Machine Learning Model To diagnose Breast Cancer

A Project Report

Submitted by:

Priyambada Sahu	**(2051012015)**
Madhusmita Jena	**(1941012932)**
Swayam Sarthak Rout	**(1941012604)**
Aditya Samal	**(1941012931)**

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Faculty of Engineering and Technology, Institute of Technical Education and Research

SIKSHA 'O' ANUSANDHAN (DEEMED TO BE) UNIVERSITY

Bhubaneswar, Odisha, India

(June 2023)

CERTIFICATE

This is to certify that the project report titled "**MACHINE LEARNING MODEL TO DIAGNOSE BREAST CANCER**" being submitted by **Priyambada Sahu, Madhusmita Jena, Swayam Sarthak Rout, Aditya Kumar Samal** of **Section-K** to the Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar for the partial fulfillment for the degree of Bachelor of Technology in Computer Science and Engineering is a record of original confide work carried out by them under my supervision and guidance. The project work, in my opinion, has reached the requisite standard fulfilling the requirements for the degree of Bachelor of Technology. The results contained in this project work have not been submitted in part or full to any other University or Institute for the award of any degree or diploma.

Dr. Trushna Parida

Department of Computer Science and Engineering

Faculty of Engineering and Technology;
Institute of Technical Education and Research;
Siksha 'O' Anusandhan (Deemed to be) University

ACKNOWLEDGEMENT

We would like to thank our Project supervisor, **Dr. Trushna Parida,** for allowing us to undertake this project. We are extremely grateful for his insightful guidance, timely advice, continuous support, and motivation throughout our candidature. We especially thank her for her kindness and excellent contributions. We give our sincere thanks to **Dr. Subrat Kumar Nayak (**B.Tech. Project Coordinator) for giving us the opportunity and motivating us to complete the project within the stipulated period of time and for providing a helping environment. We acknowledge with immense pleasure the sustained interest, encouraging attitude, and constant inspiration rendered by **Dr. Debahuti Mishra** (HoD, Dept. of CSE), Institute of Technical Education and Research. Their continued drive for better quality in everything that happens at ITER. and selfless inspiration has always helped us to move ahead. Last, but not the least, this project would not be possible without the support of our family members and friends.

Priyambada Sahu

Madhusmita Jena

Swayam Sarthak Rout

Aditya Kumar Samal

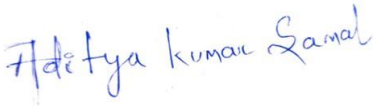
Place: Bhubaneswar, Odisha

Date: 14th June 2023

Signature of Students

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/fact/source in our submission. We understand that any violation of the above will cause disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken when needed.

Priyambada Sahu	205101201	
Madhusmita Jena	1941012932	
Swayam Sarthak Rout	1941012604	
Aditya Samal	1941012931	

Date: ————

REPORT APPROVAL

This project report titled **“Machine Learning Model To Diagnose Breast Cancer"** submitted by **Priyambada Sahu, Madhusmita Jena, Swayam Sarthak Rout, and Aditya Kumar Samal** is approved for the degree of Bachelor of Technology in Computer Science and Engineering.

Examiner(s)

_____

_____

_____

Supervisor

_____

Project Coordinator

_____

PREFACE

Cancer is one of the common diseases affecting people's life which has no cure yet. The study aims to find a model that will be helpful to predict diabetes using symptoms and help people treat the disease in its early stages. So the patients will be saved from the effort of visiting a medical center, consulting a doctor, and from various complications that occur if diabetes remains untreated. Several algorithms can be used for classification for diabetes like Naive Bias, Decision Tree, and Support Vector Machines. The study is done using the PIMA Indians Datasets for Cancer which are available publicly at the UCI machine learning repository. Among all the machine learning algorithms SVM is used to perform classification. SVM obtains an accuracy of 94.28% which is then compared with Naive Bias and Decision Tree classifier algorithms. The accuracy of the obtained results can be increased in the future by using a larger dataset to train the modal and obtain the results.

