

# **ENSEMBLE LEARNING BASED FRAMEWORK FOR BREAST CANCER PREDICTION**

Major project report submitted in partial fulfillment of the requirement for the  
degree of Bachelor of Technology

in

**Computer Science and Engineering**

By

DIVYAM GOYAL(181273)

**UNDER THE SUPERVISION OF**

Dr Aman Sharma  
Assistant Professor, Deptt. Of CSE & IT




Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Wakanaghat, 173234, Himachal Pradesh, INDIA

## DECLARATION

I hereby declare that this project has been done by me under the supervision of (Dr Aman Sharma, Assistant Professor, Deptt. Of CSE & IT), Jaypee University of Information Technology. I also declare that neither this Project nor any part of this Project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

A handwritten signature in blue ink that reads "Aman Sharma" with a horizontal line underneath.

(Dr. Aman Sharma) Assistant Professor

Department of Computer Science & Engineering and Information Technology Jaypee University of Information Technology

A handwritten signature in blue ink that reads "Divyam" with a horizontal line underneath.

Submitted by:

(Divyam Goyal -181273)

Computer Science & Engineering Department Jaypee University of Information Technology

# CERTIFICATE


## Candidate's Declaration

I herewith declare that the work bestowed during this report entitled “Breast Cancer Prediction Using Ensemble Learning” is in partial fulfillment of the wants for the award of the degree of Bachelor of Technology in Computer Science and Engineering submitted within the Department of technology & Engineering, Jaypee University of data Technology Wagnaghat is authentic record of my very own work allotted over a amount from August 2021 to Gregorian calendar month 2021 underneath the direction of Dr. Aman Sharma, prof, Department of Computer Science and Engineering.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

**Divyam Goyal, 181273**

This is to certify that the above statement made by the candidate is true to the best of my knowledge.



(Supervisor Signature)

**Professor Aman Sharma**

Assistant Professor

Department of Computer Science & Engineering

Computer Science & Engineering and Information Technology Jaypee University of Information Technology, Wagnaghat

## ACKNOWLEDGEMENT

I am very grateful and need my profound obligation to Supervisor Dr. Aman Sharma, Assistant Professor, Department of CSE Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of Machine Learning and Artificial Intelligence has helped me enormously to carry out this project. His endless patience, critical steering, continual encouragement, constant and energetic direction, constructive criticism, valuable recommendation, reading several inferior drafts and correcting them in any respect stages have created it attainable to finish this project.

I would like to express my heartiest gratitude to Dr. Aman Sharma, Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win. In this situation, I also want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I have to acknowledge with due respect the constant support and patience of my folks.

Divyam Goyal

## TABLE OF CONTENT

<b>Content</b>	<b>Page No.</b>
Certificate	3
Acknowledgement	4
Table of content	5
Abstract	6
<b>Chapter 01: Introduction</b>	7-8
<b>Chapter 02: Literature Review</b>	8-11
2.1. Our Contribution	
<b>Chapter 03: Background &amp; Preliminaries</b>	11-17
<b>Chapter 04: Proposed Framework</b>	17-30
4.1 Model Selection	
4.2. Parameter Setting	
4.3. Experimental Setup	
4.3.1 Data set	
4.3.2 Data visualization & Correlation of data attributes	
4.3.3 Proposed Methodology	
<b>Chapter 05: Results &amp; Discussion</b>	30-38
5.1 Performance Metrics	
5.2 Comparison with ML Models	
5.3 Comparison with Existing Literature	
<b>Chapter 06: Conclusion and Future Work</b>	39
<b>References</b>	39-42

## Abstract

**Background and objective:** breast cancer is a sickness wherein the breast cells as a rule develop enormous in size. breast malignant growth comes in various types and one of a kind structures. Those cells in which the breast becomes dangerous decides the sort of breast malignant growth. It is more normal in ladies than in men to have breast cancer growth. The most widely recognized kinds of breast cancer growth are: Infiltrative bosom disease cells that begin in the breast channel and spread to different pieces of the bosom tissue. Infiltrative disease cells can spread to different pieces of the body, an interaction known as metastasis. Infiltrative lobular malignant growth Cancer cells that beginning in the lobular and spread to the encompassing bosom tissue. These disease cells are extremely dynamic and can spread to different pieces of the body.

**Methods:** Analysts in the field of clinical sciences are captivated by Artificial Intelligence. Specialists utilize an assortment of AI strategies and ways to deal with foresee bosom malignant growth. In this work apparently, we have utilized the Wisconsin breast Cancer Dataset (WBCD) from UCI machine learning repository which is one of the most utilized dataset accessible on the web. For each picture, the mean, standard error, and "worst" or worst (mean of the three biggest qualities) highlights were determined, yielding 30 elements. For instance, field 3 addresses Mean Radius, field 13 addresses Radius SE, and field 23 addresses Worst Radius. To evaluate the viability and strength of the created model, a few performance measures are utilized, for example, ROC, AUC bend, explicitness, F1-score, sensitivity, and accuracy.

**Results:** In this study, we have proposed a framework with a stacked ensemble classifier using several machine learning algorithms including Decision tree classifier, AdaBoost classifier, GaussianNB, and MLP classifier. Our proposed framework attained an accuracy of 97.66% which is higher than the existing literature. So, the results show that the proposed ensemble technique outperforms the existing techniques

**Conclusion:** It is shown that malignant and benign cancer cells are successfully identified by the proposed ensemble approach which would be helpful in detecting the cancer at an early stage and treating the diseased accordingly. Moreover, this ensemble approach can be applied to other problem domains of interest.

## 1. Introduction

The complete number of ladies biting the dust in 2021 is 963,000, as anticipated by the World Health Organization. In any case, it is anticipated by the organization that the number could reach up to 2.9 million around the world. Bosom disease typically happens in ladies however seldom in men. Bosom malignant growth is an infection yet generally, until the time ladies or men get mindful of the side effect it goes past its original state Breast malignant growth is a typical and hazardous illness that influences ladies. Disease is the advancement of abnormal cells that are hereditarily adjusted and transformed. In conclusion and treatment, it spreads all through the body, bringing about death. There are two kinds of bosom disease: harmless and threatening. Threatening is delegated destructive in light of the fact that it can taint different organs and is carcinogenic, while harmless is named non-unsafe. Subsequently, we require a framework that can distinguish bosom disease before it advances with the end result of being lethal [1]. Embracing protected, sensible procedures and using current innovation can lessen the requirement for guardians while likewise bringing down generally medical services costs. A few lives could be saved if canny dynamic techniques and advances were created. AI (ML) is perhaps the most broadly involved strategy for rapidly preparing machines and creating prescient models for better direction. By breaking down the tumor size, AI helps with the early discovery of bosom malignant growth and decides the idea of disease. AI strategies are the most famous techniques for accomplishing great outcomes in order and forecast issues. ML methods used to identify malignant growth and foresee the presence or nonattendance of tumors could be helpful to bosom disease research. AI can likewise be utilized to anticipate the harm of tumors [2]. Thus, it is absolutely impossible to forestall bosom malignant growth, yet early discovery can incredibly work on the anticipation. Furthermore, the treatment expenses can be essentially diminished subsequently. Notwithstanding, in light of the fact that malignant growth side effects can be uncommon on occasion, early location can be troublesome. Mammograms and self-bosom tests are fundamental for identifying any early abnormalities before the tumor advances [3]. The primary objective of this paper is to propose a group framework for Breast Cancer. This paper analyzes existing disease recognition models inside and out and reports on the profoundly precise and proficient outcomes.

The rest of the paper is organized as follows: We provided a literature review in the second section, in which we linked to several research works and explained the viability and performance of various algorithms for heart disease prediction. We discussed many machine learning techniques in Section 3. The suggested framework, including model selection, parameter setting, experimental setup, and recommended technique, is detailed in Section 4. A comparison of the proposed framework with existing Machine Learning (ML) models and literature is explained in Section 5, performance measures. In this part, the results are compared to the existing model and literature. The conclusion and future scope are included in Section 6.

## **2. Literature Review**

In this section we have taken 10 different research works and explained how other researchers have approached the problem and their different methodologies. The cases of Breast cancer are increasing gradually. Researchers have been experimenting with a wide range of approaches and algorithms to predict Breast cancer with more accuracy. Over the years, several types of research on the prognosis of breast cancer have been going on. Some of the research are mentioned below. Authors have applied algorithms such as Bayesian network, Radial Basis Function, Back propagation network (BPN), Artificial neural network (ANN), Convolutional neural network (CNN), Support vector machine (SVM), KNN, Logistic Regression(LR), and DT. KNN, Cubic SVM(CSVM), Simple Logistic Regression, SVM, MLP, NSVC, Optimized ANN. The authors have applied the dataset firstly on an individual basis. Afterwards, the dataset is applied together to form a clear picture.



Table 1: Comparison of existing approaches for Breast cancer prediction

<b>S. No.</b>	<b>Author (s)</b>	<b>Approach</b>	<b>Dataset</b>	<b>Performance measures</b>
1.	Jabbar et al.(2021)[14]	Bayesian network, Radial Basis Function	Wisconsin Breast Cancer Data set (WBCD)	Accuracy - 97%
2.	Anastraj(2019) [15]	Back propagation network, Artificial neural network (ANN), Convolutional neural network (CNN), Support vector machine (SVM)	Wisconsin Breast Cancer (original) dataset	Accuracy- 94%
3.	Kasaudhan et al.(2015)[16]	SVM	Digital Database for Screening Mammography (DDSM)	Accuracy-91.5% Sensitivity-95% Specificity-88% MCC-83.2%
4.	Mejia et al.(2015)[17]	KNN	Federal Fluminense University Hospital	Accuracy - 94.44%

<b>5.</b>	Avramov and Si (2017)[18]	Logistic Regression(LR), DT, KNN, Cubic SVM(CSVM)	UCI	Stacking Accuracy- 98.56%
<b>6.</b>	Jiang, and Xu. (2017)[19]	RF-Recursive Feature Elimination (RF-RFE) method	Zhejiang Cancer Hospital	Accuracy- 77.05% Sensitivity- 84.21% Specificity- 65.21% AUC-0.76
<b>7.</b>	M. Ngadi et al. (2016)[20]	NSVC	UCI	Accuracy- 99%
<b>8.</b>	Assiri et al. (2020)[21]	Simple Logistic Regression, SVM, MLP	Wisconsin Breast Cancer Dataset (WBCD)	Accuracy- 99.42%
<b>9.</b>	Bevilacqua et al. (2016)[22]	Optimized ANN	Radiologists of the university of Bari Aldo Moro	Accuracy - 89.77% Sensitivity- 89.08% Specificity- 90.46%
<b>10.</b>	Salma (2015)[23]	Fast Modular Artificial Neural Network(FM-A NN)	WBCD, KDD cup 2008	WBCD Training Accuracy- 99.2% KDD Training Accuracy-99.96%

They have considered a wide range of factors of Breast cancer and used classification matrices to reach a robust solution. In [14] authors have taken Wisconsin Breast Cancer Data set (WBCD) and applied Bayesian network, Radial Basis Function approach, In [16] Digital Database for Screening Mammography (DDSM) was used as a dataset and SVM approach was applied

Similarly, in [17] Federal Fluminense University Hospital Mammography Dataset was taken and KNN approach was used, In [19] dataset was taken from Zhejiang Cancer Hospital and RF-Recursive Feature Elimination (RF-RFE) method was used to achieve 77.05% accuracy.

