**BREAST CANCER PREDICTION**

**Abstract**

With the advancements in the technology and owing to its tremendous use in the medical science has undoubtedly transformed many things leading to efficient disease diagnosis and treatment. Though the advancements in the field of medical science have surely made things easier and more effective than before but timely detection and timely treatment of a disease still remains a big challenge. Breast cancer is one of the leading causes of death in women. Breast cancer is the second most severe cancer among all of the cancers already unveiled. An automatic disease detection system aids medical staffs in disease diagnosis and offers reliable, effective, and rapid response as well as decreases the risk of death. In this paper, we compare six supervised machine learning techniques named support vector machine (SVM), K-nearest neighbours(KNN), random forests, Naïve Bayes, Decision tree and logistic regression. The dataset is Wisconsin Breast Cancer dataset taken from UCI machine learning database. The performance of the study is measured with respect to accuracy, sensitivity, specificity, precision. All the parameters of all the six ML models have been compared and then best model out of the six is decided.

1. **Introduction**

A prominent issue that can be observed in today’s world is the detection of a particular disease and that too the timely detection of a particular disease. The precise diagnosis of any disease is undoubtedly a very complicated job in medical science. Any disease prediction and diagnosis requires several parameters to be analysed. Though this analysis is mostly accurate but no method can guarantee 100% accurate results.

Breast cancer has emerged as a very serious problem in females. According to World Health Organization(WHO) in 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world’s most prevalent cancer. There is more lost disability-adjusted life years (DALYs) by women to breast cancer globally than any other type of cancer.  Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life.    
  
Breast cancer mortality changed little from the 1930s through to the 1970s.  Improvements in survival began in the 1980s in countries with early detection programmes combined with different modes of treatment to eradicate invasive disease.

The problem of this paper is to predict whether the lump with the given data will be a benign or a malignant tissue in future.

Several techniques have been implemented to detect breast cancer at an early stage but nothing has been much successful due to the small size of the lump at earlier stage. Although mammography provides a much convenient and fast method to diagnose but this can also not guarantee 100% accuracy in the results.

There are other techniques as well like ultrasound, dynamic MRI for the diagnosis. But the evolution of artificial intelligence and machine learning models have created variety of scopes and applications in recent times especially in medical science. Many machine learning models have been devised which are useful in identification of breast cancer at an early stage. It can be predicted that whether the small lump can become cancerous in near future or not. Although no method can guarantee full accuracy on any problem statement but it can be somewhat useful in the prediction of the future to some extent.

The number of deaths due to breast cancer in women have never drastically reduced. Although the treatment is available but the deaths still doesn’t reduce. The main reason being the detection of the cancer is not done at the appropriate time leading to worsening of the condition. The symptoms of breast cancer are also not that significant due to which the people tend to overlook those symptoms.

The aim of this paper is to devise some ways and models to predict the possibility of cancerous lumps at an earlier stage based on the past data available so that the diagnosis can be done at the proper time and the treatment can be started at correct time. If the possibility of cancer formation will be detected at an earlier stage, then the required treatment can be provided to the patient at the earliest and that would save many innocent lives.

Though a number of modalities have been demonstrated, none of the modalities are able to provide a correct and consistent result. In mammography, the doctors should read a high volume of imaging data which reduces the accuracy. This procedure is also time-consuming, and in some worse case, detects the disease with the wrong outcome. This paper compares some machine learning techniques to detect the disease from the input features. Five supervised machine learning approaches have been used to diagnose the disease with proper outcome.

1. **Literature Review**

Ultrasound characterisation of breast masses by S. Gokhale written by proposed a system where they found that doctors have known and experienced that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, forming a lump or mass that they may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body. Some women can be at a higher risk for breast cancer because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment. In this study we use four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, Artificial Neural Network and random forest.