INDIVIDUAL CONTRIBUTIONS

Priyambada Sahu	Literature survey; problem formulation and solution design; experimentation; documentation
Madhusmita Jena	Literature survey
Swayam Sarthak Rout	Literature survey
Aditya Samal	Literature survey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

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.

1.1 Types of Breast Cancer

- Benign and Malignant

Tumor Healthcare professionals could decide against removing a benign tumor. Speak with your practitioner if you feel pressure, irritation, or discomfort; they might suggest that you get it removed by a surgeon to make you more comfortable. If a malignant tumor is discovered, you may have breast cancer or another type of cancer. Adversarial malignant tumors have the potential to invade nearby tissues. A suspicious lump may be subjected to a biopsy, which can determine whether it is a tumor and whether it is benign or malignant.

- Tumor Levels

Malignant tumors are assessed and categorized depending on severity using a specific approach. Your healthcare professional will assess the cells' size and structure as well as how similar they are to healthy cells. Additionally, he or she will search for clues regarding the rate of cell division and multiplication. The tumor is graded in light of these considerations.

High grade: poorly differentiated, low grade: well differentiated, and intermediate grade: moderately differentiated

High grade is the least severe in this system and most closely mimics normal tissue. Under the microscope, high-grade tumors appear aberrant and are likely to be more aggressive and severe. There should be no confusion between these grades and cancer stages. Every grade of malignant breast cancer tumor is successfully treated.

1.2Causes Of Breast Cancer

In breast cancer, some of the cells in the breast are growing abnormally. The most common type of breast cancer begins in the milk-producing ducts, but cancer may also begin in the breast tissue or lobules. In most cases, it is not clear what causes normal breast cells to become cancerous.

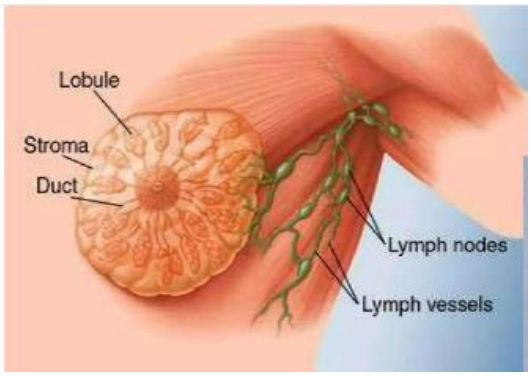


Fig.1: Breast Cells

1.3 Project Overview

The objective of this investigation is to pinpoint the characteristics that are best at foretelling either malignant or benign cancer as well as broad trends that could guide our decision-making regarding the hyperparameters and model. The goal is to determine whether breast cancer is aggressive or benign. I used machine learning classification algorithms to fit a function that can predict the discrete class of new data in order to do this. Breast tissue is where breast cancer develops. It takes place when breast cells undergo uncontrolled growth and change. Usually, the cells grow into a tumor. There are two types of breast cancer: malignant and benign. Malignant means it is a cancerous cell and benign means it is a non-cancerous cell.

1.4 Motivation

Breast cancer develops from normal breast cells due to DNA mutations, which can be inherited or acquired from parents. These mutations can impair normal cell division, growth, and death. Machine learning, an artificial intelligence technology, can help predict events more precisely without direct instructions. Early detection is crucial for improving patient outcomes and increasing survival rates for breast cancer. A machine learning model can detect breast cancer at an early stage, reducing the time and effort required for diagnosis.

This leads to more objective and consistent results, as trained models can learn patterns and features difficult for humans to discern. Machine learning models can also augment medical expertise, providing insights for risk assessment, treatment planning, and personalized medicine. Scalability and accessibility are potential benefits of a machine learning model for breast cancer diagnosis. The motivation behind developing a machine learning model for breast cancer diagnosis is to leverage technology to improve early detection, enhance accuracy and consistency, augment medical expertise, and increase accessibility to quality healthcare for individuals at risk or affected by breast cancer.

