Breast Cancer Risk Prediction

Pankaj Kumar Varshney[1], Hemant Kumar[2], Jasleen Kaur[3], Ishika Gera[4]

[1, 2, 3,4]*Department of Computer Science,*
Institute of Information Technology and Management, Janakpuri

pankaj.surir@gmail.com, hemantmbmgla@gmail.com,
kaurjasleen420.jk@gmail.com, ishikagera1998@gmail.com

Abstract.-The number and size of restorative/medical databases are expanding quickly yet the greater part of these information are not investigated for finding the significant and concealed learning. Propelled information and data mining methods can be utilized to find concealed examples and connections. Models created from these strategies are helpful for medicinal specialists to settle on right choices. The present research contemplated the utilization of information mining strategies to create prescient models for bosom (breast) malignant growth repeat in patients who were followed-up for a long time. Objective is to fabricate a model utilizing many machine learning algorithms to foresee whether bosom cell tissue is malignant (cancerous) or benign (non-cancerous).We executed machine learning techniques/algorithms, i.e., k-nearest algorithm, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Artificial Neural Network (ANN) to build up the prescient models. The primary objective of this paper is to think about the execution of these outstanding calculations on our information through affectability, explicitness, and precision to think about the execution of these outstanding calculations on our information through affectability, explicitness, and precision. Our analysis shows that accuracy of DT, RF, SVM and ANN are 0.937, 0.951, 0.965, and 0.958 respectively. The SVM classification model predicts bosom malignancy repeat with least error rate and most astounding precision. The anticipated exactness of the DT demonstrate is the most minimal of all. The results are achieved using 10-fold cross-validation for measuring the unbiased prediction accuracy of each model.

Keywords- Classification; Logistic Regression‟s-nearest algorithm; Decision tree; Random Forest; Gradient Boosting; Support vector machine; Artificial Neural Network.

I. INTRODUCTION

Breast cancer (BC) is the most common cancer in women, affecting about 10% of all women at some stages of their life. In recent years, the incidence rate keeps increasing and data show that the survival rate is 88% after five years from diagnosis and 80% after 10 years from diagnosis [1]. Breast cancer is the second leading cause of cancer death among women. It occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed. Build model to predict whether breast cell tissue is malignant or benign, we will construct a predictive model using SVM machine learning algorithm to predict the diagnosis of a breast tumor. They studied 951 breast cancer patients and used tumor size, auxiliary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age as input variables [7]. Pendharker patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes [4]. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women‟s death is estimated2. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women[3]. Early prediction of breast cancer is one of the most crucial

works in the follow-up process. Data mining methods can help to reduce the number of false positive and false negative decisions [2,3]. Consequently, new methods such as knowledge discovery in databases (KDD) has become a popular research tool for medical researchers who try to identify and exploit patterns and relationships among large number of variables, and predict the outcome of a disease using historical cases stored in datasets. Machine learning is not new to cancer research. Artificial neural networks (ANNs) and decision trees(DTs) have been used in cancer detection and diagnosis for nearly 20 years. Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images to the classification of malignancies from proteomic and genomic (microarray). According to the latest PubMed statistics, more than 1500 papers have been published on the subject of machine learning and cancer. However, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning has been used primarily as an aid to cancer diagnosis and detection [4]. In this paper, using data mining techniques, authors developed models to predict the recurrence of breast cancer by analyzing data collected from ICBC registry. The next sections of this paper review related work, describe background of this study, evaluate three classification models (DT, SVM, and ANN), explain the methodology used to conduct the prediction, present experimental results, and the last part of the paper is the conclusion. To estimate validation of the models, accuracy, sensitivity, and specificity were used as criteria, and were compared. In the present work only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

II. LITERATURE REVIEW

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using data mining approaches such as decision trees [5,6]. Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analyzing a large dataset, the SEER cancer incidence database

[6]. Linden et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumor size, auxiliary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age as input variables

[7] Pendharker patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes [4]. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

