**Abstract**

Nearly 17.9 million people died from cardiovascular disease (CVD) in 2019, accounting for 32% of all fatalities worldwide. One of the leading causes of death in the globe is cardiovascular disease. Out of these, heart attack and stroke caused 85% of the deaths. On another account, breast cancer has been the most common cancer in women in the United States. 685 000 people worldwide died in 2020 as a result of breast cancer, which affected 2.3 million women. The most common disease in the globe as of the end of 2020 was breast cancer, which had been diagnosed in 7.8 million women in the previous five years. Typically, middle-aged and older women are the ones who develop breast cancer. Recently, machine learning models have become a key methodology in detection of cancers and cardiovascular diseases. These solutions give medical practitioners diagnostic support and indicators. In this work we have compare various machine learning classification algorithms, apply them to both types of diseases’ dataset, and examine how these algorithms perform when subjected to either of the classes of these diseases in order to contribute to the research and study of cardiovascular and breast cancer through computer aided diagnosis (CAD). Our two main goals in this work are to first offer an automated machine learning ensemble model for categorizing cardiovascular and breast malignancies, and second to compare the performance of several classification algorithms to find the best classifier for the task. The proposed technique was specifically developed as a potential support for clinical care based on patient diagnostic data. Our strategy was developed and tested on one cardiovascular and one breast cancer public datasets with heterogeneous characteristics. By performing a comprehensive analysis of binary classification in both the types of diseases, the proposed approach can exceed state of the art in both the tasks, reaching an accuracy of 98.35% in detection of breast cancer and 94.28% in detection of heart diseases averaged over various iterations of model training. We are confident that incorporating the suggested ensemble methods would produce stable and dependable CAD systems. Being tested on breast malignancies and cardiovascular diseases, our strategy can be easily extended to other diseases as well. In conclusion, we found that our machine learning ensemble strategy could effectively be applied for this task with a considerable potential benefit to healthcare for patient management.

## **Introduction:**

We currently live in a time where artificial intelligence has a significant influence on our daily lives. Systems based on machine learning (ML) and deep learning (DL) have revolutionized a number of sectors, including industry, transportation, and government. Machine learning has recently demonstrated state-of-the-art performance in a variety of disciplines, including fraud detection, image identification, and product recommendation. Similarly, the healthcare domain has benefited the most from machine learning and deep learning. Healthcare, a sector that has historically been resistant to significant technological shocks, is now starting to be affected by ML algorithms. Recently, ML techniques have demonstrated outstanding results in a variety of tasks, including the identification of body organs from medical images, the classification of interstitial lung diseases, the detection of lung nodules, the reconstruction of medical images, and the heart, to name a few.

Worldwide, medical organizations gather information on a range of health-related areas. To extract insightful knowledge from these data, multiple machine learning approaches can be used. However, the amount of data gathered is enormous, and it is frequently highly noisy. These datasets can be easily investigated using a variety of machine learning techniques, even though they are too large for human minds to grasp. As a result, these algorithms have recently proven to be highly helpful in precisely predicting the existence or absence of different diseases and disorders.

The most comprehensive solution to numerous machine learning problems is ensemble methods. Improving generalization performance has always been one of the key goals in the machine learning field. Ensemble techniques involve training numerous models and integrating their predictions to enhance the predictive performance of a single model. Due to its strong generalization potential, ensemble learning has attracted a lot of study attention for more than 20 years. Instead of training just one base classifier, this method trains a number of them, and then uses a fusion strategy to merge their outputs. According to other empirical research the combination of multiple classifiers typically increases the generalized performance when compared to a single classification algorithm and this technique is followed by ensemble methods because in ensemble classifiers, predictions are made using the cumulative output of each model to increase accuracy. Compared to individual classifiers, ensemble classifiers can perform better. In this paper we have focused on doing a comparative analysis of ensemble methods in the healthcare domain. We have done the analysis on well-known healthcare datasets of breast cancer and heart disease and applied major ensemble machine learning algorithms like Random forest, AdaBoost, and GradientBoost along with a few other traditional classification algorithms like Random Forest, Decision Trees, and Support vector machines. We have also added several statistical approaches for feature engineering like SMOTE for balancing the target classes in the unbalanced datasets.

Nearly 17.9 million individuals died from cardiovascular disease (CVD) in 2019, accounting for 32% of all mortality globally. One of the biggest causes of death in the globe is cardiovascular disease. Out of these, heart attack and stroke caused 85% of the deaths. On another account, breast cancer has been the most frequent malignancy in women in the United States. 685000 women worldwide died in 2020 as a result of breast cancer, which affected 2.3 million women. The most frequent ailment in the globe as of the end of 2020 was breast cancer, which had been diagnosed in 7.8 million women in the previous five years. Typically, middle-aged and older women are the ones who acquire breast cancer.

Recently, machine learning models have become a crucial tool in identification of malignancies and cardiovascular disorders. These implementations result in diagnostic support and indicators to health care practitioners for right cure and prevention of these diseases. In this study, we selected two datasets, one with data on breast cancer and the other with data on heart disease. We then applied various ensemble machine learning algorithms and traditional machine learning algorithms to both datasets, compared the results with those from prior studies, and then presented the study that applied these algorithms to both types of diseases and examined how they performed under either of the classes of disease of each data set.