### **2.1. Our Contribution:**

- The proposed framework uses a Stacking Based Ensemble Learning approach to increase classifier diversity
- The ensemble model is trained on a large dataset, which aids in generalizing the trained model
- In order to pick the appropriate parameter for ML model training, Hyper Parameter Tuning is utilized
- On the basis of accuracy, precision, sensitivity, precision, the suggested framework is compared to the current literature
- The proposed framework gives great accuracy in less computation time
- As compared to existing deep learning frameworks it requires less computational resources

### **3. Background & Preliminaries**

Various machine learning classification algorithms employed in the proposed framework are detailed in this section. Before the final Ensembling of top performing models, other Classifier models were attempted. On the training data set, ten distinct classifiers were trained. Following the initial training, four models were chosen based on their accuracy scores.

#### **A. Decision tree classifier[4]**

Decision tree classifiers have numerous applications. Their capacity to catch elucidating decision-production data from the information gave is their most fundamental characteristic. Preparing sets can be utilized to produce decision trees. A straightforward model for ordering models is a decision tree. It is administered AI, where the information is disintegrated

persistently founded on certain boundaries. The methodology for deciding the class of a given informational index in a decision tree begins at the root hub of the tree. This calculation checks the upsides of the first quality against the upsides of the record trait (the genuine informational collection), then, at that point, follows the branch and moves to the following hub in light of the correlation. The calculation contrasts the property estimation and other kid hubs and continues on toward the following hub. It rehashes the entire interaction until it arrives at the leaf hub of the tree.

#### B. AdaBoost classifier[5]

Inspiration is a kind of manufactured AI method that consolidates predictions from numerous frail students. The feeble student is a generally essential model however has limit on the dataset. Some time before a genuine calculation could be concocted, support was a hypothetical idea, and the AdaBoost (versatile support) strategy was the primary proficient execution of the idea. The AdaBoost strategy utilizes tiny choice trees (one level) For powerless students that are presented slowly all through. Each ensuing model in the succession looks to address the predictions made by the past model. This is finished by adjusting the training dataset to zero in more on training cases that past models didn't predict accurately.

#### C. GaussianNB[6]

Gaussian Naive Bayes is a Naive Bayes variety that acknowledges nonstop information and depends on the Gaussian ordinary circulation. A regular presumption while working with persistent information is that the nonstop qualities related with each class follow an ordinary (or Gaussian) appropriation. Gaussian Naive Bayes acknowledges consistent esteemed highlights and models with (typical) Gaussian appropriations.

#### D. MLP classifier[7]

The Multilayer Perceptron (MLP) is a brain network upgrader that utilizes feedforward brain organizations. It is separated into three layers: information, yield, and covered. The information layer gets the sign to be handled. The result layer is accountable for undertakings like forecast and arrangement. A MLP's actual process motor is an erratic number of stowed away layers sandwiched among information and result layers. Information streams in the forward bearing

from the contribution to the result layer of a MLP, equivalent to a feedforward network. The neurons in the MLP are prepared utilizing the backpropagation learning procedure. MLPs are equipped for managing issues that are not directly distinguishable and can surmised any ceaseless capacity. The significant use instances of MLP are design order, acknowledgment, forecast and estimation.

#### E.K-Nearest Neighbor (KNN)[8]

KNN is an order approach that is non-parametric. Perhaps the most notable arrangement calculation. The fundamental thought is that realized information is requested in a space determined by the highlights that have been picked. At the point when new information is given to the calculation, it will analyze the classes of the k nearest information to decide the new information's class. Medjahed et al. (2013) explored the utilization of KNN calculations to arrange bosom malignant growth. The effect of qualities, for example, distance and grouping rules on arrangement results is seen during examination. The KNN order has various benefits, including its straightforwardness and productivity. Anyway Despite its productivity, calculation lengths with huge data sets can be long, deciding the quantity of neighbors to utilize (k) takes experimentation, and the calculation is powerless with anomalies, which can altogether affect its effectiveness.

#### F.XGBoost[9]

XGBoost is a managed learning method that creates right models through a boosting technique. Boosting is an outfit learning approach that involves building a few models successively, with each new model looking to amend surrenders in the past model. In tree boosting, every additional model added to the troupe is a choice tree. XGBoost offers quick and precise equal tree boosting (otherwise called GBDT, GBM) to address a wide scope of information science applications. XGBoost is one of the most incredible angle boosting machine (GBM) systems for a wide scope of issues accessible today. The H2O XGBoost execution is comprised of two particular modules. The primary module, h2o-genmodel-ext-xgboost, expands the h2o-genmodel module with a XGBoost-explicit MOJO. The module likewise incorporates all of the important

XGBoost paired libraries. The module can incorporate numerous libraries for every stage to empower various setups (e.g., with/without GPU/OMP). H2O endeavors to stack the most impressive first (right now a library with GPU and OMP support). Assuming it falls flat, the loader advances to the following in the chain. H2O incorporates a XGBoost library with a simple arrangement (supporting just a single CPU) for every stage as a backup plan in the event that different libraries can't be stacked.

#### G.Support Vector Classifier[10]

SVMs are managed AI calculations for arrangement and relapse examination. The SVM might order information in both straight and nonlinear ways. Nonlinear arrangement is achieved utilizing the Kernel work. The pieces in nonlinear characterization are homogeneous polynomial, complex polynomial, Gaussian spiral premise work, and exaggerated digression work. In SVM, the Gaussian part with a solitary boundary functions admirably. Since it beats the others, the SVM strategy is the most generally utilized AI worldview, especially in modern settings. The SVM strategy is generally viewed as the best technique for diagnosing coronary corridor illness. While examining a lot of information, the SVM procedure has potential disadvantages, for example, unreasonable memory use. The addressed SVM calculation's boundaries are hard to decipher. Prior to utilizing the SVM strategy, all information should be appropriately marked. The SVM strategy gives the upside of further developed grouping exactness and investigation execution, arrangements (e.g., with/without GPU/OMP). H2O endeavors to stack the most impressive first (as of now a library with GPU and OMP support). Assuming it falls flat, the loader advances to the following in the chain. H2O incorporates a XGBoost library with a simple arrangement (supporting just a single CPU) for every stage as a contingency plan on the off chance that different libraries can't be stacked.

#### H.Stochastic Gradient Descent[11]

A binary classifier was made utilizing the SGD approach. To do this, the SGD procedure chooses irregular models from the preparation set and registers the slope in view of that solitary event, which is the expense capacity's negligible worth. The arrangement is then performed utilizing a

straightforward binary classifier that can identify whether the cardiovascular infection is available, utilizing the boundaries decided to augment the expense work.

#### I. Random Forest[12]

The decision tree is the groundwork of arbitrary backwoods classifiers. A decision tree is a progressive design worked from the properties of an information assortment (or free factors). An action combined with a subset of the elements separates the decision tree into hubs. The arbitrary woods is an assortment of decision trees that are connected to an assortment of bootstrap tests produced from the first informational index. To parcel the hubs, the entropy (or Gini list) of a subset of the properties is used. The bootstrapped subsets of the first informational collection have similar size as the first informational collection. Breiman's articles on irregular woods classifiers are significant (Breiman, 1996, 2001). According to Suthaharan, the bootstrapping method works with in the development of arbitrary timberlands with the required number of decision trees to increment arrangement exactness through the idea of cross-over diminishing (2015). The best trees are then picked by a democratic cycle and a bagging method (bootstrap total). This average arbitrary timberland method is utilized in the proposed mental registering design.

#### J.Gradient Boosting[13]

Gradient boosting produces added substance relapse models by fitting a straightforward defined work (base student) to current "pseudo"- residuals utilizing least squares iteratively. The pseudo-residuals are the gradients of the misfortune practical that are being limited concerning the model qualities at each preparing information point assessed in the ongoing advance. Gradient boosting's guess exactness and execution speed can be extensively improved by integrating randomization into the cycle. At each cycle, a subsample of the preparation information is picked indiscriminately (without substitution) from the whole preparation informational collection. This haphazardly chosen subsample is used rather than the whole example to fit the base student and figure the model update for the ongoing emphasis.

## Ensemble Learning

Ensemble is the art of bringing together a diverse group of learners (individual models) to improve the model's stability and predictive power. Ensemble Learning is the process of combining all of the predictions.

Commonly used ensembling techniques are-

**1. Bagging:** Bagging tries to implement comparable learners on tiny sample populations and then averages the results. You can employ different learners on various populations in generalized bagging. As you may assume, this aids in the reduction of variance error.

