CANCER CELLS CLASSIFICATION AND MODEL PERFORMANCE EVALUATION USING SUPPORT VECTOR MACHINE BASED CLASSIFIER APPLYING DIVERSE KERNEL

[1] Dr. Snehal K Joshi

[1]Asst.Professor
[1]Computer Department,
[1]Dolat-Usha Institute of Applied Sciences, Valsad, India

Abstract : **Support Vector Machine algorithm is very efficient particularly in case of non-linearly separable data as well as efficiently implemented where the data is labeled or non-labeled. However, at same time, the major constraint that affect performance of Support Vector Machine algorithm is choice of Kernel. Current study is pertaining to selection of appropriate Kernel in process to obtain optimal hyperplane. The study is based on a large dataset containing feature sets of human cancer cells characteristics. Dataset is consists of attributes related to human cells and their relevant properties. It contains records; which are classified as malignant or benign in scale of one to ten with various medical diagnostic procedures. Mapping of the data in higher order of polynomial is performed using four kernels, which include Linear, polynomial, Radial basis Function (RBF) and Sigmoid. Initially Linear kernel is used for mapping the data. The model is fit using kernel type linear using train set. This model is applied on cross-verification dataset to measure the precision, recall and f1-score. Accuracy is measured for cross-verification dataset, which is followed, by measuring accuracy for test-data-set. Similar process is repeated using other three types of kernels namely Polynomial, Radial bias Function and sigmoid. Confusion matrix is used for obtaining True Positive, True Negative, False Positive and False Negative results for cross-verification-data-set as well as for test-data-set. F1-score and jaccard-similarity-score are used to find the accuracy using all four kernels. Obtained results using all four kernels are compared and the variations are measured. This comparison provides the best-fit model, which is obtained using the specific kernel. On obtaining and comparing AUC (Area under Curve) and compare for performance evaluation, very interesting and unexpected result obtained. This is very significant in evaluation of performance measurement of the model. This study determines that selection of kernel is problem specific and accuracy or AUC cannot be sole criteria to decide the performance evaluation of the model.**

Keywords: ***Support vector machine, machine learning, Kernel, Linear, RBF, Polynomial, Sigmoid, AUC***

I. INTRODUCTION

Support vector Machine is supervised machine learning algorithm, which works on data mapped to high-dimensional feature sets. It is classification algorithm and used particularly, when the data are not linearly separable. When data cannot be separated linearly, the feature sets are mapped in such a way that the data can be classified using hyper plane. Objective of the SVM algorithm is to obtain the hyperplane such that the distance between the two divided class data points and the hyperplane that is also known as margin can be maximized. Ultimate goal is to obtain this optimal hyperplane where the margin is maximized. Kernel function is applied on data instances in process to map the non-linear observations, which results into higher-dimension space that can be separable using hyperplane.

Support Vector Machine is very efficient when it is applied on data that is not separable linearly. Obtaining the best-fit margin is biggest challenge in case of SVM and it needs significantly high computations. Feature mapping is the reason to increase computational complexity of the algorithm and hence the training performance of fitting the model is increased. However, this computational complexity has been reduced by applying Kernel. At other end, choosing appropriate Kernel is again a challenge. Kernel plays role of transforming the input data in appropriate format that results in obtaining efficient margin. Various Kernels are used including Linear Kernel, Non-linear kernels, sigmoid kernel, Radial Basis Function (RBF) Kernel and Polynomial. In case of complex problems where it is essential to classify certain advance, level of kernels can be used. Selection of Kernel for given problem plays very important role.

However, choosing the Kernel is problem specific. Therefore, it is not possible to generalize the idea of Kernel selection. Although, it is possible to categorise the selection of Kernel based on the similarity of feature sets and need of problem. The dataset used is pertaining to human cells features classified as possible malignant or benign and possess eleven characteristics. These dataset is verified as benign or malignant having value two or four recorded in class attribute. To obtain the most accurate classifier for this problem, four kernels namely linear Kernel, Polynomial Kernel, Radial Basis Function (RBF) kernel and sigmoid kernel will be used. The dataset is obtained from UCI Machine learning Repository (Asuncion and Newman, 2007) which is publicly available.

