International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Comparison of K nearest Neighbours and Support Vector Machine to Build a Breast Cancer Prediction Model

Rachaell Nihalaani[1], Rohan Sawant[2], Juhi Janjua[3]
[1, 2]Undergraduate Research Scholar, Department of Computer Engineering, Thadomal Shahani Engineering College, Mumbai-50, India
[3]Assistant Professor, Department of Computer Engineering, Thadomal Shahani Engineering College, Mumbai-50, India

Abstract: Breast cancer is a greatly widespread and dangerous type of cancer with approximately 2.3 million cases in the past year. It has surpassed lung cancer which was the most common cancer. Data mining and classification of data have helped medical experts to segregate and make use of the data achieve a higher accuracy of results. In this paper, we have referred to the Wisconsin Breast Cancer dataset. We have compared SVM and kNN algorithms for training the dataset and the more accurate one is utilised to test the final model using 10-fold cross validation practice. Furthermore, the maximum accuracy was obtained by the SVM model in the training phase and was further tuned in order to achieve a final accuracy of 96.93% in the test phase.
Keywords: Machine Learning, Binary Classification, Prediction Model, Support Vector Machine, K Nearest Neighbours

I. INTRODUCTION

For long, Breast Cancer Prediction has been a cardinal research issue in the medical and healthcare communities. Breast cancer is cancer in which breast cells begin to develop abnormally. Some of the identified risk factors for breast cancer include older age (over 50), female sex, personal or familial history of breast cancer, genetic mutations to genes such as BRCA1 and BRCA2, alcohol consumption, obesity, lack of physical exercise, menstruation start at an early age (before 12), late menopause phase (after 55) and having children late or not at all.

A few familiar types of breast cancer are Ductal carcinoma in situ, Invasive ductal carcinoma, Invasive lobular carcinoma Lobular carcinoma in situ among others. Survival rates of breast cancer mainly depend on the cancer type and stage. Other factors that may influence survival rates are age, gender, and race. Mammogram, an imaging test which is the most common way to see below the surface of your breast, is useful in diagnosis. Ultrasound, a test that utilizations sound waves to make an image of the tissues somewhere down in your breast and recognizes a strong mass is likewise valuable for analysis A more advanced test is a breast biopsy, during which a tissue sample is removed, using a needle or an incision, from the suspicious area to have it tested at a lab to determine the type of cancer. [1]

Since various kinds of breast cancer exist, all with varying stages or spread, intensities and genetic makeup, a system that would allow for early detection and prevention would be extremely useful.

This paper discusses and compares two machine learning techniques, K-Nearest Neighbour (kNN) and Support Vector Machine (SVM). Our research objective is to analyse the performance of these techniques and build a model to predict whether a person has breast cancer.

The performance of these techniques will be assessed by the evaluation metric of classification accuracy, which will be found on both unstandardized and standardised data. The technique that proves to be more accurate will be further fine-tuned on the basis of certain parameters and will be applied on the dataset to observe how it would perform on the test data. Subsequently, the finishes of this paper ought to permit analysts or clinical experts to handily pick the best AI procedure out of these two that can give the ideal exhibition to future use.

The remaining of this paper is planned as follows. Section' 2 reviews the study of related literature. Section 3 illustrates the theoretical considerations. Section 4 presents the experimental methodology including the experimental procedure and exploratory analysis of the referred dataset. Section 5 discusses the experimental results. Section 6 draws the conclusion to the paper.