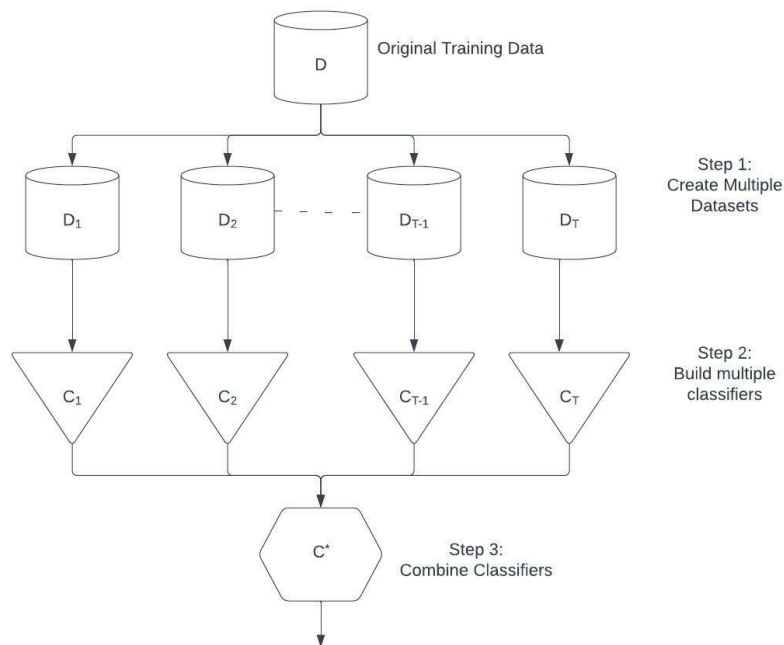


Figure 1: Bagging ensemble framework

**2. Boosting:** Boosting is an iterative strategy for adjusting an observation's weight based on the previous categorization. It seeks to raise the weight of observation if it was classified erroneously, and vice versa. Boosting reduces bias error and produces good prediction models in general. They may, however, overfit the training data on occasion.



**3. Stacking:** This is an intriguing method of mixing models. A learner is used to integrate the output of multiple learners. Depending on the combining learner we select, this can result in a reduction in either bias or variance error.

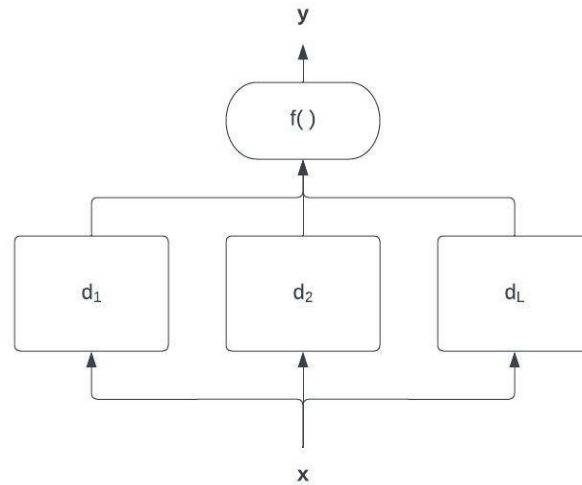


Figure 2: Stacking Ensemble framework

We have used Stacking in our proposed ensemble framework.

#### 4. Proposed Framework

In this section, we discussed model selection criteria and parameter settings for several algorithms utilized in the framework's construction. The experimental setup, as well as the proposed approach, have been detailed.

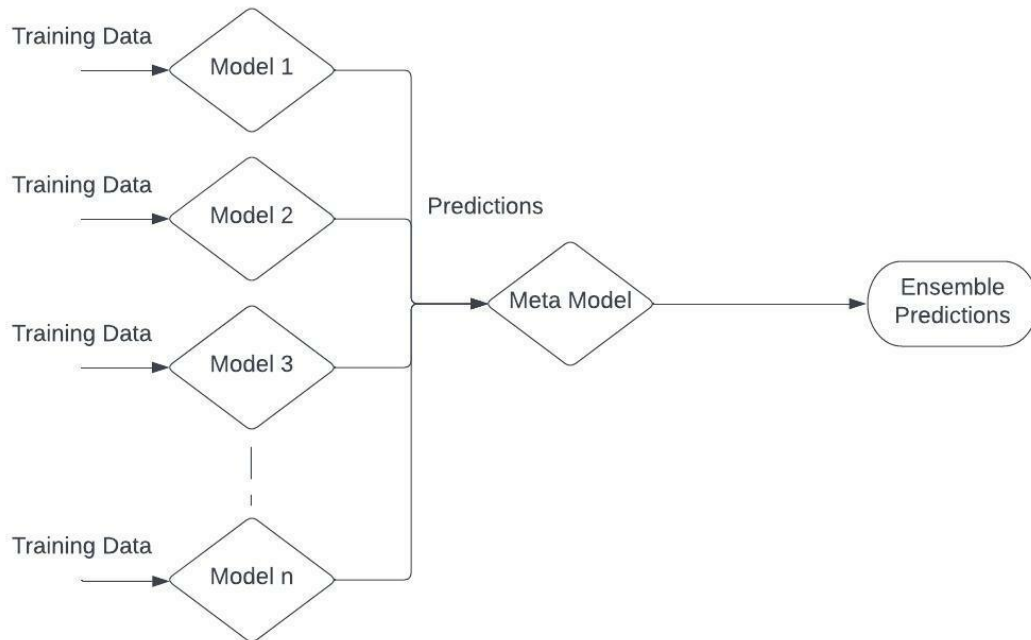


Figure 3: The Proposed ensemble framework for Breast Cancer Detection

#### 4.1 Model Selection

The proposed architecture for heart disease prediction is depicted in detail in Figure 3. To begin, we used data from the UCI Machine Learning Repository, as described in Section 3.3. We deleted outliers after thoroughly analyzing the data and discovering correlations among other parameters. Then we divided our data into two parts: training data, which contained 80% of the cases, and testing data, which had 20% of the instances.

Following the construction of our Model, we used stacking, also known as stacked regression, which is a class of algorithms that involves training a second-level meta learner to discover the best combination of base learners. Stacking differs from bagging and boosting in that the purpose is to group together strong, diverse groupings of learners. We utilized cross-validation using 10-folds to select the best models among a variety of baseline models. Then, to construct ensemble predictions, we feed our pre-processed data into the base learners. We choose the best performing baseline models for the stacked ensemble based on their cross-validation accuracy so that their ensemble will outperform individual machine learning models. Finally, we compare our

findings to those of the baseline learners and other current approaches to Breast cancer prediction.

## 4.2. Parameter Setting

In this section, we have explained diverse parameters used to boost accuracy across our Stacked Ensemble model. In AdaBoost we have tried different values of  $n$  estimators which are 100, 500, 1000, and 2000 out of which we received the best accuracy from the value of  $n$  estimator as 500, KNN we have run our model with different values of  $K$  and we received the best accuracy for  $k=9$

In GaussianNB we have taken exponential scale =0.5, In MLPClassifier there are hidden layers and we have used Adam optimizer, SGD and lbfgs as solvers.

## 4.3. Experimental Setup

### 4.3.1 Data set

The UCI machine learning repository's Wisconsin-Breast Cancer (Diagnostics) dataset (WBC)[24] is a classification dataset that records breast malignant growth case measurements. Harmless tumors are classified into two types: harmless and dangerous. The assortment includes 569 instances and 32 attributes, including an ID number and a diagnosis (M = harmful, B = harmless). 3-32) For every cell nucleus, the accompanying ten genuine esteemed features are registered: radius (mean of distances from the middle to points on the edge), surface (standard deviation of dim scale values), border, region, smoothness (neighborhood variety in radius lengths), compactness (perimeter<sup>2</sup>/region - 1.0), concavity (severity of inward portions of the form), curved points (number of sunken portions of the shape), symmetry ("coastline guess" - 1).

Table 2: Description of Nominal Attributes

Attributes	Description
ID number	Specifies the unique ID of a patient
Diagnosis	It is categorised into two types M = malignant, B = benign
radius	It is the mean of distances from center to points on the perimeter
texture	(standard deviation of grey-scale values)
perimeter	It defines the parameter of the nucleus of cell
area	It defines the area of the nucleus of cell
smoothness	It is the local variation in radius lengths
compactness	$(\text{perimeter}^2 / \text{area} - 1.0)$
concavity	(severity of concave portions of the contour)
concave points	(number of concave portions of the contour)
symmetry	Consists of symmetry mean
fractal dimension	("coastline approximation" - 1)

#### 4.3.2 Data visualization & Correlation of data attributes

In this section, we visualised the data, which is a key aspect of developing a model. Data visualisation facilitates story telling by translating data into a more intelligible format and displaying trends and outliers. A good visualisation conveys a story by removing data noise and

emphasising the most relevant facts. In Figure 4, we utilised a heatmap to compare the features to the data. To depict the distribution of numerical data in Figure 5, we utilised a violin plot, which is a mix of a box plot and a kernel density plot that shows peaks in the data. Figure 6 depicts a boxplot of characteristics to value to summarise and display data from multiple sources. Figure 7 depicts the swarmplot, which is a method of showing the attribute distribution. Figure 8 depicts the mean and standard deviation of the classifier's precision values for the three training sets. Figure 9 depicts the ensembled model's classifiers and their best prediction accuracy, while Figure 10 depicts the confusion matrix.