Harmonic imaging and real-time compounding has been shown to enhance image resolution and lesion characterisation. More recently, USG elastography seems to be quite encouraging. Initial results show that it can improve the specificity and positive predictive value of USG within the characterisation of breast masses. The reason why any lesion is visible on mammography or USG is that the relative difference within the density and acoustic resistance of the lesion, respectively, as compared to the encompassing breast tissue. [1] Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach written by Pragya Chauhan and Amit Swami proposed a system where they found that Breast cancer prediction is an open area of research. In this paper dierent machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest, support vector machine, neural network, linear model, adabost, naive bayes methods are used for prediction. An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which over- came the limitations of the classical weighted average method. Genetic algorithm based weighted average method is used forth prediction of multiple models. The comparison between Particle swarm optimisation(PSO), Dierential evolution(DE) and Genetic algorithm(GA) and it is concluded that the genetic algorithm outperforms for weighted average methods. One more comparison between classical ensemble method and GA based weighted average method and it is concluded that GA based weighted average method outperforms. On Breast Cancer Detection: An Application of Machine Learn ing Algorithms on the Wisconsin Diagnostic Dataset by the Abien Fred M. Agarap. In this paper, six machine learning algorithms are used for detection of cancer. GRUSVMmodel is used for the diagnosis of breast cancer GRUSVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbour (NN) search, Softmax Regression, and Support Vector Ma- chine (SVM) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by measuring their classification test accuracy, and their sensitivity and specificity values. The said dataset consists of features which were computed from digitised images of FNA tests on a breast mass. For the implementation of the ML algorithms, the dataset was partitioned in the following fashion 70 percent for training phase, and 30 percent for the testing phase. Their results were that all presented ML algorithms exhibited high performance on the binary classification of carcinoma, i.e. determining whether benign tumour or malignant tumour. Therefore, the statistical measures on the classification problem were also satisfactory. To further corroborate the results of this study, a CV technique such as k-fold cross-validation should be used. The appliance of such a way won't only provide a more accurate measure of model prediction performance, but it'll also assist in determining the foremost optimal hyper-parameters for the ML algorithms.

Breast Cancer Diagnosis by Die rent Machine Learning Methods Using Blood Analysis Data by the Muhammet Fatih Aslan, Yunus Celik, Kadir Sabanci, and Akif Durdu for carcinoma early diagnosis. During this paper, four divergent machine learning algorithms are used for the early detection of carcinoma. The aim of this project is to process the results of routine blood analysis with divergent ML methods. Methods used are Artificial Neural Network (ANN), Extreme Learning Machine (ELM), Support Vector Machine (SVM) and Nearest Neighbour (k-NN). Dataset is taken from the UCI library. In this dataset age, BMI, glucose, insulin, homeostasis model assessment (HOMA), leptin, adiponectin, resistin, and chemokine monocyte chemoattractant protein (MCP1) attributes were used. Parameters that have the best accuracy values were found by using four divergent Machine Learning techniques. This dataset includes age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resisting and MCP1 features that can be acquired in routine blood analysis. The significance of these data in breast cancer detection was investigated by ML methods. The analysis was performed with four divergent ML methods. k-NN and SVM methods are determined using Hyper parameter optimization technique. The highest accuracy and lowest training time were given by ELM which was 80%. and 0.42 seconds. [5] Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction by Yixuan Li and Zixuan Chen used two datasets in the study. The study firstly collects the data of the BCCD dataset which contains 116 volunteers with 9 attributes and data of WBCD dataset which contains 699 volunteers and 11 attributes. Then we pre-process the raw data of WBCD dataset and obtained the info that contains 683 volunteers with nine attributes and therefore the index indicating whether the volunteer has the malignant tumour. After comparing the accuracy, Fmeasure metric and ROC curve of 5 classification models, the result has shown that RF is chosen as the primary classification model during this study. Therefore, the results of this study provide a reference for experts to distinguish the character of carcinoma. In this study, there are still some limitations that ought to be solved in further work. For instance, though there also exist some indices people haven't found yet, this study only collects the info of 10 attributes during this experiment. The limited data has an impact on the accuracy of results. Additionally, the RF can also be combined with other data mining technologies to get more accurate and efficient results in the longer term work. [6] The purpose of the paper "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study” by Mumine Kaya Keles was to predict and detect breast cancer early even if the tumour size is petite with non-invasive and painless methods that use data mining classification algorithms. Therefore, a comparison of data mining classification algorithms was made with the Weka tool. In this paper, the Weka data mining software was applied to an antenna dataset so as to examine the efficacy of data mining methods in the detection of breast cancer. The dataset that was created had 6006 rows/values, 5405 of which were used as the training dataset, while 601 were used as the test data set. The dataset was then converted to the arff format, which is the file type used by the Weka tool. The 10-fold cross-validation was then used to obtain the most authentic results using the Knowledge Extraction based on Evolutionary Learning data mining software tool. Random forest performed the best during the 10 fold cross-validation giving an average accuracy of 92.2 percent. [7] The project “Breast Cancer Prediction Using Data Mining Method” by Haifeng Wang and Sang Won Yoon is used to test the influence of feature space reduction, a hybrid between principal component analysis (PCA) and related data mining models is proposed, which applies a principle component analysis method to reduce the feature space. To evaluate the performance of these models, two widely used test data sets are used, Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995). 10- fold cross-validation method is implemented to estimate the test error of each model. PCs-SVM is the highest for WBC data that is a 97.47 percent, and PCi-ANN is the best considering accuracy for WDBC data that is 99.63%. The reason for better results from PCA pre-processing is because the principal components only represent a large part of the information in the complete data space, which to some extent can reduce data noise, as a result, feature space is enriched (elite effect).