The sections of this paper are divided into different categories, and they all primarily focus on using the ensemble methods indicated above to predict heart disease and breast cancer. In section 2 we have discussed the fundamentals of heart disease and breast cancer research, as well reviewed some earlier pertinent literature and a brief literature review of classic ML algorithms. The research methodologies are discussed in section 3. Section 4 provides performance evaluation metrics of both the selected datasets which includes metrics like accuracy, recall, precision and AUC score. Following section 5, we have discussed and analyzed the outcomes of prediction results from both the datasets. Finally in section 6, we have concluded the research and discussed the future scope of this work. *[This paragraph is subject to change as we proceed in writing the paper.]*

## **Literature Survey:**

Methodology

**Table 1.** Statistical details of Heart Disease Dataset

**Dataset**:

HEART DISEASE:

We utilized data from two separate domains, heart disease and breast cancer. In this study, we obtain the heart disease dataset from the UCI machine learning repository regarding heart illness. We have a total of 303 instances, 164 of which belong to the healthy and 139 to those with cardiac disease. 14 clinical features are collected for each case. The dataset contains 165 cases of heart disease and 138 records with no heart disease, resulting in a minor imbalance. It consists of four distinct databases, although only the

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Features 🡪***  ***Statistics ↓*** | ***age*** | ***sex*** | ***cp*** | ***trestbps*** |  | ***chol*** | ***fbs*** | ***restecg*** | ***thalach*** | ***exang*** | ***oldpeak*** | ***slope*** | ***ca*** | ***thal*** | ***target*** |
| **count** | 303 | 303 | 303 | 303 |  | 303 | 303 | 303 | 303 | 303 | 303 | 303 | 303 | 303 | 303 |
| **mean** | 54.3 | 0.68 | 0.96 | 131.62 |  | 246.26 | 0.14 | 0.52 | 149.64 | 0.32 | 1.03 | 1.39 | 0.72 | 2.31 | 0.54 |
| **std** | 9.08 | 0.46 | 1.03 | 17.53 |  | 51.83 | 0.35 | 0.52 | 22.90 | 0.46 | 1.16 | 0.61 | 1.02 | 0.61 | 0.49 |
| **min** | 29.0 | 0.00 | 0.00 | 94.00 |  | 126.00 | 0.00 | 0.00 | 71.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **25%** | 47.5 | 0.00 | 0.00 | 120.00 |  | 211.00 | 0.00 | 0.00 | 133.50 | 0.00 | 0.00 | 1.00 | 0.00 | 2.00 | 0.00 |
| **50%** | 55.0 | 1.00 | 1.00 | 130.00 |  | 240.00 | 0.00 | 1.00 | 153.00 | 0.00 | 0.80 | 1.00 | 0.00 | 2.00 | 1.00 |
| **75%** | 61.0 | 1.00 | 2.00 | 140.00 |  | 274.50 | 0.00 | 1.00 | 166.00 | 1.00 | 1.60 | 2.00 | 1.00 | 3.00 | 1.00 |
| **max** | 77.0 | 1.00 | 3.00 | 200.00 |  | 564.00 | 1.00 | 2.00 | 202.00 | 1.00 | 6.20 | 2.00 | 4.00 | 3.00 | 1.00 |

UCI Cleveland dataset was utilized. This database has a total of 76 properties, although all published experiments utilize only 14 features. For our study, we utilized the already-processed UCI Cleveland dataset available on the Kaggle website. Table 1 provides a comprehensive statistical overview of the 14 qualities used in the proposed work. Dataset used was the Heart disease Dataset which is a combination of 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features [9]. Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis.

**Table 2.** Heart Disease Dataset Attribute Description

|  |  |  |
| --- | --- | --- |
| ***Feature*** | ***Description*** | ***Value Type*** |
| **Age** | Represents the age of the person | Numerical |
| **Sex** | Represents the person's gender; 0 denotes a female and 1 a male. | Categorical |
| **Cp** | Symbolizes the patient's level of chest pain. | Categorical |
| **Trestbps** | It displays the BP of the patient. | Numerical |
| **Chol** | It displays the patient's cholesterol level of the patient. | Numerical |
| **Fbs** | It represents the patient's fasting blood sugar level. | Categorical |
| **Restecg** | It displays the ECG's result. | Categorical |
| **Thalach** | Accomplishment of maximum heart rate | Numerical |
| **Exang** | Used to determine whether an exercise-induced angina is present. If yes, then 1; if not, 0 | Categorical |
| **Oldpeak** | Describes the patient's level of depression. | Numerical |
| **Slope** | Describes the state of the patient at the height of exercise. It is categorized into three values (Unsloping, Flat, and Down Sloping). | Categorical |
| **Ca** | It represents the result of fluoroscopy. | Categorical |
| **Thal** | Test require for patient suffering from pain in chest or difficulty in breathing. There are 4 values which represent the Thallium test. | Categorical |
| **Target** | It is the dataset's last column. There are only two classifications in this dataset's binary classification (0, 1). Class "0" indicates that there is a low likelihood of heart disease, while class "1" indicates a significant likelihood of heart disease. The other 13 attributes determine whether "0" or "1" is used. | Categorical |