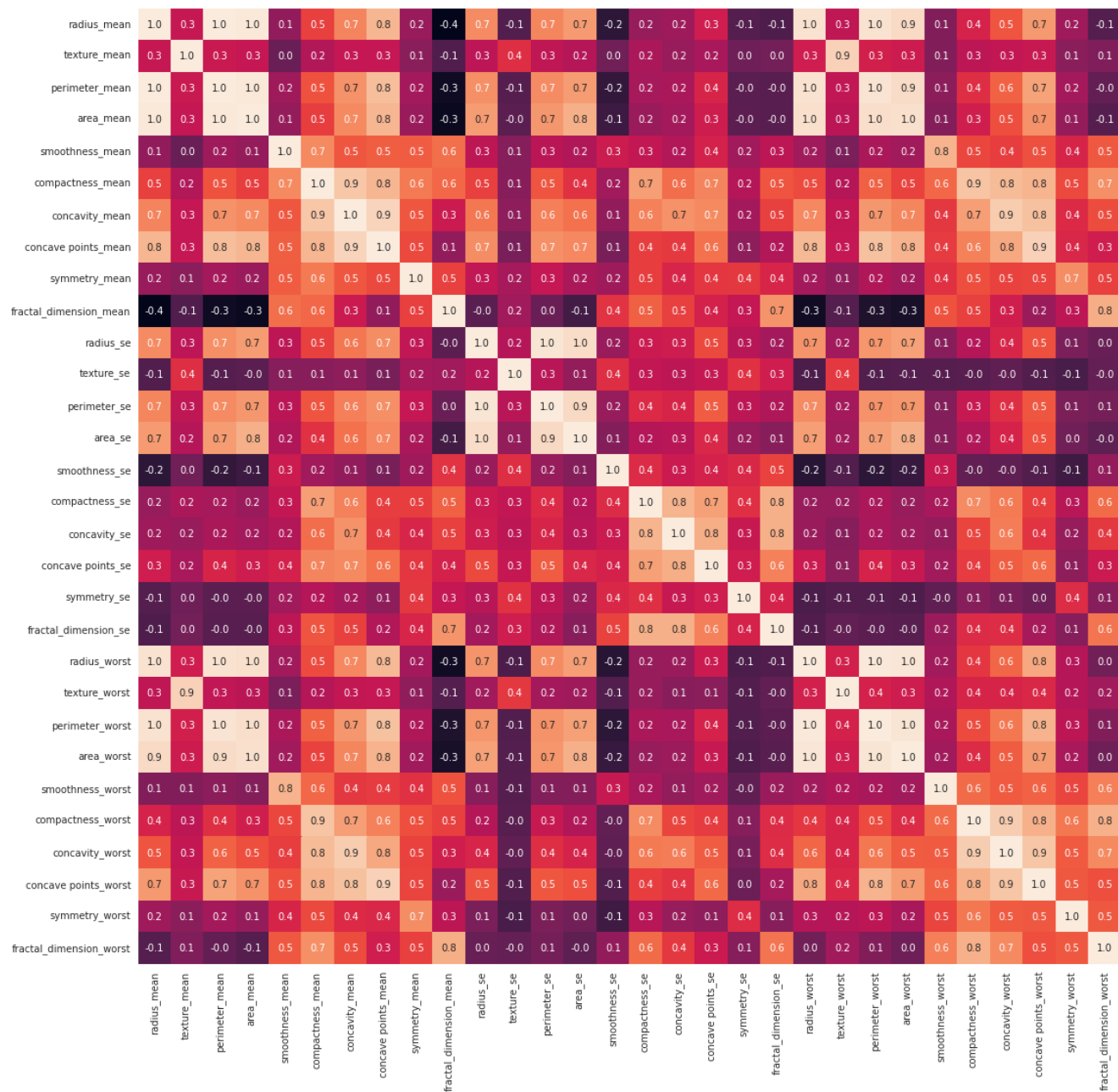


Figure 4: Cross-correlation values through Heat map(X-axis: Features of Breast cancer dataset, Y-axis: Features of Breast cancer dataset)

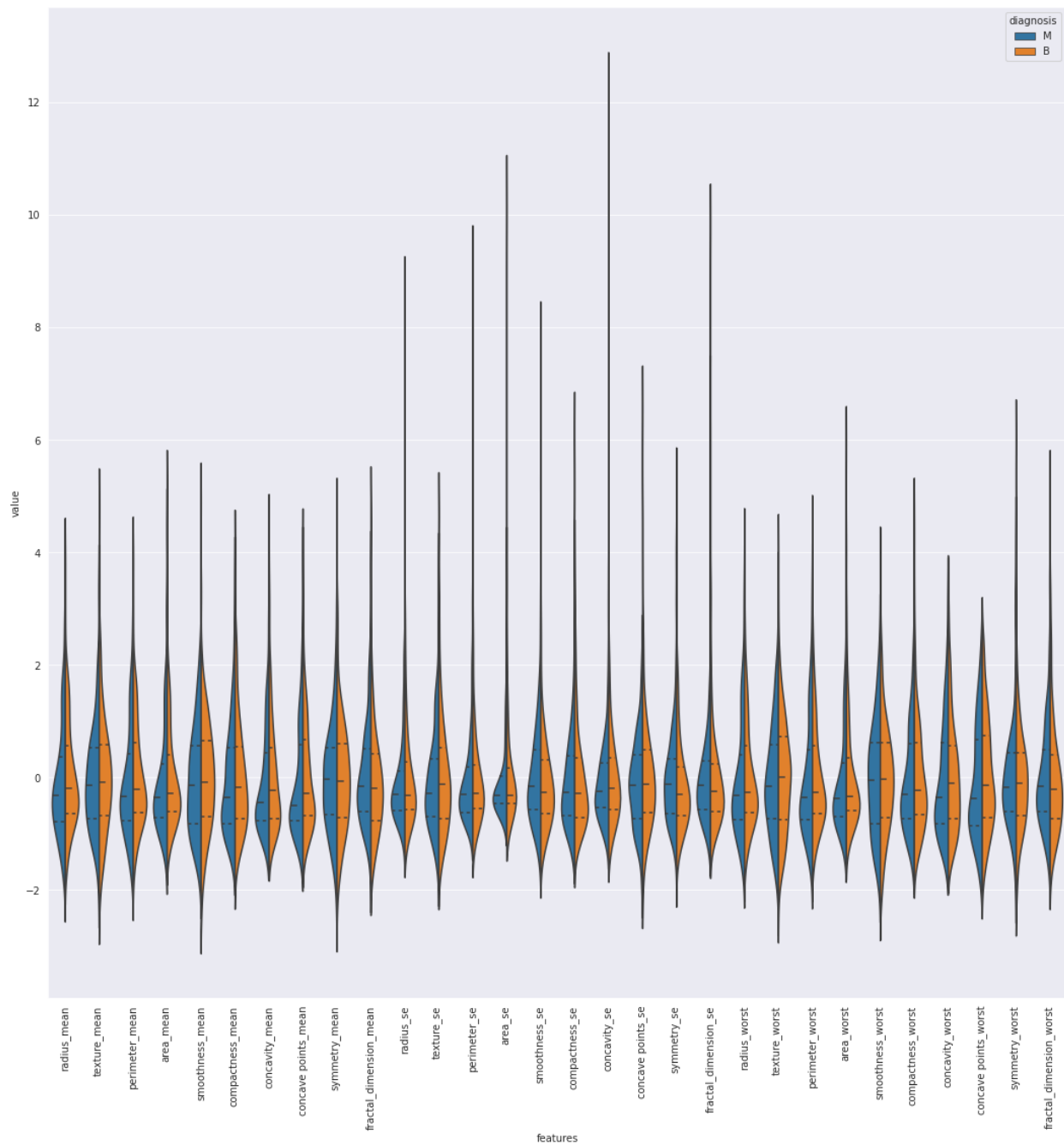


Figure 5: Features to value violinplot(X-axis: Features of Breast cancer dataset, Y-axis: Corresponding Value of each column in Breast cancer Dataset)

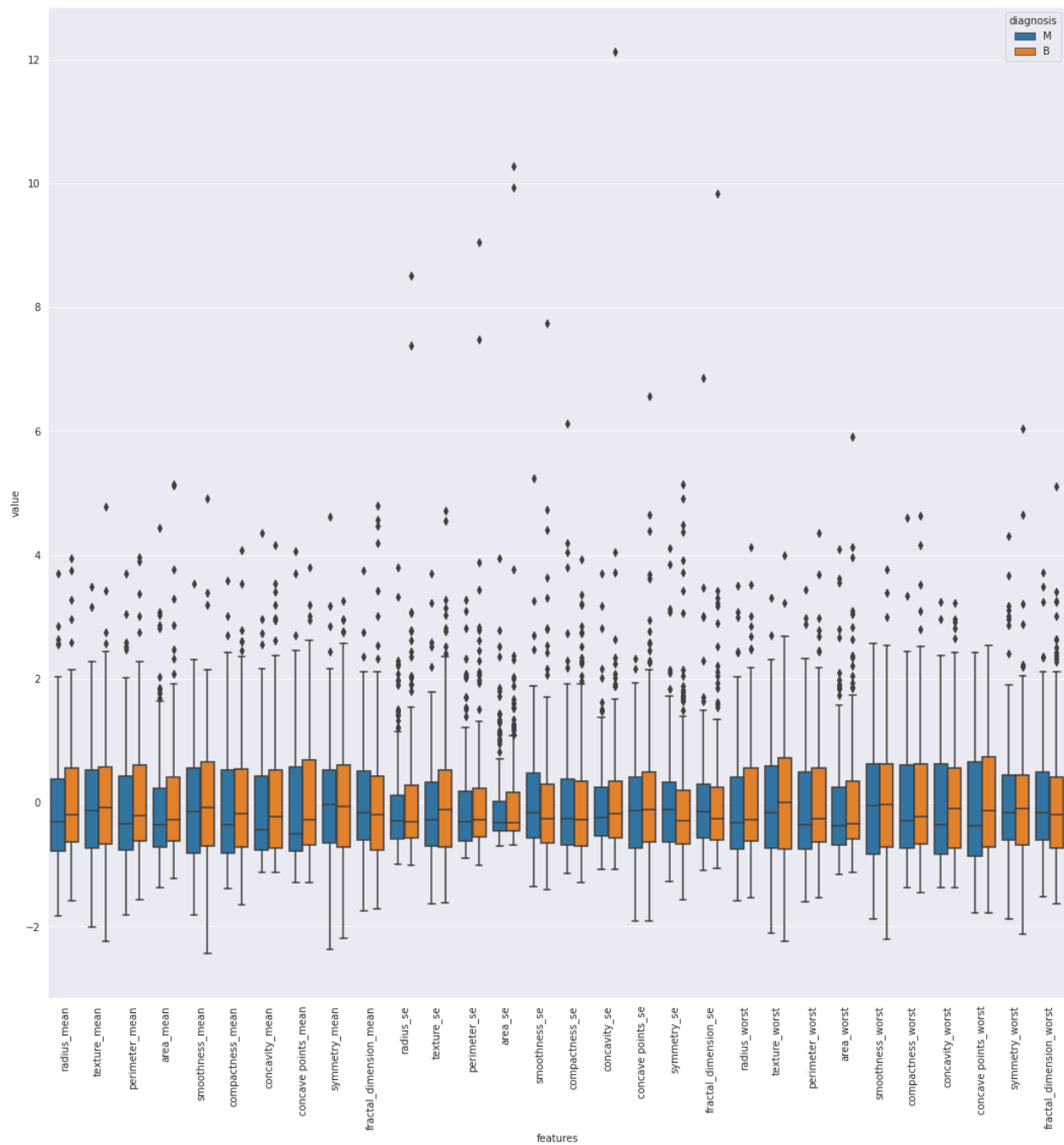


Figure 6: Features to value boxplot(X-axis: Features of Breast cancer dataset, Y-axis: Corresponding Value of each column in Breast cancer Dataset)