II. REVIEWS AND OBJECTIVES

As per the study on Support Vector Machine classifier and empirical comparison of kernel selection on text-independent speaker identification, linear, Radial Basis Function, Linear kernel and Polynomial kernel are compared and optimum performance is observed to 82.47% speaker identification rate using polynomial kernel over linear kernel and Radial Basis Function (RBF) [1]. As per the study by S.Amari and S.Wu (1999) proposed improvement of Support Vector Machine by enlarging spatial resolution around the separating boundary which is based on Riemannian geometry [2]. They proposed modification in Gaussian Radial Basis Function kernels that resulted in significant improvement in generalization errors. Study and findings by Keerthi, Chapelle and Decboste observes that Support Vector Machine is highly accurate, however it is not performing in case of high volume datasets due to its speed as numbers of vectors are very large [3]. They suggested measures to overcome these problems by proposed system, which is based on three properties. First suggestion as per their study is to decouple basis functions from the support vector. Second approach

is to obtain set of kernel basis functions of specified maximum size that approximate Support vector Machine cost function and the third approach is to scale using the training datasets.

Various works in the field of cancer classification using different approaches are performed. Different classifiers and accuracy measurements are used based on certain important features. Ayer T et al. ([5]) used ANN based classifier model for identifying breast cancer using dataset of size 62,219. Type of data contains demographic and mammogram features included important features that they considered include, Age and mammography findings. They used k-fold cross validation method where k=10 and obtained AUC as performance measurement which is 0.965. In another study carried out by Waddell M et al.([6]) based on multiple myeloma classification using features SNPs(Single Nucleotide Polymorphism) which is genetic variation; used leave-One-out validation method for dataset of 80 patients using SVM based classifier(Kernel=RBF type). They obtained 71% accuracy for this classifier. Another study of similar nature carried out by Listgarten J et al.([7]) using SVM classifier obtained 69% accuracy for Breast cancer classification having dataset volume of 174 patients which used 20-fold cross validation method. Stajadinovic et al.([8]) obtained AUC=0.71 as performance measure for their work on Colon carcinomatosis classification for dataset of 53 patients using Naïve Bayes based classifier based classifier which used cross validation. Work carried out by Exarchos K et al.([9]), Park C et al.([10]) and Eshlaghy A et al.([11]) used 10-fold cross validation method for the Support Vector Machine based classifier model and dataset of size 86, 437 and 547 in order to classify clinical and imagining tissue for oral cancer, Colon cancer and Cervical cancer respectively. They obtained accuracy as performance measurement of their classifier based model 100%, 76.5% and 95% respectively. For dataset of 440 patients having features based on clinical and gene expression for classification of Lung cancer, accuracy obtained 83.5% by Chen Y-C et al. ([12]). They used ANN based classifier and used Cross validation as validation method. Study carried out by Chang S-W et al. ([13]) to classify oral cancer used Support vector machine based classifier and obtained 75% accuracy for dataset of 31 size used cross validation as validation method. In another study carried out by Xu X et al. ([14]) to classify Breast cancer for dataset of size 295 using classifier based on Support vector machine yield 97% accuracy that used Leave-one-out cross validation method. Another study carried out by the Rosado P et al.([15]) observed accuracy of 98% for the classifier model based on Support vector machine to classify Oral cancer for database size of 69 and using validation method cross validation.

It is interesting to observe in these all studies that Support vector machine based classifier is more popular in classifying the cancer-based studies. However, some studies used ANN based classifier too. While looking at the dataset volume for the study it was ranging from two figures to few hundreds and yield accuracy ranging from 68% to 100% depending on type of classification problem. One more interesting observation noticed from this study is about the kernel that use to fit the classifier. All SVM based classifiers used RBF (Radial basis Function) kernels. No studies have specified the reason behind choosing the RBF kernel. Validation method used in majority of studies are either k-fold (k-5 or 10 mainly) or Leave-one out cross validation methods. Performance measurements used are either Accuracy or AUC (Area under curve) but none of the study has used both measures and compared them or depicted the difference in observation between these two measures.

Considering these past studies and observations obtained current study on mainly focus on three things. (i) Apply diverse kernels to fit the SVM based model instead of using default RBF (Radial Basis Function) kernel and identify most appropriate kernel by assessing the performance in terms of accuracy and other measures. (ii) Considering F1-score and Jaccard-similarity score apart from Accuracy as measure to assess the classifier's performance and (iii) Obtain AUC (Area under Curve) as performance measure for the classifier and assess the classifier's performance apart from accuracy and other measures.