1. **Problem Statement**

The problem statement is to predict on the basis of the past data available that whether the lump with a given data is benign i.e. not cancerous or is it malignant i.e. cancerous. This prediction will help to predict at an early age whether the lump will become cancerous in near future or not. And if there are chances of lump being cancerous in near future, then appropriate treatment can be provided to the patient for speedy recovery and to reduce the chances of deaths.

1. **Dataset used**

The data set I used in this paper is Breast Cancer Wisconsin (Diagnostic) Data Set which is available on Kaggle where the cases are either benign or malignant. The data set contains 569 rows and 32 columns. There are 33 different attributes in the data set. There are intances of benign tumours and instances of malignant tumours in the data.

**Features in the data set**

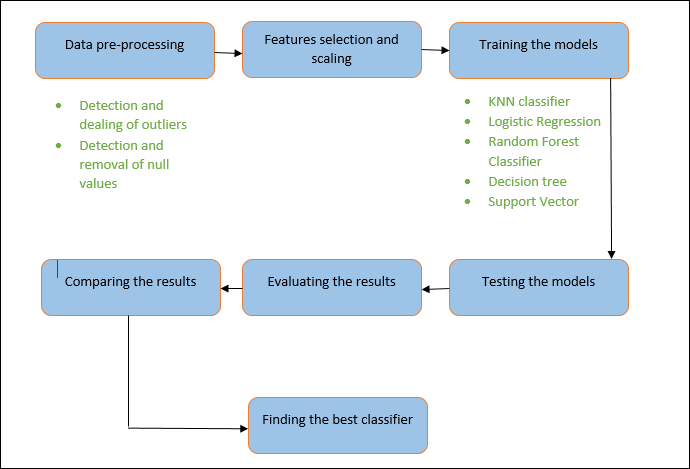
There are 32 different features in the data set. To name a few ‘radius\_mean’, ‘conavity\_mean’,’area\_mean’,’smoothness\_mean’ are some of the features present in the data set. The features are categorized into 2 categories one is mean category and the other is the worst category. The features are the information regarding the lump like the mean radius and mean area of the lump and many other parameters.

The benign cases are identified as a positive class, and the malignant cases are identified as a negative class in my research.

The reason why I chose this data set is that this data set is from a recognized source and this data set provides many features which are requires to make accurate predictions. Moreover, the data set provides binary classification which is benign and malignant in this case which makes easy for us to make predictions for binary classes.

1. **Methodology**

The main objective of our problem is to identify whether a given lump is benign or malignant and which classification models has the best accuracy in the prediction out of 6 used models i.e. Decision tree, Logistic regression, Naïve Bayes, Support Vector Machine and Random forest classifier and KNN. The methodology adopted for solving the problem is as follows:



**Data Pre-processing**

The first and the foremost step in the process is to process the data. The processing of data is needed in order to remove unwanted data which are not required in the prediction and which can hamper the prediction process.

The data pre-processing includes detection and removal of outliers, detection and removal of null values and to make data clean which can be directly used for training the models.

1. **Removal of outliers:** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.in simpler words we can say that outliers are the data points which lie far away from most of the data points. Removal of these outliers is necessary as these extreme values can hinder the training and prediction process and can lead to false predictions and less accurate results.

In order to remove outliers many methods have been devised. In this paper, I have used the interquartile range for the detection of outliers which is the most standard way. I have detected the outliers in every feature of the data set and replaced the minimum values with the minimum quartile and maximum values with the maximum quartile.

1. **Removal of null values:** There were no values in the data set. So this step of null values removal was not required.

**Features Selection and Scaling**

The second step in the prediction is to select the features. There is only 1 attribute that defines our classification class i.e. ‘diagnosis’.

