International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 02 | Feb 2021 www.irjet.net p-ISSN: 2395-0072

Prediction of Breast Cancer using Support Vector Machines

Apratim Sadhu

UG Student, Dept. of Computer Science Engineering, Chandigarh University, Mohali, India

---

Abstract - *Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer deaths worldwide. It is also the second primary cause of cancer deaths among women continue to suffer from it. The early diagnosis of the disease can improve the chance of survival significantly as it can help with timely clinical treatment to patients. The use of statistical and machine learning algorithms can be useful for the initial prediction of breast cancer. One of those techniques is Support Vector Machines(SVM). This paper presents the support vector classification algorithm for the early detection of breast cancer on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Breast cancer diagnosis differentiates benign tumors from malignant tumors. The classification accuracy, ROC and F-Score of various kernel functions is presented. The experimental results show that radial basis function(RBF) kernel-based SVM is a better choice for classification of the given dataset. This paper discusses the performance of SVM with RBF kernel by measuring its classification test accuracy and its sensitivity and specificity values.*

Key Words: **Breast Cancer, Classification Accuracy, Machine Learning, prediction ,Support vector machines, WDBC.**

1.INTRODUCTION

Thousands of women around the globe fall victim to cancer per annum. The physical body comprises of many cells each with its own unique function. Around 42000 women die from cancer yearly, that's 1 woman every 13 minutes is dying from this disease a day. Cancer is usually caused by a genetic disease. However, only 5-10% of cancers are inherited from parents. Instead, 85-90% of breast cancers are thanks to genetic abnormalities that happen as a result of the ageing process and therefore the "wear and tear" of life generally. Tumours could also be cancerous (malignant) or not cancerous (benign). Mammograms can detect cancer early, possibly before it's spread. [1] The rate of the latest cases of female cancer was 128.5 per 100,000 women per annum. The death rate was 20.1 per 100,000 women per annum. These rates are age-adjusted and supported 2013–2017 cases and 2014–2018 deaths. Statistics reveal that there'll be an estimated 42,170 female deaths and 272,480 new cases recorded within the US in 2020. [2]. Recent years have seen an intense improvement in survival rates for ladies with cancer, which may be mainly attributed to an in-depth screening and enhanced treatment. [3]

The literature discusses support vector machines machine learning technique that is applied to develop models for cancer classification. The recent advances in data collection and storage techniques have made it possible for various medical companies and hospitals to stay vast amounts of knowledge concerning their medical records concerning medication and symptoms of a disease. The uses and potentials of those methodologies have found its scope in medical data.

Furthermore, a comparative study of the above-mentioned machine learning methods, shows that SVM provides comparatively better performance in terms of both accuracy and computation time. It is, however, important to work out the acceptable kernel functions when constructing the SVM model. the utilization specific kernel like linear, polynomial, RBF and sigmoid leads to varied accuracy. The comparative experimentation of those kernel methods during this paper shows that RBF kernel may be a better kernel for classification on the cancer Wisconsin(Diagnostic) dataset.

A complicating factor is that the collected dataset for cancer prediction is typically class imbalanced, with the minority class containing a little number of patients with cancer and therefore the majority class containing an outsized number of patients without cancer which suggests that using only prediction accuracy or classification accuracy to gauge the prediction models is insufficient [4]. Other evaluation

metrics that use different types of classification errors, like the world under the curve (AUC) or the receiver operating characteristic (ROC) curve [5], should even be examined to completely understand the performance of the prediction model.. Therefore the objective of this research is to match SVM using various kernel functions(i.e, linear, polynomial, RBF and sigmoid). Their performance is going to be assessed using different evaluation metrics, including classification accuracy, ROC and F-measure. Alongside that, the precision, recall and F-score of RBF kernel function is calculated alongside its ROC curve to completely summarize the prediction performance of the kernel. Further discussion of the performance of the best kernel on the dataset is presented in this paper. The findings of this paper should allow future researchers to simply choose the foremost effective baseline technique which will provide the optimal prediction performance for future comparison.

2. LITERATURE REVIEW

Classification is among the most common methods that go under supervised learning. It uses historical labelled data to develop a model that is then used for future predictions.