II. LITERATURE REVIEW

A variety of machine learning algorithms and techniques have been used to predict an accurate model. [2] attempts to use a 10-fold cross validation system to get an accurate results on SVM and kNN machine learning techniques. The following methodology provides better results for training as well as testing. The performance is assessed on the basis of accuracy, recall, false discovery and omission rate, and Matthews correlation coefficient. This model obtained an accuracy of 98.57% for the SVM technique and 97.14% for kNN technique. [2] led to the conclusion that SVM was the far more accurate and low in error machine learning technique for the Breast cancer prediction. [3] in their approach to finding an accurate model, they have worked with Machine learning algorithms such as k nearest neighbours (kNN), Support Vector machines (SVM), Decision Trees and naïve Bayes on the Wisconsin Breast cancer dataset in order to find the most accurate and effective model. Their experiments were conducted using a WEKA mining tool in a simulated environment. They found that SVM outperforms the other algorithms with the highest accuracy of 97.13% and the lowest error rate. In [4], they have used kNN, Linear Regression and SVM algorithms to exclusively produce a model with a high accuracy and low error rate. They found that if the most predictive variables were provided to all the three algorithms, kNN yielded the highest classification accuracy of 99.28%. In [5], they observed in previous papers that SVM always managed to outperform other techniques but a very few studies managed to focus on inspecting the prediction performances based on the different kernel functions. Moreover, it was unclear if SVM classifier ensembles could outperform single SVM classifiers as they were alleged to improve the performance of SVM. [5] attempts to compare the training of SVM and SVM ensemble on the basis of classification accuracy, ROC, F-measure and computational timings. For a small scale dataset, they have found with the bagging method, linear kernel based SVM ensembles and with the boosting method , RBF kernel based SVM ensembles to be the most applicable choices. Similarly, RBF kernel based method accomplish better than other classification methods on a larger dataset. 10-fold cross-validation is a widely used technique used to estimate predictive models by dividing the data into training and testing sets. We can see this technique in [6] used by Liu, Ya-qin, C.wang and Lu Zhang. The primary advantage of this method is that each observation is used exactly once, and all observations can be used for both the training and validation of data. A new algorithm has been proposed by Thongkam, Xu and Zhang, [7] that combines AdaBoost algorithm and random forests for predicting the survival of breast cancer patients. The algorithm uses random forests as an enervated learner of AdaBoost in order to select the high weight instances throughout the boosting process to enhance accuracy, stability and reduce overfitting problems.[8] have used artificial neural networks to test the accuracy of 5-year , 10-year and 15-year breast cancer specific survival. They concluded that without even training without information of nodal status. Consistently high accuracy and a better predictive performance than other algorithms is important to predict cancer survival.

III.THEORETICAL CONSIDERATIONS

An easy way to comply with IJRASET paper formatting requirements is to use this document as a template and simply type your text into it. Classification uses machine learning algorithms that learn how to assign a class label to examples from the problem domain. This literature illustrates a binary classification problem. Typically, binary classification tasks involve two classes - the normal state and the abnormal state. The class for the normal state is allocated the 0 class label and the class with the abnormal state is allocated 1 class label [9].

A. Support Vector Machine (SVM)

'Support Vector Machine (SVM) is a supervised machine learning model used for classification and prediction of unknown data. It is also widely used in binary classification.

SVM models are trained on a dataset, to be analysed and to classify unknown data into the same two classes that were present in the training data. For instance, if in a dataset we have data which is pre-labelled into two classes: positive and negative, then we can train a model to classify new data into these two classes. SVM is said to be a linear learning method, that is, it can be defined by a linearly separable hyper plane, the lack of which will fail to form a linear classifier.

Being a supervised classification model, it tries to make the distance between the closest training point and either class maximum, in order to achieve better performance on test set. The classification process is as follows:

1) It takes the labelled sample of data, and draws a line separating the two classes. This line is called the decision boundary. The solution is based only on the training data points, called support vectors, that are really close to the decision boundary.

2) Now when new data needs to be classified, it goes either into the left or right side of the decision boundary. It is classified under the category corresponding to the side data enters. To classify our data with the best precision, we need to split the two categories such that the decision boundary separates the two classes with maximum space between them [10].

B. K-Nearest Neighbours

K-Nearest Neighbours (kNN) algorithm is also a supervised machine learning model used for categorisation problems. It does not make any postulations as to the the fundamental data distribution. It is known to perform well in pattern recognition and predictive analysis. The classification process is as follows:

1) kNN first gathers data points that are close to the new data point taken into consideration. The distance between data points is effectively impacted by any attributes that can vary on a large scale.
2) Then, the algorithm sorts data points on the basis of their Euclidian distance from the new data point.
3) Next, the algorithm takes a specific number of data points, with lesser distance among all and then categorizes those data points. The number of closest data points is usually chosen as an odd number if the number of classes is 2. The class of the new data point will be the class with the highest number of data points.

