Cancer Prediction using Machine Learning

Sawan Verma, Saloni Chaudhary, Sunil Kumar, Pranav Singh Rana
Department of Computer Science & Engineering
Meerut Institute of Engineering and Technology, Meerut

Abstract:- The breast cancer is very common and a dominant cancer in women over the world. It is increasing in countries which are developing and where most of the cases are prognosed in the late stages. Some of the strategies which were already suggested or proposed and shows a comparability between the ML algo by using various approach such as the ensemble approach , using blood analysis or Data mining algo ,etc.In this research paper there is a comparability of the two machine leaning algo RF(Random Forest) and Decision Tree. The data set was divided into the two stages that is training stage and the testing stage. The algo will be used in this application which gives the best results and then the approach of the model will be classifies that the cancer just as malignant or benign.

Keywords:- *Machine Learning, Breast Cancer, Identification, Classification, Prediction, Random Forest, Decision Tree, Malignant, Benign.*

I. INTRODUCTION

Breast Cancer is a very usual and a dominant Cancer among women over the world[4].According to the global statistics that represent the preponderance of new cancer's patients and cancer-relevant deaths and it makes a serious health issue of the public in the societies[2].The initial diagnose of the breast cancer it can be improves by the methods of predictions & chance of survival substantially, as it can encourage clinical treatment of the patients on time. In addition the exact classification of being cancer can avoid the people were going for treatments which are not necessary. So the subject about much research is the breast cancer's correct diagnosis and classification of patients that is the patients belongs to the group of malignant or benign[1].Due to its different benefits in overcritical factors detection from a breast cancer datasets ,ML is universally accepted as the technique of alternative in breast cancer classification. The effective ways to classify the data are the methods of classification and data mining . Especially in the field of medical, where those methods are extensively used in examination and diagnosis to make the conclusion. The analysis focus to detect the features that are better helpful in predicting malignant or benign cancer and to see the usual trends that may aids us in the selection of model and selection of hyper parameters. The main goal is to classifying that the cancer of breast is belongs to the group of benign or malignant. We have used the classification of machine learning methods to achieve this and fit the function which can be predict the distinct class of new information/inputs.

II. BACKGROUND /LITERATURE REVIEW

On the Detection of Breast Cancer: It is an operation/Application of ML Algorithms at the Wisconsin Diagnostic Dataset via way of means of the Abien Fred M. Agarap. There are 6 Machine Learning algorithm's that are used for detection of most cancers in this paper. GRUSVM model is used for the prognosis of breast most cancers GRUSVM, Softmaxregression, K-NN, LR (LinearRegression), Multilayer Perceptron,search and SVM at the Wisconsin Diagnostic Breast Cancer dataset via way of means of measuring their type check accuracy, and there specificity and sensitivity values. A stated dataset includes functions that have been estimated from digitized pictures of FNA checks on a mass of breast. So that Machine learning algorithms implemented, the Dataset become isolated within side the following style 70 percentage for education stage, and 30 percentage for the trying out stage. Their effects have been that every one offered machine learning algorithm's displayed excessive overall achievement at the binary type of tumor, i.e. figuring out whether or not benign cancer or malignant cancer. Hence, the analytical measures at the type trouble have been additionally satisfying. To similarly strengthen the effects for this research, the approach of CV including the k-fold & cross-validation need to be use. A equipment of one of these manner may not handiest offer a extra correct degree of version prediction overall performance, however it will additionally help in figuring out the most top of the line hyper-para-meters for the machine learning algorithm's[3].

A ML approach analysis for a Breast Cancer Prediction with the aid of using in VIT university,vellore by Priyanka Gandhi and Prof.Shalini L. In this research paper, ML strategies are observed to be able to increase a accuracy of diagnosis. Approach along with CART ,KNN,RF(Random Forest) are compared. The dataset used is received from UC Irvine ML Repository. It is discovered that KNN set of rules has tons higher overall implementation than the alternative strategies utilized in comparison. The maximum correct version changed into K-Nearest Neighbour. The type version along with RF algo and BT(Boosted Trees) confirmed the same certainty. Hence, the maximum correct classier may be used to discover the cancer in order that the remedy may be discovered in initial phase[4].

A breast Cancer Diagnosis via way of means of Dierent ML approaches by Using Blood Analysis Data via way of means of a Akif Durdu,Muhammet Faith Aslam, Kadir Sabanci and Yunus Celik for tumor initial diagnosis. During this paper, 4 dierent ML algo have been used for the initial recognition of tumor. A purpose of this undertaking is to procedure the consequences of habitual blood evaluation with dierent Machine Learning strategies.

