Model Selection for Predicting Breast Cancer using Supervised Machine Learning Algorithms

Ajit Kumar
Dept. of Computer Science & Engineering
Netaji Subhash Engineering College
Kolkata, India
ajit.kum0404@gmail.com

Rajkumar Patra
Dept. of Computer Science & Engineering
Netaji Subhash Engineering College
Kolkata, India
rajkr.patra@gmail.com

Anupam Ghosh
Dept. of Computer Science & Engineering
Netaji Subhash Engineering College
Kolkata, India
anupam.ghosh@rediffmail.com

***Abstract*—Breast Cancer is the most common malignancy in women affecting 2.1 million women every year and causing the maximum number of deaths in women due to cancer. It occurs as a result of the unusual development of cells in the breast tissue, which is generally referred to as a Tumor. A tumor does not signify cancer. It may be not cancerous (benign), pre-cancerous (pre-malignant), or cancerous (malignant). Various types of tests such as mammograms, MRIs, ultrasound, and biopsy are frequently used to identify breast cancer. Early detection and treatment will help to improve breast cancer outcomes as well as survival. Therefore, this paper consists of a relative study of the breast cancer prediction using different supervised machine learning algorithms like Logistics Regression, K-Nearest Neighbors, Decision Tree Classifier, Gaussian NB, and Support Vector Machine on the UCI repository dataset. Concerning the performance of all the models, the accuracy score, precision, recall, and F-score of each model have been compared. After using various models, we got to see that Logistic Regression is a well-suited algorithm for Breast cancer prediction and came up with better accuracy and other performance indices as compared with other models.**

Keywords—Data Mining, Breast cancer prediction, Supervised machine learning, Classification models, Logistic Regression, Support Vector Machine.

I. Introduction

According to the report of the World Health Organization (WHO), Breast Cancer is the second fatal cancer among women following lung cancer. As per the statistics[1] provided by the WHO in the year 2018, it has been estimated that out of 2.1 million new cases 627,000 women will lose their life due to breast cancer which is around 15% of all the cancer deaths in women. This rate is higher with women in more developed countries. The rate is rising in nearly every county globally. People visit oncologists in case of any symptoms of breast cancer. The oncologist can easily identify breast cancer by performing tests such as Magnetic resonance imaging (MRI), mammogram, ultrasound and tissue biopsy, etc. Based on the test results, the doctor may further recommend the tests, therapy, and patient regularly undergo sentinel node biopsy which helps to determine whether cancer has spread beyond a primary tumor into the lymphatic system. It is very crucial to detect breast cancer in the early stage to boost the prediction and reduce the death rate notably [2]. The patient can also take medical treatments as early as possible. A tumor can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Classification of tumors can help the patients to avoid irrational fear of developing cancer and avoid undertaking needless treatments. The application of machine learning with its proper advancements widely helps in the prediction of breast cancer from critical features of UCI datasets. Breast cancer is mainly a classification problem where the dataset can be classified into benign and malignant tumors using machine learning algorithms which further helps in predicting the report whether the patient has a cancerous tumor or not i.e., benign and malignant tumors. Thus, the various researches are being carried out for the proper diagnosis and categorization of patients into benign and malignant tumors group. Our aim is to develop a prediction system that can predict breast cancer on the Wisconsin breast cancer dataset using supervised machine learning models.

II. Related Work

While going through the research papers many of them show remarkable classification accuracy. Data mining has been applied to various medical datasets of the past and current research papers where classification is one of the most significant techniques to deal with breast cancer problems. Vikas Chaurasia, BB Tiwari, and Saurabh Pal [3] compared the performance of three common data mining methods i.e., Naive Bayes, RBF Network, and J48 in order to build a predictive model to identify the best classifier for breast cancer on the Wisconsin breast cancer (original) datasets. The final result shows that the performance of the Naive Bayes model is the best among the three models with a classification accuracy of 97.36%. Li Rong Sun Yuan [4] has explored SVM and KNN and finally came up with the SVM-KNN algorithm which is an enhanced classification algorithm of the SVM. Experimental results show that SVM achieves a classification accuracy of 96.09% on the test subset while on the other hand; SVM-KNN achieves 98.03% accuracy. U Ojha and Dr. S Goel [5] used different data mining algorithms on the Wisconsin prognostic breast cancer dataset which consists of four clustering algorithms namely EM, PAM, K-means, and fuzzy c-means and four classification algorithms namely Naive Bayes, C5.0, SVM, and KNN using an R programming tool. Results show that the Decision Tree (C5.0) and SVM are the best predictors with an accuracy of 81% on the proposed sample. Haifeng Wang and Sang Won Yoon [6] compared SVM, ANN, Naive Bayes, and AdaBoost classifier in order to identify a powerful machine learning model for predicting breast cancer. They worked on the two extensively used datasets, Wisconsin breast cancer database, and Wisconsin diagnostic breast cancer to evaluate the performance of these models. Further, they implemented PCA for dimensionality reduction.