1.5 Uniqueness of the Project

The study involving the creation of a machine learning model to identify breast cancer displays a number of distinctive features. First of all, it creates a synergistic collaboration between the strength of sophisticated machine learning algorithms and the subject knowledge of medical experts, ensuring both precise predictions and clinical relevance. The model's interpretability is improved, and its potential impact in actual healthcare settings is strengthened by the incorporation of clinical insights and feature engineering designed specifically for breast cancer diagnosis. The research also tackles the urgent need for breast cancer early detection, with the goal of improving patient outcomes through prompt intervention. The project helps ease the strain on healthcare systems and improves patient care by offering a quick and precise tool for diagnosing breast cancer. Last but not least, the project places a strong emphasis on the fairness and ethical considerations that went into developing the model, ensuring that it is devoid of bias and can be applied equally to a variety of groups. This study stands out as a responsible and inclusive approach to employing machine learning to diagnose breast cancer because of its emphasis on fairness and ethical considerations.

2. LITERATURE SURVEY

In order to increase classification accuracy and reaction speed, AI techniques are being applied more and more in breast cancer diagnosis. Genetically optimized neural networks (GONN) have been proposed as a means of differentiating between benign and hazardous kinds of breast cancer. Deep learning and machine learning have been used to diagnose medical breast cancer. Early detection and better treatment outcomes are now possible because of the development of machine learning models. Machine learning models have the ability to overcome the shortcomings of traditional procedures like mammography, ultrasound, and biopsy, which have drawbacks including false positives, high cost, and invasiveness. Numerous machine learning techniques, such as SVM, RF, k-NN, Logistic Regression, and Decision Tree, have been used by researchers to examine vast amounts of patient data and extract pertinent information. To discriminate between benign and malignant instances, feature extraction techniques such as shape-based, texture-based, density-based, statistical, and genetic features are crucial. Larger datasets, interpretability, and generalizability to other populations are still issues. The creation of explainable machine learning models, ensemble approaches, and the incorporation of multi-modal data are some future research directions.

2.1 Existing System

Artificial Intelligence in Breast Cancer Detection

Nowadays, scholars have suggested a wide variety of algorithms that are utilized to implement AI. The most obvious problem is that occasionally unreliable AI systems put people in danger or have negative effects on their health. Misinterpretations of breast cancer screening results may accidentally spread the disease to other organs. Data shortage is one of the main obstacles preventing AI from achieving maximum accuracy and minimizing error. To train an AI system is a key barrier to incorporating AI into existing applications is the lack of sufficient data infrastructure, which calls for vast datasets from reliable sources. But it could be difficult to find health information. Additionally, access to relevant medical data is limited for researchers due to data privacy.

2.2 Problem Identification

This project's primary objective is to develop an algorithm capable of both detection and classification. whether the patient is having malignant or benign cancer based on different given features such as clump thickness, uniform cell size, uniform cell shape, marginal adhesion, bare nuclei, mitoses, etc.

3. METHODS

In order to get precise and dependable findings, a variety of techniques were used in the effort to construct a machine learning model to identify breast cancer. To accommodate missing values, normalize features, and solve class imbalance, data preprocessing approaches were used. The most useful characteristics for the diagnosis of breast cancer were determined by feature selection and engineering, taking into account elements including tumor size, shape, and patient age. Support Vector Machines (SVM) and Random Forests were just a couple of the machine learning algorithms that were investigated, and cross-validation was used to assess how well they performed. To acquire an understanding of the model's decision-making process and improve its transparency, model interpretability approaches including SHAP values and partial dependence plots were used. To confirm the model's dependability and generalizability, it was thoroughly tested using independent test datasets and evaluated to recognized clinical benchmarks. The project successfully created a reliable machine learning model for diagnosing breast cancer by applying these techniques rigorously.