C. Terminologies

1) *Cross Validation:* With the aim of testing the model, we will need to use a part of our dataset. However, reducing the training data may lead to loss of patterns and trends, and underfitting, both of which leads to errors and high bias. Cross validation is used to prevent this, by providing sufficient data to train the model but also leaving enough data for validation purposes. It is also used to judge the predictive performance of the models and to assess their performance outside the training sample, to an unknown data, that is, the test data. When we fit our model, we only fit it onto a training dataset. Without performing cross validation, we would remain in the dark about how our model performs on new data in terms of prediction accuracy [10].
2) *Standardisation of Data:* Data is standardised to make sure that data is uniform and consistent in terms of format and content. Standardization gives a greater meaning to data point and data set. It is a popular scaling technique that measures each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution so that it has a mean of 0 and a standard deviation of 1. It assumes that the observations fit a Gaussian distribution (bell curve) with a well-behaved mean and standard deviation, to get reliable results [11].
3) *SVM Parameter Tuning:* We have tuned two key parameters of the SVM algorithm. First, the value of C, which is also known as the penalty parameter. It lets our algorithm know how much we care about misclassified points. A high value for C indicates that we care about classifying all data accurately. If the C parameter is increased, we expect the future data to be further away from the points that the model has been trained on. A higher C value creates finer boundaries between classification areas. In the RBF kernel and sigmoid model, a larger C value improves the accuracy of the untuned RBF kernel model [10]. Second, the type of Kernel is another tuning parameter. Kernel plays an important role in classification and is used to analyze some patterns in the given dataset. SVM uses kernels to transform the data points and create an optimal decision boundary. Kernels deal with high dimensional data in an efficient manner. Some popular kernel types are Gaussian Radial Basis Function (RBF), linear kernel, Gaussian kernel, polynomial kernel, Sigmoid kernel, Bessel Function kernel, Gaussian kernel, and Anova kernel [12].
4) *Confusion Matrix:* It is a matrix for summarizing the classification algorithm performance. The number of correct and incorrect predictions are summarized with count values and are broken down by each class. It shows how the classification model is confused when it makes predictions. This breakdown overcomes the limitation of using classification accuracy alone [13].
5) *Accuracy:* It is the fraction of predictions that the model gets right [14].
6) *Precision:* It is the fraction of relevant information among the extracted information. High precision indicates that more relevant instances are being extracted by the model [14].
7) *Recall / Sensitivity:* It is the fraction of number of relevant instances with a known positive for which the result is positive [14].
8) *F1 Score:* It is equal to the weighted average of Precision and Recall. It considers false positives and false negatives. This measure is more beneficial than accuracy, especially in an uneven class distribution [14].
9) *Specificity:* It is the fraction of number of relevant instances with a known negative for which the result is negative [15].
10) *False Discovery Rate:* It is the fraction of number of relevant instances with a positive test result for which the true condition is negative [15]
11) *False Omission Rate:* It is the fraction of number of relevant instances with a negative test result for which the true condition is positive [15].
12) *Matthews Correlation Coefficient (MCC):* It is a statistical rate which generates a high score only if the estimate achieved good results in all of the four categories of confusion matrixs, in proportion to both to the size of positive elements and the size of negative elements in the dataset. The closer MCC is to 1, the higher the possibility of the model being a pure binary classifier.

IV. PROPOSED FRAMEWORK

A. Methodology

This section explains the general process of binary classification and offers a detailed explanation of the flow of the experiment performed.