III. Proposed Methodology

Our proposed methodology includes the use of supervised machine learning algorithms and various

978-1-7281-7340-5/20/$31.00 ©2020 IEEE

classification techniques like Logistic Regression, Decision Tree Classifier (CART), Support Vector Machine, Gaussian Naive Bayes, and K-Nearest Neighbor with Principal Component Analysis based feature selection approach in the model building.

A. *Exploratory Data Analysis (EDA)*

Visually inspecting the data is very necessary to understand the dataset. Since the labels in the data are discrete, the prediction falls into the two categories, i.e. Malignant or Benign. There are 569 samples of malignant and benign tumor cells in the dataset. Here, we have identified the problem statement and saved the cleaner version of the data frame for the Exploratory Data Analysis.

The purpose of EDA is to find the clues about the tendencies of data, its qualities and formulate the assumptions and hypothesis of our analysis which bring Data processing to be successful, it is essential to have a basic statistical description and to opt which data values should be treated as noise or outliers.

After plotting the histogram, as shown in Fig. 1, there may be exponential distribution in the attributes like concavity and concavity point, whereas texture, symmetry attributes, and smoothness may have a nearly Gaussian distribution. On the other hand, while visualizing the distribution of data using the density plots, we observed that there may be exponential distribution in the attributes like perimeter, area, concavity, radius, and compactness. Attributes like texture, smooth, and symmetry have Gaussian distribution which is quite interesting since many data mining methods assume a Gaussian univariate distribution on the input attributes.

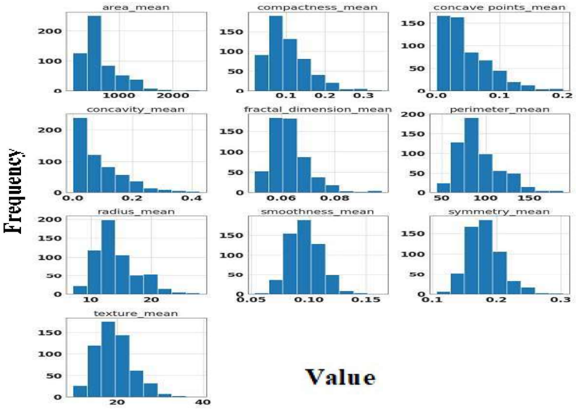

Fig. 1. Histograms of mean variables

B. *Data Preprocessing*

The data preprocessing phase is considered one of the essential phases for any data mining problem which involves the transformation of raw or real-world data into a cleaned and understandable format. For data preprocessing to be successful, it is essential to have an overall picture of exploratory data analysis. It involves a number of activities which are:

- Handling missing values
- Assigning numerical values to categorical data
- Normalizing the attributes

In the Wisconsin breast cancer dataset, all the attributes contain non-null values which show that there is not any missing value that is to be handled. Since dataset contains categorical variable in diagnosis column which is to be converted by calling the transform method of Label Encorder on two dummy variables as:

- 0 = Benign (B, Non-Cancerous)
- 1 = Malignant (M, Cancerous)

After that, the feature standardization technique is used in the dataset which transforms features that have the Gaussian distribution and standard deviation to a standard Gaussian distribution with 0 as the mean value and 1 as the standard deviation value. Further on, to estimate the performance of a machine learning algorithm, different training, and testing datasets need to be used. Here, we have split up the dataset into a training and testing subset (80% training, 20% testing dataset) and trained the algorithm on the 80% training dataset. We have used the 20% testing dataset to predict 114 observations and evaluate the prediction values against the expected results which will further help in algorithm comparison.

C. *Feature Selection*

A lot of feature pairs divide nicely the data to a similar extent, so it makes sense to use one of the dimensionality reduction methods to try to use as many features as possible and maintain as much information as possible. We normally select a subset of features that have a high correlation with the class labels. The PCA based feature selection strategy helps us to find the essential features depending on the covariance matrix of the dataset. PCA plots the actual data into a dimensional space with fewer features in such a way that variance is maximized. After applying PCA as shown in Fig. 2, we get two principal components PCA_1 and PCA_2 which provides the highest variance, and second highest variance respectively. Thereafter, we usually split the cleaned dataset into the training and testing sets in the ratio of 4:1.

