Breast Cancer Prognosis Using Machine Learning Models

Gopika G Kini[1], Indrani P[2], Niharika Patil[3], Nirusha N Nayak[4] and Dr. Shabana Sultana[5] *[1,2,3,4]The National Institute of Engineering, Mysuru, KA 570008 India*

[5]Assistant Professor, Dept. of Computer Science and Engineering, NIE, Mysuru, KA 570008 India (E-mail: [1]gopika1697@gmail.com , [2]nirushanayak31@gmail.com, [3]patil.niharika09@gmail.com, [4]indranipanchangam@gmail.com, [5]shabana@nie.ac.in)

Abstract**—** Breast Cancer has been a common cause of death among women worldwide. Timely diagnosis of the disease is very essential for proper treatment and decision making. Human limitations affect the rate of correct diagnosis. Machine learning models are very effective in classifying the data. The classification algorithms also help experts minimize possible errors. In this study, we have used Wisconsin diagnostic breast cancer (WDBC) dataset to classify tumors as benign and malignant. Performance analysis of the six classification algorithms – KNN, Naïve Bayes, Random Forest, Logistic Regression, SVM and Decision tree is done in terms of effectiveness and efficiency of the model

Keywords**—***Machine learning, Classification, Breast Cancer Classification, kNN, Decision Tree, Naïve Bayes, SVM, Random Forest, Logistic Regression*

I. INTRODUCTION

Cancer is a type of disease that involves abnormal growth of cells that can rapidly spread to other parts of the body. The erratic growth of cells causing tumors can be cancerous or non-cancerous. The non-cancerous tumors are called benign tumors. They are harmless and they do not spread to other parts of the body. The cancerous tumors, called malignant tumors can easily spread to other parts and thus need immediate attention. Breast cancer is the second most common cancer diagnosed in women in the United States, recording up to 1 million cases a year and the leading cause of death among women between the age group 40-45 years. If the diagnosis is not done in the earlier stages, breast cancer could be fatal. To facilitate the early diagnosis by physicians, a precise, reliable and consistent system is required that can distinguish between benign and malignant tumors without having to undergo surgical biopsy. Over the past few decades due to advances made in science and technology, Information and Communication technologies like artificial intelligence play a major role in the field of medical sciences.

Machine learning is a set of tools utilized for the creation and evaluation of algorithms that facilitate prediction, pattern recognition, and classification. Machine Learning is applied to the non-trivial process of extracting and presenting implicit knowledge, previously unknown, useful and comprehensible, from large data sets, with an objective to predict automated form tendencies and behaviors. It includes steps like collecting data, cleaning data, selecting the model, training the model and finally testing the model. The models created can either be descriptive or predictive. A predictive model learns from the known data and predicts the needed results. Classification is a type of supervised learning algorithms which classifies or maps the data into predefined groups which they belong to. Pattern recognition, image recognition, medical diagnosis, etc. are some of the applications of classifiers. The application of classifiers in the medical field is increasing gradually. It helps the physicians in the process of decision making. It is extensively used to classify tumors and other malignancies, predict sequences of genes responsible for cancer and determine the prognostic. In this paper classification algorithms are used to classify the cancer cells into benign and malignant.

Data mining and machine learning models can be used together to build classifying models for efficient diagnosis and prognosis of breast cancer. In this paper, we implement four such classifiers; Naïve Bayes, k-Nearest Neighbor, Support Vector Machine and Decision tree for the classification of the tumor as benign or malignant, and compare their efficiency and effectiveness in terms of accuracy, precision, and specificity. We make use of the Wisconsin Diagnostic Breast Cancer dataset for this purpose.

II. RELATED WORK

Naïve Bayes is the algorithm used in paper [1] which uses the insight of data and tools are used to discover the relationship between data . The paper aims at building a user interface which screens the patient record and try to detect the probability of having breast cancer by Naive Bayes classifier. As the breast cancer is classified as one of the most dangerous disease early detection with much more less effort with this classifier is approached with Wimbledon dataset.

In the paper [2] the SVM algorithm is implemented where the dimensionality of the features is reduced into 2 features with ICA, which resulted in increase in accuracy . The SVM is implemented with reduced dimensionality and with RGB kernel and polynomial kernel and accuracy of these systems is compared .SVM gave accuracy of 94.40% and it increased with dimensionality reduction.