2.1.Support Vector Machine(SVM)

Support Vector Machines(SVM) was first introduced by Vapnik[6]. SVM is one of the supervised ML classification techniques that is widely applied in the field of cancer diagnosis and prognosis. SVM works by selecting critical samples from all classes known as support vectors and separating the classes by generating a function that divides them as broadly as possible using these support vectors. Therefore, it can be said that a mapping between an input vector to a high dimensionality space is made using SVM that aims to find the most suitable hyperplane that divides the data set into classes [7]. This linear classifier aims to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance, by finding the best suited hyperplane [8].

An SVM classifier performs binary classification, i.e., it separates a set of training vectors for two different classes (x_1, y_1), (x_2, y_2),..., (x_m, y_m), where $x_i \rightarrow R^d$ denotes vectors in a d-dimensional feature space and y_i {-1, +1} is a class label. The SVM model is generated by mapping the input vectors onto a new higher dimensional feature space denoted as F: $R^d \rightarrow H^f$ where d < f. An optimal separating hyperplane in the new feature space is constructed by a kernel function $K(x_i,x_j)$, which is the product of input vectors x_i and x_j and where $K(x_i,x_j) = \Phi(x_i).\Phi(x_j)$. [9]

The Kernel Trick[10] :This is a mathematical trick that allows us to learn a classifier in higher dimensional space. It works by directly computing the distance(more precisely, the scalar products) of the data for the expanded feature representation, without actually ever actually computing the expansion.

There are basically four kernel functions. They are linear, polynomial, RBF and sigmoid with the first three being the most common. They are as follows: [11]

linear: $K_{linear}(x_i , x_j) = \langle x_i,x_j \rangle$.

polynomial: $K_{poly}(x_i , x_j) = (\gamma\langle x_i ,x_j\rangle + 1)^d$, where d is the degree of polynomial.

rbf: $K_{rbf}(x_i , x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, where γ is specified by parameter gamma, must be greater than 0.

Sigmoid: $K_{sigmoid}(x_i , x_j) = \tanh(\gamma\langle x_i,x_j\rangle + r)$, where r is specified by coef0.

This paper upon comparing various kernel, used Radial Basis function(RBF) kernel to classify the data, also known as the Gaussian kernel. When training an SVM with the *Radial Basis Function* (RBF) kernel, two hyperparameters must be considered: C and gamma. The hyperparameter C, common to all SVM kernels, trades off misclassification of training examples against simplicity of the decision surface. A low value of C smoothens the decision surface, while a high value of C aims at classifying all training examples correctly. The extent of influence of a single training example is defined by gamma. The larger the value of gamma is, the closer other examples must be to be affected.

The distance between data points is measured by the Gaussian kernel:

$$K_{rbf}(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (1)$$

Here, x_i and x_j are data points, $\| xi - xj \|$ denotes Euclidean distance.

Related works show that there is no formal way of selecting kernel functions. The choice of kernel functions is dependent on the respective data and specific domain problem.

2.2 Related Works

Several studies have been conducted on the implementation of ML techniques on detection of Breast Cancer detection and diagnosis to increase accuracy.

Avramov and Si [12] worked on feature extraction and the impact of the selection on performance. They have used 4 ways of correlation selection (PCA, T-Test Significance and Random feature selection) and 5 classification models (LR, DT, KNN, LSVM, and CSVM). The best result was achieved by stacking the logistic, SVM and CSVM improve accuracy to 98.56%.

Ayeldeen et al. [13] used AI and its techniques for breast cancer detection. They used 5 different methods(Bayesian Network, Multi CC, Decision Tree Radial Basis Function and Random Forest) for performance comparison. Random Forest algorithm showed the highest result with 99% performance.

In a study conducted by Aminikhanghahi et al. [14], wireless cyber mammography images were explored. After selecting and extracting features, the researcher has chosen two different ML techniques, SVM and GMM to check their accuracy. Their findings showed that SVM is more accurate(80-90%) if there is no noise or error, else GMM is better, and safer having 70-80% accuracy.

Hafizah et al. [15] compared two algorithms, SVM and ANN using four different datasets of breast and liver cancer including WBCD, BUPA JNC, Data, Ovarian. The researchers have demonstrated that both methods are having high performance but still, SVM with accuracy 99.5% was better than ANN having an accuracy of 98.54%.

Min-Wei Huang , Chih-Wen Chen, Wei-Chao Lin , Shih-Wen Ke , Chih-Fong Tsai in [9] used SVM and ensemble of SVM for prediction on WDBC(Original) dataset comparing linear, polynomial and RBF kernel using different evaluation metrics to compare the predictions. They found that the RBF SVM ensemble has the maximum accuracy of 99.52% and ROC value of 0.876 on the model.