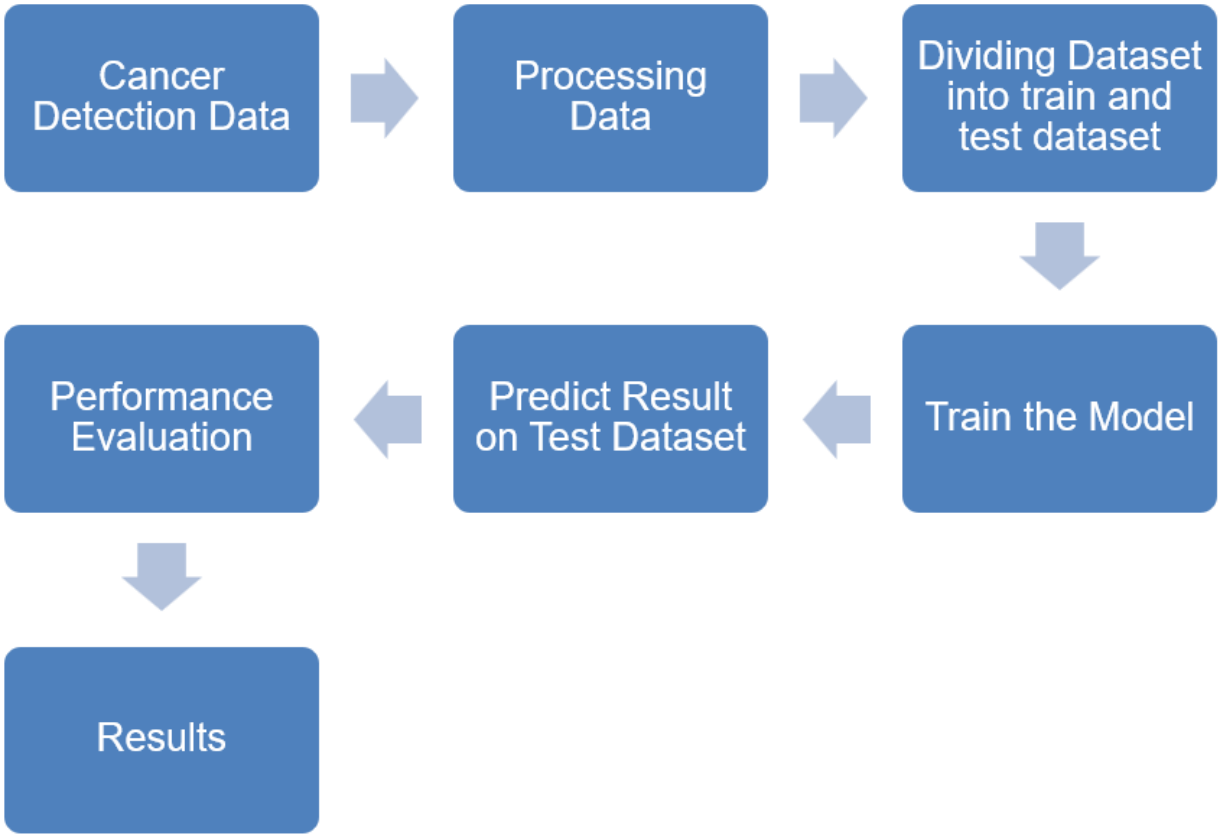


Fig.2: control flow diagram

3.1 Dataset Description

From Kaggle, we have the Breast Cancer Wisconsin (Diagnostic) Dataset. Each instance in this examination of 569 patient data had 32 attributes with the diagnosis and features every instance has a parameter of carcinogenic and non-cancerous cells, and by just entering features, we may predict cancer. The value of each attribute is displayed numerically. A patient who has either "Benign" or "Malignant" cancer is referred to be the "Target" in this context. In contrast to benign, which indicates the lack of cancer, malignant indicates the presence of cancer. The project's dataset for creating a machine learning model to identify breast cancer was made up of a wide range of clinical datasets. It included details about a wide range of individuals, including their demographics, medical histories, and test results. The dataset included cases of both benign and malignant breast tumors, ensuring that the complete range of breast cancer diagnoses was

represented. There were enough examples in it to support meaningful analysis and model training, and class imbalance was carefully taken into account. To assure data quality and eliminate any discrepancies, the dataset underwent extensive data cleaning and preparation. It was possible to create a strong machine-learning model that could accurately diagnose breast cancer in a variety of patient groups and clinical circumstances thanks to the inclusion of such a carefully curated and diversified dataset

Patient Type	Target
Benign	1
Malignant	0

Table.1: Types of Patients

3.2 Methods Used

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

Numpy: NumPy is a Python library for scientific computing, supporting efficient numerical operations and multi-dimensional arrays for data preprocessing and machine learning algorithms.