Approaches used are Extreme Learning Machine, ANN, k-Nearest Neighbor and Support Vector Machine. UCI library provides this dataset. In this dataset age, chemokine monocyte chemoattractant protein (MCP1), resistin, adiponectin, leptin, (HOMA), insulin, glucose and BML attributes had been used. Parameters which have the high-quality accuracy values had been determined via way of means of the use of four dierent Machine Learning techniques. This dataset consists of adiponectin, HOMA, resistin, leptin, insulin, glucose, BMI, age and MCP1 capabilities that may be obtained in habitual blood evaluation. The importance of those statistics in breast most cancers detection turned into investigated via way of means of ML strategies. The evaluation turned into accomplished with four dierent strategies of ML.KNN & SVM strategies are decided a use of Hyper-para-meter optimization technique. The maximum accuracy and minimum schooling time had been given via way of means of ELM which turned into 80 percent & 0.42 sec[5].

Estimation of the work of ML approach for the Breast Cancer forecast/Prediction with a aid of using Zixuan Chen & Yixuan Li used a datasets withinside the examine. A examine first of all collects the statistics of a BCCD dataset that includes 116 patient with nine features and statistics of WBCD dataset that includes 699 patient & eleven features. After that we preprocess a uncooked statistics of WBCD dataset & received a information that incorporates 683 patient with 9 features & consequently the index distinguishing even if or not the patient has the malignant cancer. After evaluating a accuracy, Fmeasure metric & ROC curve of five type models, the end result has proven that Random Forest is selected because the number one type version all through this study.Hence, effects of this examine offer a reference for specialists to differentiate the man or woman of carcinoma .In this examine, there are nonetheless a few obstacles that have to be decode in addition effort. For present, after all additionally exist a few indices human beings haven't discovered yet, this examine best gathered a information of ten attributes all through this analysis. The restricted statistics has an effect at the accuracy of effects. additionally , the Random Forest also can be mixed up with a different statistics datamining strategies to get extra correct and green effects withinside the long term work[6].

A motive of this research prospectus "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: a comparative Study"by Mumine Kaya Keles changed into to expect & hit upon breast most cancers early even supposing a cancer length is petite hit upon non-invasive and pain-free strategies that use facts mining class algo's. Hence, the contrast of facts mining class algo changed into create with a Weka tool.This research prospectus, a Weka facts mining software program changed into implemented to an antenna dataset if you want to have a look at the efficacy of facts mining strategies withinside the diagnosis of breast most cancers. The dataset that changed into that created had 6006 values, 5405 of which had been used because the schooling dataset, at the same time as 601 had been used because the take a look at facts set. The dataset changed into then transformed to a arff pattern, that's a document kind utilized by a Weka tool. A ten-fold cross-validation changed into so it use to achieve the maximum true consequences the use of the Extraction of the knowledge primarily based totally on Progressive Studying facts mining software program tool. RF finished a nice at some stage in the ten fold cross-validation providing a median accuracy of 92.2 %[7].

This project "Breast Cancer Prediction Using Data Mining Method" with the aid of using Sang Won Yoon and Hafeng Wang is used to check the impact of characteristic area reduction, a hybrid among main factor evaluation (PCA) & associated data mining models is proposed, that execute the precept factor evaluation approach to lessen the characteristic area. To examine the overall performance of those models, broadly used check data units are used, Wisconsin Breast Cancer Database (1991) & Wisconsin Diagnostic Breast Cancer (1995). 10- fold cross-validation approach is applied to measure a check mistakess of every model. PCs-SVM is best for WBC data that could be the 97.forty seven percent, and PCi-ANN is the first-class thinking about accuracy for WDBC data this is 99.63%. A purpose for higher effects from PCA preprocessing is due to the data the main additives handiest constitute a huge a part of the data withinside a entire data space , which to a point can decrease data noise, as a outcome, characteristic space is enriched[8].

"Machine Learning with Application in breast most cancers Diagnosis and Prognosis" through Webin Yue and Zidong Wang In this prospectus, they furnished explainations of diverse ML strategies and their utility in BC analysis will not to examine the information in the benchmark database WBCD. ML strategies have proven their incredible cappotential to beautify class and prediction accuracy. However many algo's have finished very excessive accuracy in WBCD, a occasion of progressed algo's stays necessary. Classification accuracy can be a important evaluation standards however it is now no longer the only one. Different algorithms keep in mind exclusive aspects, and feature exclusive mechanisms. Although for numerous a long time artificial neural network have ruled BC analysis and diagnosis, it is clean that greater currently opportunity Machine learning techniques are implemented to intelligent healthcare system to apply the variety of alternatives to medical practitioner[9].