Recent studies have shown that RBF kernel is the most common kernel function for the prediction and the majority of research is carried out on the Wisconsin Breast Cancer Dataset (Original) [16]. These studies have used classification accuracy as the primary evaluation metric. But this may not be sufficient for complete and accurate prediction. Various other metrics such as specificity and sensitivity should also be taken into account.

Senturk Z.K., Kara R [17] used various data mining techniques to predict accuracy, sensitivity and specificity. They concluded that SVM is a better technique than other classification techniques with an accuracy of 96.79%.

You H., Rumbe G [18] have done a comparative study on different classification technique for breast cancer classification and have concluded that knn has shown the highest accuracy score of 100% on the Wisconsin Breast Cancer Dataset.

Aruna S., Rajagopalan S.P., Nandakishore L.V[19] have done a comparative study on Wisconsin Breast Cancer dataset using Naïve Bayes, RBF Networks, Trees-J48, Trees-CART, SVM-RBF kernel and concluded that SVM kernel performs the best in terms of accuracy, sensitivity and specificity on the dataset.

The above-related works indicate that SVM-RBF kernel function is comparatively a better classification technique for the prediction of breast cancer. Consequently, other kernel function might perform quite well. This paper thus discusses the comparative performance of the kernel functions in terms of accuracy score, F-score and ROC value and compares the effect of scaling the data and setting of proper hyperparameter values on the respective evaluation metrics of the kernel functions of SVM.

3. METHODOLOGY

3.1 The Dataset

The various SVM kernel functions machine learning algorithm were trained to predict breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset[20]. The dataset consists of features that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The mentioned features describe the characteristics of the cell nuclei found in the image[20].

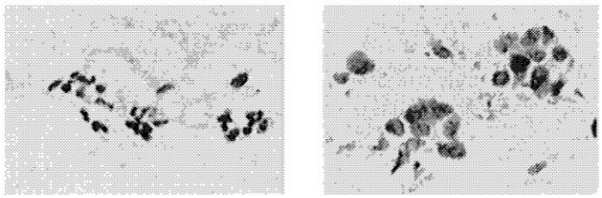

Fig-1: Image from [20] as cited by [21]. Digitized images of FNA: (a) Benign, (b) Malignant.

There are a total of 569 data points in the dataset: 212 – Malignant, 357 – Benign. The dataset has features as follows: (1) radius, (2) texture, (3) perimeter, (4) area, (5) smoothness, (6) compactness, (7) concavity, (8) concave points, (9) symmetry, and (10) fractal dimension. Each

feature constitutes of three information [20]: (1) mean, (2) standard error, and (3) "worst" or largest (mean of the three largest values) computed. Thus, having a total of 30 dataset features.

3.2 Training Set

The classifier will be tested using the k − fold cross validation method. This validation technique will randomly separate the training set into k subsets where 1 of the k − 1 subsets will be used for testing and the rest for training. 10- fold cross-validation is the preferred k value used in most validation in ML and will be used in this paper [22][23]. This implies 9 subsets are going be used for training of the classifier and therefore the remaining 1 for the testing. This approach is used to avoid over fitting of the training set, which is most likely to occur in small data sets and large number of attributes. The hyper-parameters used for all the classifiers were assigned manually for the best results. All experiments are executed in the Jupyter platform.

3.3 Dataset Pre-processing

To avoid inappropriate assignment of relevance, the dataset was standardized using (2).

$$z = \frac{X-\mu}{\sigma} \quad (2)$$

where X is the feature the needs to be standardized, μ is the mean value of the feature, and σ is the standard deviation of the feature.

3.4 Experimental Procedure

All experiments in this study were conducted on a laptop computer with Intel® Core™ i5-8250U CPU @ 1.60GHz, 8GB of DDR5 RAM, and Radeon (TM) 530 Discrete/Hybrid 2048 MB GDDR5 1125 MHz GPU.

The experimental procedure is carried out in the following steps. The dataset was partitioned by 80%(training set)/ 20%(testing set) based on 10-fold cross-validation strategy. The training set is used to train the four SVM kernel functions and test set is fed to the classifiers before the performance is evaluated using accuracy score, ROC score and F-score. Also the evaluation metrics is calculated on both general, standardized data and applying appropriate hyperparameters to find the best classification.

