Breast Cancer Detection Based on Machine Learning Feature Selection And Extraction Algorithm

Pranali Shirsat[1], Dr. Surendra Bhosale[2]
[1, 2] Dept of Computer Science
[1, 2] EED VJTI,Mumbai, India

Abstract- *Breast cancer is cancer that forms in the cells of the breasts.After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. Breast cancer can occur in both men and women, but it's far more common in women. The data science and machine learning algorithm is used for classification. For classification using machine learning algorithm all of the features present in the dataset might not be useful in building a machine learning model to make the necessary prediction. Using too many features might even make the predictions worse. So, feature selection plays a huge role in building a machine learning model. Feature selection and extraction can make use of less number of data feature to give best possible accuracy.Especially when dealing with a large number of variables there is a need for dimensionality reduction.Feature Selection can significantly improve a learning algorithm's performance. The goal is to find an optimal feature-subset that maximizes the accuracy of prediction*

Keywords- Data science, feature selection,feature extraction, machine learning, Random forest.

I. INTRODUCTION

About 1 in 8 U.S. women (about 12%) will develop invasive breast cancer over the course of her lifetime.In 2019, an estimated 268,600 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 62,930 new cases of non-invasive (in situ) breast cancer. About 2,670 new cases of invasive breast cancer are expected to be diagnosed in men in 2019. A man's lifetime risk of breast cancer is about 1 in 883.About 41,760 women in the U.S. are expected to die in 2019 from breast cancer, though death rates have been decreasing since 1989. Women under 50 have experienced larger decreases. These decreases are thought to be the result of treatment advances, earlier detection through screening, and increased awareness.For women in the U.S., breast cancer death rates are higher than those for any other cancer, besides lung cancer. Substantial support for breast cancer awareness and research funding has helped create advances in the diagnosis and treatment of breast cancer. Breast cancer survival rates have increased, and the number of deaths associated with this disease is steadily declining, largely due to factors such as earlier detection, a new personalized approach to treatment and a better understanding of the disease.The utilization of data science andmachine learning approaches in medical fields proves to be prolific as such approaches may be considered of great assistance in the decision making process of medical practitioners.With an unfortunate increasing trend of breast cancer cases, comes also a big deal of data which is of significant use in furthering clinical and medical research, and much more to the application of data science and machine learning in the aforementioned domain.

Machine learning, a sub discipline in the field of Artificial Intelligence, explores the study and design of algorithms that can learn from data.Machine Learning provides methods/algorithms that make system computationally intelligent. Such algorithms build models based on input and then use these models to make predictions or decisions. Machine Learning is mainly useful in cases where algorithmic/deterministic solutions are not available i.e. there is a lack of formal models or the knowledge about the application domain is scarce. The algorithms have been developed in diverse set of disciplines such as statistics, computer science, robotics, computer vision, physics, and applied mathematics. Advantages of machine learning over statistical models are accuracy, automation, speed, customizability and scalability. As medicine plays a great role in human life, automated knowledge extraction from medical data sets has become an immense issue. Research on knowledge extraction from medical data is growing fast The present study aims to investigate five kinds of feature selection methods for classification of breast cancers from normal. A feature extraction method is used to extract the required features. Random forest supervised learning is used for classification and efficiency calculation.

II. METHODOLOGY

A. Data set

The dataset used here is the Breast Cancer Wisconsin (Diagnostic) Data Set. This dataset contains 569 records of and 32features (including the Id and diagnosis). The features

represent various parameter that are useful in predicting if a tumoris malignant or benign.Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Featuresarenothing but the characteristics of the image suchasthe cells radius, texture, perimeter,area, smoothness , compactness, concavity, concave points, symmetry, fractal dimension and their respective mean and worst values.The features are fed into the machine learning model.This dataset can be found on UCI Machine Learning Repository.

B. Data vizualization

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data. To communicate information clearly and efficiently, datavisualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. Data visualization is used to know the data before feature selection. Data standardization is a data processing workflow that converts the structure of disparate datasets into a common data format.

$$standard\ data = \frac{(data - mean)}{standard\ deviation} \quad (1)$$

The data features are divided into three groups of ten features each and then plotted using seaborn swarm plot.