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling.

**Training the models**

The next step in the prediction is to split the data into test and train data and then applying machine learning algorithm to train the data set for making predictions.

In the machine learning strategies, the learning procedure can be parted into two principal classifications such as supervised and unsupervised learning. In supervised learning, an arrangement of information cases is utilized to prepare the machine and is marked to give the right outcome. But in case of unsupervised learning, there are no pre-decided informational indexes, no idea of the usual result, which implies that the objective is harder to accomplish. Regression and classification are the most common methods that go under supervised learning. In case of regression, the target variable is continuous, and for classification, the target variable that is used for prediction is discrete.

In this paper, I have used 6 distinctive classification algorithms which are as follows:

* K-Nearest Neighbours
* Support Vector Machine
* Logistic Regression
* Decision Tree
* Naïve Bayes
* Random Forest Classifier

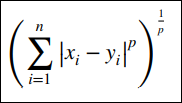
**K-Nearest Neighbours**

The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

KNN is a supervised learning technique that means the label of the data is identified before making predictions. Clustering and regression are two purposes to use it. K represents a numerical value for the nearest neighbours. KNN algorithm does not have a training phase. Predictions are made based on the Euclidean distance to k-nearest neighbours. This technique is applied to the prediction of breast cancer dataset since it already has labels such as malignant and benign. The label is classified according to the nearest neighbour to the class labels of its neighbours.

If the number of neighbours is denoted by N in K-NNs, then N samples are considered using the following distance metric value:

Minkowski Distance: Dist(x, y) =  -(1)

where if p=1, then it is Manhattan distance, if p=2, then it is Euclidean distance, and if p=∞, then it is Chebyshev distance. Among many choices, Euclidean distance is globally used. Among these K neighbors, the calculation will then check the quantity of information that focuses on every class, and afterward, it will relegate the new information point to the classifcation which frames the more signifcant part.

**Support Vector Machine**

Support vector machine is one of the most common machine learning techniques. The objective of the algorithm is to find a hyperplane in N-dimensions that classifies the data points. The major part of this algorithm is finding the plane that maximizes the margin. N dimension diversifies based on the feature numbers. Comparing two features could be done smoothly. However, if there are several features for classification, it is not always that straightforward. Maximizing the margin provides more accurate prediction results.

SVM has a small trade-off between large margin and accurate classification. If the exact classification without sacrificing any individual sample is applied, the margin could be very narrow, which could lead to a lower accuracy level. On the other hand, by maximizing the margin between classes to get a better accuracy, support vectors that are closest to the hyperplane could be considered with other class members.

Maximal edge classifier accompanies the meaning of hyperplane which expresses in an n-dimensional space. The hyperplane is of (n − 1) dimensions with level subspace that need not go through the root. It is difficult to draw a hyperplane in a higher dimension, so (n − 1) dimensional level subspace is still used. An SVM classifier can be constructed easily if there exists a separating hyperplane. The dataset categories cannot be divided using hyperplane, so feature space has to be enlarged using Gaussian radial basis function (RBF) or sigmoid function, cubic, quadratic or even higher order polynomial function. The hyperplane that is used in p-dimensions is as follows:

β0 + β1X1 + β2X2 + ……… + βpXp = 0 -(2)

where X1, X2,…, and Xp are the data points in the sample space of p-dimension and β0, β1, β2,…, and βp are the hypothetical values.

**Decision Tree**

A decision tree (DT) is one of the most common supervised learning techniques. Regression and classification are two main goals to use it. It seeks to solve problems by drawing a tree figure. Features are known as decision nodes, and outputs are leaf nodes. Feature values are considered as categorical in the decision tree algorithm. At the very beginning of this algorithm, it is essential to choose the best attribute and place it at the top on tree figure and then split the tree. Gini index and information gain are two methods for the selection of features.

Randomness or uncertainty of feature x is defined as entropy and can be calculated as follows:

H(x)=Ex[I(x)]=−∑p(x)logp(x) -(3)

Entropy values for each variable are calculated, and by subtracting these values from one, information values can be obtained. A higher information gain makes an attribute better and places it on top of the tree.

