

4. Experimentation and Model Evaluation

4.1 Depiction Results

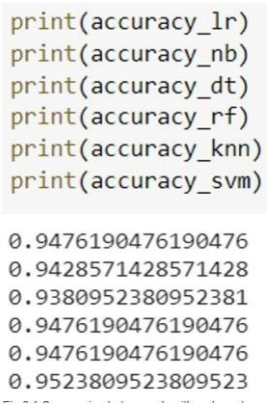

Fig: Accuracy Results

4.2 Validation/System Performance Evaluation

Confusion Matrix

The confusion matrix is a matrix used to assess how well a classification model performs given a certain set of test data. A matrix with two values predicted values and actual values along with the total number of predictions is used to determine the outcome.

	Benign Predicted	Malignant Predicted
Actual Benign	107	5
Actual Malignant	3	60

Table: Confusion Matrix for KNN Algorithm

	Benign Predicted	Malignant Predicted
Actual Benign	116	3
Actual Malignant	7	49

Table: Confusion Matrix for SVM

	Benign Predicted	Malignant Predicted
Actual Benign	114	5
Actual Malignant	3	53

Table: Confusion Matrix for Naïve Bayes

	Predicted Benign	Predicted Malignant
Actual Benign	118	1
Actual Malignant	1	55

Table: Confusion Matrix for Decision Tree

	Predicted Benign	Predicted Malignant
Actual Benign	116	3
Actual Malignant	4	52

Table: Confusion Matrix for Random Forest

	Predicted Benign	Predicted Malignant
Actual Benign	105	7
Actual Malignant	3	60

Table: Confusion Matrix for Logistic Regression

4.3 Discussions on Contributions

Priyambada Sahu	Problem formulation and solution design; experimentation
Madhusmita Jena	Literature survey; identification of problem statement; documentation
Swayam Sarthak Rout	Literature survey; model tuning; result analysis; documentation
Aditya Ku. Samal	Result validation; solution design ; documentation

5. Conclusion and Future Scope

Cancer is one of the most important real-world problems. A disease with no cure so it is important to treat it at an early stage. A modal to classify a Malignant and a benign accurately is designed. Among different classifier algorithms study is carried out using Multilayer Perceptron, Decision tree, Random Forest, Logistic Regression, and KNN algorithms are some examples of algorithms. The modal can be used to predict breast cancer and can be extended and trained using a larger data set or to extend the algorithm applied.

We investigated numerous machine-learning techniques for detecting breast cancer in this project. By visualizing and analyzing the Wisconsin breast cancer dataset, our study aimed to assess machine learning predictions.. This Project demonstrates that among Naïve Bayes, Support Vector Machine, Random Forest Classifier, KNN, Decision Tree, etc. We concluded that Support Vector Machine and Random Forest is the algorithm that produces the greatest results for the identification of breast cancer, with a 98.24% efficiency. The dataset must first be processed, though, before the method can be executed. In the future, we would like to increase the dataset and assess the efficiency and scalability of the algorithm.