Paper [3] techniques such as fuzzy, clustering, SDC techniques are used to classify breast cancer with feature selection which indicates SVM has greater accuracy rate. It contains % sensible feature which are correlated and the accuracy obtained was 99.51 %.

Paper [4] a lump in a breast is of great concern for the breast cancer , USG is the one which helps in the evaluation .

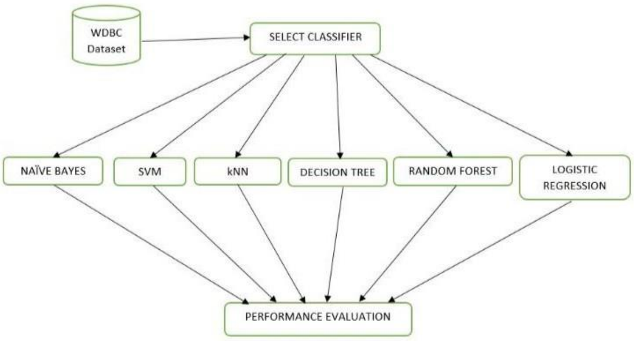
III. IMPLEMENTATION

A. *Dataset information*

WBDC diagnosis dataset consists of 569 instances which are 357 benign and 212 malignant. Every instance includes ID, diagnosis (B= benign, M= malignant) and other features like clump thickness, radius, cell size, cell thickness, marginal adhesion, etc. This dataset can be divided and used for training and testing the different classification models.

B. *Machine Learning Approaches*

FIG 1: OUTLINE OF THE PROPOSED WORK



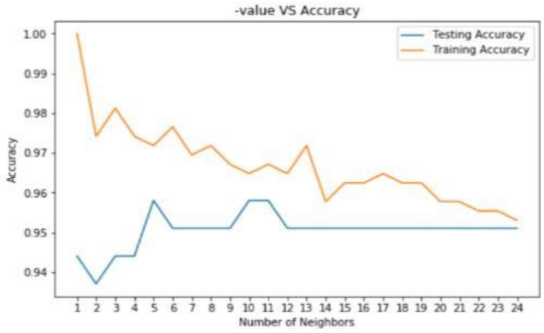
1) k-Nearest Neighbors

kNN is a very popular algorithm used for both classification and regression problems. In classification, the objective is to map the data into any of the classes using their ‘k’ nearest neighbors. An element is classified by a plurality vote of its neighbors, with the element being assigned to the class most common among its k nearest neighbors. K value can vary with the experiment.

For example, when k = 1 the object is simply assigned to the class of the single nearest neighbor. The distance is calculated using Euclidean formula.

‘k’ values can be chosen using different methods

FIG 2: CHOOSING K VALUE



For our dataset, k=5 showed the highest accuracy,

WDBC dataset was passed to kNN classifier with k value being 5, and we got an accuracy of 95.80%.

2) Naïve Bayes Classifier

Naïve Bayes classifier is a probabilistic classification model that makes predictions based on class membership. This is a framework that uses statistics and probabilities of each individual attributes to make predictions. In NBC, it is assumed that attributes of a dataset are mutually independent. The Naïve Bayes Classifier is based on the Bayes’ Theorem as given below:

$$p(c|x) = p(x|c)\, p(c) / p(x) \qquad (1)$$

which can be rewritten as:

$$p(c|x)=p(c1|x)p(c2|x)p(c3|x).........p(cn|x) \text{ x } p(c) \quad (2)$$

Where, p(c|x) is the posterior probability of class c given predictor x, p(c) is the class prior probability, p(x) is predictor prior probability and p(ci|x) is the likelihood for i = 1 to n. The probabilities of each feature of each class are found and their product will be used to determine posterior probability and hence the membership to that particular class.

Training data is passed to the NB classifier. The mean and standard deviation of each class is calculated. The probability of each feature is calculated using probability density function. The predictions in the dataset are used to train the model. To predict the class of an instance from the testing set, calculate the probability of each class.

When the model was tested with the test data, we got an accuracy of 93.70%

3) Support Vector Machine

SVM finds a hyper plane which is used to classify the data points. Mathematically in the ambient space, a subspace whose dimension is one less is a hyper plane, SVM finds the hyper plane which has a maximum distance between two classes.