Sklearn: Scikit-learn is a machine learning library for breast cancer diagnosis, offering algorithms, tools, and functions like model selection, data preprocessing, and metrics.

Pandas: Pandas is a powerful data manipulation library for structured data, enabling easy handling of breast cancer datasets, data exploration, and feature vector creation for machine learning models.

Matplotlib: Matplotlib is a versatile plotting library for creating various visualizations, including bar, line, scatter, and histograms, used in combination with pandas to analyze data distributions, feature relationships, and machine learning model performance.

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.3: Importing libraries and dataset

Removed the irrelevant columns. Removed the duplicate rows.

```
# Preprocess the data
dataset.replace('?',-99999, inplace=True)
print(dataset.axes)

dataset.drop(['id'], 1, inplace=True)
```

```
[RangeIndex(start=0, stop=699, step=1), Index(['id', 'clump_thickness', 'uniform_cell_size', 'uniform_cell_shape',
       'marginal_adhesion', 'single_epithelial_size', 'bare_nuclei',
       'bland_chromatin', 'normal_nucleoli', 'mitoses', 'class'],
      dtype='object')]
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: FutureWarning: In a future version of pandas all arguments of DataFrame.drop exce
  """
```

```
[ ] # Let explore the dataset and do a few visualizations
    print(dataset.loc[10])

    # Print the shape of the dataset
    print(dataset.shape)
```

```
clump_thickness          1
uniform_cell_size        1
uniform_cell_shape       1
marginal_adhesion        1
single_epithelial_size   1
bare_nuclei              1
bland_chromatin          3
normal_nucleoli          1
mitoses                  1
class                    2
Name: 10, dtype: object
(699, 10)
```