Gini index is a measure of how often a randomly chosen element would be incorrectly identified. Therefore, a lower Gini index value means better attributes. Gini index can be found with the given formula:

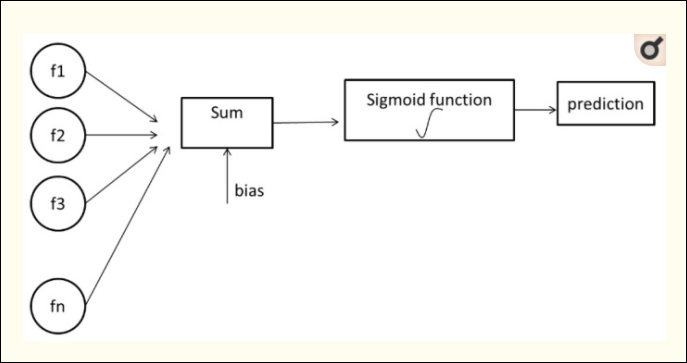
G=∑pi∗(1−pi)

for i=1,…n -(4)

A decision tree is easy to understand. However, if data contain various features it might cause problems that are called overfitting. Therefore, it is crucial to know when to stop growing trees. Two methods are typical for restricting the model from overfitting: pre-pruning, which stops growing early, but it is hard to choose a stopping point; and post-pruning, which is a cross-validation used to check whether expanding the tree will make improvements or lead to overfitting. DT structure consists of a root node, splitting, decision node, terminal node, sub-tree, and parent node. There are two main phases of the DT induction process: the growth phase and the pruning phase. The growth phase involves a recursive partitioning of the training data resulting in a DT where decision trees have a natural “if”, “then”, “else” construction that makes it fit easily into a programmatic structure.

**Logistic Regression**

Logistic regression is a technique that firstly used for biological studies in the early twentieth century. It has become widespread for social studies too. Logistic regression is also one of the predictive analyses. Logistic regression is appropriate to use when there is one binary dependent variable and other independent variables. Linear and logistic regressions are different in terms of the dependent variable. Linear regression is a more appropriate technique for continuous variables.



**Naïve Bayes**

Naïve Bayes is a straightforward and also fast algorithm for classification. Its working process is based on Bayes theorem. It is represented below:

P(X|Y)=P(Y|X)P(X)P(Y) -(5)

The fundamentals of this algorithm assume that each variable contributes to the outcome independently and equally. In this case, each feature will not be dependent on each other and will affect the output with the same weight. Therefore, the naïve Bayes theorem does not apply to real-life problems, and it is possible to get low accuracies while using this algorithm. Gaussian Naïve Bayes is one kind of naïve Bayes application. It assumes that features follow a normal distribution. The possibility of features is considered to be Gaussian and has a conditional probability. Gaussian naïve Bayes theorem is given below:

P(xi|y)=12πσ2ye^(−(xi−μy)22σ2y) -(6)

**Random Forest Classifier**

Random forest is an ensemble learning model that can be used for both regression and classification. Indeed, a random forest consists of many decision trees. Therefore, in some cases, it is more logical to use random forest rather than a decision tree.

The rotation forest algorithm consists of generating a classifier that is based on the extraction of attributes. The attribute set is randomly grouped into K different subsets. It aims to create accurate and significant classifiers.

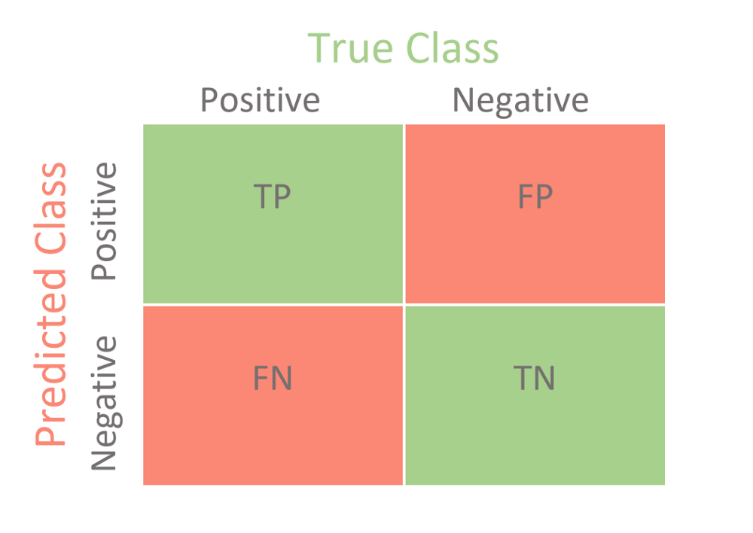
In the decision tree, feature selection is the main problem, and there are different approaches for that. Furthermore, random forest searches the best feature among a random subset of features, instead of searching for the most prominent feature while splitting nodes. It is possible to make it even more arbitrary by using random thresholds for each feature rather than seeking the best one.