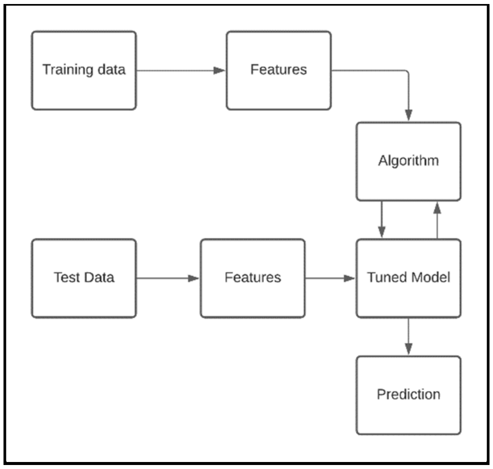


Fig. 1 Binary Classification Process

Data collection is the first and foremost step. We have used a benchmark dataset [16] in this paper. Then, in the second step, we familiarise ourselves with the data. Exploratory analysis of the data includes reading the contents of the dataset; exploring the shape of data. The statistical details, such as count, standard deviation, mean, minimum, lower quartile, median, upper quartile, maximum and interquartile range, of data are also analysed to get a rough estimate of the contents. Next, we visualise the data and pre-process it. We convert the diagnosis column to numerical values as machine learning models can only work with numerical values. We also visualise the data using density plots to understand the data distribution; and also check the correlations between the attributes. Truly understanding your data and familiarising yourself with it, is key.

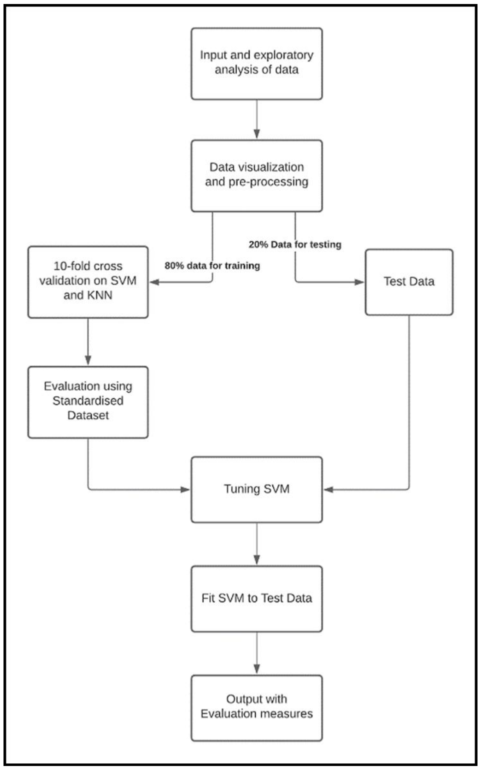

Input and exploratory analysis of data
Data visualization and pre-processing
20% Data for testing
80% data for training
10-fold cross validation on SVM and KNN
Test Data
Evaluation using Standardised Dataset
Tuning SVM
Fit SVM to Test Data
Output with Evaluation measures

Fig. 2 Flowchart for Experimental Process

Next, data is split into predictor variables (input variables, used to determine the output) and target variables (variables that need to be predicted). Data is split into train and test sets in a 80-20 ratio. Binary Classification is done on two powerful machine learning algorithms - k-nearest neighbours (kNN) and linear support vector machine (SVM). Firstly, we do a test with default setting, on both the algorithms, to get an initial estimation of how each of them might perform. We use 10-fold cross validation for each testing, i.e. the data is split into 10 groups. Secondly, we standardise the input dataset to check if its performance improves. Pipelines are used for the standardisation of data, and models are built for each fold to get a fair indication of how each model with standardised data might perform on new and unfamiliar data. Next, we choose the algorithm that performs better on standardised data, and will go ahead with obtaining the best model. We will tune the best performing model and recognise the optimal hyper parameters that result in the best model. We use the grid search technique using 10-fold cross validation with a standardised copy of the sample training dataset. Cross-validation techniques are used when trying to fit a model into a training dataset. Grid search method takes a dictionary that describes the parameters that could be tried on a model to train it. Thus it helps to determine the best parameters or coefficients for a given model. If SVM performs better on standardised dataset, we can tune the kernel, and C value parameters, whereas if kNN performs better, we can tune the Euclidean distance. The final step is to fit the better performing algorithm on the test set and observe its performance.

B. Dataset

For this research purpose, we have used the Wisconsin Breast Cancer (WBC) Dataset which is retrieved from the UCI machine learning repository dataset [16].

This dataset has a total of 699 instances, in which 458 (65.50%) cases are labelled as benign and 241 (34.50%) of the cases are labelled as malignant.

The data is broadly categorized into two classes in the ratio of 2:4, in which 2 is the benign class and 4 is the malignant class. This dataset has 11 integer-valued attributes. There are also up to 16 missing values of the attributes in the data which are substituted by mean for that attribute.