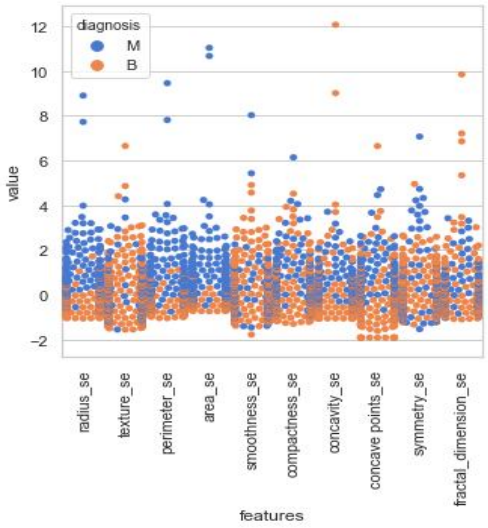

Fig.1 Data visualization of ten features

The swarm plot is the categorical plot which represents individual feature values and its respective target values (M-Malignant, B-Benign). The plot makes easy to understand which data feature classifies the target values more accurately. Here area_se and radius_se features classify the target values more clearly.

C. Feature selection

Feature selection can be termed as the process of selecting a particular feature from a huge collection of features. All the features are residing within the data. So we need to select a particular feature from huge set of features that are residing within the dataset .Feature selection plays an important role in the machine learning and data mining. Machine learning is a subfield of computer science that evolved from the study ofthe pattern recognition and computational learning theory in artificial intelligence. In machine learning, feature selection is also termed as variable selection or attribute selection.In this part we will select feature with different methods that are feature selection with correlation, univariate feature selection, recursive feature elimination (RFE), recursive feature elimination with cross validation (RFECV) and tree based feature selection. We will use random forest classification in order to train our model and predict.

Feature selection with correlation and random forest classification.

Correlation is a statistical term which in common usage refers to how close two variables are to having a linear relationship with each other. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features.

$$Corr = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \quad (2)$$

Where X_i = feature value, $\overline{X}$ = mean of feature values, Y_i = target value and $\overline{Y}$=mean of target values.

Univariate feature selection and random forest classification.

Univariate feature selection method uses statistical tests can be used to select those features that have the strongest relationship with the output variable. The chi_square distribution is used to selec best five features for classification.The scikit-learn library provides

the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.We have selected best 5 features and the accuracy is almost 96% .

$$x_c = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

Where O_i=observed value and E_i=Expected value

Recursive feature elimination (RFE) with random forest

TheRecursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remains. It uses the model accuracy to identify which attributes and combination of attributes contribute the most to predicting the target attribute using p-value statistics.

$$z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (4)$$

Where $\bar{X}$=population mean. μ_0= hypothesized mean, σ=standard deviation and n=number of samples

- **Recursive feature elimination with cross validation and random forest classification**

Recursive feature elimination (RFE) is a feature selection meathod that fits a model and removes the weakest feature until the specified number of features is reached. Features are ranked by the model's coefficients or feature importance attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. **Recursive feature elimination with cross validation estimates the number of features required for highest accuracy using cross validation method.**

- **Tree based feature selection and random forest classification**

Tree based estimators can be used to compute feature importance, which in turn can be used to discard irrelevant features using featureimportanceattribute of random forest classification method.

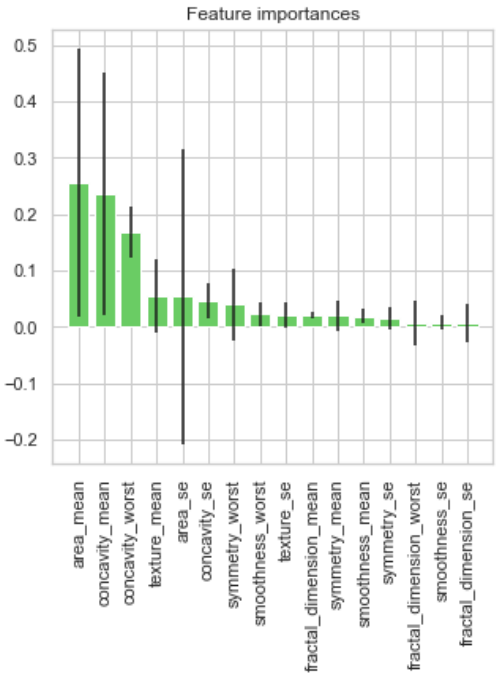

Fig.2 Feature imporatance plot

Using above plot we have selected the five highest important features.

Machine learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed**.** Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learnautomatically without human intervention or assistance and adjust actions accordingly[2]. There are four types of machine learning algorithm namely supervised machine learning algorithm, unsupervised machine learning algorithm, semi-supervised machine learning algorithm and Reinforcement machine learning.