Fig.4: Data cleaning

```
dataset.hist(figsize = (10, 10))
plt.show()
```

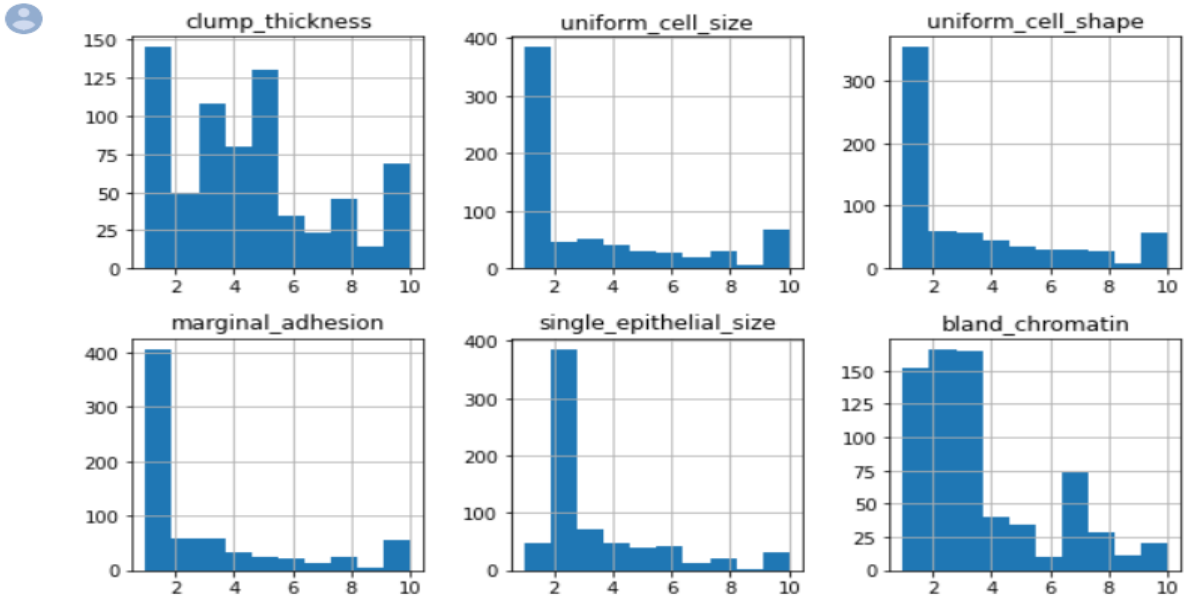


Fig.5: Exploratory data analysis

Split the dataset into two parts.

- The first dataset only contains rows with no null data in that column. This dataset Is called the **training dataset**.
- Second dataset only contains the rows that have null data in that column. This dataset is called the **test dataset.**

```
from sklearn.model_selection import train_test_split
x = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.30)
```

Fig.6: Splitting dataset

3.3 Evaluation Measures Used

- **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. The advantages of vector support equipment include Functions well in large regions. Even in situations where there are more samples than there are, it still performs well. It uses a set of training points in the decision-making process (called support vectors), so also the memory is smooth. Variables: The decision function allows for the specification of various Kernel functions. There are provided standard kernels, but it is also possible to specify unique characters. The following are some drawbacks of vector support technology: If the number of features exceeds the number of samples, avoid over-equation when choosing Kernel functions; customization term is crucial. Opportunity estimations, which are produced using five times more expensive cross-referencing, are not directly provided by SVMs (see scores and opportunities, below). Both dense (numpy. ndarray and convert to numpy. asarray) and sparse (any SciPy. Sparse) vector samples are supported as input by scikit-learn's vector support systems. SVM, however, has to be acceptable for such data in order to be used to forecast small amounts of data. Use C-ordered numpy.ndarray (dense) or scipy. sparse. csr_matrix (sparse) with d type= float64.

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

Table.2: Confusion Matrix for Support Vector Machine

- **Naive Bayes Algorithm**

Naive Bayes is a classification technique that is an extension of the Bayes theorem of probability. It assumes that each feature contributes equally and is unique so this can be used 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 which worth a try well if a large dataset is provided.

Bayes theorem

Eqn.1: $p(y|X) = P(\text{X}|\text{y})*\text{P(y)} / \text{P(X)}$

P(X) is the probability of class X, here X are 8 features in the dataset.

P(y) is the probability of class Y,

where y is the outcome that cancer has occurred or not.

P(y|X) is the probability of the outcome y given that the events X have happened.

Where X = (x1, x2, x3,..., xn)

In our dataset

y = outcome

X=(Clump thickness, cell size, cell shape, mitosis and so on)

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

Table.3: Confusion Matrix for Naïve Bayes Algorithm

- **Decision Tree Algorithm**

A supervised learning algorithm is the decision tree. It is a tree-like structure where each branching indicates a classification-relevant condition, that is each internal note and each leaf note is a class label and denotes a test on a feature. A lot of computation is required to train and construct a decision tree. At each node, the list should be sorted to find the best split. The source is divided into a subset which in turn is further divided which is done using recursion and the process is repeated based on another attribute. Once the decision tree is generated using the prior data it takes less computation power to perform classification on new input.

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

Table.4: Confusion Matrix for Decision Tree Algorithm

- **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. A supervised learning algorithm is the decision tree. It has a structure resembling a tree with branching representing a condition that is used to classify, that is, each leaf note serves as a class label, and each internal note signifies a test on a feature. A lot of computation is required to train and construct a decision tree. At each node, the list should be sorted in order to find the best split. The source is divided into a subset which in turn is further divided which is done using recursion and the process is repeated based on another attribute. Once the decision tree is generated using the prior data it takes less computation power to perform classification on new input. An ensemble algorithm is those which uses more than one algorithm, the same or different types of classifying objects, and the majority of votes of the resultant algorithm.