The dataset is divided in the ratio of 80:20 with 80% for training phase and 20% for testing phase. We have used a 10-fold cross validation technique in which the data is distributed into 10 equal sized parts. Two of the parts are used for testing and eight parts are used for training.

C. Performance Measure Indices

Various performance measure indices are used to measure the implementation of some machine learning techniques. A confusion matrix is formed and TP, TN, FP, and FN are used to evaluate the parameters. These terms are:

TP = True Positive (Correctly Identified)
TN = True Negative (Incorrectly Identified)
FP = False Positive (Correctly Rejected)
FN = False Negative (Incorrectly Rejected)

The formulas used to measure the performance of the specified system are:

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall or Sensitivity} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{5}$$

$$\text{False Discovery Rate (FDR)} = \frac{FP}{FP+TP} \tag{6}$$

$$\text{False Omission Rate (FOR)} = \frac{FN}{FN+TN} \tag{7}$$

$$\text{Matthews Correlation Coefficient} = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{8}$$

V. EXPERIMENTAL RESULTS

This section discusses the tables and graphs obtained on performing binary classification on the chosen dataset [16] using and k-nearest neighbours (kNN) and linear support vector machine (SVM).

The diagnosis column in the dataset is converted to numerical values such that M (Malignant) =1 and B (Benign) =0. We find the count of malignant and benign cases from the dataset and can infer from the output in Table 1 that the majority of cases are benign.

Table 1
Data Diagnosis

0	357
1	212

We visualise the data using density plots, as shown in Fig. 3, and can infer that the data shows a general Gaussian (normal) distribution i.e. it is symmetric about the mean and data near the mean is more frequent in occurrence.

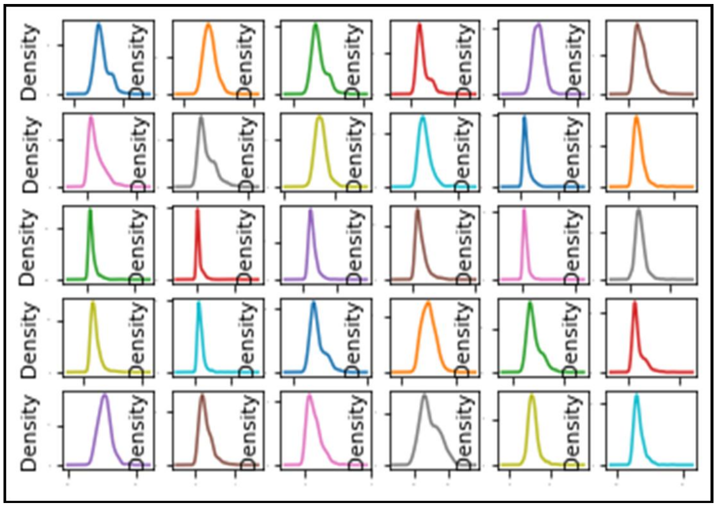


Fig. 3 Density Plot of Data

We also check the correlation between attributes, as shown in Fig. 4. The blue boxes indicate a negative correlation i.e. one increases and the other decreases, the yellow and green boxes indicate only a moderate correlation and the red boxes indicate that the attributes are correlated with each other.

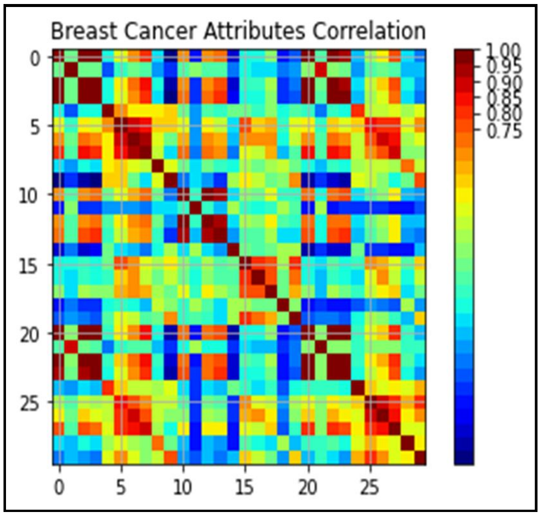


Fig. 4 Breast Cancer Attributes Correlation

Once the data is fragmented into train and test sets in a 80-20 ratio, we perform a test on both the algorithms, k-nearest neighbours (kNN) and linear support vector machine (SVM), with their default setting, using 10-fold cross validation.