BREAST CANCER:

The Wisconsin breast cancer database was used and analyzed in this study. Dr. William H. Wolberg gathered them at the hospitals affiliated with the University of Wisconsin Madison. This database contains 699 records. The WBCD data consists of 699 patient records as of July 15, 1992. This collection provides example code numbers and 10 other attribute for each case. Clump, Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses are the independent characteristics. All of them are represented by integers between 1 and 10. Class is the dependent attribute, which is represented by the integers 2 and 4, where 2 represents malignant and 4 representing benign breast. There are 458 (65.5%) benign entries and 241 (34.5%) malignant records in this database, which shows the imbalance in the data. The dataset is downloaded from the UCI machine learning repository's official website. Table 3 provides the statistical details of each feature in the dataset.

**Table 3.** Statistical details of Breast Cancer Dataset

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Features ->*** | ***Sample code number*** | ***Clump Thickness*** | ***Uniformity of Cell Size*** | ***Uniformity of Cell Shape*** | ***Marginal Adhesion*** | ***Single Epithelial Cell Size*** | ***Bare Nuclei*** | ***Bland Chromatin*** | ***Normal Nucleoli*** | ***Mitoses*** | ***Class*** |
| **count** | 6.83e+02 | 683 | 683 | 683 | 683 | 683 | 683 | 683 | 683 | 683 | 683 |
| **mean** | 1.08e+06 | 4.442167 | 3.150805 | 3.215227 | 2.830161 | 3.234261 | 3.544656 | 3.445095 | 2.869693 | 1.603221 | 2.699854 |
| **std** | 6.21e+05 | 2.820761 | 3.065145 | 2.988581 | 2.864562 | 2.223085 | 3.643857 | 2.449697 | 3.052666 | 1.732674 | 0.954592 |
| **min** | 6.34e+04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| **25%** | 8.78e+05 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| **50%** | 1.17e+06 | 4 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 2 |
| **75%** | 1.24e+06 | 6 | 5 | 5 | 4 | 4 | 6 | 5 | 4 | 1 | 4 |
| **max** | 1.35e+07 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 4 |

**Data preprocessing**:

In addition to the methods employed, the quality of the dataset and the preprocessing techniques also influence the performance and precision of the prediction model. Preprocessing refers to the procedures performed on a dataset before machine learning techniques are deployed. The preprocessing phase is crucial because it prepares the dataset and transforms it into a format that the algorithm can interpret.

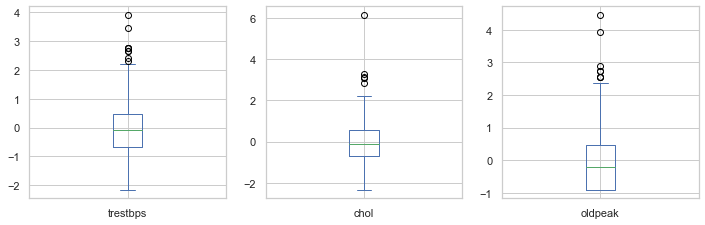
It is possible for datasets to contain errors, missing data, redundancy, noise, and other issues that render the data unsuitable for direct usage by the machine learning algorithm. The size of the dataset is an additional issue. Some datasets have a large number of attributes, which makes it more difficult for the algorithm to examine the data, detect patterns, and generate correct predictions. By examining the dataset and employing the appropriate data pretreatment procedures, it is possible to resolve such issues. Data preprocessing steps consist of data cleansing, data transformation, missing values imputation, data normalization, and feature selection, as well as other steps depending on the nature of the dataset.

The first and most critical criterion for the development of effective machine learning models is the quality of the data. Even on large data sets, the speed, generalization ability, and execution rate of machine learning algorithms are improved by the use of pre-processing methods since they help reduce the noise in the data. Feature selection, the removal of missing data, and outlier removal are three of the most common methods that are utilized in applications for data extraction and monitoring respectively. There were no null or NaN values among the 303 data rows for heart disease dataset. Due to the fact that 16 out of 699 data records in the Wisconsin breast cancer dataset had null records, we eliminated those data rows, leaving us with 683 data rows from which to conduct our analysis. The nature of the missing attributes, which are all related to the size of the bare nuclei, leads one to the initial conclusion that it is not a viable option to replace them with the mean value from the training set. This is due to the fact that the size of an individual cell is not related to the size of the mean of other cells, therefore removing those null records was the better option than filling it with mean values.