Some built-in function parameters could make the model faster or more accurate. Max features, n estimators, and min sample leaf are used for increasing the power of prediction. N jobs and random state are generally used for making models faster. In this study, n estimators, which determines the number of trees to grow, and random state parameters are used to increase the accuracy and speed of the model.

The workflow of random forest is given below.

1. From the training set, picked K data points randomly.
2. From these K data points, generate the decision trees.
3. From generated trees, choose the number of N-tree and repeat steps (i) and (ii).
4. Form the N-tree that predicts the category to which the data points relate for a new data point, and assign the new data point via the category with the highest probability.
5. **Performance criteria**

The performance of the machine learning algorithms has been assessed on the basis of some performance parameters. I have developed a confusion matrix for analysing the different criteria for the assessment of the best algorithm. The confusion matrix is a measure of the performance of any classification algorithm. The confusion matrix consists of a 2D matrix consisting of 2 rows and 2 columns denoting the actual and predicted results respectively.



TP denotes True positive

FP denotes False positive

FN denotes False negative

TN denotes True negative

The criteria for measuring the efficiency of the models are as follows:

1. Accuracy: Accuracy determines how accurate the model has predicted both positive and negative results and to know whether the model is over fitted or not or is there any biasedness in the model towards a particular class or not.

Accuracy =

1. Sensitivity: Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall.

Sensitivity=

1. Specificity: Specificity gives the measure of the proportion of those values that got predicted as negatives or true negatives.

Sensitivity=

1. Precision: Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

Precision=

1. **Observations**

Now, the next step is to compile the results of the various machine learning algorithm and compare their outcomes based on our performance criteria to determine which model has best served our purpose.

Let’s compare the confusion matrices of various models and their accuracies.

**Confusion matrix for KNN classifier**

|  |  |  |
| --- | --- | --- |
|  | Benign | Malignant |
| Benign | 111 | 0 |
| Malignant | 4 | 55 |

**Confusion matrix for Random Forest classifier**

|  |  |  |
| --- | --- | --- |
|  | Benign | Malignant |
| Benign | 111 | 1 |
| Malignant | 2 | 59 |

**Confusion matrix for Logistic Regression**

|  |  |  |
| --- | --- | --- |
|  | Benign | Malignant |
| Benign | 109 | 1 |
| Malignant | 2 | 59 |

**Confusion matrix for Support Vector Machine(SVM) classifier**

|  |  |  |
| --- | --- | --- |
|  | Benign | Malignant |
| Benign | 111 | 0 |
| Malignant | 1 | 59 |

**Confusion matrix for Decision Tree classifier**

|  |  |  |
| --- | --- | --- |
|  | Benign | Malignant |
| Benign | 101 | 9 |
| Malignant | 5 | 56 |

**Confusion matrix for Naïve Bayes classifier**

|  |  |  |
| --- | --- | --- |
|  | Benign | Malignant |
| Benign | 104 | 6 |
| Malignant | 4 | 57 |

1. **Results and comparison**

In order to find out the results, we have used all the 529 data items from the data set. We had divided the data into two parts: 70% of the data items have been taken for training the model and remaining 30% of the data items have been taken for the testing of the model.

I have used 6 different machine learning models and evaluated their performance on the basis of 4 criteria mentioned above.

The confusion matrix results for all 6 models have been shown in the following graph:

On the basis of the results, we can infer that KNN, random forest classifier and SVM predicts the most number of true positive results which are 111 out of 170 results which is followed by random forest and logistic regression having 109 out of 170 respectively.

Random forest classifier, SVM and logistic regression have predicted the most number of true negatives which are 59 followed by Naïve Bayes and decision tree respectively.

Decision tree has predicted most number of false negatives which are 5 followed by Naïve bayes and KNN respectively.

Decision tree has predicted most number of false positives which are 9 followed by Naïve Bayes respectively. Both KNN and SVM has predicted no false positive values which is commendable.