After that the best kernel function is used further on the standardized data using appropriate hyperparameter values to find out the train and test set accuracy, AUC under ROC curve , precision and recall value and the misclassification.

4. RESULT AND DISCUSSIONS

4.1 SVM Classifiers on the non-scaled dataset

Figure 2 shows the performance of the SVM classifier with linear, RBF, polynomial and sigmoid kernel functions on the non-scaled dataset and without setting any hyperparameter in terms of classification accuracy, F-measure and ROC value.

As we see, the performance of polynomial kernel is better in terms of accuracy score with an accuracy of 91.12% , followed by RBF(90.76%). The best F-score is obtained by polynomial kernel(0.9307), the best ROC is obtained by Linear kernel(0.9760). Moreover, there is no big performance difference between polynomial and RBF SVM. Table-1 summarizes the performance comparison of all kernel functions for unscaled data.

Table-1: Performance of SVM Classifiers on the non-scaled dataset

METRICS	Linear	RBF	Polynomial	Sigmoid
Classification Accuracy	90.5700	90.7681	91.2126	41.9758
F-Score	0.930430	0.927437	0.930722	0.550780
ROC	0.974683	0.975556	0.969897	0.225963

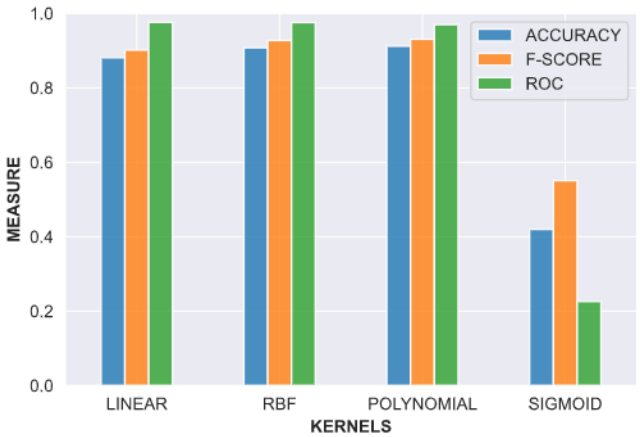

Fig -2: Performance of SVM Classifiers on the non-scaled dataset

4.2 SVM classifiers on the scaled dataset

Figure 3, shows the performance of the SVM classifier with linear, RBF, polynomial and sigmoid kernel functions on the scaled dataset and without setting any hyperparameter in terms of classification accuracy, F-

measure and ROC. Standardization scaling technique used to scale the data.

As we see, the performance of RBF kernel is better in terms of accuracy score, F-Score and ROC value with an accuracy of 96.05%, F-score and ROC value of RBF kernel is 0.9716 and 0.9921 respectively which are the best among other kernels for the standardized data. Moreover there is very little performance difference between linear and RBF SVM in terms of accuracy score and F-Score. Table-2 summarizes the performance comparison of all kernel functions for scaled data.

Moreover, there can be a large increase of performance accuracies after performing feature scaling with hyperparameter optimization of the kernel functions as compared to SVM classifiers with only feature scaling.

Table-2: Performance of SVM classifiers on the scaled dataset

METRICS	Linear	RBF	Polynomial	Sigmoid
Classification Accuracy	96.0531	96.7101	89.0048	95.1836
F-Score	0.966820	0.971681	0.916656	0.959407
ROC	0.980602	0.992137	0.989898	0.983938

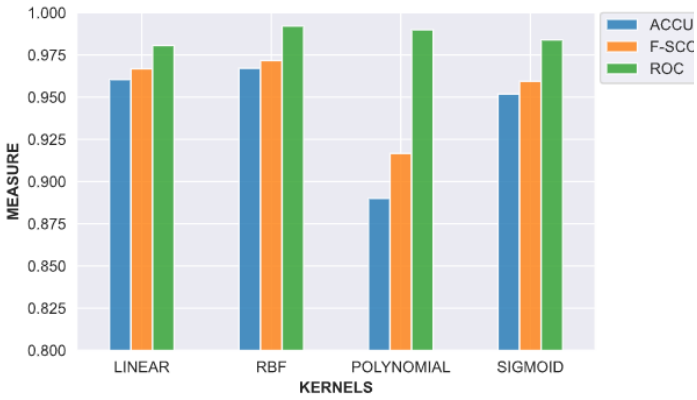

Fig- 3: Performance of SVM classifiers on the scaled dataset

4.3 SVM classifiers on the scaled dataset with optimized hyperparameters

Figure 4, shows the performance of the SVM classifier with linear, RBF, polynomial and sigmoid kernel functions on the scaled dataset with hyperparameters in terms of classification accuracy, F-measure and ROC. Standardization scaling technique used to scale the data.