Many ML models work better when the data is distributed normally and worse when it is skewed. It is crucial to recognize the skewness that is present in the features and to carry out appropriate transformations and mappings in order to convert the skewed distribution into a normal distribution. In the heart disease dataset, chol, trestbps, oldpeak, and thalach are the only numerical values as rest of them are categorical, and there was small skewness in three of these four features, with chol and trestbps being rightly skewed, trestbps being left skewed, and the oldpeak attribute displaying a minimal level of skewness. In the Wisconsin breast cancer dataset, all values fell within the range of 1 to 10, and none of the features, with the exception of the Bland Chromatin attribute, displayed significant skewness. We applied logarithmic transformation to the skewed attributes of both datasets, excluding those whose skewness value was minimal and majority of the data was near mean value. As a result, the majority of the data for each feature on which logarithmic transformation were applied shifted closer to its respective mean, which had a substantial effect on the skewness value.

Periodically, a dataset may contain outlier values that deviate from the other data and exceed the expected range. These are referred to as outliers, and by comprehending and maybe removing them, machine learning modelling and model quality can often be enhanced which is due to the data being shifted to a given range. In the heart disease dataset, as it can be seen from the boxplot in figure 1, majority of the outliers are from trestbps, chol and oldpeak given they are of numerical type as they contains continuous values, In the Wisconsin breast cancer dataset none of numerical features showed any significant outlier which may be caused as present of most of categorical values in the breast cancer dataset. Thus we performed outlier removal only on heart disease dataset on the basis of attributes trestbps and chol which were not in the range of Q1 and Q3 of specified features. In the heart disease dataset, as shown by the boxplot in figure x, the majority of outliers come from the numerical features trestbps, chol, and oldpeak, which contain continuous values. In contrast, in the Wisconsin breast cancer dataset, none of the numerical features displayed a significant outlier as it may be due to the no presence of any continuous valued feature. Thus, we only removed outliers from the heart disease dataset based on the parameters trestbps and chol that did not fall within the ranges Q1 and Q3 of the given features.

**Table 1.** Outlier in Heart Disease dataset



Even after outlier removal, both datasets retain a small number of outliers; therefore, if the data in any condition contains data points that are far apart, scaling is a technique used to bring them closer together, or in simpler terms, scaling is used to generalize data points so that the distance between them is reduced. Standardization functions by centering data points around the mean of all data points displayed in a feature with a unit standard deviation. This means that the data point will have a mean of zero and a standard deviation of one. The StandardScaler() function from the Sklearn library was used to normalize the continuous value features of both datasets, chol, trestbps, thalach, oldpeak from the heart disease dataset and Bare Nuclei ,Clump Thickness from Wisconsin breast cancer dataset were standardized based on the remaining outliers after eliminating the major outliers based on the features respective ranges. The data sets was standardized using Equation 1. given below.



Occasionally, the ratio of classes in a dataset may not be optimal, resulting in data imbalances that are addressed by various undersampling strategies. Data imbalance can lead to the loss of essential information about the data itself, which may be required for constructing rule-based classifiers such as Random Forests and a few other ensemble machine learning models. The class imbalance problem often arises when some classes are significantly more prevalent than others. In such situations, traditional classifiers tend to be overpowered by the huge classes and disregard the little ones. The sample selected through random undersampling may be skewed. And in that scenario, it will not be an accurate picture of the people. In such situations, we typically employ strategies that replicate minority data points at random to raise their count. This may frequently result in overfitting. Synthetic Minority Oversampling Technique (SMOTE) [18] is a more advanced oversampling technique that develops new minority class samples to balance the data distribution. We have applied SMOTE to balance the under sampled class in our case. In the heart disease dataset, there are 138 samples representing a low likelihood of heart disease and 165 samples showing a high likelihood of heart illness. Here we can observe that there is a little data imbalance between classes 1 and 0 in the proportion of around 54:46. Similarly, the count of class 2 in the Wisconsin breast cancer dataset is 458, representing benign breasts, and the count of class 4 is 241, representing a malignant class of cancer type. Consequently, the imbalance ratio between class 2 and class 4 is around 65:35, indicating that class 2 is in majority and class 4 is under sampled. As a result, we applied SMOTE to the Wisconsin breast cancer dataset, and table 4 displays the result of applying SMOTE to balance the under sampled class.

Applying SMOTE on WBCD

|  |  |  |
| --- | --- | --- |
| **Class** | **Before SMOTE** | **After SMOTE** |
| ***2 - Benign*** | 444 | 444 |
| ***4 - Malignant*** | 239 | 444 |

**Feature Engineering**

Feature selection is an important step in building a classification model. It is advantageous to limit the number of input attributes in a classifier in order to have good predictive and less computationally intensive models [17].

Two of the fourteen features in the data set, age and sex, are used to identify the patient's personal information; as a result, we eliminated them from our dataset. The remaining eleven features are considered vital since they contain crucial clinical records. Clinical data are essential for diagnosing cardiac disease and determining its severity. Similarly, in the WBC dataset, the Sample Code number was merely a unique identifier for the data samples and did not transmit any crucial breast cancer-related information; hence, it was eliminated.