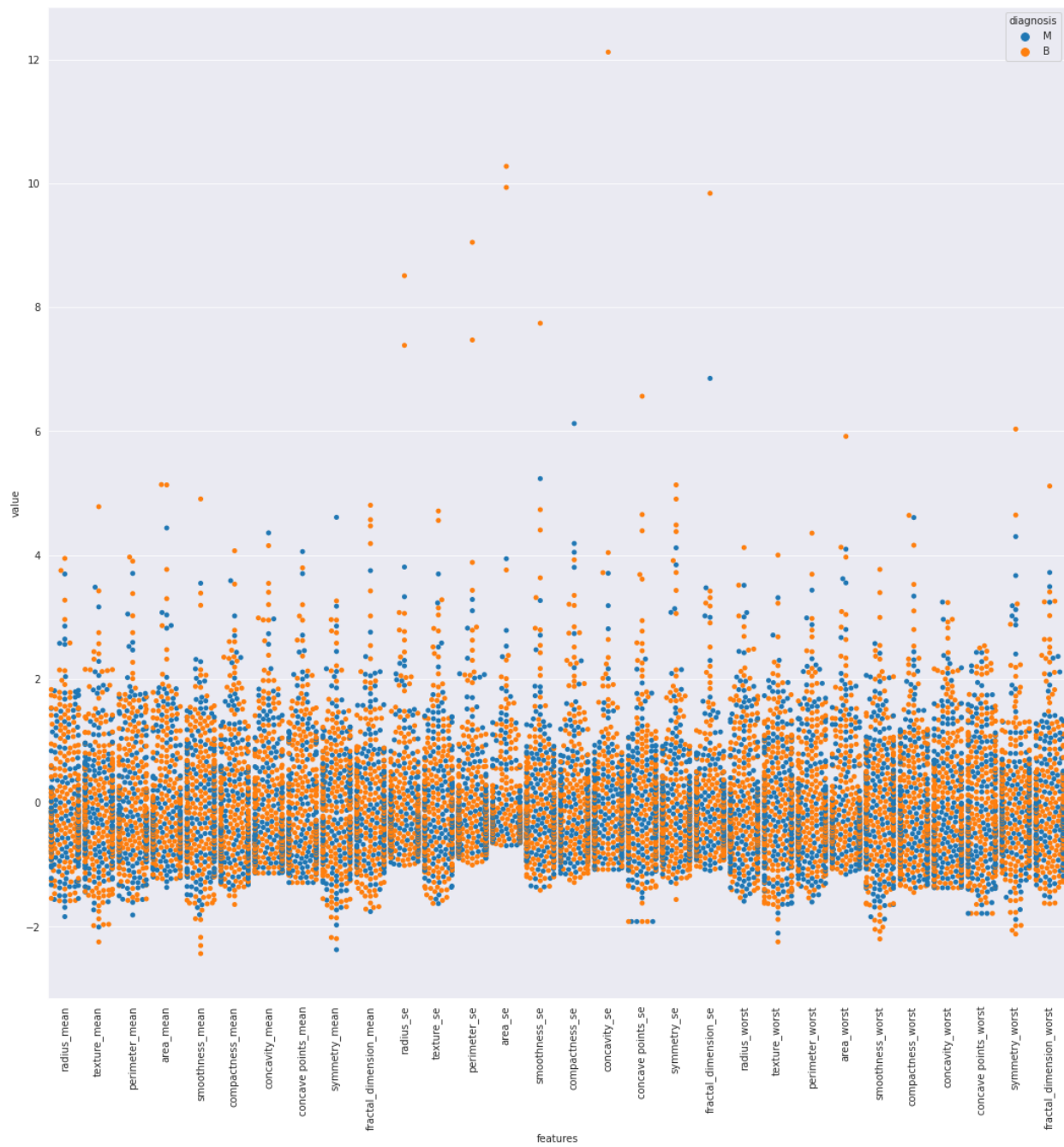


Figure 7: Features to value swarmplot(X-axis: Features of Breast cancer dataset, Y-axis: Corresponding Value of each column in Breast cancer Dataset)

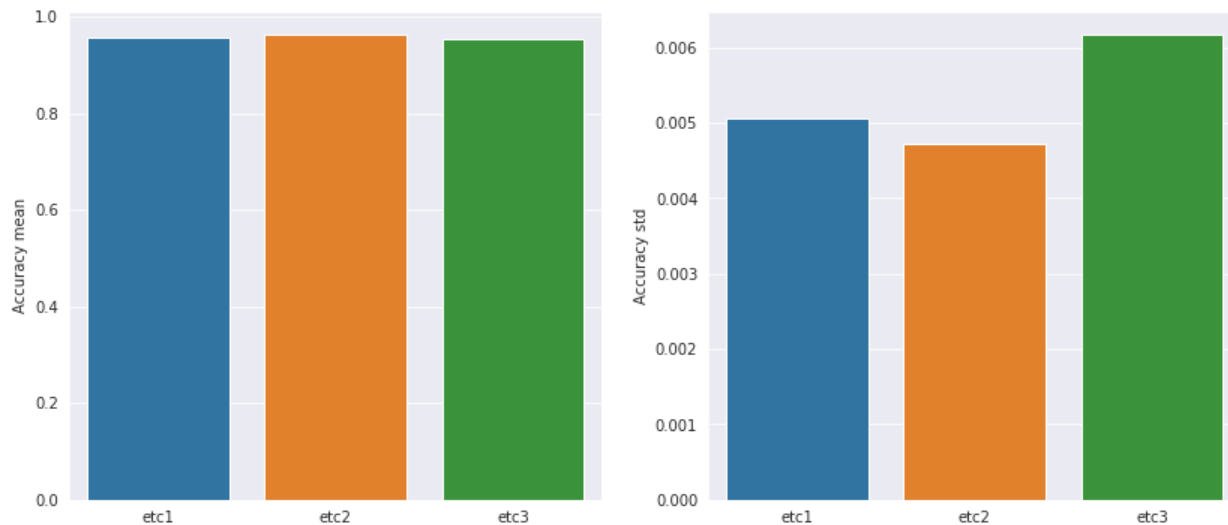


Figure 8: Mean and deviation of the classifier's precision values for each of the three training sets(X-axis: Training sets 1,2,3, Y-axis: Accuracy mean & Standard Deviation)

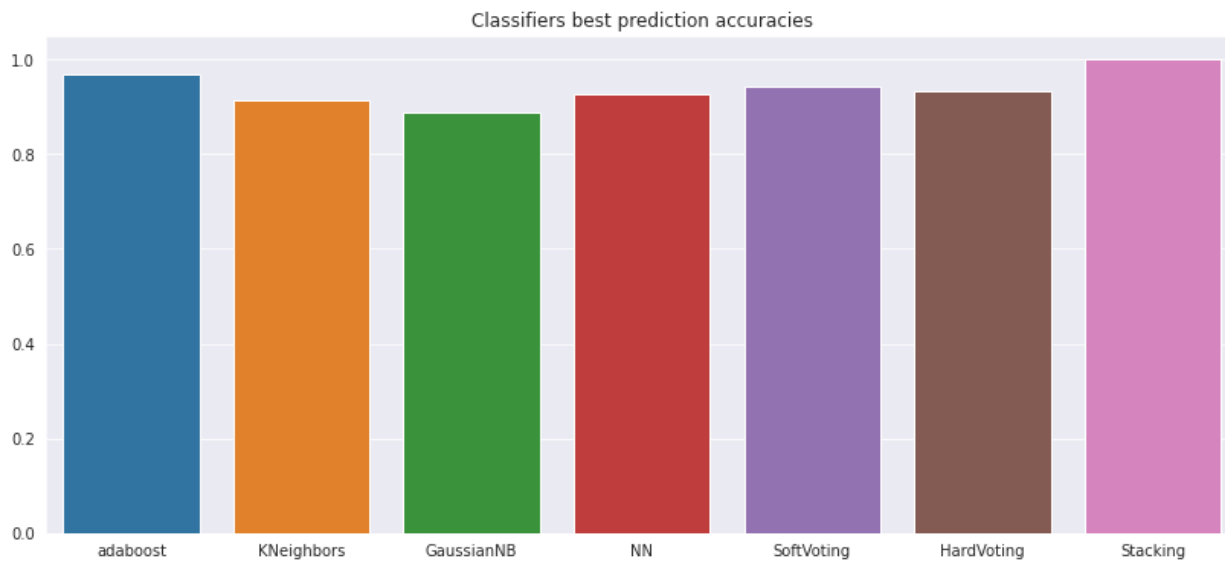


Figure 9: Classifiers best prediction accuracy(X-axis: Classifiers , Y-axis: Prediction Accuracy)

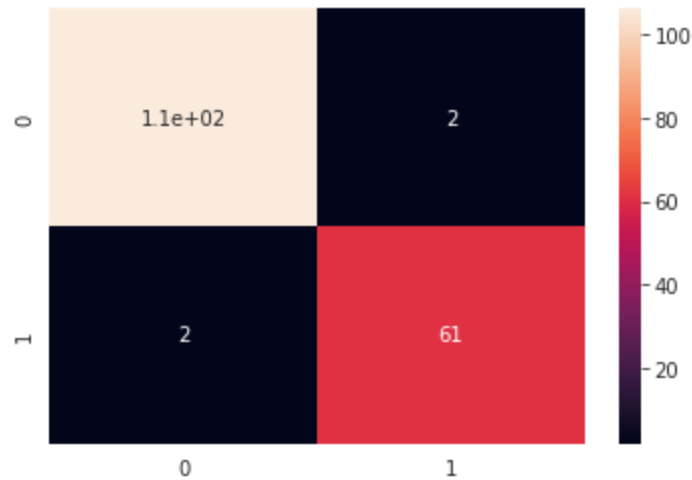


Figure 10: Confusion Matrix

### 4.3.3 Proposed Methodology

Stacking, also known as Super Learning, is an ensemble technique in which a "meta learner" is trained using a mix of classification models. The goal of stacking is to bring together a diverse group of strong learners. Algorithm 1 shows the stacking algorithm, while Table 4 defines the variables.