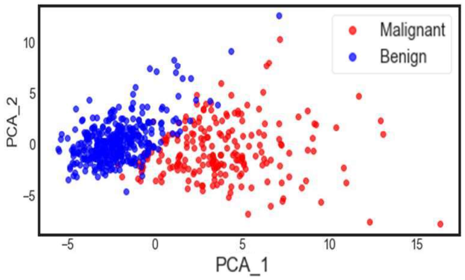

Fig. 2. PCA transformation

D. *Model Selection*

Since the labels in the data are discrete, the prediction falls into the two categories i.e. Malignant or Benign. This type of problem in machine learning is called a classification problem. Thus, the goal of our study is to categorize whether breast cancer is benign or malignant besides predicting the recurrence and non-recurrence of malignant cases after a certain time. The Wisconsin dataset contains the known set of input datasets and its known responses to the data (output) which consists of two categories. So supervised classification algorithms are used to prepare the model. Here, we will go through five different classification techniques used in supervised machine learning.

- Logistic Regression
- Decision Tree Classifier
- Gaussian Naive Bayes
- K-Nearest Neighbors
- Support Vector Machine

i. Logistic Regression

The Logistic Regression is a supervised classification algorithm. It is used to predict the probability of the target variable. It is based on linear regression to evaluate the output and minimize the error. It uses a complex cost function which can be defined as a sigmoid function or logistic function. We have used Label Encoder to label the categorical data and then after using Standard Scalar a part of preprocessing which transforms the data in such a manner that it has a mean as 0 and standard deviation as 1.

ii. Decision Tree

The Decision Tree is a non-parametric supervised learning technique. It can be used for the task of classification as well as regression. It is well known and widely used methodology. A decision tree algorithm consists of an internal node that represents a check on a feature, each leaf node in the tree represents a class label, whereas conjunctions of features that lead to those class labels are represented by the branches. The path from the root to the leaf represents classification rules. In this problem statement, leaf nodes are categorized as benign and malignant. Then certain rules are established to check if the tumor is of which category.

iii. Gaussian Naive Bayes

The Gaussian Naive Bayes method is used for estimating the probability of a dataset to belong to a class using Bayes' rule. It is normally used when the features have continuous values. Since we have seen in Figure I, features like texture, smooth, and symmetry which follows the Gaussian or nearly Gaussian distribution and in density plot of the attributes perimeter, radius, area, concavity, compactness follows the same because of which the training data features fit well in the model.

iv. K-Nearest Neighbor

The K-Nearest Neighbor method is a supervised algorithm where an object is classified depending on the majority vote of its k nearest neighbors. It is non–parametric techniques because the classification of the test data point depends on the nearest training data point without considering the parameters of the data point. The accuracy of this model might increase with an increase in the number of nearest neighbors.

v. Support Vector Machine

The Support Vector Machine is used for both classification as well as the regression problem. This algorithm largely used for image reorganization besides the aspect-based and color-based classification. It tries to find a classifier which maximizes the margin between positive and negative data points. The SVM is mostly used for the dataset where the number of features and instances is high. The following two parameters of the SVM algorithm can be tuned to increase performance:

The kernel function, and

The value of c (amount of the margin relaxation)

The Radial Basis Function is used as the default kernel function of SVM with a c value set to 1.0.

The data obtained from the patient's digital image of the fine needle aspiration of a breast mass used to form the dataset. This dataset will further help to predict breast cancer with the help of the selected model. The complete working process of the proposed methodology is shown as a flowchart in Fig. 3.

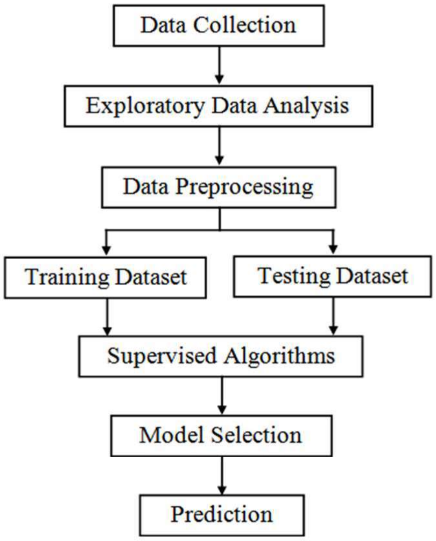

Fig. 3. Flowchart of the proposed methodology

IV. Results and Discussion

A. Dataset Description

The features in the Wisconsin breast cancer dataset contain the following columns:

- ID number (Unique ID number of samples)
- Diagnosis (M=malignant, B=benign where 212 malignant and 357 benign)

The dataset contains ten real values unique features that are obtained from digitized images of cell nuclei where features values are recorded in the data frame up to four significant digits. These are as follows:

- Texture (standard deviation of gray-scale values)
- Radius (mean of distances from the center to points on the perimeter)
- Perimeter (core tumor's size)
- Area (inside the boundary of core tumor)
- Compactness (perimeter^2 / area – 1.0)
- Smoothness (local variation in radius lengths)
- Concavity (severity of concave portions of the contour)
- Fractal dimension ("coastline approximation" – 1)
- Concave points (number of contour's concave portions)

- Symmetry (similar area of tumor parts that matches)

Columns 3-32 in the dataset are filled by the mean, standard error, and worst/largest of these 10 features which were obtained from each image.