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in

order to modify the model accordingly.In contrast, **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Random Forest Algorithm

Random Forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable predictionOne big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems[5].Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction.

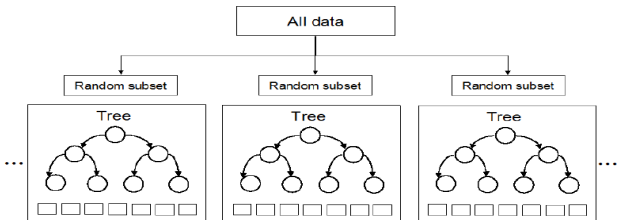

Fig.3 Random forest algorithm

Accuracy is calculated after implementing random forest algorithm.

Feature Extraction using Principal Component Analysis(PCA)

Principle component analysis is a statistical process that transforms data, which is basically a group of observations, into orthogonally related new variables. These new variables are called principal components[8]. Principal Component Analysis uses linear algebra to transform the dataset into a compressed form. Generally this is called a data reduction techniques. A property of PCA is that you can choose the number of dimensions or principal component in the transformed result. Before implementing PCA the data is normalized for improving the performance of PCA.

$$X_{nor} = \frac{X - X_{mean}}{X_{max} - X_{min}} \quad (5)$$

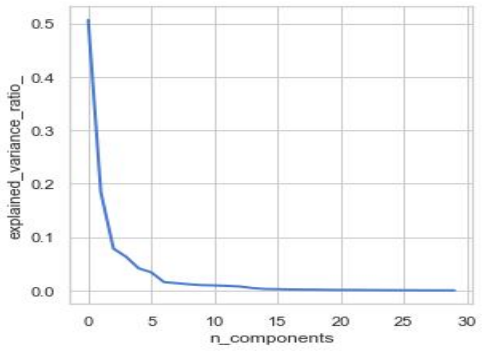

Fig.4 Principal component analysis

The plot shows variance ratio of each feature with respect to the number of features selected. The plot indicates that the varoiance ratio is high if the number of features selected is minimum. As the number of features increase the variance ratio starts decreasing.

III. PERFORMANCE EVALUATION

1. Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table). In predictive analytics, a table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than mere proportion of correct classifications (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (that is, when the numbers of observations in different classes vary greatly).

Table 1: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.

- **True positives (TP):** These are cases in which we predicted yes (they have breast cancer), and they do have the breast cancer.
- **True negatives (TN):** We predicted no, and they don't have the breast cancer.
- **False positives (FP):** We predicted yes, but they don't actually have the breast cancer. (Also known as a "Type I error.")
- **False negatives (FN):** We predicted no, but they actually do have the breast cancer. (Also known as a "Type II error.")

2. Accuracy : Accuracy can be calculated as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of instances}}$$

$$\textbf{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \qquad \textbf{(6)}$$

As a heuristic, accuracy can immediately tell us whether a model is being trained correctly and how it may perform generally. However, it does not give detailed information regarding its application to the problem. The problem with using accuracy as your main performance metric is that it does do well when you have a severe class imbalance.

3. Classification report

The classification report visualizer displays the precision, recall, F1, and support scores for the model.

Precision: The Precision is the proportion of the positive cases that were predicted correctly. Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives. Precision is used with recall, the percentage of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad \textbf{(7)}$$

Precision is the probability that a (randomly selected) retrieved document is relevant. Precision is a good measure to determine, when the costs of False Positive is high. For instance, email spam detection. In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

Recall: The Recall or TP-Rate is the proportion of the correctly identified positive cases

.

$$\textbf{Recall} = \frac{TP}{TP+FN} \qquad \textbf{(8)}$$

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.To fully evaluate the effectiveness of a model, you must examine both precision and recall. Unfortunately, precision and recall are often in tension. Recall is the probability that a (randomly selected) relevant document is retrieved in a search. Recall actually calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative.

f1 score: The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. A measure that combines precision and recall is the

harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F1 = 2*\frac{Precision*Recall}{Precision+Recall}$$

$$F1 = 2*\frac{TP}{TP+FN+FP} \quad (9)$$

F1 Score is needed when you want to seek a balance between Precision and Recall. Right…so what is the difference between F1 Score and Accuracy then? We have previously seen that accuracy can be largely contributed by a large number of True Negatives which in most business circumstances, we do not focus on much whereas False Negative and False Positive usually has business costs (tangible & intangible) thus F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives).

Support: Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn‘t change between models but instead diagnoses the evaluation process.