As we see, the performance of RBF kernel is better in terms of accuracy score, F-Score and ROC value with an accuracy of 96.71%, F-score and ROC value of RBF kernel is 0.9716 and 0.9937 respectively which are the best among other kernels for the standardized data. Moreover, there is very little performance difference between linear and RBF SVM in terms of accuracy score and F-Score. Table-3 summarizes the performance comparison of all kernel functions for standardized data with optimized hyperparameters.

Table-3: Performance of SVM classifiers on the scaled dataset with optimized hyperparameters

METRICS	Linear	RBF	Polynomial	Sigmoid
Classification Accuracy	96.4976	96.7101	96.0483	94.7295
F-Score	0.971372	0.971681	0.966863	0.956463
ROC	0.992884	0.993777	0.992141	0.984349

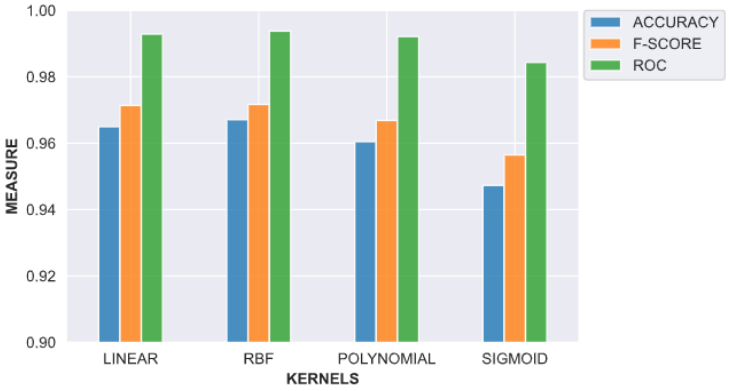

Fig- 4: Performance of SVM classifiers on the scaled dataset and hyperparameter optimization

Evidently, there is a significant improvement of classification accuracy of all the kernel functions on optimizing the hyperparameters. The F-Score, ROC values of the kernel have improved upon hyperparameter optimization.

RBF kernel stands out with the best classification accuracy, F-score and ROC value. In the rest of the paper, the performance of RBF kernel shall be discussed in detail and this kernel function will be applied on the test set to find the performance in terms of classification accuracy, F-Score and ROC values.

4.4. RBF Kernel Function

Radial Basis Function(RBF) kernel function consists of two hyperparameters: C and gamma. The gamma parameter (1) which controls the width of the Gaussian kernel. It determines the size of what it means for points to be approximate . The C parameter may be a regularization parameter, similar to that utilized in the linear models. It limits the importance of every point.

The various values of C is checked over the train and test accuracies for four different values of gamma(0.0001,0.001,0.01,0.1). The appropriate value of C for best accuracy on the test set is selected based on figure 5. It illustrates the test and train score for gamma=0.1 and different values of C. For values of C 1.00 afterwards, there is no significant overfitting and the algorithm achieved highest accuracy. Whereas, in rest of the cases, either there the test and train set overfit or highest accuracy is not achieved. So, gamma=0.1 is the best choice.