III. MATERIAL AND METHODS

Machine Learning, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference [5]. Every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system. It has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different techniques and algorithms [2]. There are two main common types of Machine learning methods known as (i) supervised learning and (ii) unsupervised learning. In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning scheme/model to find patterns or discover the groups of the input data. In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be

estimated, based on this process. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share. Suppose for example that we have collected medical records relevant to breast cancer and we try to predict if a tumor is malignant or benign based on its size. The ML question would be referred to the estimation of the probability that the tumor is malignant or no (1 = Yes, 0=No). It depicts the classification process of a tumor being malignant or not. The circled records depict any misclassification of the type of a tumor produced by the procedure. Another type of ML methods that have been widely applied is semi-supervised learning, which is a combination of supervised and unsupervised learning. It combines labeled and unlabeled data in order to construct an accurate learning model. Usually, this type of learning is used when there are more unlabeled datasets than labeled. When applying a ML method, data samples constitute the basic components [4]. Every sample is described with several features and every feature consists of different types of values. Furthermore, knowing in advance the specific type of data being used allows the right selection of tools and techniques that can be used for their analysis. Some data-related issues refer to the quality of the data and the preprocessing steps to make them more suitable for ML. Data quality issues include the presence of noise, outliers, missing or duplicate data and data that is biased-unrepresentative. When improving the data quality, typically the quality of the resulting analysis is also improved. In addition, in order to make the raw data more suitable for further analysis, preprocessing steps should be applied that focus on the modification of the data. A number of different techniques and strategies exist, relevant to data preprocessing that focus on modifying the data for better fitting in a specific ML method. Among these techniques some of the most important approaches include (i) dimensionality reduction (ii) feature selection and (iii) feature extraction. Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance. Feature selection and Data cleaning should be the first and most important step of your model designing [1,2].In this post, you will discover feature selection techniques that you can use in Machine Learning. Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features. There are many benefits regarding the dimensionality reduction when the datasets have a large number of features. ML algorithms work better when the dimensionality is lower [10]. Additionally, the reduction of dimensionality can eliminate irrelevant features, reduce noise and can produce more robust learning models due to the involvement of fewer features. In general, the dimensionality reduction by selecting new features which are a subset of the old ones is known as feature selection. Three main approaches exist for feature selection namely embedded, filter and wrapper approaches[8,9]. In the case of feature extraction, a new set of features can be created from the initial set that captures all the significant information in a dataset. The creation of new sets of features allows for gathering the described benefits of dimensionality reduction. However, the application of feature selection techniques may result in specific fluctuations concerning the creation of predictive feature lists. Several studies in the literature discuss the phenomenon of lack of agreement between the predictive gene lists discovered by different groups, the need of thousands of samples in order to achieve the desired outcomes, the lack of biological interpretation of predictive signatures and the dangers of information leak recorded in published studies. Fig 1 depicts the feature correlation/selection of diagnosis dataset.

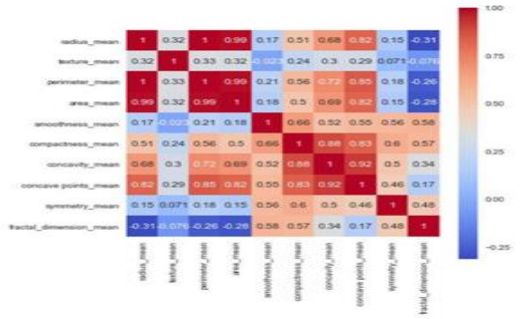

Fig 1 Feature Correlation i.e, red dots correspond to malignant diagnosis and blue to benign.

However, the application of feature selection techniques may result in specific fluctuations concerning the creation of predictive feature lists. Several studies in the literature discuss the phenomenon of lack of agreement between the predictive gene lists discovered by different groups, the need of thousands of samples in order to achieve the desired outcomes, the lack of biological interpretation of predictive signatures and the dangers of information leak recorded in published studies. Fig 2 depicts count of diagnosis in which green color shows benign tumor and blue color shows malignant tumor.