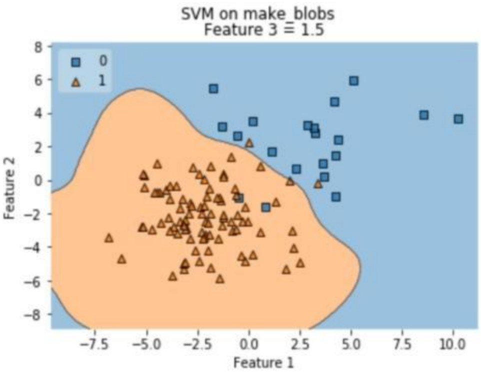
An n-dimensional space is formed by the numerical inputted data (x), this input variable space is split by the hyper plane. Based on their class SVM selects the hyper plane which best separates the data points in the space.

$$(Z0) + (Z1*X1) = 0 \qquad (3)$$

If this line completely separates the data set into two class, the intercept Z0 and Z0 the coefficient which determines the slope of the line are determined by the algorithm where X1 is the input variable. For the cancer data set we use RGB-Gaussian RBF kernel on cancer dataset. SVC model is trained and fit. The accuracy result for training was 1.0 and for the testing was 0.629 which showed over fitting because for the scaling of data as SVM is sensitive to it.

SVM looks for the largest margin or distance between points on the side of decision lines and there are the support vectors. This model is more accurate as this adds a more level of complexity.

FIG 3: DECISION REGION



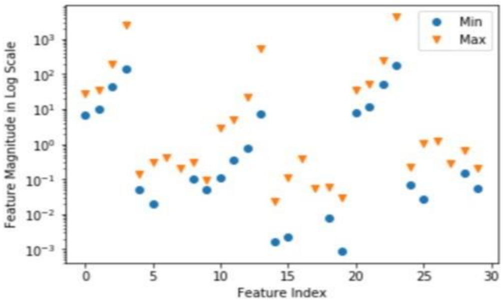
.

The graph is plotted to find min and max of each value with log as a scale. It shows the difference between min and max values, so we brought the features to same scale and evaluate. We used standard preprocessing module to scale the data.

The performance was improved and over fitting is solved and accuracy on training set was 0.948 and on testing set is 0.951 which is under fitting.

The performance of algorithm can be improved by adjusting c and gamma, when c is increased to 1000 the performance is much better with 0.988 accuracy of training set and 0.972 testing.

FIG 4: FEATURE INDEX



4) Random Forest

Random forest is a type of supervised learning algorithm which can be used for both classification and regression. It is an ensemble algorithm, i.e., it uses one or more algorithms of the same or different kind for classification of objects. This classifier randomly selects subsets of training data and creates a set of decision trees for them. Then, it selects the best tree to find the final class of the test object b aggregating the votes from different trees. With our dataset, we got an accuracy of

5) Decision Tree

The decision tree is a non-parametric supervised learning algorithm which can be used for both classification and regression. The objective of this algorithm is to predict the value of the target variable by learning simple decision rules deduced from the given dataset. In decision tree algorithm given data set is divided to form the homogeneous subsets based on the input variable. Every node in the tree represents the test case and the arc descending from the node represents one of the possible answers for the test case.

6) Logistic Regression

The logistic regression classification model is used for classifying problems where the dependent variable is categorical, i.e., it has two possible outcomes. It makes use of the logistic regression function that is as given below:

$$\text{logistic}(\eta) = 1/1+\exp(-\eta) \qquad (4)$$

In the logistic regression model, it is assumed that the dependent variable is dichotomous in nature. It is also assumed that there are no outlier in the data and that there is no high correlations among the predictors. With our dataset, we got an accuracy of 97.20%.

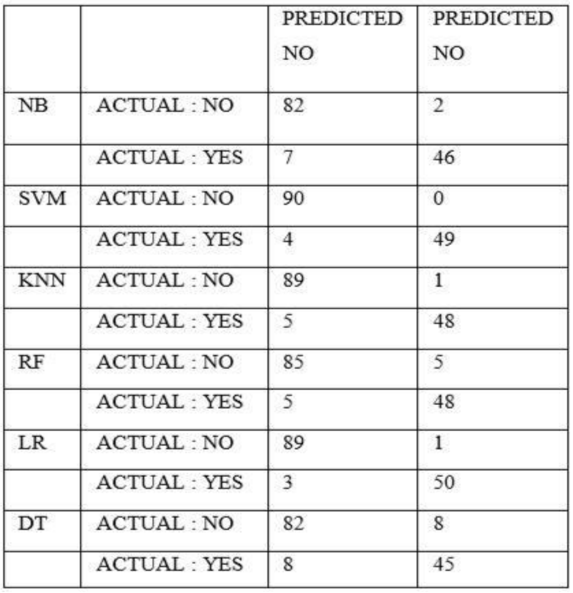
IV. RESULTS

Choosing the right model is a very important factor in any classification problem. A slight change in accuracy or precision will have greater impact on the patient diagnosis. The classification models are evaluated based on accuracy, precision, specificity and sensitivity. It helps us in determining the best suitable classifier for WDBC dataset.