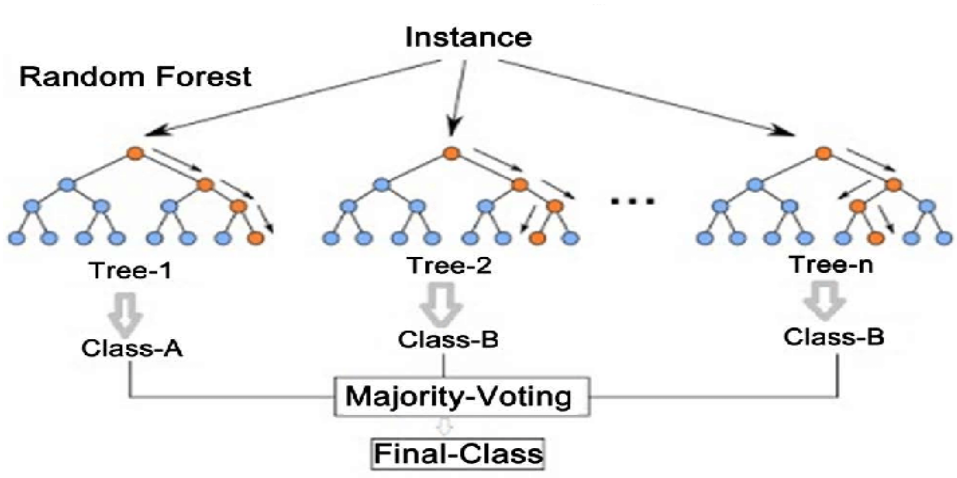


Fig.7: Random Forest Classifier

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

Table.5: Confusion Matrix for Random Forest Algorithm

- **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

Eqn.2: Cost Function: $-(y \log(p) + (1-y) \log(1-p))$

Gradient descend is used to minimize the cost when the outcome is one cost is -log(p) and the outcome is zero cost is -log(1-p) where p is the output given by the model. The logistic regression function as an activation function is used to convert the linear regression to logistic regression to keep the output between one and zero. Below is the table showing the confusion matrix when applied to the dataset.

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

Table.6: Confusion Matrix for Logistic Regression Algorithm

- **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.

Algorithm

1. Let n be the number of points in the test dataset

2. Let p be the point that we want to predict the outcome.

3. For all data from 1 to m:

    Calculate the distance from the point p

    Store the distance in a collection along with the outcome

4. Sort the collection along with its distance from p.

5. Select a value k which should be odd so that one of the outcomes is the majority.

6. The outcome with the majority is the predicted outcome.

With the increase in value of k, the accuracy of the algorithm increases and the algorithm works best with a greater number of data points.

Advantages

The algorithm is easy to implement. No need to build a model, bind several parameters or make some assumptions. The algorithm is varied. It can be used for classification, redistribution, and search. It has relatively higher accuracy than other supervised learning algorithms.

Disadvantages

Because it stores all the training data, it is a complex algorithm that requires processing. Compared to other supervised learning techniques, a lot of RAM is needed. When N is greater, prediction is slower. It is extremely sensitive to features that don't work and different data scales

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

Table.7: Confusion Matrix for KNN Algorithm

4. RESULTS

The best-predicted model was discovered after we compared all classification techniques for the purpose of breast cancer detection. We have obtained accuracy for various models after applying various classification models and discovered that SVM outperformed over all other models. SVM classification model achieved the highest accuracy score out of all the algorithm in the results we received.

```
print(accuracy_lr)
print(accuracy_nb)
print(accuracy_dt)
print(accuracy_rf)
print(accuracy_knn)
print(accuracy_svm)
```

```
0.9476190476190476
0.9428571428571428
0.9380952380952381
0.9476190476190476
0.9476190476190476
0.9523809523809523
```

Fig.8: Accuracy Results

4.1 Parameters Used

These are the fallowing features are used in our project to develop the model.

1. **Clump thickness:** Blood clots are clumps that occur when blood hardens from a liquid to a solid.