Figure – distplot of important (high correlated features)

The relationship between each characteristic in the two datasets has been visualized using Pearson's correlation analysis. This helps to identify the feature in the datasets that is closely associated with the class feature. Figure 3 and figure 4 depict the Pearson's correlation matrix for each feature of the heart disease and breast cancer dataset. As shown in Figure 3, which is the correlation matrix of the heart disease dataset, cp, thalach, and chol have the highest magnitude of correlation values, with cp and thalach being positively correlated and chol being negatively correlated; therefore, they are the factors that have the greatest impact on the likelihood of developing heart disease. None of the features showed a high correlation with one another, and they all had a decent correlation with the target class; therefore, we did not exclude any of the features.

Similarly, in Figure 4, Uniformity of Cell Size, Uniformity of Cell Shape, and Bare Nuclei are the features affecting breast cancer detection the most with a positive correlation value of 0.82; all other features also show a strong correlation value with Class features; therefore, none of the features have been eliminated. However, Uniformity of Cell Size and Uniformity of Cell Shape are significantly connected with each other with a strong positive correlation value of 0.91; as a result, we have removed the Uniformity Cell Shape attribute from the target class i.e. Class.

Figure – breast cancer correlation matrix and heart disease correlation matrix

Correlation table can also be included

**Model Training and Disease Classification**

This is the most critical step, during which a model for predicting both the disease class, that is, whether a person has heart disease or not, and whether a woman has breast cancer or not, is constructed. For both diseases class prediction, we have implemented a number of machine learning methods. Our problem statement is a binary class classification problem since we have to predict whether a person has heart disease or not, or a woman has breast cancer or not, as there is only two possibility. The Classification algorithm is a Supervised Learning method for classifying incoming observations according to previously established criteria. Classification is the process through which a computer learns from a dataset or set of observations and then uses that knowledge to sort new observations into predetermined categories. Examples include "yes" or "no," "zero" or "one," "spam" or "not spam," "cat" or "dog," etc. Depending on the context, classes may also be referred to as labels or goals. Classification differs from regression in that its outcome is not a numeric result but rather a categorical one: "Green" or "Blue," "Fruit" or "Animal," etc. The Classification method is a form of supervised learning, thus it requires labelled input data—that is, data with both an input and an expected output, which, in our case, is the target label in the heart disease dataset and the class label in the breast cancer dataset. In the section that follows, we will explore all of the models that are used to forecast the AQI class and their characteristics, which they utilize to predict a specific class from a given dataset.

Ensemble-based classification algorithms are one of the most extensively used classification techniques for data streams. Their popularity is a result of their superior performance in compared to single-learner systems and their ease of deployment in real-world applications. Ensemble algorithms are particularly advantageous for data stream learning because they may be combined with algorithms for drift detection and dynamic updates, such as the selective removal or addition of classifiers. Our work for both datasets is also focused on categorization utilizing various ensemble models and comparing them to traditional machine learning methods. We specifically used 10 machine learning models to classify the presence of disease in both datasets, including ensemble models such as Adaboost, GradientBoost, XgBoost, LightGBM, Random forest, and some other classical machine learning models such as Support Vector Classifier, Decision Tree, and K-nearest neighbors.

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Ensemble modelling is a process in which numerous varied models are developed to predict an outcome, either by employing a variety of modelling techniques or by employing a variety of training data sets. The ensemble model then aggregates each base model's forecast, yielding a single final prediction for the unseen data. The goal of employing ensemble models is to reduce prediction generalization error. When the ensemble approach is utilized, the prediction error of the model lowers as long as the basis models are diverse and independent. In creating a prediction, the technique seeks the wisdom of the multitude. Despite the fact that the ensemble model has numerous base models, it functions and performs as a single model. The majority of practical classification problems uses ensemble model.

*Bagging*:

In Bagging which is also called as bootstrap aggregation, we build a variety of base models, such as M1, M2, M3, ..., MK. If a data collection containing n records is given, row sampling with replacement is the process of extracting a selection of rows from the provided data set for each model, which may also be duplicated, where Di is the data sample for each base model, where Di ⊆ n. Once all models have been trained using the supplied data, then majority of the results from all the base models using test model are taken into consideration as the final result. Random Forest is an example of bagging technique. (Bagging diagram paste)

*Boosting*:

Boosting is an ensemble modelling technique that attempts to construct a strong classifier from a collection of weak classifiers. It is accomplished by developing a model in series utilizing weak models. First, a model is constructed using the training data. The second model is then constructed in an attempt to address the errors in the previous model. This approach is repeated until either the whole training data set is properly predicted or the maximum number of models are added. Adaboost, Xgboost are some of the examples of boosting algorithms. Boosting is an ensemble modelling method that is frequently used to solve binary classification issues. By transforming a number of weak learners into strong learners, these methods increase prediction ability. (paste boosting diagram)

Logistic Regression:

Logistic regression is one of the most common supervised technique out of all the machine learning algorithms. Using a given collection of independent factors, it is used to predict the categorical dependent variable. The outcome of a categorical dependent variable is predicted by logistic regression. The result must therefore be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc., but instead of providing the exact value like 0 or 1, probabilistic values between 0 and 1 are provided. Logistic Regression is quite similar to Linear Regression, with the exception of their respective applications. Linear Regression is used to solve problems involving Regression, while Logistic Regression is used to solve problems involving Classification.