A. Confusion Matrix

Confusion matrix, which describes the performance of the model, is used to evaluate the algorithm performance which consist of the following data: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). With the help of this matrix, it is easy to determine the error rate.

TABLE I. CONFUSION MATRIX



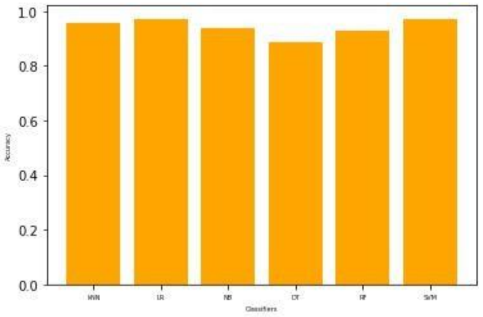
		PREDICTED NO	PREDICTED NO
NB	ACTUAL : NO	82	2
	ACTUAL : YES	7	46
SVM	ACTUAL : NO	90	0
	ACTUAL : YES	4	49
KNN	ACTUAL : NO	89	1
	ACTUAL : YES	5	48
RF	ACTUAL : NO	85	5
	ACTUAL : YES	5	48
LR	ACTUAL : NO	89	1
	ACTUAL : YES	3	50
DT	ACTUAL : NO	82	8
	ACTUAL : YES	8	45

B. Accuracy

Accuracy is the one which indicates how accurate the model is.

```
Accuracy =(TP + TN)/(TP + TN + FP + FN)
```

FIG 5: ACCURACY OF CLASSIFIERS

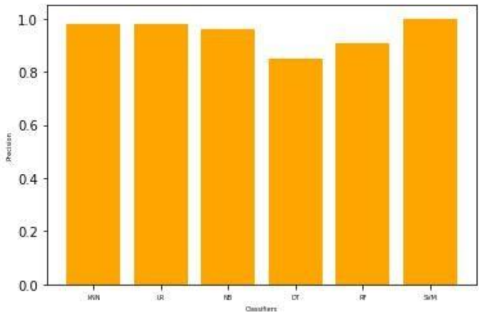


C. Precision

It tells how often a predicted true value is actually true.

```
Precision = TP / (TP + FP)
```

FIG 6: PRECISION OF CLASSIFIERS

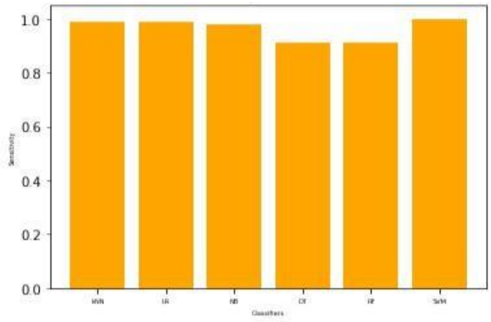


D. Sensitivity

Sensitivity analysis is the one determining the true positive rates.it determines the impact of independent variable on dependent variable.

```
Sensitivity = TP / (FN + TP)
```

FIG 7: SENSITIVITY OF CLASSIFIERS

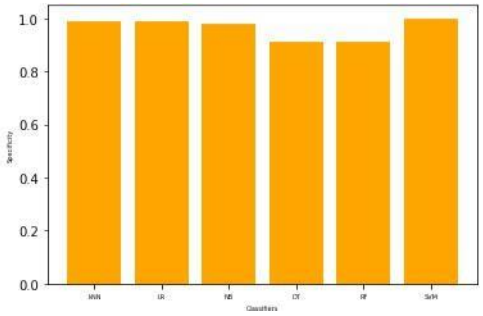


E. Specificity

Specificity analysis is the one determining the true positive rate.

```
Specificity = TN / (TN + FP)
```

FIG 8: SPECIFICITY OF CLASSIFIERS



F. False Positive Rate

It tells us how often the prediction is incorrect when the actual value is False

```
False positive rate= FP / (TN + FP)
```