For current study, dataset used is pertaining to human cell and its nine features. Class is the attribute, which contain actual category of cell, which is either Benign or Malignant. The classifier model is derived using classification algorithm, Support Vector Machine and training dataset. Since the classifier classify the test-datasets either benign or Malignant, the model's performance evaluation can be assessed based on four parameters, accuracy, specificity, sensitivity and precision. Since the classification is not linearly separable, the approach is used to transfer the dataset in higher dimension and by doing this; the dataset can be classified using hyper-plane. Higher dimensionality is achieved by applying kernel. This process is called kernelling. Four different kernels namely Liner Kernel, polynomial Kernel, Radial Basis Function (RBF) kernel and sigmoid kernel are used. Objective is to optimize the model performance and identify the kernel that is most appropriate for given problem. For the purpose of performance evaluation of kernel, True positive (TP), True Negative (TN), False positive (FP) and False Negative (FN) observations are obtained using confusion matrix. True Positive (TP) is those observations for which actual value is malignant and predicted values match with it. True Negative (TN) is those observations for which actual value is benign and predicted values match with it. False positive (FP) is those observations which were benign but the model classify them as malignant during test. This is Type-I error. Finally, False negative (FN) is those observations which were Malignant but falsely classify by the model as Benign. This is Type-II error.

Criteria for performance evaluation are observing Accuracy, precision, sensitivity and specificity. F1-score and jacquard-similarity-score are also considered.

$$\text{Accuracy} = (TP + TN)/\text{Total Observations} \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Sensitivity} = TP / (TP + FN) \quad (3)$$

$$\text{Specificity} = TN / (TN + FP) \quad (4)$$

Measurement of F1-score is significant when the Precision and Recall values need to balance. It gives good idea and it is used to compare among the results obtained using four different kernels. F1-score is obtained using weighted average.

$$F1 = 2 \times (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

Jaccard-similarity-score is also obtained for the classifier, which is modeled for different kernels. This similarity index observes provides ratio among the intersection of two sets and union of both sets. Here, we use Precision and Recall to obtain the Jaccard-similarity-score. It measures the similarity between the actual and predicted observations when the classifier classifies test-dataset. Using Jaccard-similarity-score and f1-score, we evaluate the performance of the classifier models, which used four different kernels.

$$\text{Jaccard-score} = ((A \cap B) / (A \cup B)) * 100 \quad (6)$$

Performance of models is also evaluated using AUC (Area under the Curve) of ROC (Receiver Operating Characteristics) which signify the model performance and depicts the performance for all possible thresholds between True Positive Rate (TPR) and False Positive Rate (FPR). It is one of the important measures to assess the performance. Please embed all fonts, in particular symbol fonts, as well, for math, etc.

2.1 IMPLEMENTING KERNEL FUNCTION

Several Kernels are popularly used for Support Vector Machine. It includes Polynomial Kernel, Gaussian Kernel, Gaussian Radial Basis function (RBF) Kernel, Laplace RBF Kernel, Hyperbolic Tangent Kernel, Sigmoid Kernel, Linear Kernel are some of the important Kernels. We are using four Kernels for current problem.

2.1.1 Linear Kernel

Linear Kernel is most appropriate when the dataset is having large number of features and are linearly separable. It is particularly very efficient when the dataset is pertaining to Text Classification.

$$k(X, Y) = 1+xy+xy \min(x, y) - ((x+y)/2) \min(x,y)^2 + 1/3(\min(x,y)^3) \quad (7)$$

2.1.2 Polynomial Kernel

Polynomial Kernel is popularly used in case of Image processing problems. Parameter d signifies the degree of polynomial.

$$k(X_i, X_j) = (X_i \cdot X_j + 1)^d \quad (8)$$

2.1.3 Radial Basis Function (RBF)

Radial Basis Functions are having two versions, which include Laplace Radial Bias Function and Gaussian Radial Bias Function. Gaussian Radial Bias function is general-purpose function that is applied when the prior knowledge of data is not available.

$$k(X_i, X_j) = \exp(-\gamma \| X_i - X_j \|^2) \quad (9)$$

Where $\gamma > 0$.