Decision Tree Classifier:

Random Forest Classifier:

Random Forest is a popular machine learning method that belongs to the supervised learning technique. It is applicable to both Classification and Regression problems in machine learning but majorly used for classification problem. It is based on ensemble learning, which is the process of integrating numerous classifiers to solve a complex problem and improve the model's performance. Random Forest is a classifier that takes the average of a number of decision trees on various subsets of a given dataset to enhance the predicted accuracy of that dataset. The increasing number of trees in the forest increases precision and eliminates the issue of overfitting. Before producing an output, the Random Forest approach takes into account several decision trees. As a result, it is essentially an ensemble of decision trees. This strategy is based on the assumption that more trees will lead to the correct decision. In classification, it employs a voting mechanism to determine the class, whereas in regression, it takes the mean of all the decision tree outputs. It performs effectively with huge, high-dimensional datasets.

Support Vector Classifier:

The support vector machine, also known as SVM for short, is a sophisticated supervised algorithm that can deal with both regression and classification problems, but it performs classification work better than regression. It is able to deal with a number of continuous and categorical cases simultaneously. For classification problems the purpose of the Support Vector Machine (SVM) technique is to generate the best line or decision boundary that can divide an n-dimensional space into classes. This will allow us to easily place any new data points in the appropriate category in the future. A hyperplane is the term used to describe this optimal decision boundary. The extreme points and vectors that contribute to the creation of the hyperplane are selected using SVM. These extreme cases are referred to as support vectors, which is how the method got its name: Support Vector Machine.

Naïve Bayes Classifier:

The Naive Bayes Algorithm is a form of supervised learning that is used to solve classification issues; it is based on Bayes' theorem. Its primary application is in high-dimensional training datasets for text classification. One of the simplest and most effective Classification algorithms, Naive Bayes Classifier facilitates the development of rapid, predictive machine learning models. It makes predictions based on the object's likelihood because it is a probabilistic classifier.

K Nearest Neighbors:

The k-nearest neighbors algorithm (k-NN) in statistics is a non-parametric supervised learning approach invented by Evelyn Fix and Joseph Hodges in 1951 and later expanded by Thomas Cover. Its applications include classification and regression. The input in both cases consists of the k closest training examples in a data collection. The outcome of k-NN classification is a described. A plurality vote of its neighbors classifies an object, with the object assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, the item is simply assigned to the class of the object's single nearest neighbor.

Adaboost:

The statistical classification meta-algorithm known as AdaBoost, or "adaptive boosting," was developed by Yoav Freund and Robert Schapire in 1995. It can be used in conjunction with many different learning algorithms to enhance performance. The results of the other learning algorithms, or "weak learners," are merged to create a weighted total that represents the boosted classifier's final results. Although AdaBoost can be used to many classes or bounded intervals on the real line, it is often used for binary classification. AdaBoost is adaptive in that it modifies future weak learners in favour of instances that prior classifiers incorrectly classified. It may be less prone to the overfitting issue than other learning algorithms in particular situations. It can be demonstrated that the final model converges to a strong learner even if the performance of each individual learner is just marginally better than random guessing. AdaBoost is frequently used to combine weak base learners (like decision stumps), but it has been demonstrated that it can also combine strong base learners (like deep decision trees) well, leading to an even more precise model.

Gradient Boost:

A popular boosting approach is gradient boosting. Each predictor in gradient boosting corrects the error of its predecessor. Unlike Adaboost, each predictor is trained using the residual errors of the predecessor as labels rather than adjusting the weights of the training instances. The Gradient Boosted Trees approach uses CART as its base learner (Classification and Regression Trees).

It provides a prediction model in the form of an ensemble of decision trees-like weak prediction models. The resulting technique, known as gradient-boosted trees, typically beats random forest when a decision tree is the weak learner. The construction of a gradient-boosted trees model follows the same stage-wise process as previous boosting techniques, but it generalizes other techniques by enabling the optimization of any differentiable loss function.

XgBoost:

Extreme Gradient Boosting is the abbreviation for XGBoost. XGBoost is a distributed gradient boosting library that has been developed to be very effective, adaptable, and portable. To prevent over-fitting in XGBoost, regularization helps to smooth the final learned weights. The regularized objective will typically choose a model that makes use of straightforward and predictive functions. Two further methods are utilized in addition to the regularized target to reduce overfitting even more. The first method is shrinkage, which Friedman first proposed. After each stage of tree boosting, shrinkage scales newly acquired weights by an amount equal to (Nita). Shrinkage reduces the impact of each tree and makes room for new trees to enhance the model, much like a learning rate in stochastic optimization.