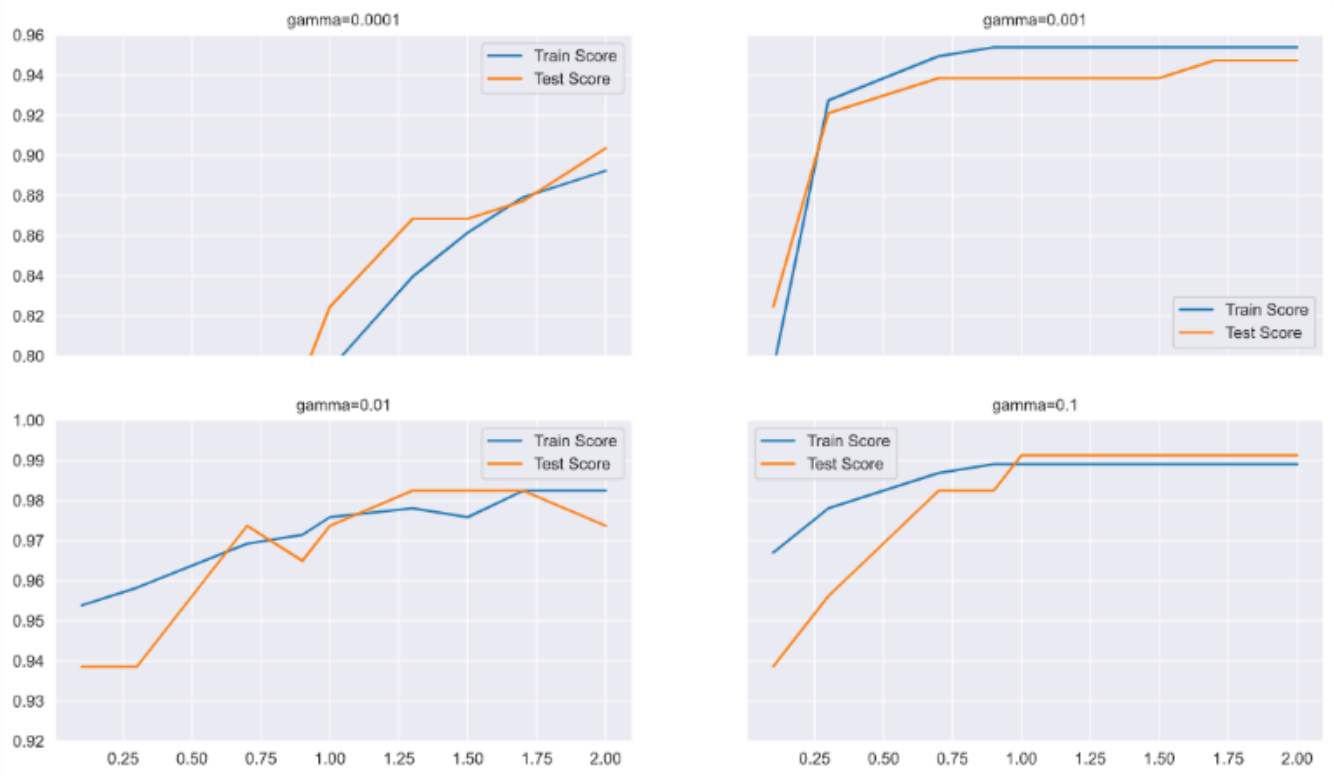

Fig-5: Comparison of train and test set accuracies for various values of C and gamma

Classification accuracy, F-Score, Precision, Recall and ROC value of RBF kernel is calculated on the test set which is 20% split of the WDBC dataset and presented here.

4.4.1. Accuracy

The classification accuracy is a measure of how well the classifier can correctly predict cases into their correct category. Accuracy can be calculated using the following equation:

$$\text{Accuracy} = \left(\frac{TP+TN}{P+N}\right) \times 100\% \quad (3)$$

Where TP and TN represents the True Positive and True Negative values respectively. Similarly, P and N represents the Positive and Negative population of Malignant and Benign cases. The results show a classification accuracy of 99% in the test.

4.4.2. Precision

Precision is the measure of the number of true positives among all the true measures including true positives and true negatives. It is also known as confidence.

$$\text{Precision} = \left(\frac{TP}{TP+FP}\right) \times 100\% \quad (4)$$

Where TP represents True Positive and FP False Positive. Precision values have been summarized in table-4.

4.4.3. Recall

Recall is the measure of the number of positive samples captured by positive predictions. It is also known as sensitivity. This measure is desirable, especially in the medical field because the number of observations that are correctly diagnosed. In this study, it is more important to correctly identify a malignant tumor than it is to

incorrectly identify a benign one. It is known as true positive rate (TPR).

$$\text{Recall} = \left(\frac{\text{TP}}{\text{TP + FN}}\right) \times 100\% \quad (5)$$

Where TP and FP represent True Positive and False Negative respectively. Recall values have been summarized in table-4.

4.4.4. F-Score

Looking at only one of precision or recall will not provide the full picture. The f-score or f-measure is one way to summarize them, which is with the harmonic mean of precision and recall:

$$\text{F} - \text{Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \quad (6)$$

This particular variant is called the f1 -score. It can be a better measure than accuracy on imbalanced binary classification datasets as it takes precision and recall into account.

The F-Score of the RBF kernel on the test set is 99% which have been summarized in table-4.