A. *Breast Cancer Classification*

Breast most cancers class is a class which divides the carcinoma into classes relying on how they have spread at all. Classification algo's offers the prediction approximately one or greater discrete variables and help the alternative features in a dataset. To run the classification algorithms the data processing software program is required. The purpose of analysis is to pick out a best remedy. Analysis lets scientists to find, group, and nicely call organisms through a uniform device that' why it's far necessary. There are notably used techniques with inside the data processing are classification and clustering. Clustering is locate to extract records from the fixed of understanding to get businesses or clusters and illustrate a set of records itself. Classification is likewise known as supervised learning of with inside the ML of, it use to categorise the unexplained conditions supported learning of current styles and classes from the set of records and finally deliver the prediction at the destiny conditions. The training set, that's hired to create the structure which is classifying structure, and consequently the check set, that has a tendency to evaluate a classifier, are usually stated in a class responsibilities classification can be a pretty complicated optimization problem. There are many machine learning of techniqes are carried out with the aid of using researchers to clear up this classification problem. The artificial neural network, random forest, aid vector device, and so on are the maximum well-known set of rules this is used for breast most cancers class or prediction. Scientists attempt to discover the best set of rules to recognize the main correct class result, however, facts of variable best will also have an effect on the class result. Further, the uncommonness of understanding will have an effect on the variety of set of rules packages also. If the early observation is done of carcinoma is, there are greater remedy alternatives and a miles higher hazard for the survival. A women whose cancer is identified at an initial degree have a ninety three percentage or better survival charge in the first five years. You can positioned your thoughts snug with the aid of using getting checked regularly. Finding most cancers early stage also can save the life[1].

B. Machine learning algorithms

A ML is an utility of AI which deliver a strength to the model to routinely examine and enhance from revel in without being programmed manually. ML emphasis and relies upon at the phase of computer applications a good way to get provided the data furnished and use that data to examine. The approach of learning of starts with datasets, specimen, training, rules.So that you can then determine out a sample & capable of make a upgrades withinside the close to future, if necessary.

III. PROPOSED METHODOLOGY

Her we proposed a methodology that contains some steps those are data preprocessing, data preparation, feature selection, feature projection, feature scaling, model selection and prediction. Now discuss about these steps. First step is data preprocessing, it is very important step in data mining. It explained that the data is manipulated or modified before its use in the model to make the process easier. The second step is data preparation, basically it refers to the cleaning of the data and it ensures that the given data is accurate. Now third step is feature selection, it refers to the variable or attribute selection, It refers to the variable or attribute selection. We can explained it as a process or technique to reduce the number of input variables during the development of predictive model or we can say that selection of most compelling features from a given set of data. Next step is feature projection, it is also know as feature extraction. Basically feature projection is used to reduce the dimensionality of space. It covert the higher dimensional space into the fewer dimensional space. Next step is feature scaling, it refers to the standardization of independent feature that are present in the data of fixed range. After the feature selection, the next step is model selection. Model selection is defined as the technique or method of selection of model from the candidate models for a the dataset of training. Now the last step is prediction. Prediction refers to the predict the output after the model has been trained on the previous dataset and applied to the new dataset[1].

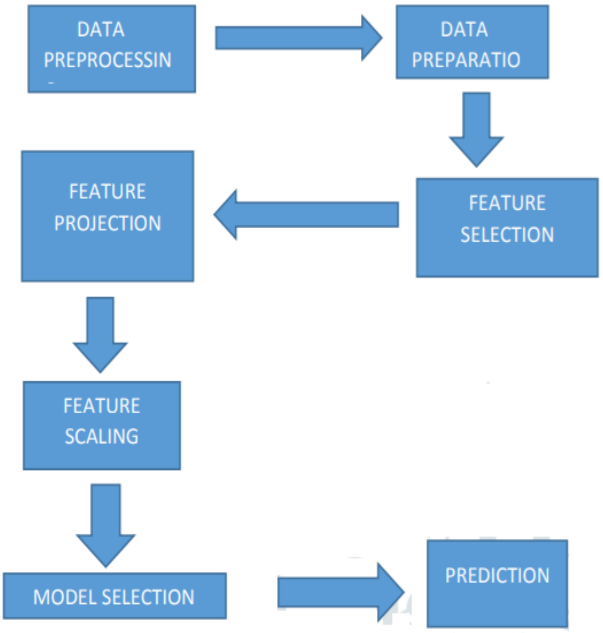

Fig. 1: Proposed Methodology

A. Random forest

RF is a learning algorithm which fit to a learning category of supervised. It is used for both regression as well as classification. Random forests additionally referred to as RDF("random decision forests") which makes a massive quantity of trees that obtain their output through whole study of techniques for category and regression. There are two features bagging and feature randomness uses to construct those trees. RF is more better than the DT because it does not overfit the data[1].