Light Gradient Boosting Machines:

A gradient boosting system called Light GBM makes use of a tree-based learning technique. While other algorithms grow trees horizontally, Light GBM grows trees vertically, which translates to Light GBM growing trees leaf-wise while other algorithms grow levels-wise. The leaf with the greatest delta loss will be chosen to grow. Leaf-wise method can reduce loss more than a level-wise strategy when expanding the same leaf. Large amounts of data can be handled using Light GBM, which requires less memory to operate. The popularity of Light GBM is also attributed to its emphasis on precise outcomes. Data scientists frequently use LGBM to build data science applications since it also enables GPU learning.

Cross Validation:

A statistical technique called cross-validation is used to assess the effectiveness (or accuracy) of machine learning models. It serves as a safeguard against overfitting in predictive models, especially when the available data may be scarce. As part of cross-validation, we created a fixed number of folds (or partitions) of the data, called k, for training each model on each dataset. The model will be trained on (k-1) fold data, and 1 fold will be used for testing for each number of fold. In this way, all the folds of the data will be used for testing and training on data, improving the models' predictability.

GridSearchCV:

In machine learning, machine undertake this exploration and select the optimal model architecture automatically. Alternatively, we can manually tune each parameter for each machine learning model. The search for the optimal model architecture is referred to as hyper parameter tuning, as the parameters that form the model architecture are known as hyper parameters.

GridSearchCV is a method for adjusting hyper parameters to find the best values for a particular model, as  a model's performance is strongly affected by the value of its hyper parameters. Furthermore, it is necessary to attempt all potential values in order to determine the optimal values because there is no way to predict in advance what the best values for hyper parameters would be. We utilized GridSearchCV to automate the tweaking of hyper parameters because doing it manually could take a lot of time and resources. The model selection package of Scikit-learn (or SK-learn) contains a method called GridSearchCV which we have used to perform hyper parameter tuning. We were able to fit the estimator model with the best score on our training data set using GridSearchCV in looping over predefined hyper parameters. We assigned all the list of chosen parameters in the param\_grid attribute of GridSearchCV function, and the attributes which produced the highest prediction score can be seen in table 7.

**Table 7.** Hyperparameters used for each model

|  |  |  |
| --- | --- | --- |
| **Model** | **Dataset** | |
| **Heart Disease** | **Breast Cancer** |
| *Logistic Regression* | max\_iter=5000, intercept\_scaling=0.6 | max\_iter=4500, intercept\_scaling=0.9 |
| *Decision Tree Classifier* | 'criterion': 'entropy',  'max\_depth': 5,  'min\_samples\_leaf': 4,  'min\_samples\_split': 6 | 'criterion': 'entropy',  'max\_depth': 7,  'min\_samples\_leaf': 2,  'min\_samples\_split': 5 |
| *Random Forest Classifier* | 'bootstrap': False,  'max\_depth': 2,  'max\_features': 'sqrt',  'min\_samples\_leaf': 2,  'min\_samples\_split': 2,  'n\_estimators': 72 | 'bootstrap': True,  'max\_depth': 4,  'max\_features': 'auto',  'min\_samples\_leaf': 2,  'min\_samples\_split': 2,  'n\_estimators': 64 |
| *Support Vector Classifier* | kernel='linear', probability=True, shrinking = False, cache\_size=200, verbose=False | kernel='linear', probability=True, shrinking = False, cache\_size=200, verbose=False |
| *Naïve Bayes* | var\_smoothing= 2e-9 | var\_smoothing= 2e-11 |
| *KNN* | n\_neighbors = 5, weights = 'uniform',algorithm = 'brute',metric = 'minkowski' | n\_neighbors = 4, weights = 'uniform',algorithm = 'brute',metric = 'minkowski' |
| *AdaBoost* | 'algorithm': 'SAMME.R', 'learning\_rate': 0.1, 'n\_estimators': 50 | 'algorithm': 'SAMME.R', 'learning\_rate': 0.1, 'n\_estimators': 150 |
| *Gradient Boosting* | 'learning\_rate': 0.05, 'n\_estimators': 75, criterion: ’friedman\_mse’, subsample=1.0 | 'learning\_rate': 0.5, 'n\_estimators': 150, criterion: ’friedman\_mse’, subsample=1.0 |
| *XgBoost* | 'learning\_rate': 0.1, 'max\_depth': 1, 'n\_estimators': 50 | 'learning\_rate': 0.1, 'max\_depth': 4, 'n\_estimators': 150 |
| *LightGBM* | 'learning\_rate': 0.1, 'n\_estimators': 100, 'num\_leaves': 20 | 'learning\_rate': 0.05, 'n\_estimators': 200, 'num\_leaves': 20 |

***Model Performance Evaluation*:**

The performance of the suggested machine learning model framework for AQI prediction class is evaluated using the five criteria listed below:

Accuracy: The predicted value resulting from dividing the total of True Positive and True Negative by the sum of True Positive, False positive, False Negative, and True Negative values of a confusion matrix.