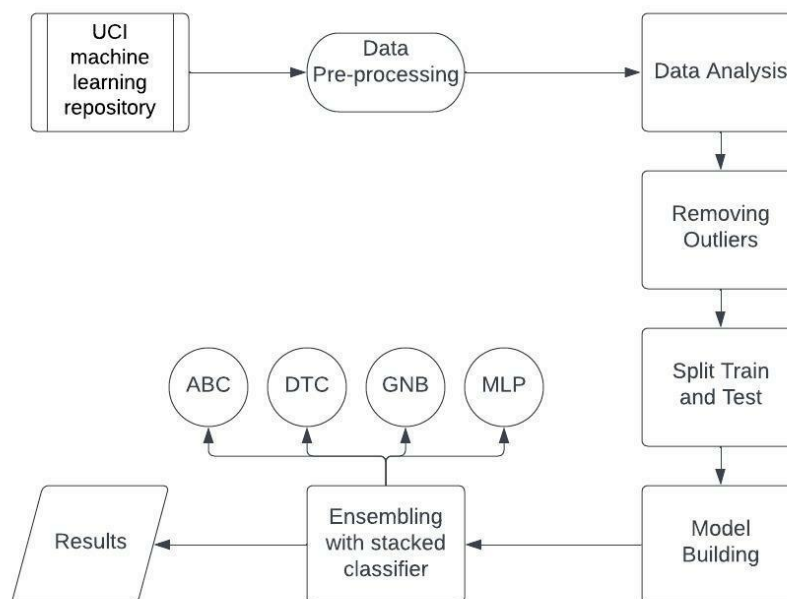


Figure 11: The architecture of proposed Ensemble Model

**Algorithm 1:** Proposed Ensemble framework for heart disease prediction

```

T' = N(T)                /*normalization of dataset*/
Let H = {h1, h2, h3 ... hn} /*the given dataset*/
E = {E1, E2, E3, ... En} /*the set of Machine learning ensemble classifiers*/
X= the 80% dataset for training, X ∈ H /* 80% of dataset is used for training*/
Y= the 20% dataset for testing, Y ∈ H /* 20% of dataset is used for testing*/
Z = meta level classifier
D = n(H)                /*where D is no. of attributes of dataset*/
Begin
  M(j) = E(j)           /*used for training the Model on X*/
  Next j                /*loop where j is iterating variable*/
  M = M ∪ Z             /*union of model and meta level classifier*/
End
Result = M classifies Y

```

**4.3.3.1 Data Preprocessing Phase**

In Data Preprocessing we have performed normalization on the dataset

$T' = N(T)$  normalization of dataset

**4.3.3.2 Training Phase**

In Training Phase we have taken multiple datasets namely  $h_1, h_2, h_3, \dots, h_n$  then we have taken multiple Machine learning algorithms  $E_1, E_2, E_3, \dots, E_n$  for ensembling, Furthermore we have divided the dataset into 80% training dataset and 20% testing dataset and we have taken  $Z$  as meta-level classifier.

Let  $H = \{h_1, h_2, h_3 \dots h_n\}$  be the given dataset

$E = \{E_1, E_2, E_3, \dots, E_n\}$ , the set of Machine learning ensemble classifiers

$X =$  the 80% dataset for training,  $X \in H$

$Y =$  the 20% dataset for testing,  $Y \in H$

$Z =$  meta level classifier

$D = n(H)$

#### 4.3.3.3 Testing Phase

In Testing phase we have used our training model on the testing dataset with the metalevel classifier and given the result that the model classifies the testing set.

Begin

$M(j) = E(j)$  used for training the Model on  $X$

Next  $j$

$M = M \cup Z$

End

Result =  $M$  classifies  $Y$

Table 4: Symbols used in Algorithm 1

S. No.	Symbols	Meaning
1.	H	Attributes of dataset
2.	E	Machine learning classifiers
3.	X	Training set
4.	Y	Testing set
5.	Z	Meta level classifier
6.	D	Number of attributes in dataset

7.	j	Iterator Variable
8.	M	Model
9.	T	Dataset
10.	T'	Normalized Dataset
11.	N	Normalization

In the proposed ensemble model, firstly, the original data is fed into several different models. The Meta classifier is then used to estimate each model's input and output, as well as the weights. The best models are picked, while the remainder are discarded. Stacking is the merging of many base classifiers learned on a single dataset using distinct learning methods using a meta classifier. To make ensemble predictions, train the base learners and pass the predictions to the meta learner. Figure 11 summarizes our work: first, we gathered data from the UCI Machine Learning Repository, as described in Data Collection & Pre-processing; next, after thoroughly examining the data and identifying correlations among various attributes, we eliminated outliers. Our data was then given out in 80 percent and 20 percent values.

## 5. Results & Discussion

We detailed the outcomes and analysis of our proposed framework in this section. The algorithms were evaluated using a variety of performance indicators. In addition, we compared our model to various current models in terms of accuracy, precision, sensitivity, precision, F1 Score, ROC, and MCC. In Section 2, we also discussed the proposed model in relation to other algorithms and models.

### 5.1 Performance Metrics

True positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) are the performance measures mentioned here, as stated below:

True positive rate (TPR) or sensitivity: When the disease is present, it describes the chance of a classifier correctly anticipating a positive result. The formula is as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

-----Eqn1[25]

Specificity or True negative rate (TNR): It is a classifier's likelihood of predicting a poor outcome when there is no sickness. The formula is as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{-----Eqn2[26]}$$

Accuracy: It is one of the most widely used metrics for assessing the performance of a classifier. It's stated as: It's calculated as a percentage of correctly identified samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \text{-----Eqn3[27]}$$

AU-ROC (area under the receiver operating characteristic curve): It is also a helpful and extensively used performance statistic for classification issues. TPR vs FPR at various threshold values are plotted. The AU-ROC is an excellent metric for performance comparison because it evaluates performance across a wide range of class distributions and error levels. This is how it's defined:

$$\text{AU-ROC} = 1/2 \left( \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) \quad \text{-----Eqn4[28]}$$

F1 score: It is defined as the weighted average of precision and recall (or harmonic mean). A score of 1 is considered the best, while a score of 0 is regarded as the worst. In F-measures, the TNs are not taken into account. The following formula can be used to compute the F1 score:

$$\text{F1 score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{-----Eqn5[29]}$$

## 5.2 Comparison with ML Models

We built numerous baseline models and used cross-validation with 10-folds to choose the best models. Models with high accuracies are used in the stacking approach. The accuracy of the baseline models is shown in Table 5 and their graph is plotted as shown in Figure 12. As we have

observed from Table 5 and its graph, the best performing algorithms are Decision Tree Classifier, AdaBoost Classifier, GaussianNB & MLP Classifier. These algorithms are stacked as shown in Figure 12. As shown in Table 5, we have tested the algorithms on the basis of accuracy. From Figure 12, the stacked classifier has a greater classification accuracy of 97.66 percent than the other classifiers. Hence we can conclude that the algorithms chosen by us to ensemble our model gives the highest accuracy of the 10 algorithms we have compared it with.

Reasons for using Decision Tree Classifier: Decision trees produce reasonable principles, Decision trees perform grouping without requiring a lot of calculation, Decision trees are equipped for dealing with both consistent and downright factors, Decision trees give an obvious sign of which fields are generally significant for expectation or order. Reasons for using AdaBoost: Adaboost is less inclined to overfitting as the information boundaries are not together advanced, The precision of frail classifiers can be improved by utilizing Adaboost. Reasons for using GaussianNB: Fast and adaptable model gives exceptionally solid outcomes, It functions admirably with enormous dataset, There is compelling reason need to invest a lot of energy for preparing, It gives better evaluating execution by wiping out unimportant details. Reasons for using MLP Classifier: Can be applied to complex non-linear problems, Works well with large input data, Provides quick predictions after training.

In machine learning, sensitivity and specificity are two proportions of the exhibition of a model. Sensitivity is the extent of true up-sides that are accurately anticipated by the model, while specificity is the extent of true negatives that are accurately anticipated by the model. As a rule, sensitivity is a higher priority than specificity when the goal is to augment the quantity of positive models that are accurately arranged. In any case, specificity is a higher priority than sensitivity when the goal is to limit the quantity of negative models that are inaccurately ordered. It's vital to take note that both sensitivity and specificity can be impacted by thresholding. At the end of the day, changing the endpoint for what considers a positive forecast can influence both sensitivity and specificity. Thus, it's typically best to report the two measurements while assessing the exhibition of a machine learning model.