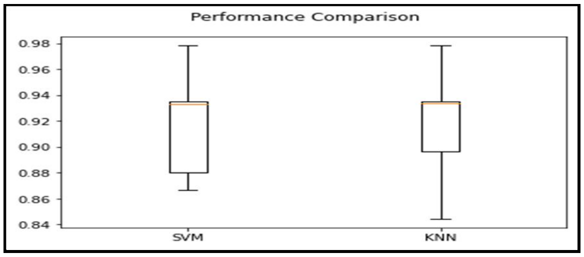


Fig. 5 Performance Comparison of SVM and KNN with default setting

From the output shown in Fig. 5, we can infer that kNN performed better, with a ~92% mean accuracy, while SVM did not perform as well, with only a ~91% accuracy. But, now we standardise the input dataset, using pipelines, to check if it's performance improves.

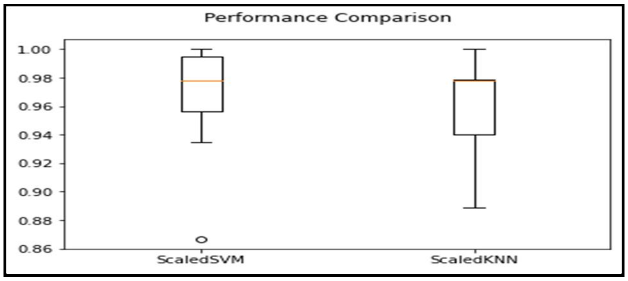


Fig. 6 Performance Comparison of SVM and kNN on Standardised Data

From the output shown in Fig. 6, we can infer that the performance of SVM, with a ~96% mean accuracy, has improved drastically after using scaled data, and now performs better than kNN, with a mean accuracy of ~95%.

Now that it is determined that SVM performs better than kNN, we go ahead and tune 2 key parameters of the SVM, the C value, for which default is 1.0; and the type of kernel, for which default is radial basis function (rbf). We try tuning over a combination of C values and 'linear', 'poly', 'rbf' and 'sigmoid' kernel types. After using the grid search method using 10-fold cross validation, we observe that the most accurate configuration was SVM with a value of C=2.0 and an RBF kernel, with an accuracy of 96.93%.

SVM is fit to the dataset and its performance on the test set is observed. As long as the symptoms of breast cancer exhibit clearly visible patterns, we are in a position to expect a high performance from this algorithm.

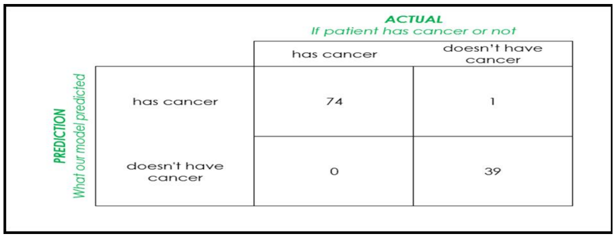
A. Confusion Matrix

1) True Positive = 74

2) True Negative = 39

3) False Positive = 1

4) False Negative = 0



ACTUAL
If patient has cancer or not

PREDICTION
What our model predicted

	has cancer	doesn't have cancer
has cancer	74	1
doesn't have cancer	0	39

Fig. 7 Result Confusion Matrix

Using these values, shown in Fig. 7, we compute the performance measure indices for the model, to gain a better understanding of how the model performed. In Table 2, the Matthews Correlation Coefficient value is ~1, and that indicates a high possibility for it to be a pure binary classifier.