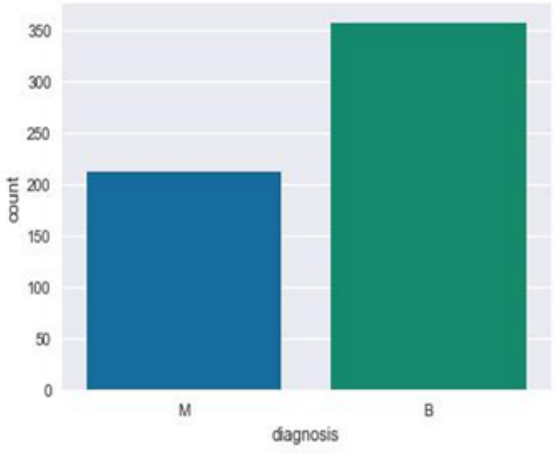

Fig 2 Diagnosis:- Benign(B)-357, Malignant(M)-219

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets. Training set: A data scientist uses a training set to train a model and define its optimal parameters, a subset to train a model. Test set: A test set is needed for an evaluation of the trained model and its capability for generalization, a subset to test the trained model. Validation set: The purpose of a validation set is to tweak a model‟s hyper parameters higher-level structural settings that can‟t be directly learned from data. These settings can express, for instance, how complex a model is and how fast it finds patterns in data. The k-NN algorithm is arguably the simplest machine learning algorithm. To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset—it‟s “nearest neighbors.” Fig 3 depicts the training dataset and test dataset accuracy when the n_neighbors value is 3.The figure shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x-axis. Considering a single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy drops, indicating that using the single nearest neighbor leads to a model that is too complex. This suggests that we should choose n_neighbors=3.Then the Accuracy of K-NN classifier on training set and test set is 0.96 and 0.92 respectively. Logistic Regression is one of the most common linear classification algorithm is logistic regression. Logistic regression examines the relationship between a binary outcome (dependent) variable such as presence or absence of disease and predictor (explanatory or independent) variables such as patient demographics or imaging findings. Fig 4 depicts the coefficient magnitude of features and the accuracy level change when we set value of C i.e., C=1 provides quite good performance, with 96% (0.955) accuracy on training and 0.94 accuracy on test set. C=100 provides higher accuracy on both training set (0.967) and test set (0.972). C=0.01 provides lower accuracy on the training set (0.948) and much lower accuracy on the test set (0.895).

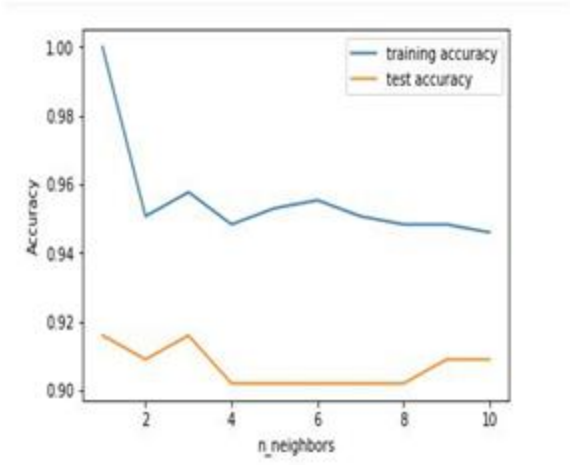

Fig 3 Training and test set accuracy

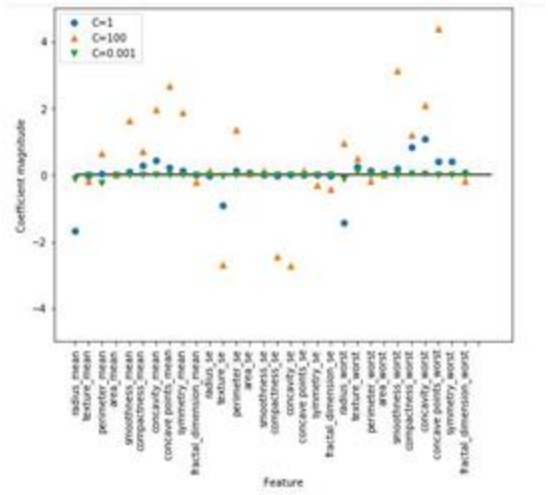

Fig 4 Coefficient Magnitude of features

IV. MACHINE LEARNING TECHNIQUES

Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making [7]. These techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [5,6].Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time [2]. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results — in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer [1, 3]. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient Boosting Classifier is used to find out accurate prediction. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane in an N-dimensional space (N—the number of features) that distinctly classifies the data points either side. Support vector machine (SVM) is an emerging powerful machine learning technique to classify cases. SVM has been used in a range of problems and they have already been successful in pattern recognition in bioinformatics, cancer diagnosis [6]. SVM is a maximum margin classification algorithm rooted in statistical learning theory. It is the method for classifying both linear and non-linear data. It uses a non-linear mapping technique to transform the original training data into a higher dimension. It performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors[9].Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input [5]. The patterns they recognize are numerical, contained in vectors, into which all real-world data, are it images, sound, text or time series, and must be translated.