III. DATABASE CLEANING AND WRANGLING

Dataset is used for the problem is pertaining to human cell features which is available publicly from UCI Machine learning repository by Asuncion and Newman [4]. Dataset contain more than 20 thousand records of human cell samples. In total, there are nine features pertaining to the cell and the last field called class is having value 2 or 4, which represent Benign or Malignant respectively. These features include Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Bare nuclei , Bland chromatin, Normal Nucleoli and Mitoses which are ranged from 1 to 10 , where 10 signify highest possibility close to Malignant and 1 signify close to Benign. These datasets are medically verified datasets available publicly for academic purpose. First step is to clean records and remove the one having missing data.

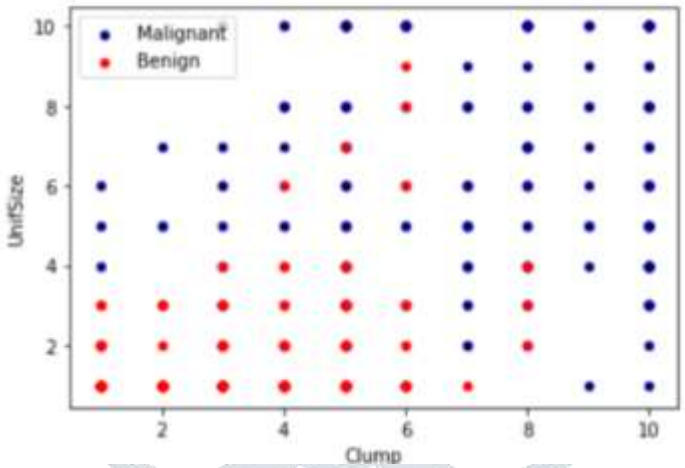

Fig.1: Uniformity of size and Clump Features

As shown in Figure-1, the class distribution Benign and Malignant are pertaining to feature sets Uniformity and Clump. It is visible that malignant and Benign classes are not linearly separable. Bare Nuclei is attribute that contain some values that are non-numeric. Hence, those records, which contain non-numeric data, are eliminated from the dataset. As result, the dataset contain all attributes that are of numeric type. Prediction variable is Class attribute, hence rest all attributes excluding ID are predictors. Predictors are obtained in array form and used as predictors to fit the model in process to obtain target class as Malignant or Benign.

IV. SUPPORT VECTOR MODEL DESIGN

Support Vector Machine algorithm implementation is by fitting the model in process to obtain target Y which is class attribute and X represents feature sets which include Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Bare nuclei , Bland chromatin, Normal Nucleoli and Mitoses. Dataset contain 699 records after cleaning process. N-fold approach is used to obtain Train-set and Test-set. Dataset is divided in four parts each contain 25% part of total dataset selected. This partition is obtained randomly. First 25% part is used as Test-Set for model testing which is fitted using the remaining three parts, which is 75% of the dataset. Hence, the testing of model is implemented four times iteratively; using 25% part each as Test-set and remaining 75% part as training set to fit the model. Accuracy, Specificity and sensitivity are obtained iteratively for each Test-set.

It is important to note that final observations are averaged using all four performance measurements obtained for four folds. It is also important to note that for each fold; model is trained using four different kernels to make sure that performance comparison is possible on similar ground by using uniform training dataset and testing dataset.

Table-1: Train-set and Test-set using four-fold

Fold	Train-Set	Test-Set
Fold-1	(512,9) (512,)	(171,9) (171,)
Fold-II	(512,9) (512,)	(171,9) (171,)
Fold-III	(512,9) (512,)	(171,9) (171,)
Fold-IV	(513,9) (513,)	(170,9)(170,)

4.1 Implementation of Linear Kernel:

Initially the model is trained using Linear Kernel. After training the model using Fold-I which contain train-set consists of 512 records and 9 features sets and tested using Test-set having 171 records and 9 feature sets. This process is repeated for rest three folds, Fold-II, Fold-III and Fold-IV.

Table-2: Confusion Matrix for Linear Kernel

Label	Precision	Recall	F1-Score	Support
2 (Benign)	0.99	0.93	0.96	110
4 (Malignant)	0.88	0.98	0.93	61

Confusion matrix obtained as shown in Table-2 for Linear Kernel is without normalizing. F1-Score obtained is 0.96 and 0.93 for Cl ass 2(Benign) and 4(Malignant) respectively.