B. Decision Tree

DT is also a supervised learning algorithm. The aim of using this algorithm is to make the training model that can be use to predict a value of a target variable. It use a top-down technique to data in order which give a knowledge set, they conflict to institution and label conclusion which might be comparable among them, and look for the simplest guidelines that cut up the observations that aren't the equal among them till they attain a positive quantity of similarity. They use a process that is known as layered splitting, in which at every layer they conflict to split the data into or greater groups, simply so data fall below an equivalent group is maximum just like each other, and groups are as die-rent as feasible from each other[1].

IV. RESULT AND DISCUSSION

We discuss the result through the different diagrams those represents that how much people belongs to the group of malignant or benign. There are four representations of the result in the form of diagrams as barplot, matrix, histogram and pairplot. These shows the result:-

In fig 2, there are some graphs that show the result about the cancer is malignant or benign. These graphs shows that how much people belongs to the group of malignant and how much people belongs to the people belongs to the group of benign. In this figure, there are two target values that are 0.0 and 1.0. Here 0.0 means the person belongs to the group of malignant and 1.0 means the person belong to the group of benign.

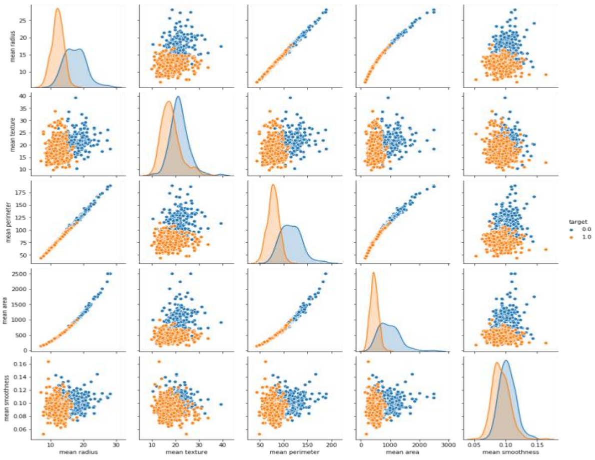

Fig. 2: Pairplot of Cancer Dataframe

In Fig. 3, the result represents in the form of correlation barplot. Now we discuss about the correlation barplot. Basically correlation barplot shows the result in the form of barplot by creating the figure of correlation coefficient. In above figure Fig. 3, there are two groups positive and negative. Maximum bars shows the negative result of these features but some shows the positive result of these features. In this figure Fig. 3, if we remove these particular features (mean fractal dimension, texture error and symmetry error) then the accuracy in result will be increase because these features having the less data that is not capable for correlation with the target value.

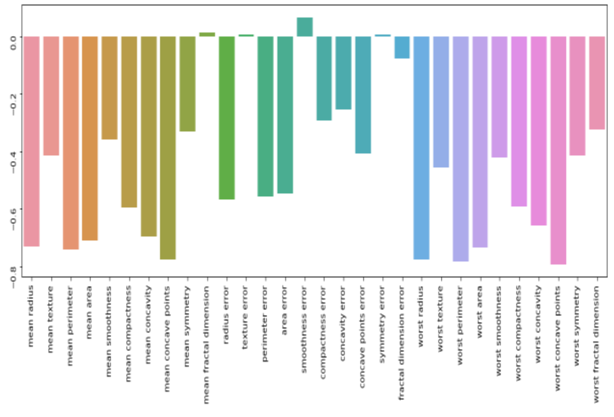

Fig. 3: Correlation Barplot

The figure 4 gives the data of people who are suffering from cancer or not as the result. Clearly, we can see that 212 people belong to the group of malignant and 357 people belongs to the group of benign. As the result we found that Random forest algorithm gives the best result as compare to the DT algorithm. DT algorithm gives the accuracy about 84% and Random forest algorithm gives the accuracy about 98% which is more than the accuracy of Decision Tree. On the comparison between these two algorithms, we got the higher accuracy with the random forest algorithm.