Table 5: Comparison of ML algorithms &amp; their respective accuracies

<b>S. No.</b>	<b>Algorithm</b>	<b>Accuracy</b>
1.	Decision Tree Classifier	94.71%
2.	AdaBoost	96.47%
3.	GaussianNB	92.10%
4.	MLP	97.57%
5.	KNN	95.32%
6.	XGBoost	96.49%
7.	SVM	91.88%
8.	SGD	85.32%
9.	Random Forest	96.47%
10.	Gradient Boosting	95.60%

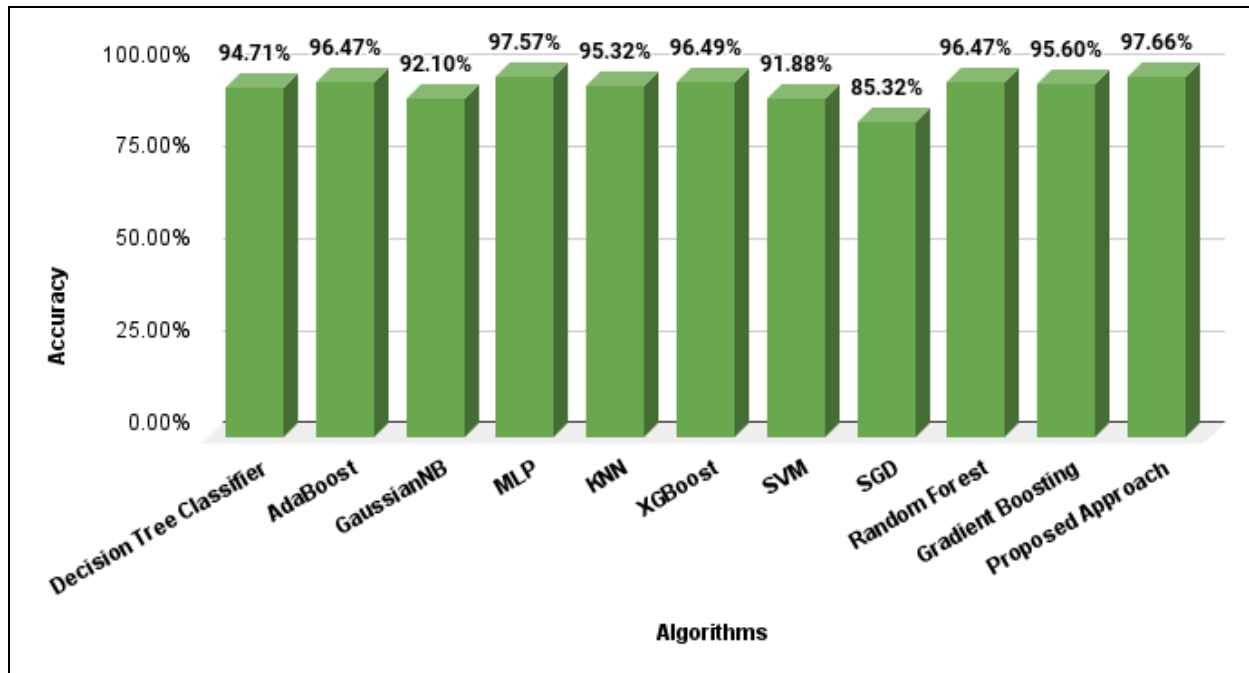


Figure 12: Comparison of accuracy of the proposed framework with different ML models

Table 6: Comparison of proposed Framework with existing ML Models

Model	Accuracy	Precision	Sensitivity	Specificity
Proposed Approach	<b>97.66%</b>	<b>92.00%</b>	<b>93.49%</b>	<b>91.07%</b>
Decision Tree Classifier	94.71%	87.31%	95.12%	84.82%
AdaBoost	96.47%	82.08%	89.43%	78.57%
GaussianNB	92.10%	78.67%	86.99%	74.10%
MLP	97.57%	88.54%	94.30%	86.60%
KNN	95.32%	90.62%	94.30%	89.28%
XGBoost	96.49%	80.14%	88.61%	75.89%
SVM	91.88%	80.00%	87.80%	75.89%
SGD	85.32%	81.43%	88.61%	77.67%
Random Forest	96.47%	83.59%	86.99%	81.25%
Gradient Boosting	95.60%	81.61%	90.24%	77.67%

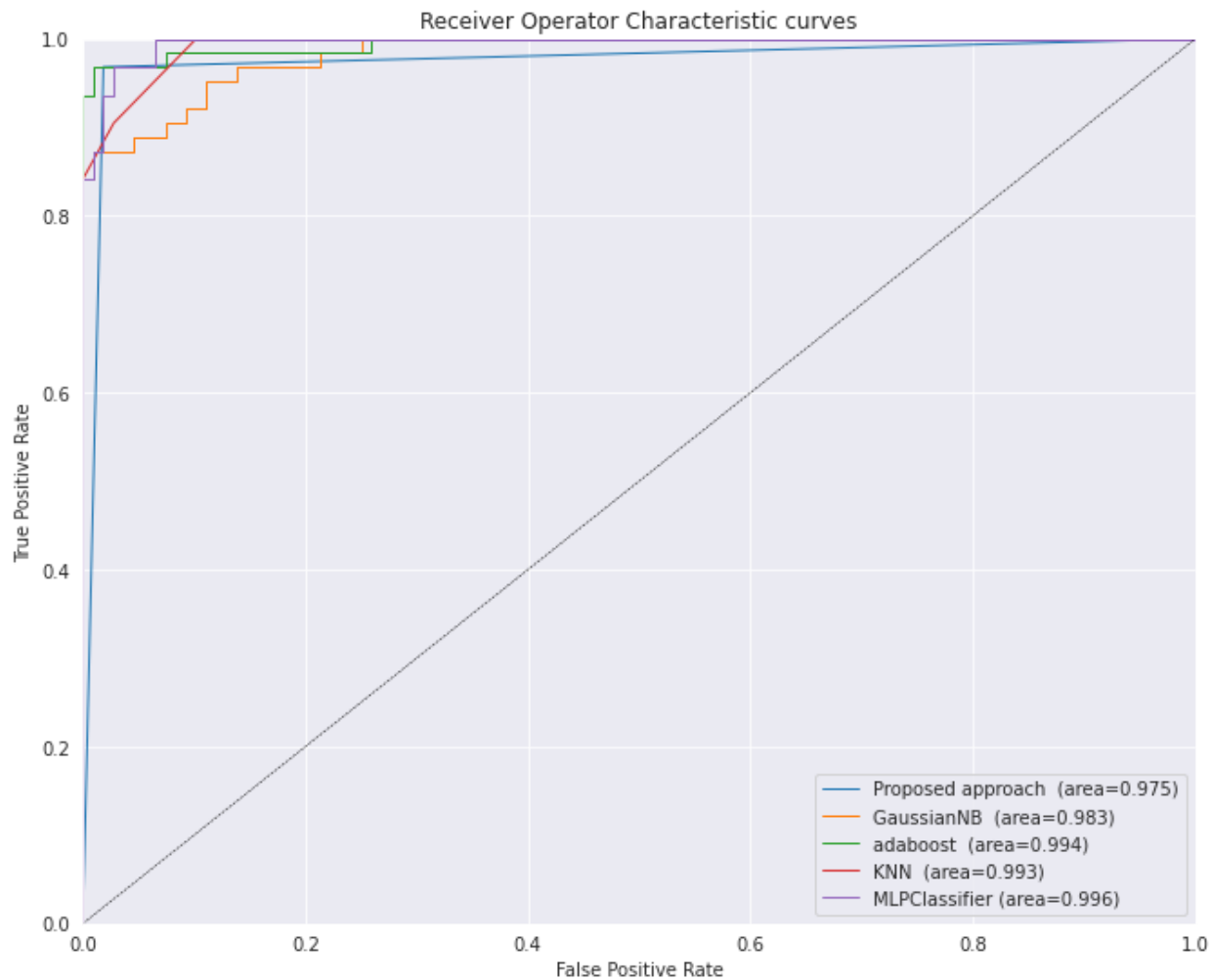


Figure13: AUC-ROC curves for Proposed Framework and other classifiers  
(X-axis: False Positive Rate, Y-axis: True Positive Rate)

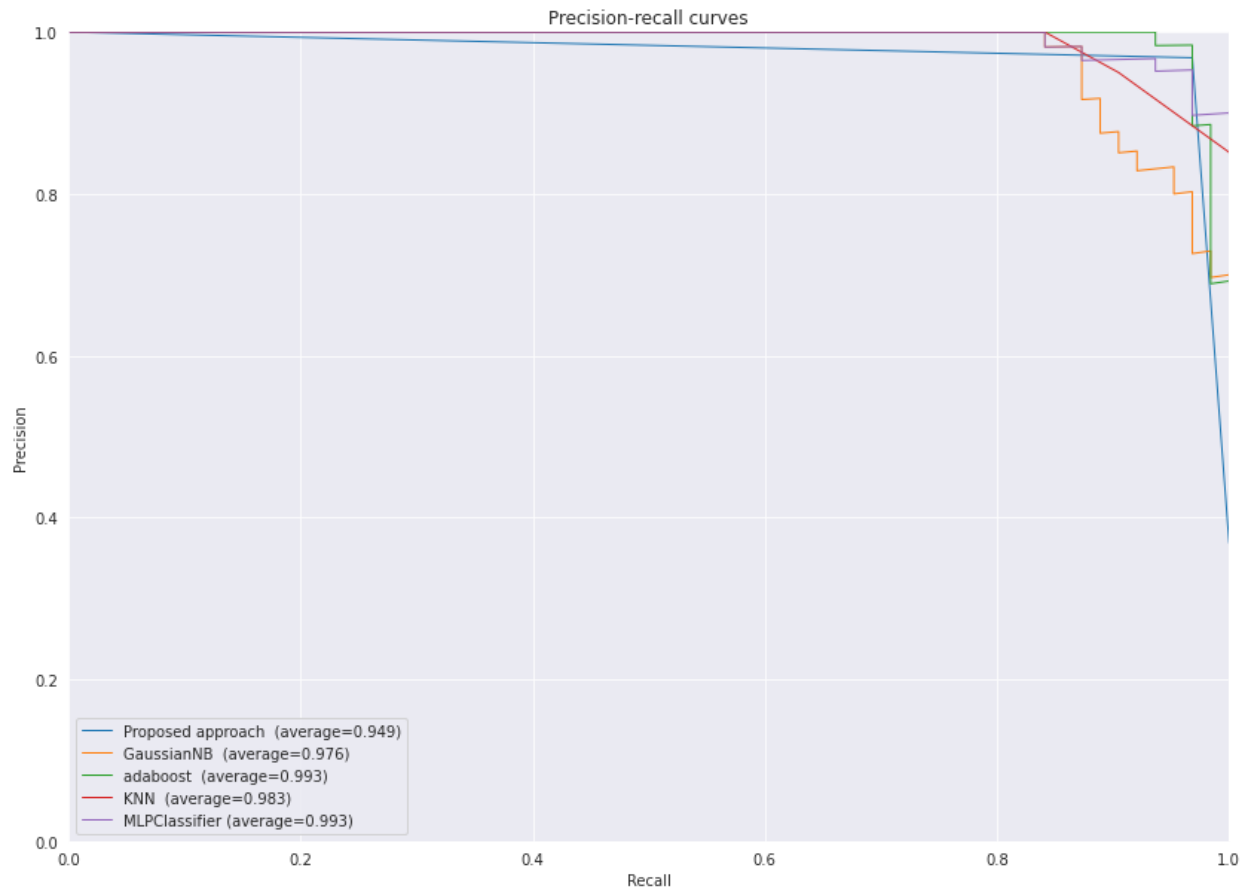


Figure14: Precision-Recall Curve for the Breast cancer predicting classifiers(X-axis: Recall, Y-axis: Precision)