FIG 9: FALSE POSITIVE RATE OF CLASSIFIERS

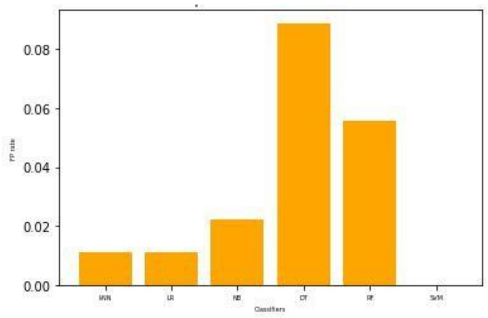
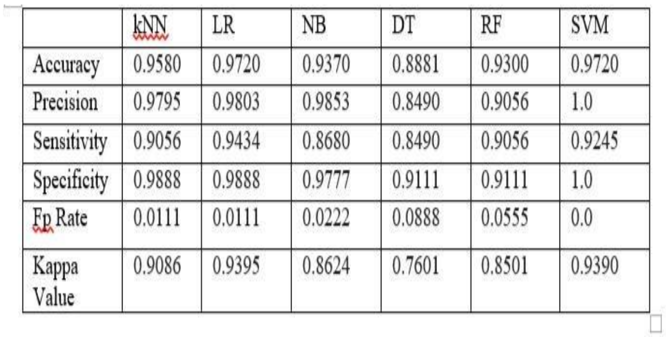


TABLE II. COMPARISON OF CLASSIFIERS



	kNN	LR	NB	DT	RF	SVM
Accuracy	0.9580	0.9720	0.9370	0.8881	0.9300	0.9720
Precision	0.9795	0.9803	0.9853	0.8490	0.9056	1.0
Sensitivity	0.9056	0.9434	0.8680	0.8490	0.9056	0.9245
Specificity	0.9888	0.9888	0.9777	0.9111	0.9111	1.0
Fp Rate	0.0111	0.0111	0.0222	0.0888	0.0555	0.0
Kappa Value	0.9086	0.9395	0.8624	0.7601	0.8501	0.9390

The above table compares the performance of the six classifiers used.

V. CONCLUSION

An important task in data mining and machine learning areas is to build accurate, computationally efficient and reliable model. On the Wisconsin Breast Cancer dataset, we used six classifiers, which are: NB, KNN, Random Forest, Decision Tree, SVM and Logistic Regression. We tried to compare efficiency of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. From the experiment, we observe that Logistic Regression and SVM have highest accuracy of 97.2% among other classifiers, however SVM outperforms other models in terms of precision and low error rate, thus achieving best performance among all models.

ACKNOWLEDGMENT

The authors would like to thank the Principal, Dr. G Ravi, the head of the department of Computer Science and Engineering, Dr. H D Phaneendra and the faculty of The National Institute of Engineering, Mysuru for providing us constant support thru out the project.

REFERENCES

[1] Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer Shweta Kharya Bhilai Institute of technology, Durg, C.G. – India Shika Agrawal CSIT Durg,

[2] Breast Cancer Classification by Using Support Vector Machines with Reduced Dimension Ahmet Mert1 , Niyazi Kilic2 , Aydn Akan2 1 Dept. of Navigation Eng. Piri Reis University, 34940, Tuzla Istanbul, Turkey 2 Dept. of Electrical and Electronics Eng. Istanbul University, 34320, Avcilar Istanbul, Turkey amert@pirireis.edu.tr

[3] Breast Cancer Classification Using Machine Learning Meriem AMRANE1 Saliha OUKID2 Computer Science Department, LRDSI Laboratory, University of Blida 1, Blida, Algeria 1 amrane.meriem@outlook.fr , 2 osalyha@yahoo.com Ikram GAGAOUA3 Tolga ENSARø 4 Computer Engineering, Istanbul University, Istanbul, Turkey 3 i.gagaoua@gmail.com , 4 ensari@istanbul.edu.tr

[4] S.Chakraborty, "Bayesian kernel probit model for microarray based cancer classification", Computational Statistics and Data Analysis, Vol. 12, pp. 4198–4209, 2009

[5] http://www.imaginis.com/breast-health/what-is-breast-cancer-2, March 2011

[6] http://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[7] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[8] Chuan Liu, Wenyong Wang, Meng Wang, Fengmao Lv and Martin Konan, An efficient instance selection algorithm to reconstruct training set for support vector machine, Knowledge-Based Systems, 116, (58), (2017).