V. DISCUSSION AND RESULTS

This section presents the result of all the machine learning algorithms that we have used to do prediction breast cancer risk. This paper has explored risk factors for predicting breast cancer by using machine learning techniques. Each technique has its

own limitations and strengths specific to the type of application. Our results show that SVM is the best predictor and indicator of breast cancer because it gives the higher accuracy in predicting data.

VI. DECISION TREE

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [5, 6]. Decision Tree Classifier is a class capable of performing multi-class classification on a dataset. As with other classifiers, it takes as input two arrays: an array X, sparse or dense, of size [n_samples, n_features] holding the training samples, and an array Y of integer values, size [n_samples], holding the class labels for the training samples. The maximum depth of the tree „max_depth" is an argument of Decision Tree Classifier. If the max_depth of tree is none, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples and then accuracy on training set is 1.000 and accuracy on test set are 0.937. If the max_depth is set a value of 4 then the accuracy on training set is 0.986 and accuracy on test set is 0.937.Feature importance in decision trees it rates how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means "not used at all" and 1 means "perfectly predicts the target." Fig 5 illustrates feature perimeter worst is by far the most important feature. This confirms our observation in analyzing the tree that the first level already separates the two classes fairly well.

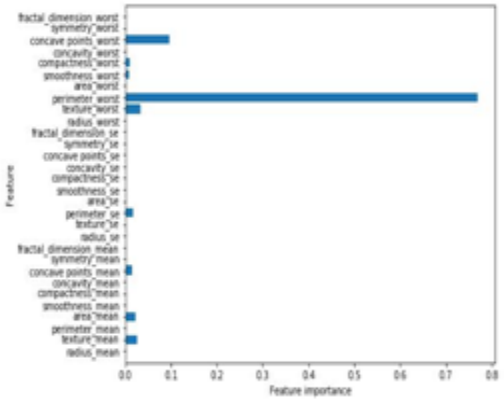

Fig 5 Feature importance of decision tree

A. Random Forest

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time [2]. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results — in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer [1, 3]. Random forest classifier creates a set of decision trees from randomly selected subset of training set and the accuracy on training set is 0.995 and accuracy on test set is 0.951. The random forest gives us an accuracy of 95.8%, better than a single decision tree, without tuning any parameters. Similarly to the single decision tree, the random forest also gives a lot of importance to the "worst radius" feature, but it also chooses "perimeter worst" to be the most informative feature overall. Fig 6 illustrates similarly to the single decision tree, the random forest also gives a lot of importance to the "worst radius" feature, but it also chooses "perimeter worst" to be the most informative feature overall.

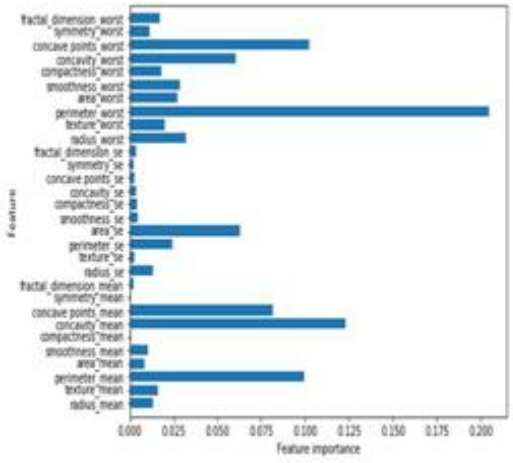

Fig 6 Feature importance of random forest

B. Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an

ensemble of weak prediction models, typically decision trees. Gradient Boosting Classifier is used to find out accurate prediction. When „random state" argument of GB Classifier is zero (0) then accuracy on training set is 1.000 and accuracy on test set is 0.944. As the training set accuracy is 100%, we are likely to be over fitting. To reduce over fitting, we could either apply stronger pre-pruning by limiting the maximum depth or lower the learning rate. When „max_depth" argument of GB Classifier is one (1) then accuracy on training set is 0.988 and accuracy on test set is 0.937. When „learning rate" argument of GB Classifier is (0.01) then accuracy on training set is 0.984 and accuracy on test set is 0.930.Both methods of decreasing the model complexity reduced the training set accuracy, as expected. In this case, none of these methods increased the generalization performance of the test set. Fig 7 depicts that the feature importance's of the gradient boosted trees are somewhat similar to the feature importance of the random forests, though the gradient boosting completely ignored some of the features. It chooses "perimeter worst" to be the most informative feature overall.