The various measures of evaluations are summarized in the table 4:

Table-4: Evaluation Report of RBF kernel

	Precision	Recall	f1-Score
Malignant	97%	100%	99%
Benign	100%	99%	99%
Average	99%	99%	99%

The confusion matrix of the classification is shown in table 5:

Table-5: Confusion Matrix

	Malignant	Benign
Malignant	39	0
Benign	1	74

It is evident that we can achieve an accuracy of 99.12% on the held-out test dataset. From the confusion matrix, there is only 1 misclassification among 114 test samples. The performance of this algorithm is expected to be high given the symptoms for breast cancer should exhibit certain clear pattern.

4.4.5. ROC Area

A receiver operating characteristics (ROC) graph is a way to visualize a classifier's performance by showing the trade-off between the cost and benefit of that classifier. It is one of the most common and useful performance measure. This 2-D graph plots the TPR(benefit) on the y-axis and the FPR on the x-axis (cost).[24]. The Area under a ROC graph shows the performance of the classifier. This is obtained by dividing the area under the plot with the total area of the graph. Values that are closer to 1 show a higher performance of the classifier.

The AUC of the calculated ROC curve is 0.993 which is the best possible value which shows that the performance of the classification on all thresholds is good. The ROC curve of the classifier is shown in fig. 6.

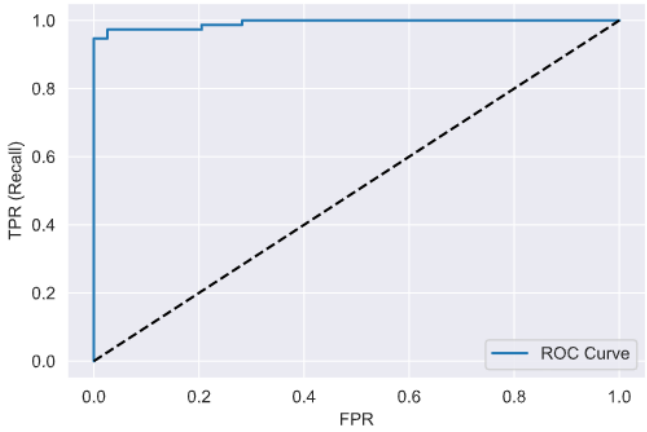

Fig-6. ROC Curve for RBF kernel

4.4.6. Discussion

The results present in table IV and V show that the support vector classification has a very good performance in terms of recall and precision. Also, this classification technique has the optimum ROC performance as the AUC is equal to 1.00. This shows that SVM has higher chance of discriminating between malignant and benign cases. The selection of accurate hyperparameters have increased the accuracy, specificity and precision of the classification of malignant and benign tumor. Numerous ML algorithms can be applied to get more accurate result in the prediction of breast cancer. Proper adjustments of the hyperparameters will enhance the classification accuracy. In future more data must be collected from across the world for a more precise and accurate classification of the disease. Future study will concentrate on finding more factors that have the potential to cause breast cancer and to include those potential factors in the dataset for a better classification. This can help in the enhancement and

automation of diagnosis of the disease. Future studies on the disease and application of various data mining and ML algorithm can help in better prediction of breast cancer.

5. CONCLUSIONS

In this paper, comparisons of results of different evaluation metrics for the four kernel functions namely linear, RBF, polynomial and sigmoid have shown that Radial Basis Function(RBF) kernel is the best among the four in terms of all the evaluation metrics discussed in the paper. RBF kernel showed maximum comparison accuracy of , F-score, ROC value on the validation set of the dataset choice for binary classification on the dataset.

The RBF kernel function is used on the test that that is 20% of the whole dataset and it has given the best result in the binary classification of benign and malignant tumor with only a single misclassification and accuracy of 99.12% . Thus, it can be concluded that RBF kernel function is preferred and is the best choice for an optimum result of the binary classification of breast cancer on WDBC dataset.

This experimental setting has never been shown before will allow to fully understand the performance of different kernel functions on both scaled and unscaled data and a better prediction model can be identified as a classifier for further studies.

REFERENCES

[1] Breast Cancer Organization https://www.breastcancer.org/symptoms/understand_bc/what_is_bc

[2] National Cancer Organization **https://seer.cancer.gov/statfacts/html/breast.html**

[3] Breast Cancer Organization **https://www.breastcancer.org/symptoms/types**

[4] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[5] Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Volume 27, Issue 8, 2006, Pages 861-874, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2005.10.010. (https://www.sciencedirect.com/science/article/pii/S016786550500303X) .