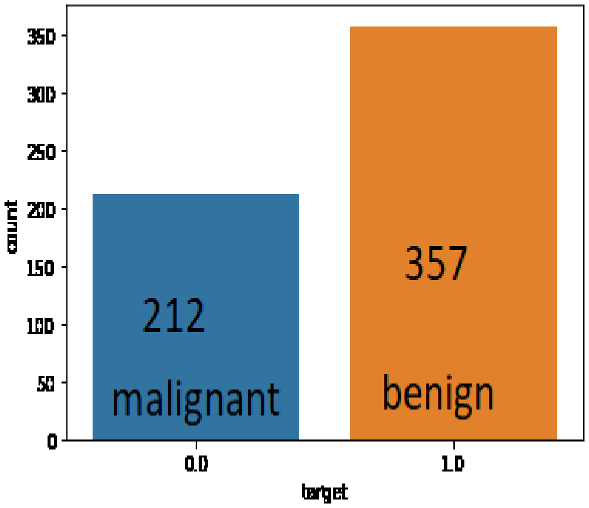

Fig. 4: Data Visualization Countplot of Cancer and non cancer

In Fig. 5, the result represents in the form of correlation matrix. Now we discuss about the correlation matrix. Basically correlation matrix shows the correlation between the two different variables in the form of table. In above figure Fig. 5, 1 represents that the person belongs to the group of benign and 0 represents that the person belongs to the group of malignant. Here malignant means person suffering from cancer and benign means person is not suffering from cancer. This figure clearly show that how much person belongs to the group of benign or malignant.

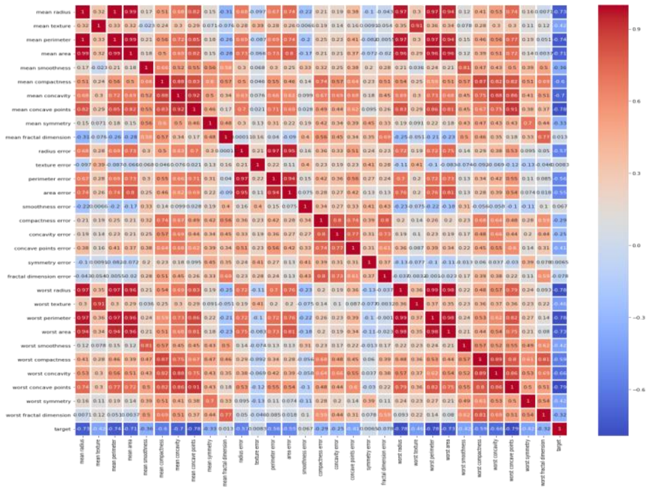

Fig. 5: Correlation Matrix

V. CONCLUSION

If breast most cancers located at the stage of early it will help the thousands of persons to save their lives. This undertaking will assist the actual global sufferers and docs to acquire the lot of data as they. The studies on 9 research documentation has helped us to acquire a information for a undertaking prospective via way of means of us. We may be capable of classify and are expecting the most cancers into being or malignant via way of means of the use of the ML algorithms. ML algo may be used for scientific orientated studies, it speed up's the system, reduces the human mistakes & decreases the guide mistakes. And it will very helpful for the human beings because it can saves the life of the people by diagnosis at earlier stage of the cancer.

REFERENCES

[1.] ”Ultrasound characterization of breast masses”, The Indian journal of radiology imaging by S. Gokhale, Vol. 19, pp. 242-249, 2009. K. Elissa, “Title of paper if known,” unpublished.

[2.] Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach” by Pragya Chauhan and Amit Swami, 18 October 2018.

[3.] “On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset” by Abien Fred M. Agarap, 7 February 2019.

[4.] “Analysis of Machine Learning Techniques for Breast Cancer Prediction” by the Priyanka Gupta and Prof. shalini L of VIT university, vellore, 5 May 2018.

[5.] “Breast Cancer Diagnosis by Dierent Machine Learning Methods Using Blood Analysis Data“ by the Muhammet Fatih Aslan, Yunus Celik , Kadir Sabanci and Akif Durdu, 31 December, 2018.

[6.] ”Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction”, by Yixuan Li, Zixuan Chen October 18, 2018.

[7.] “Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study” by Mumine Kaya Keles, Feb 2019.

[8.] “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.

[9.] “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis” by Wenbin Yue, Zidong Wang, 9 May 2018.