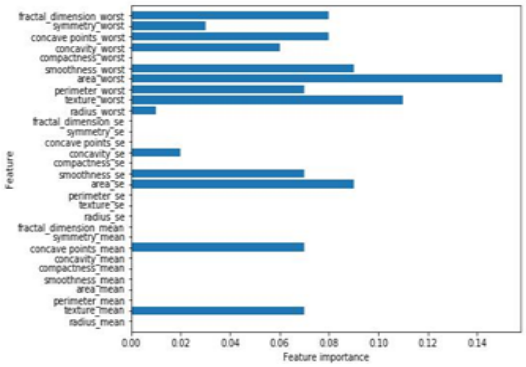

Fig 6 Feature importance of random forest

C. Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient Boosting Classifier is used to find out accurate prediction. When „random state" argument of GB Classifier is zero (0) then accuracy on training set is 1.000 and accuracy on test set is 0.944. As the training set accuracy is 100%, we are likely to be over fitting. To reduce over fitting, we could either apply stronger pre-pruning by limiting the maximum depth or lower the learning rate. When „max_depth" argument of GB Classifier is one (1) then accuracy on training set is 0.988 and accuracy on test set is 0.937. When „learning rate" argument of GB Classifier is (0.01) then accuracy on training set is 0.984 and accuracy on test set is 0.930.Both methods of decreasing the model complexity reduced the training set accuracy, as expected. In this case, none of these methods increased the generalization performance of the test set. Fig 7 depicts that the feature importances of the gradient boosted trees are somewhat similar to the feature importances of the random forests, though the gradient boosting completely ignored some of the features. It chooses "perimeter worst" to be the most informative feature overall.

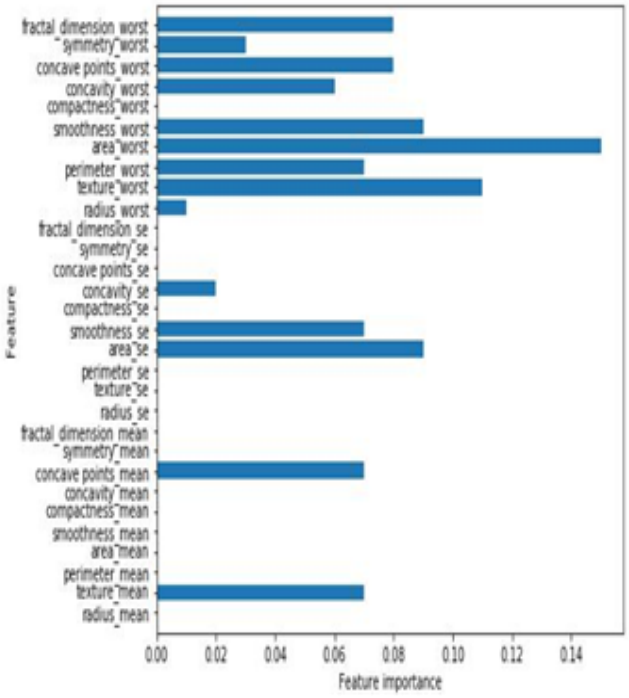

Fig 7 Feature importance of Gradient Boosting

D. Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane in an N-dimensional space (N—the number of features) that distinctly classifies the data points either side. Support vector machine (SVM) is an emerging powerful machine learning technique to classify cases. SVM has been used in a

range of problems and they have already been successful in pattern recognition in bioinformatics, cancer diagnosis [6]. SVM is a maximum margin classification algorithm rooted in statistical learning theory. It is the method for classifying both linear and non-linear data. It uses a non-linear mapping technique to transform the original training data into a higher dimension. It performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors [9]. By importing Support Vector Machine the accuracy on training set is 1.00 and accuracy on test set is 0.63 the model over fits quite substantially, with a perfect score on the training set and only 63% accuracy on the test set. SVM requires all the features to vary on a similar scale. We will need to rescale our data that all the features are approximately on the same scale. Min Max Scaler is used to rescale the data so that all features will vary on same scale. By importing Min Max Scaler the accuracy on training set is 0.95 and accuracy on test set is 0.94. Scaling the data made a huge difference as training and test set performance are quite similar but less close to 100% accuracy. C and gamma are the parameters of Radial Basis Function (RBF) kernel SVM. Now, we can try increasing either C or gamma to fit a more complex model. By increasing C=1000 in SVM the accuracy on training set is 0.986 and accuracy on test set is 0.965 Here, increasing C allows us to improve the model significantly, resulting in 96.5% test set accuracy.