[6] Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

[7] Williams G. (2011) Descriptive and Predictive Analytics. In: Data Mining with Rattle and R. Use R. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9890-3_8.

[8] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Volume 13, 2015, Pages 8-17, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2014.11.005. (https://www.sciencedirect.com/science/article/pii/S2001037014000464).

[9] 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.

[10] Mller, Andreas C., and Sarah Guido. *Introduction to Machine Learning with Python: a Guide for Data Scientists*. O'Reilly, 2017.

[11] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[12] Todor K. Avramov and Dong Si. 2017. Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis. In Proceedings of the International Conference on Compute and Data Analysis (ICCDA '17). Association for Computing Machinery, New York, NY, USA, 69–74. DOI:https://doi.org/10.1145/3093241.3093290.

[13] H. Ayeldeen, M. A. Elfattah, O. Shaker, A. E. Hassanien and T. Kim, "Case-Based Retrieval Approach of Clinical Breast Cancer Patients," 2015 3rd International Conference on Computer, Information and Application, Yeosu, 2015, pp. 38-41, doi: 10.1109/CIA.2015.17.

[14] Samaneh Aminikhanghahi, Sung Shin, Wei Wang, Soon I. Jeon, Seong H. Son, and Chulwoo Pack. 2015. Study of wireless mammography image transmission impacts on robust cyber-aided diagnosis systems. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC '15). Association for Computing Machinery, New York, NY, USA, 2252–2256. DOI:https://doi.org/10.1145/2695664.2695832.

[15] Sy Ahmad Ubaidillah, Sharifah Hafizah & Sallehuddin, Roselina & Ali, Nor Azizah. (2013). Cancer Detection Using Aritifical Neural Network and Support Vector Machine: A Comparative Study. Jurnal Teknologi. 65. 10.11113/jt.v65.1788.

[16] William H Wolberg. 1992. Breast cancer Wisconsin (Original) data set. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/] (1992).

[17] Karapinar Senturk, Zehra & Kara, Resul. (2014). Breast Cancer Diagnosis Via Data Mining: Performance Analysis of Seven Different Algorithms. Computer Science & Engineering: An International Journal. 4. 35-46. 10.5121/cseij.2014.4104.

[18] You, H., & Rumbe, G. (2010). Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data. *Int. J. Interact. Multim. Artif. Intell., 1*, 5-12.

[19] Aruna, S & Rajagopalan, Dr & Nandakishore, L. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology. 2. 10.5121/csit.2011.1205.

[20] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/] (1992).

[21] Zafiropoulos E., Maglogiannis I., Anagnostopoulos I. (2006) A Support Vector Machine Approach to Breast Cancer Diagnosis and Prognosis. In: Maglogiannis I., Karpouzis K., Bramer M. (eds) Artificial Intelligence Applications and Innovations. AIAI 2006. IFIP International Federation for Information Processing, vol 204. Springer, Boston, MA . https://doi.org/10.1007/0-387-34224-9_58.

[22] Fushiki, Tadayoshi. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing. 21. 137-146. 10.1007/s11222-009-9153-8.

[23] Powers, David & Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229-3981. L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proenc¸a, "Digital signature of network segment for healthcare environments support," Irbm, vol. 35, no. 6, pp. 299-309, 2014.

[24] A. Simons, "Using artificial intelligence to improve early breast cancer detection, "2017. Retrieved on April 10, 2018, from https://www.csail.mit.edu/news/using-artificial-intelligence-improveearly-breast-cancer-detection.

[25] Medjahed, S.A., Saadi, T.A., & Benyettou, A. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications, 62*, 1-5.

[26] Sumbaly, Ronak & Vishnusri, N. & Jeyalatha, s. (2014). Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. International Journal of Computer Applications. 98. 16-24. 10.5120/17219-7456.

[27] Hazra, A., Mandal, S.K., & Gupta, A. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. *International Journal of Computer Applications, 145*, 39-45.

BIOGRAPHY

Apratim Sadhu is currently pursuing B.Tech in Computer Science Engineering from Chandigarh University, Mohali, India. His area of specialization in the under-graduate degree is Artificial Intelligence and Machine Learning. He is a rank holder in 19th National Children Science Congress. He has written 1 research article on Machine Learning.