False Positive Rate (FPR) : False positive rate represents what fraction of all negative instances does the classifier incorrectly identify as positive. FPR is calculated as:

$$FPR = \frac{FP}{FP+TN} \quad (10)$$

4. K-fold Cross validation: Evaluating a Machine Learning model can be quite tricky. Usually, we split the data set into training and testing sets and use the training set to train the model and testing set to test the model. We then evaluate the model performance based on an error metric to determine the accuracy of the model.

This method however, is not very reliable as the accuracy obtained for one test set can be very different to the accuracy obtained for a different test set. K-fold Cross Validation(CV) provides a solution to this problem by dividing the data into folds and ensuring that each fold is used as a testing set at some point. This article will explain in simple terms what K-Fold CV is and how to use the sklearn library to perform K-Fold CV.K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. Lets take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

IV. RESULTS

Default dataset consists of 30 features. The data is visualized using seabornswarmplot. The feature selection mthods successfully optimized the number of features used for detection of breast cancer using the Random forest algorithm. The table shows the methods of feature selection and the respective accuracies with the number of features used.

TABLE.2 Features selection methods and accuracy

Sr. no	Feature Selection method	Number of features	Accuracy
1	Feature selection with correlation	16	93%
2	Univariate feature selection	16	93%
3	Recursive feature selection	5	95.8%
4	Recursive feature selection with cross validation	5	97.34%
5	Tree based feature selection	5	95.8%

Page | 6

V. CONCLUSION AND FUTURE SCOPE

In this project we have successfully compared the methods of feature selection and the respective accuracy using random forest machine learning algorithm to find the accuracy of the model. By comparing each result we can conclude that maximum accuracy with minimum number of features is calculated using Recursive feature selection with cross validation having maximum cross validation score 97.34% using only 5 features. The second highest accuracy is acquired by two methods namely Recursive feature selection and Tree based feature selection with five features and accuracy of 95.80%.

The data visualization helped to analyze the festures which can accurately classify the data to benign or malignant type. The principal component analysis feature extraction

methods shows that the maximum variance is achieved when number of features are less than or equal to five. The accuracy can be further improved using different machine learning algorithm.

VI. ACKNOWLEDGMENT

I would like to express gratitude to all those people whose support and cooperation has been an invaluable asset during this project. I would also like to thank our guide Dr.Surendra Bhosale for guiding me throughout this project.
I would also like to thank all other teaching and non-teaching staff members of the Electrical Engineering Department, VJTI for directly or indirectly helping me for the completion of the project and the resources provided.

REFERENCES

[1] Youness Khourdifi,Mohamed Bahaj,"Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification", 978-1-5386-4225-2/18/$31.00 ©2018 IEEE.
[2] Kang, M., & Jameson, N. J, "Machine Learning: Fundamentals",2018 Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things.
[3] Siyabend Turgut , Mustafa Dagtekin and Tolga Ensari ,"Microarray Breast Cancer Data Classification Using Machine Learning Methods", 978-1-5386-5135-3/18/$31.00 ©2018 IEEE.
[4] R.D. Ghongade, D.G. Wakde, "Computer-aided Diagnosis System for Breast Cancer Using RF Classifier", IEEE WiSPNET 2017 conference.
[5] Shajib Ghosh, Jubaer Hossain, Dr. Shaikh Anowarul Fattah, Dr. Celia Shahnaz, Asir Intisar Khan, "Efficient Approaches for Accuracy Improvement of Breast Cancer Classification Using Wisconsin Database", 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) 21 - 23 Dec 2017, Dhaka, Bangladesh.
[6] Vincent F. Adegoke, Daqing Chen, Ebad Banissi, Safia Barikzai, "Prediction of Breast Cancer Survivability using Ensemble Algorithms", 978-1-5386-2101-1/17/$31.00 ©2017 IEEE.
[7] Moh'd Rasoul Al-hadidi,Abdulsalam Alarabeyyat,Mohannad Alhanahnah, "Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm", 2016 9th International Conference on Developments in eSystems Engineering
[8] Dana Bazazeh1 and Raed Shubair, "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", 978-1-5090-5306-3/16/$31.00c 2016 IEEE
[9] Marwa Farouk Ibrahim,Adel Ali AI-Jumaily,"PCA indexing based feature learning and feature selection ",978-1-5090-2987-7/16/$31.00 ©2016 IEEE
[10] Aparna.U.R ,Shaiju Paul,"Feature Selection and Extraction in Data mining", 2016 Online International Conference on Green Engineering and Technologies (IC-GET)