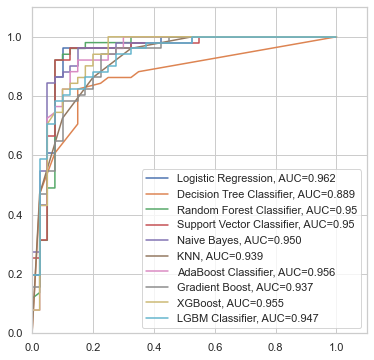
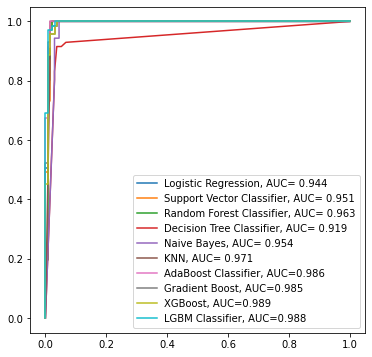
Precision: The result of dividing True Positive by the total of True Positive and False Positive of a confusion matrix.

Recall: Sometimes, Sensitivity is also known as Recall. It is the result of dividing True Positive by the total of True Positive and False Negative in a confusion matrix.

F-Measure: The F1 score can be calculated by multiplying the recall and precision of a confusion matrix and then dividing that result by the total recall and precision of the matrix. After that, the result is multiplied by two.

Performance metrics score of trained models are depicted in table 9 for specified models.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset ->  Models | | Heart Disease | | | | Breast Cancer | | | |
| Class | Performance Metrics | | | Class | Breast Cancer | | |
| Precision | F1-Score | Accuracy | Precision | F1-Score | Accuracy |
| Logistic  Regression | 0 | 0.94 | 0.93 | 94.16 | 2 | 0.94 | 0.94 | 94.5 |
| 1 | 0.94 | 0.95 | 4 | 0.93 | 0.95 |
| Decision Tree Classifier | 0 | 0.78 | 0.85 | 86.63 | 2 | 0.92 | 0.91 | 90.56 |
| 1 | 0.87 | 0.86 | 4 | 0.90 | 0.91 |
| RFC | 0 | 0.92 | 0.93 | 94.01 | 2 | 0.98 | 0.97 | 96.7 |
| 1 | 0.96 | 0.94 | 4 | 0.96 | 0.97 |
| SVC | 0 | 0.94 | 0.92 | 93.66 | 2 | 0.94 | 0.95 | 95.32 |
| 1 | 0.92 | 0.67 | 4 | 0.93 | 0.95 |
| Naïve  Bayes | 0 | 0.90 | 0.91 | 93.39 | 2 | 0.92 | 0.92 | 93.56 |
| 1 | 0.95 | 0.94 | 4 | 0.93 | 0.94 |
| KNN | 0 | 0.89 | 0.88 | 91.18 | 2 | 0.97 | 0.96 | 96.65 |
| 1 | 0.92 | 0.92 | 4 | 0.96 | 0.98 |
| Adaboost | 0 | 0.91 | 0.92 | 93.66 | 2 | 1.00 | 0.97 | 97.06 |
| 1 | 0.94 | 0.94 | 4 | 0.98 | 0.99 |
| Gradient  Boost | 0 | 0.93 | 0.90 | 93.71 | 2 | 1.00 | 0.99 | 97.05 |
| 1 | 0.94 | 0.96 | 4 | 0.97 | 0.99 |
| XgBoost | 0 | 0.91 | 0.92 | 94.28 | 2 | 0.99 | 0.97 | 98.35 |
| 1 | 0.96 | 0.95 | 4 | 0.97 | 0.98 |
| LightGBM | 0 | 0.89 | 0.90 | 92.27 | 2 | 0.96 | 0.99 | 97.78 |
| 1 | 0.93 | 0.93 | 4 | 0.99 | 0.97 |

****

## fdsfdsf

## Conclusion and Future Work

In this paper, we have offered a comparative analysis of machine learning-based computational frameworks in the application of predicting breast cancers and cardiovascular diseases. The proposed framework successfully classifies both the diseases and the type (in case of breast cancers).

Prediction of heart disease is difficult and crucial in the medical industry. However, if the disease is discovered in its early stages and preventative measures are implemented as soon as feasible, the fatality rate can be significantly reduced. In order to provide the best care, breast cancers must be correctly classified. These frameworks, when used in real-world diagnostics, will assist to reduce the number of incorrect diagnoses and identify promising treatment options.

The primary contribution of our framework is handling skewness and imbalance in data with the application of SMOTE for synthetic data augmentation for the minority class. Additionally, we have discovered that ensemble learning promotes diversity among base learners, which enhances prediction accuracy. Precision, specificity, recall, and F1 score are used to evaluate the performance of all approaches. The comparative analysis has aided us to identify XGBoost and Random Forest Classifier work best in these scenarios and are proved to be quite accurate in prediction of both the types of diseases.

In order to gather further proof that a certain classification method is appropriate for these disorders, the research might be conducted in the future on a variety of real-world datasets pertaining to different ailments. Additionally, other feature selection techniques may be used in an effort to improve these outcomes and provide a more comprehensive understanding of important characteristics in certain illnesses.

## References

[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[17] Chen, Y., Abraham, A., Yang, B.(2006), Feature Selection and Classification using Flexible Neural Tree. Journal of Neurocomputing 70(1-3): 305–313.