2. **Uniform cell size:** It is employed to assess the consistency of cell size across the sample. Sizes of cancer cells typically vary. Because of this, this metric is crucial for identifying whether or not the cells are malignant.
3. **Uniform cell shape:** The cancerous cell is characterized by a large nucleus, having an irregular shape.
4. **Marginal adhesion:** Normal cells have a propensity to adhere. This skill tends to be lost in cancer cells. Therefore, a marker of malignancy is the lack of adhesion.
5. **Single epithelial size:** It has to do with consistency. Significantly expanded epithelial cells may be cancerous cells.
6. **Bare nuclei:** : The area of the cell's nucleus that isn't encircled by cytoplasm is referred to by this expression. The majority of the time, benign tumors exhibit them.

7. **Bland chromatin:** Describes the nucleus of benign cells as having a homogeneous texture. The chromatin is typically more agglomerated in cancer cells.
8. **Normal nucleoli:** The nucleus contains tiny structures called nucleoli. The nucleolus is typically barely noticeable, if at all, in normal cells. The nucleoli grow noticeably more pronounced and occasionally more numerous in cancer cells.
9. **Mitoses:** It is a rough estimate of how many mitoses have occurred. The likelihood of cancer increases with value.

4.2 Experimental Outcomes

- Confusion Matrix

An N x N matrix called a "Confusion matrix," where N is the total number of target classes, is used to assess the effectiveness of a classification model. In the matrix, the actual target values are contrasted with those that the machine learning model predicted. This gives us a comprehensive understanding of the effectiveness of our classification model and the types of mistakes it is committing. A matrix with two values—predicted values and actual values—along with the total number of predictions is used to determine the outcome.

5. CONCLUSION

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.

6. REFERENCE

[1] Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ. 2019 Jan 28;7:e6201.

[2] Philip, John, W. Graham Harris, Camille Flaherty, and Charles Albert Frederick Joslin. "Clinical measures to assess the practice and efficiency of breast self-examination." Cancer 58, no. 4 (1986): 973-977.

[3] Chiang, C.J., You, S.L., Chen, C.J., Yang, Y.W., Lo, W.C. and Lai, M.S., 2015. Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. Japanese journal of clinical oncology, 45(3), pp.291-296.

[4] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 149(4), 778-789.

[5] UK StatisticsAuthority. Cancer statistics registrations: registrations of cancer diagnosed in 2004, England. London: UK StatisticsAuthority; 2007.

[6] Department of Health. The national cancer registration system, 2008.

[7] Breast Cancer Clinical Outcome Measures Project. BCCOM: Analysis of the management of symptomatic breast cancers diagnosed in 2004, 3 rd-year report. 2007

[8] https://www.mdpi.com/applsci/applsci-11-10753/article_deploy/html/images/applsci11-10753-g013-550.jpg.

[9] https://www.mdpi.com/applsci/applsci-11-10753/article_deploy/html/images/applsci11-10753-g014-550.jpg.

[10] https://www.mdpi.com/applsci/applsci-11-10753/article_deploy/html/images/applsci11-10753-g015-550.jpg https://www.mdpi.com/applsci/applsci-11-10753/article_deploy/html/images/applsci11-10753-g016-550.jp

7. REFLECTION OF THE TEAM MEMBERS ON THE PROJECT

SUMMARY OF TEAM WORK

1. Attends group meetings regularly and arrives on time.
2. Contributes meaningfully to group discussions.
3. Prepares work in a quality manner.
4. Demonstrates a cooperative and supportive attitude.
5. Contributes significantly to the success of the project.

SCORE

1=strongly disagree; 2=disagree; 3=agree; 4=strongly agree

Student 1: PRIYAMBADA SAHU (2051012015)

Student 2: MADHUSMITA JENA (1941012932)

Student 3: SWAYAM SARTHAK ROUT (1941012604)

Student 4: ADITYA KUMAR SAMAL (1941012931)

Attributes	Student 1	Student 2	Student 3	Student 4
1	4	1	1	1
2	4	1	1	1
3	4	1	1	1
4	4	1	1	1
5	4	1	1	1
6	4	1	1	1
Grand Total	24	6	6	6

8. SIMILARITY REPORT