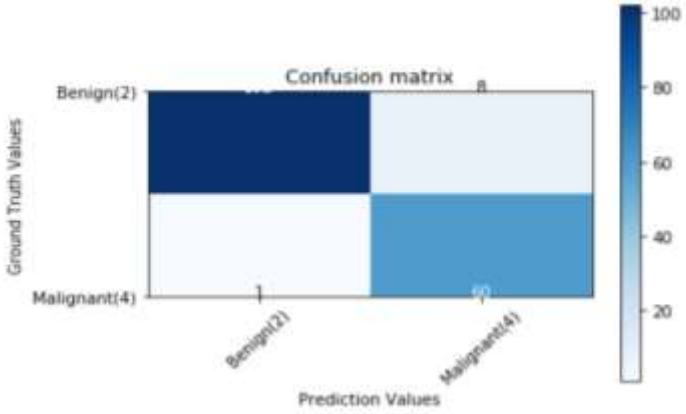

Fig.2. Confusion Matrix for Linear Kernel

F1-Score and Jaccard-score are obtained for Linear Kernel implementation as 0.9479 and 0.9473 respectively.

4.2 Implementation of Polynomial kernel:

Now, the model is trained using Polynomial Kernel. Polynomial Kernel is normally implemented in area of Image processing where feature sets are not very large. After training the model using Fold-I that contain train-set consists of 512 records and 9 features sets and tested using Test-set having 171 records and 9 feature sets. This process is repeated for rest three folds, Fold-II, Fold-III and Fold-IV.

Table-3: Confusion Matrix for Polynomial Kernel

Label	Precision	Recall	F1-Score	Support
2 (Benign)	1.00	0.93	0.96	110
4(Malignant)	0.88	1.00	0.94	61

Average F1-score obtained using sklearn library of Python for Polynomial Kernel is 0.9537. Whereas, obtained Jaccard-similarity-score is 0.9532.

4.3 Implementation of Radial Bias Function(RBF):

For the model training and testing, Laplace Radial Bias Function is used over Gaussian Radial Bias Function. As such, significance of Gaussian Radial Bias Function and implementation area is where the prior knowledge of data is not known. Hence, the Laplace Radial Bias Function (RBF) is implemented here. After training, the model using Fold-I that contain train-set consists of 512 records and 9 features sets and tested using Test-set having 171 records and 9 feature sets. This process is repeated for rest three folds, Fold-II, Fold-III and Fold-IV. Obtained average confusion matrix for all four folds test-set is as shown Table-IV.

Table-4: Confusion Matrix for Radial Bias Function (RBF) Kernel

Label	Precision	Recall	F1-Score	Support
2 (Benign)	0.99	0.91	0.95	110
4(Malignant)	0.86	0.98	0.92	61

Average F1-score obtained using sklearn library of Python for Polynomial Kernel is 0.9365. Whereas, obtained Jaccard-similarity-score is 0.9356.

4.4 Implementation of Sigmoid Kernel:

Finally, the last kernel which we use for obtaining the f1-score and Jaccard-similarity-score is sigmoid. Sigmoid Kernel application is widely in Artificial Neural Netowrk(ANN) when other kernels are insignificant. Using the sigmoid kernel, the Train-set is used to fit the model. After training the model using Fold-I that contain train-set consists of 512 records and 9 features sets and tested using Test-set having 171 records and 9 feature sets. This process is repeated for rest three folds, Fold-II, Fold-III and Fold-IV. Obtained average confusion matrix for all four folds test-set is as shown Table-V.

Table-5: Confusion Matrix for Sigmoid Kernel

Label	Precision	Recall	F1-Score	Support
2 (Benign)	0.48	0.51	0.49	110
4(Malignant)	0.00	0.00	0.00	61

Average F1-score obtained for sigmoid Kernel is 0.3173. Whereas, obtained Jaccard-similarity-score is 0.3274. It is evident from Table-5 that precision and Recall values are very low for Label 2 which is Benign. It signifies that True Positive is very less as well as False Positive is also higher. F1-score and Jaccard-similarity-index also observed very low. Accuracy obtained using f1-score is also observed to be very low in case of sigmoid kernel application.

V. OBSERVATIONS AND ANALYSIS

Once the Support Vector machine classifier model is fit, it is tested over the test-dataset for four folds using Linear, Polynomial, Radial Basis Function and Sigmoid kernels. All parameters are kept constants and default when all kernels are implemented. Precision, Sensitivity, specificity and Accuracy calculated using equations (1) to (4) from confusion matrix observations obtained from model implemented using Linear Kernel. Accuracy obtained for Linear kernel is 0.9473 which is 94.73% and significantly high. Precision observed to be 92.73%, again it is significantly high. Sensitivity and specificity is 0.99029 and 0.88235 respectively and it infers True positive rates and True Negative rates respectively. Sensitivity is 99.03% and significantly high, whereas Specificity is 88.23% which is also high.