Below is the graph comparison of all 6 algorithms on the basis of our 4 performance parameters which are accuracy, sensitivity, specificity and precision respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision |
| KNN | 97.076 | 96.52174 | 100 | 100 |
| Random Forest | 97.6608 | 98.23009 | 98.33333 | 99.10714 |
| Logistic Regression | 98.246 | 98.1982 | 98.33333 | 99.09091 |
| Support Vector Machine | 99.4152 | 99.10714 | 100 | 100 |
| Decision Tree | 91.8128 | 95.28302 | 86.15385 | 91.81818 |
| Naïve Bayes | 94.152 | 96.2963 | 90.47619 | 94.54545 |

From the table and graph, we can infer that the highest accuracy is of SVM which is 99.415% followed by logistic regression having 98% and random forest classifier having accuracy of 97% and KNN with 96% respectively.

In terms of sensitivity, the highest sensitive models are SVM having 99% and random forest classifier having 98.23% followed by logistic regression with 98.19% respectively.

This denotes that these 2 algorithm are more sensitive towards data values that are true positive data values.

In terms of specificity, we have KNN and SVM with the highest 100% among all the models which means that these 2 models have detected 0 false positives among the test data sets and have good prediction regarding true negative values.

For precision, we have KNN and SVM with 100% precision followed by random forest classifier with 99% specificity. KNN and SVM have detected 0 false positive values and have good precision over detecting the true positive data values and they can easily separate the true positives from the false positives.

One thing that is evident from the analysis is that the decision tree algorithm has got the lowest score among all the performance parameters which suggests that many improvisation methods can be applied on the algorithm in order to increase its accuracy and other features so that it can lead to more accurate predictions and less false predictions.

**Variation of accuracy in different models**

The variation of accuracy in various machine learning models has been plotted in the following graph which gives a better picture of how accuracy has shown variations among different models.

Among all the models decision tree has the lowest accuracy which shows that some improvements can be done with the model and algorithm in order to increase its overall accuracy.

Though there can be many aspects behind how any algorithm works on a particular data set and there can be many ways to improvise the results of the algorithm. There are many other algorithm as well which can work better on this data set and some algorithms which can work better on other data sets. One must be careful while choosing the machine learning algorithm to predict on a data set after comparative analysis of at least 2-4 models as the aim of our study is very crucial i.e. to predict the possibility of cancer. So, the kind of model and predictions must be taken into consideration.

Moreover, accuracy cannot be the only factor for better performance of the model other parameters must be taken into account before reaching to the solution.

So, overall in our study Support Vector Machine(SVM) outperforms all other machine learning algorithms followed by random forest classifier and then logistic regression in terms of all the performance parameters.

1. **References**
2. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiol Soc N Am. 2018;286(3):800–9.
3. Breast Cancer: Statistics, Approved by the Cancer.Net Editorial Board, 04/2017. [Online]. Available: http://www.cancer.net/cance r-types/breast-cancer/statistics. Accessed 26 Aug 2018.
4. Mori M, Akashi-Tanaka S, Suzuki S, Daniels MI, Watanabe C, Hirose M, Nakamura S. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-feld digital mammography in a population of women with dense breasts. Springer. 2016;24(1):104–10.
5. Kurihara H, Shimizu C, Miyakita Y, Yoshida M, Hamada A, Kanayama Y, Tamura K. Molecular imaging using PET for breast cancer. Springer. 2015;23(1):24–32.
6. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. Neural Comput Appl. 2013;23
7. 6. Nagashima T, Suzuki M, Yagata H, Hashimoto H, Shishikura T, Imanaka N, Miyazaki M. Dynamic-enhanced MRI predicts metastatic potential of invasive ductal breast cancer. Springer. 2002;9(3):226–30.
8. Park CS, Kim SH, Jung NY, Choi JJ, Kang BJ, Jung HS. Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. Springer. 2013;22(2):153–60.
9. Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res. 2020; https://doi.org/10.1080/03772 063.2020.1713916.
10. Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients’ recovery. SN Comput Sci. 2020;1(4):206.
11. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbours. In: Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226–229.
12. Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK. Performance evaluation of random forests and artificial neural networks for the classification of liver disorder. In: Proc. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1–5.
13. Cancer Prediction”, by Yixuan Li, Zixuan Chen October 18, 2018
14. “Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study” by Mumine Kaya Keles, Feb 2019
15. “Breast Cancer Prediction Using Data Mining Method” by Haifeng Wang and Sang Won Yoon, Department of Systems Science and Industrial Engineering State University of New York at Binghamton Binghamton, May 2015.
16. “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis” by Wenbin Yue , Zidong Wang, 9 May 2018