B. Performance Measurement

We have implemented all the five models on the dataset and measure the performance of each model.

i. Performance of Logistic Regression

The accuracy of the Logistic Regression model is 98.24% on the test dataset and AUC (Area under the ROC curve) is 0.98115. Table I shows the comparative study of predicted vs. actual output on test dataset where only 2 observations are misclassified which is least among all the models.

TABLE I. CONFUSION MATRIX OF LOGISTIC REGRESSION

		Predicted	
		Benign	***Malignant***
Actual	***Benign***	71	1
	Malignant	1	41

ii. Performance of Decision Tree

Initially, Decision Tree Classifier gives the accuracy score of 91.22% but later on with the help of Bagging Classifier which helps to reduce the variance of Decision Tree the accuracy climbs up to 94.73% on the test dataset with better confusion matrix as only 6 observations are misclassified as shown in Table II.

TABLE II. CONFUSION MATRIX OF DECISION TREE

		Predicted	
		Benign	***Malignant***
Actual	***Benign***	73	2
	Malignant	4	35

iii. Performance of Gaussian Naive Bayes

The Gaussian Naïve Bayes gives an accuracy score of 93.85% on the test dataset. Below Table III shows the comparative study of predicted vs. actual output on test dataset where 7 observations are misclassified out of which 4 are being malignant and 3 are benign.

TABLE III. CONFUSION MATRIX OF GAUSSIAN NAÏVE BAYES

		Predicted	
		Benign	***Malignant***
Actual	***Benign***	72	3
	Malignant	4	35

iv. Performance of K-Nearest Neighbor

K-Nearest Neighbors Classifier predicts the accuracy score of 92.10% but after preprocessing and using principal component analysis there are only 5 observations that are misclassified as shown in Table IV. This model shows 95.61% accuracy score with the Minkowski metric.

TABLE IV. CONFUSION MATRIX OF K-NEAREST NEIGHBOR

		Predicted	
		Benign	***Malignant***
Actual	***Benign***	72	0
	Malignant	5	37

v. Performance of Support Vector Machine

Without data preprocessing SVM predicts 65.78% of the score which on after Standard Scalar data transformation and PCA the model shows the best value of 98.02% score on Grid search having c = 0.1 in the linear kernel which on after proper tuning the model show 97.36% of accuracy score on test subsets. There are only 3 misclassified observations that are malignant as shown in Table V.

TABLE V. CONFUSION MATRIX OF SUPPORT VECTOR MACHINE

		Predicted	
		Benign	***Malignant***
Actual	***Benign***	72	0
	Malignant	3	39

C. Performance Analysis

After a brief description of the concepts of data preparation, exploratory data analysis for understanding the behavior of the dataset, data pre-processing has been performed in different ways for different models as per the requirement so that features of the dataset with higher values do not dominate when fitting the model on small scale. Since the breast cancer dataset contains 32 attributes so, principal component analysis helps in finding the most impactful features of the dataset based on the covariance matrix. Further, split the cleaned dataset into 4:1 for training and testing subsets. In the present review, different machine learning models are prepared in order to come with the best model for the prognosis of breast cancer. Of all the five applied algorithms namely Logistics Regression (LR), Decision Tree (CART), Support Vector Machine (SVM), Gaussian Naive Bayes (NB) and K-Nearest Neighbors (KNN) k-fold cross-validation method have been performed with 10fold technique which means the data set segregated in ten different chunks, nine chunks help in training subset and remaining one for testing subset whose accuracy plot has shown in Fig. 4.