Table-6: Accuracy, Sensitivity, Specificity and Precision

Kernel Type	Precision	Sensitivity	Specificity	Accuracy
Linear	0.92727	0.99029	0.88235	0.94737
Polynomial	0.92727	1.00000	0.88406	0.95322
RBF	0.90909	0.99010	0.85714	0.93567
Sigmoid	0.50909	0.47863	0.00000	0.32749

Accuracy obtained for model implementation using Polynomial kernel is 95.32% and significantly high. Precision observed to be 92.73% , again it is significantly high. Sensitivity and specificity is 1.0000 and 0.8841 respectively and it infers True positive rates and True Negative rates respectively. Sensitivity is 100.00% and optimum, whereas Specificity is 88.41% that is also high. It is important to note that Sensitivity is 100% and hence True Negative is obtained with utmost accuracy.

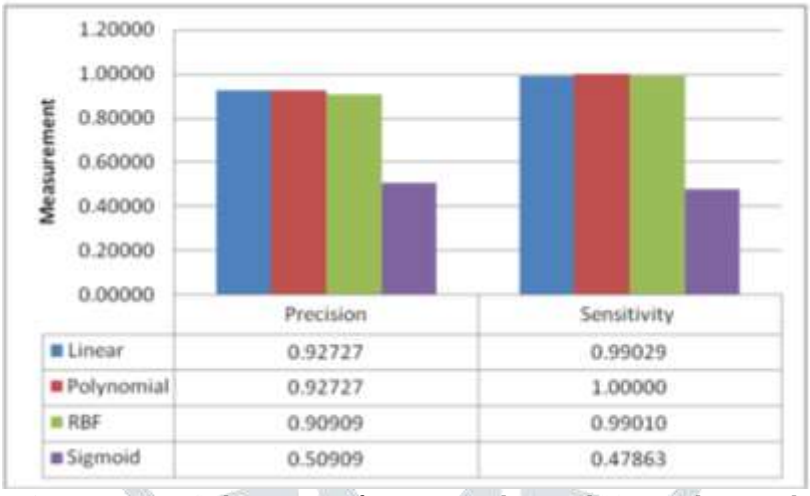

Fig.3: Precision and Sensitivity for Linear, polynomial, RBF and Sigmoid

In case of model implementation using Radial Basis Function (RBF), Accuracy obtained is 93.57% that is significantly high. Precision observed to be 90.91%; again, it is significantly high. Sensitivity and specificity is 0.9901 and 0.8571 respectively and it infers True positive rates and True Negative rates respectively. Sensitivity is 99.01% and significantly high, whereas Specificity is 85.71% which is also high.

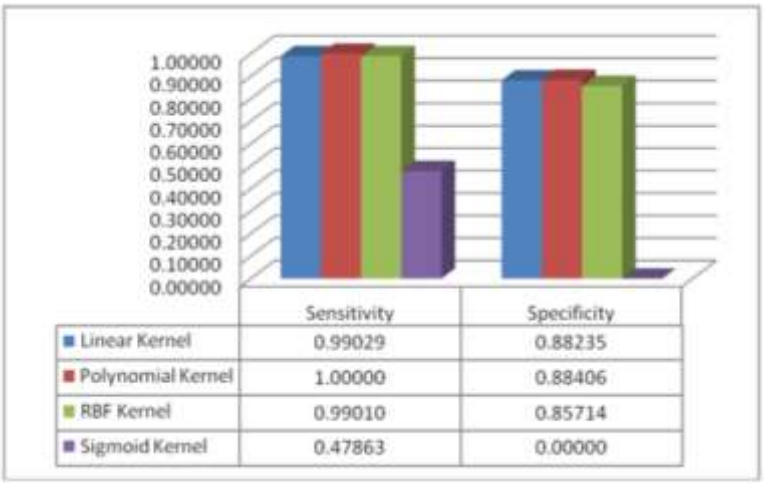

Fig.4: Sensitivity and Specificity for all four kernels.