Table 2. Performance Measure Indices

Performance Measure Index	Value
Accuracy	0.9912280701754386
Precision	0.9866666666666667
Sensitivity/ Recall	1.0
F1 Score	0.9932885906040269
Specificity	0.975
False Omission Rate	0.0
False Discovery Rate	0.013333333333333334
Matthews Correlation Coefficient	0.9808159868191383

VI. CONCLUSIONS

Breast cancer has proved to be the most widespread type of cancer in women, with statistics showing one in every nine women to be affected by it during their lifetime. It is known to cause numerous deaths among females across the world. Detection at an initial stage can help save a lot of lives. Our paper attempts to build an accurate classifier using the better performing of two algorithms, K-Nearest Neighbours and Support Vector machines. We compare both the models in our testing stage to find out which model is more accurate and conclude that the Support vector machine has an accuracy of approximately 96% after standardizing the input. The SVM model is tuned in order to achieve maximum accuracy in the testing phase. The tuned model is further tested and achieves an accuracy of 96.93 % with the testing data. Hence, our projected model will be of great help to both medical professionals and the general public for proper diagnosing of the illness particularly at an initial stage.

VII. ACKNOWLEDGMENT

We would like to express our gratitude to our professor Juhi Janjua for her immense support and guidance. We would also like to express our gratitude to our Head of Department, Dr. Tanuja Sarode and our Principal, Dr. G.T. Thampi.

REFERENCES

[1] 2021. [online] Available at: <https://www.healthline.com/health/breast-cancer#diagnosis> [Accessed 13 May 2021].

[2] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017, pp. 226-229, doi: 10.1109/R10-HTC.2017.8288944.

[3] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis”, Procedia Computer Science, Volume 83, 2016, Pages 1064-1069, ISSN 1877-0509.
https://doi.org/10.1016/j.procs.2016.04.224.
(https://www.sciencedirect.com/science/article/pii/S1877050916302575)

[4] Madhu Kumari, Vijendra Singh, “Breast Cancer Prediction system”, Procedia Computer Science, Volume 132, 2018, Pages 371-376, ISSN 1877-0509.
https://doi.org/10.1016/j.procs.2018.05.197.
(https://www.sciencedirect.com/science/article/pii/S1877050918309323)

[5] Huang M-W, Chen C-W, Lin W-C, Ke S-W, Tsai C-F (2017),” SVM and SVM Ensembles in Breast Cancer Prediction”. PLoS ONE 12(1): e0161501. https://doi.org/10.1371/journal.pone.0161501

[6] Liu, Ya-qin, C. Wang and Lu Zhang. “Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data.”, 3rd International Conference on Bioinformatics and Biomedical Engineering (2009): 1-4.

[7] J. Thongkam, G. Xu and Y. Zhang, "AdaBoost algorithm with random forests for predicting breast cancer survivability," IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 3062-3069, doi: 10.1109/IJCNN.2008.4634231.

[8] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkänen L, Joensuu H.,” Artificial neural networks applied to survival prediction in breast cancer”, Oncology. 1999 Nov;57(4):281-6. Doi: 10.1159/000012061. PMID: 10575312.

[9] Brownlee, J., 2021. 4 Types of Classification Tasks in Machine Learning. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/types-of-classification-in-machine-learning/> [Accessed 13 May 2021].

[10] Irjet.net. 2021. [online] Available at: <https://www.irjet.net/archives/V7/i7/IRJET-V7I7623.pdf> [Accessed 14 May 2021].

[11] Ferguson, K. and Ferguson, K., 2021. Why It’s Important to Standardize Your Data - Atlan | Humans of Data. [online] Atlan | Humans of Data. Available at: <https://humansofdata.atlan.com/2018/12/data-standardization/> [Accessed 14 May 2021].

[12] Dataaspirant. 2021. Seven Most Popular SVM Kernels. [online] Available at: <https://dataaspirant.com/svm-kernels/#t-1608054630726> [Accessed 14 May 2021].

[13] Brownlee, J., 2021. What is a Confusion Matrix in Machine Learning. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/confusion-matrix-machine-learning/> [Accessed 14 May 2021].

[14] Medium. 2021. Accuracy, Precision, Recall or F1?. [online] Available at: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9> [Accessed 14 May 2021].

[15] Ncss-wpengine.netdna-ssl.com. 2021. [online] Available at: <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Binary_Diagnostic_Tests-Single_Sample.pdf> [Accessed 14 May 2021].

[16] “UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. [Accessed: 18-May-2021].