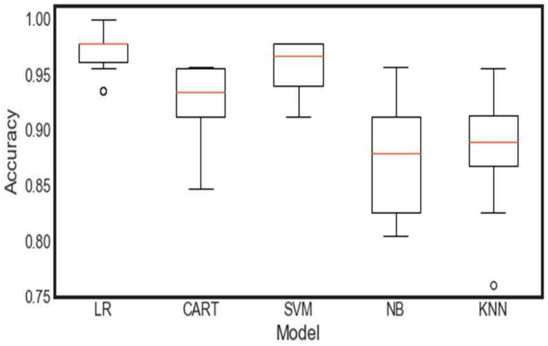

Fig. 4. Performance comparison

The result presented in Table VI shows that Logistic Regression has the best precision, recall, F-score as well as accuracy on the test dataset which is far better than remaining models. On the whole, after comparing all the five models on different parameters the results show that the modeling of breast cancer as a classification task performs better using Logistic Regression as compared to other models.

TABLE VI. PERFORMANCE MEASURE INDICES

	LR	CART	SVM	NB	KNN
Accuracy (%)	98.24	94.73	97.37	93.85	95.61
Precision (%)	98.00	95.00	97.00	94.00	96.00
Recall (%)	98.00	95.00	97.00	94.00	96.00
F-score (%)	98.00	95.00	97.00	94.00	96.00

However, on comparing the performance of the models with the related studies done in the recent past [7-9] on the Wisconsin Breast Cancer dataset, we have identified that all models except CART and KNN perform better in terms of accuracy, precision, recall, and f-score. The comparison table VII clearly shows that the accuracy of Decision Tree (CART) in the past is little bit impressive, but data analysis and data preprocessing helps Logistic Regression to be more superior as compared to other models.

TABLE VII. RESULT COMPARISION WITH RECENT PAST

Models / **Metric**		LR	CART	SVM	NB	KNN
Accuracy (%)		98.2	94.7	97.4	93.8	95.6
	Past	95.8	95.8	97.2	91.2	95.8
Precision (%)		98.0	95.0	97.0	94.0	96.0
	Past	93.5	96.5	97.5	87.5	93.5
Recall (%)		98.0	95.0	97.0	94.0	96.0
	Past	93.5	95.0	97.0	88.9	93.5
F-score (%)		98.0	95.0	97.0	94.0	96.0
	Past	93.0	95.5	97.0	-	93.0

Thus, Logistic Regression outperformed all other models in all the metrics for which we performed comparative analysis. SVM also performed well with 97.4% accuracy that is close to the accuracy of the LR model as it is efficient for complex as well as the multi-dimensional dataset.

V. CONCLUSION

We came up with five different supervised machine learning models namely Logistics Regression (LR), Decision Tree (CART), Gaussian Naive Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) are reviewed and their accuracies are compared with each other in order to find the best-supervised model which is suitable for the prediction of breast cancer using Wisconsin dataset. Since the dataset contains 32 features so, dimensional reduction helps in decreasing the multidimensional data into few dimensions. On the whole, the above study proposed that Logistic Regression is efficient for the detection of breast cancer as compared to all the other models while dealing with the complex dataset. Further ensemble learning can be explored for better performance of the classification techniques. On the other hand optimization techniques for different models will also be taken care of as consideration and the SEER dataset can also be studied in the future.

REFERENCES

[1] WHO breast cancer statistics (https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/).

[2] Yi-Sheng Sun, Zhao Zhao, Han-Ping-Zhu, "Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences, Hangzhou, 2017, pp. 1387-1397, doi: 10.7150/ijbs.21635.

[3] VikasChaurasia, BB Tiwari and Saurabh Pal, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms and Computational Technology, 2018, pp. 119-126, DOI: 10.1177/1748301818756225.

[4] L. Rong and S. Yuan, "Diagnosis of Breast Tumor Using SVM-KNN Classifier," 2010 Second WRI Global Congress on Intelligent Systems, Wuhan, 2010, pp. 95-97, doi: 10.1109/GCIS.2010.278.

[5] U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, Noida, 2017, pp. 527-530, doi: 10.1109/CONFLUENCE.2017.7943207.

[6] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction using Data Mining Method, Industrial and Systems Engineering Research Conference, 2015.

[7] K. Chakradeo, S. Vyawahare and P. Pawar, "Breast Cancer Recurrence Prediction using Machine Learning," 2019 IEEE Conference on Information and Communication Technology, Allahabad,India,2019,pp.1-7,doi: 10.1109/CICT48419.2019.9066248.

[8] David A. Omondiagbe, Shanmugam Veeramani, Amandeep S. Sidhu. "Machine Learning Classification Techniques for Breast Cancer Diagnosis", IOP Conference Series: Materials Science and Engineering, 2019, Volume: 495, Number 1, doi: 10.1088/1757-899X/495/1/012033.

[9] Puja Gupta, Shruti Garg. "Breast Cancer Prediction using varying Parameters of Machine Learning Models", Procedia Computer Science 171:593-601, Jan 2020, doi: 10.1016/j.procs.2020.04.064.