### 5.3 Comparison with Existing Literature

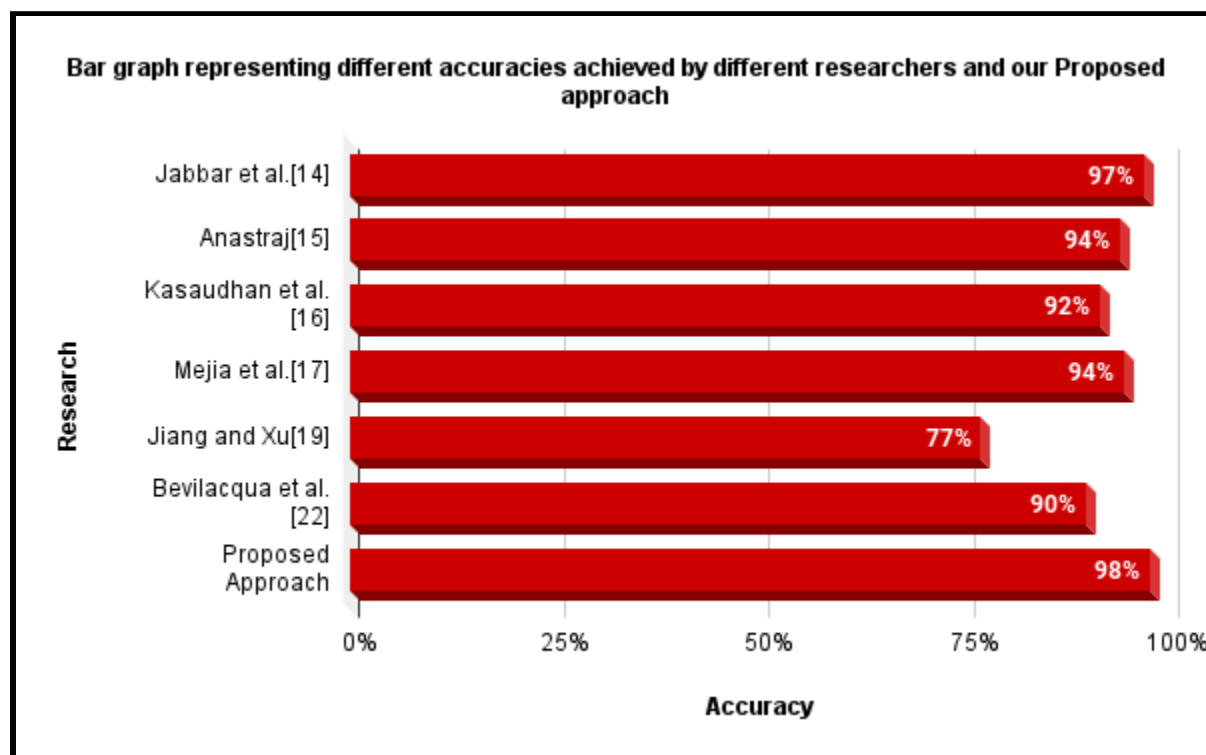


Figure 15: Bar graph representing different accuracies achieved by different researchers and our Proposed approach (X-axis: Accuracy, Y-axis: Research)

Because of the relative character of all the algorithms, the comparison in this example reveals an important aspect that can aid medical research and diagnosis. All of the algorithms' various natures and performances are depicted in Figure 15. When we use this layered method in practice, the medical community can profit from this diversity by understanding how an algorithm works in various Breast cancer research. Numerous researchers have utilized various calculations and datasets however what makes our model remarkable is that we have accomplished the most noteworthy exactness of all and on top of that we have additionally involved Multilayer perceptron classifier which has stowed away layers in it and is a piece of profound realizing which makes our ensemble model more accurate and precise.

## 6. Conclusion and Future Work

In nations like India, where resources are scarce and the population is growing, this is a problem. Better health care is urgently needed. In this worrying condition, the proposed framework can aid in a patient's early diagnosis and can also help the healthcare domain anticipate Breast cancer early. Because of its stacking machine learning technique, the epistemic evidence and findings of our proposed framework are resilient. As shown in the findings section, our suggested framework outperforms the existing state-of-the-art literature. We'd like to test our approach on a larger dataset in the future, applying deep learning concepts and stacking more algorithms for improved accuracy.

## References

- [1]Mangukiya, Manav & Vaghani, Anuj & Savani, Meet. (2022). Breast Cancer Detection with Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*. 10. 10.22214/ijraset.2022.40204.
- [2]Amethiya, Yash, et al. "Comparative Analysis of Breast Cancer Detection using Machine Learning and Biosensors." *Intelligent Medicine* (2021) <https://www.sciencedirect.com/science/article/pii/S2667102621000887>.
- [3]Vaka, Anji Reddy, Badal Soni, and Sudheer Reddy. "Breast cancer detection by leveraging Machine Learning." *ICT Express* 6.4 (2020): 320-324 <https://www.sciencedirect.com/science/article/pii/S2405959520300801>.
- [4]Szczerbicki E. Management of Complexity and Information Flow. *Agile Manufacturing: The 21st Century Competitive Strategy*. Published online 2001:247-263. doi:10.1016/b978-008043567-1/50013-9
- [5]How to Develop an AdaBoost Ensemble in Python. *Machine Learning Mastery*. Published April 30, 2020. Accessed May 5, 2022. <https://machinelearningmastery.com/adaboost-ensemble-in-python/>

- [6]Majumder P. Gaussian Naive Bayes. OpenGenus IQ: Computing Expertise & Legacy. Published February 23, 2020. Accessed May 5, 2022. <https://iq.opengenus.org/gaussian-naive-bayes/>
- [7]Abirami S, Chitra P. Energy-efficient edge based real-time healthcare support system. *Advances in Computers*. Published online 2020:339-368. doi:10.1016/bs.adcom.2019.09.007
- [8]Forina M, Casale M, Oliveri P. Application of Chemometrics to Food Chemistry. *Comprehensive Chemometrics*. Published online 2009:75-128. doi:10.1016/b978-044452701-1.00124-1
- [9]XGBoost — H2O 3.36.1.1 documentation. Docs.h2o.ai. Published 2016. Accessed May 5, 2022. <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/xgboost.html>
- [10]Mushtaq MS, Mellouk A. Methodologies for Subjective Video Streaming QoE Assessment. *Quality of Experience Paradigm in Multimedia Services*. Published online 2017:27-57. doi:10.1016/b978-1-78548-109-3.50002-3
- [11]Christian Dorion & Yoshua Bengio, 2003. "Stochastic Gradient Descent on a Portfolio Management Training Criterion Using the IPA Gradient Estimator," CIRANO Working Papers 2003s-23, CIRANO.
- [12]Caie PD, Dimitriou N, Arandjelović O. Precision medicine in digital pathology via image analysis and machine learning. *Artificial Intelligence and Deep Learning in Pathology*. Published online 2021:149-173. doi:10.1016/b978-0-323-67538-3.00008-7
- [13]Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002;38(4):367-378. doi:10.1016/s0167-9473(01)00065-2
- [14]Jabbar, M. A. . (2021). Breast Cancer Data Classification Using Ensemble Machine Learning. *Engineering and Applied Science Research*, 48(1), 65–72. Retrieved from <https://ph01.tci-thaijo.org/index.php/easr/article/view/234959>
- [15]Anastraj, K., et al. "BREAST CANCER DETECTION EITHER BENIGN OR MALIGNANT TUMOR USING DEEP CONVULSIONAL NEURAL NETWORK WITH MACHINE LEARNING TECHNIQUES."



- [16]S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, “Variability Measurement for Breast Cancer Classification of Mammographic Masses,” in Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS), Prague, Czech Republic, 2015, pp.177–182
- [17]T. M. Mejia, M. G. Perez, V. H. Andaluz, and A. Conci, “Automatic Segmentation and Analysis of Thermograms Using Texture Descriptors for Breast Cancer Detection,” 2015 Asia-Pacific Conf.Comput. Aided Syst. Eng., pp. 24 – 29, 2015.
- [18]T. K. Avramov and D. Si, “Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis,” Proc. Int.Conf. Comput. Data Anal. - ICCDA '17, pp. 69 – 74, 2017.
- [19]Z. Jiang, and W. Xu, “Classification of benign and malignant breastcancer based on DWI texture features,” ICBCI 2017 Proceedings of the International Conference on Bioinformatics and Computational Intelligence 2017.
- [20]M. Ngadi, A. Amine, and B. Nassih, “A Robust Approach forMammographic Image Classification Using NSVC Algorithm,” Proc. Mediterr. Conf. Pattern Recognit. Artif. Intell. - MedPRAI-2016 , pp.44 – 49, 2016.
- [21]Assiri, Adel S et al. “Breast Tumor Classification Using an Ensemble Machine Learning Method.” *Journal of imaging* vol. 6,6 39. 29 May. 2020, doi:10.3390/jimaging6060039
- [22]V. Bevilacqua, A. Brunetti, M. Triggiani, D. Magaletti, M. Telegrafo, and M. Moschetta, “An Optimized Feed-forward Artificial Neural Network Topology to Support Radiologists in Breast Lesions Classification,” Proc. 2016 Genet. Evol. Comput. Conf. Companion - GECCO '16 Companion, pp. 1385 – 1392, 2016.
- [23]M. U. Salma, “Fast Modular Artificial Neural Network for theClassification of Breast Cancer Data,” Proc. Third Int. Symp. WomenComput. Informatics - WCI '15 , pp. 66 – 72, 2015.
- [24]UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. Uci.edu. Published 2022. Accessed May 9, 2022. [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))

[25][26][27]Baratloo, Alireza et al. “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.” *Emergency (Tehran, Iran)* vol. 3,2 (2015): 48-9.

[28]How to Use ROC Curves and Precision-Recall Curves for Classification in Python. Machine Learning Mastery. Published August 30, 2018. Accessed May 11, 2022. <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>

[29]Joos Korstanje. The F1 score | Towards Data Science. Medium. Published August 31, 2021. Accessed May 11, 2022. <https://towardsdatascience.com/the-f1-score-bec2bbc38aa6>