E. Neural Networks

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns.

By importing MLP Classifier the accuracy on training set is 0.63 and accuracy on test set is 0.63.Results are not good. Neural networks also expect all input features to vary in a similar way, and ideally to have a mean of 0, and a variance of 1. Now we need to scale our data by importing Min Max Scaler then the accuracy on training set is 0.962 and accuracy on test set is 0.958.The results are much better after scaling, and already quite competitive. After scaling the data now we will again classify the data with MLP Classifier. After scaling the data, MLP Classifier is used then the accuracy on training set is 0.923 and accuracy on test set is 0.895.

Fig 8 shows the weights that were learned connecting the input to the first hidden layer. The rows in this fig correspond to the 30 input features, while the columns correspond to the 100 hidden units. Light colors represent large positive values, while dark colors represent negative values. One possible inference we can make is that features that have very small weights for all of the hidden units are “less important” to the model. We can see that “mean smoothness” and “mean compactness,” in addition to the features found between “smoothness error” and “fractal dimension error,” have relatively low weights compared to other features. This could mean that these are less important features or possibly that we didn’t represent them in a way that the neural network could use.

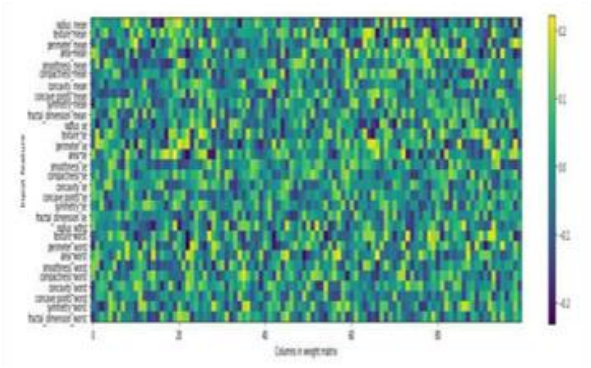

Fig 8 Weight Matrix

VII. CONCLUSION

To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed many algorithms: SVM, ANN, DT, RF, GB, and k-NN on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy of SVM reaches and accuracy of 97.13% and out performs, therefore, all other algorithms. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate. The results indicated that SVM are the best classifier predictor with the test dataset, followed by

ANN and DT. Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

Breast cancer has created a terrible situation in almost all over the world according to this study and discussion. It has been observed that the death rate is gradually coming down in some developed countries like the UK and US because of the developed technologies used in diagnosis and awareness. But in developing countries like India the situation is not good and some effective steps should be taken in this direction without any delay [3]. This study has been made on methodologies by which the breast cancer can be detected at early stages by using the breast cancer data set. It is clear from this study that the Association Rule Mining, Classification, and Clustering and Evolutionary Algorithms are good at detection and classification of breast cancer data. It is also observed that if the properties of the symptoms are identified correctly, the chances of accurate detection will improve [8]. It is also observed by the results of the previous methods that the classification algorithm increases the possibility or improved detection accuracy. The characteristics of breast cancer symptoms are different, so the chances of good results by using single algorithm are less. But by the use of combined algorithms at different levels will produce good results. So it is concluded that the framework based on data mining and evolutionary algorithms can be a milestone in case of breast cancer detection [10].

REFERENCES

[1] https://www.omicsonline.org/using-three-machine-learning-techniques-for-predictingbreast-cancer-2157 7420.1000124.php?aid=13087

[2] https://www.sciencedirect.com/science/article/pii/S2001037014000464https://www.sciencedirect.com/science/article/pi i/S1877050916302575

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/

[4] https://pubs.rsna.org/doi/pdf/10.1148/rg.301095057

[5] http://journal.waocp.org/article_31073_4ac28ea9398b1d19335b6f44a0e79afd.pdf

[6] https://breast-cancer-esearch.biomedcentral.com/track/pdf/10.1186/bc r3110.

[7] https://s3.amazonaws.com/academia.edu.documents/32919287/V3I1201402.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1555299017&Signature=rVk6dRW4pUDsqPv6J Xmm3zL%2B2Sg%3D&response-content-ition=inline%3B%20filename%3DV3I1201402. Pdf

[8] https://pdfs.semanticscholar.org/7bf7/3b15b7fd64c2b01a718a2848b4a3d35b939.pdfhttp://cancerpreventionresearch.aacrjournals.org/content/canprevres/9/1/13.full.pdf