Classifier model implementation using sigmoid kernel results is exceptionally low. Accuracy obtained is 32.75% that is significantly low. Precision observed is 50.91%; again, it is significantly low. Sensitivity and specificity is 0.4786 and 0.0000 respectively and it infers True positive rates and True Negative rates respectively. Sensitivity is 47.86% and very low, whereas Specificity is 00.00% which is exceptionally at bottom low.

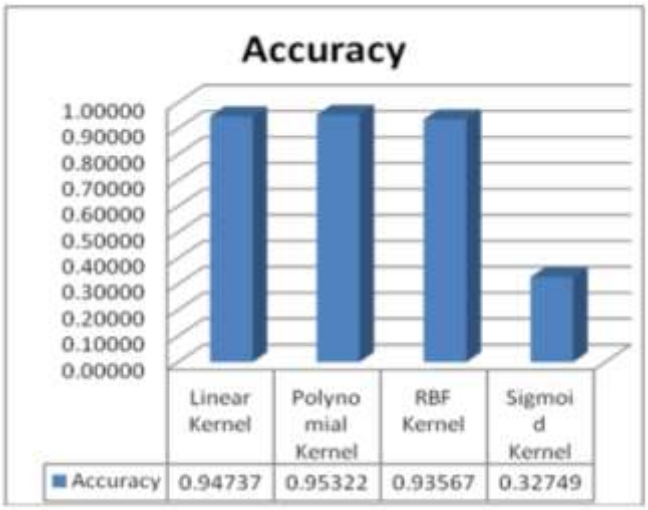

Fig.5: Accuracy for Linear, Polynomial, RBF and Sigmoid Kernel

F1-score obtained for classifier model implementation using linear kernel is 0.9479. Polynomial kernel f1-score is 0.9537. F1-score for Radial Bias Function (RBF) and Sigmoid Kernel are 0.9365 and 0.3173 respectively.

Table-7: F1-Score and Jaccard-Score for all Kernels

Kernel	F1-Score	Jaccard-Score
Linear Kernel	0.9479	0.9473
Polynomial Kernel	0.9537	0.9532
Radial Bias Function (RBF)	0.9365	0.9356
Sigmoid Kernel	0.3173	0.3274

Jaccard-similarity score for Linear Kernel is 0.9473 that is significantly high. Polynomial kernel observation is 0.9532 and in case of Radial Basis Function (RBF) is 0.9356. Jaccard-similarity score for sigmoid kernel is observed to be 0.3274, which is significantly low.

Finally, to verify the performance and to cross check the performance of the models, AUC (Area under the Curve) is also obtained. As shown in Table-VIII, the AUC observations depict performances of all four kernel based models. Very surprising observation is obtained for sigmoid kernel based model.

Table-8: AUC score comparison with Accuracy

Kernel	AUC	Accuracy
Linear Kernel	0.9981	0.9473
Polynomial Kernel	0.9992	0.9532
Radial Bias Function (RBF)	0.9823	0.9356
Sigmoid Kernel	0.9964	0.3274

It is observed that Accuracy obtained 32.74% while implementing sigmoid kernel but in contradictory measures obtained in case of AUC. It is observed that AUC score obtained for sigmoid kernel based model is significantly high with 99.64%. This is very surprising result and normally, if the observations are considered based on AUC, sigmoid based kernel model can be considered. However, its accuracy is contradictory and it does not suggest having model based on sigmoid kernel.

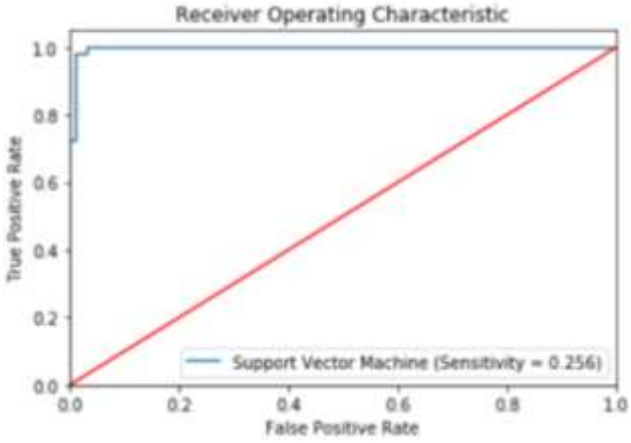

Fig.7: AUC-ROC curve for Sigmoid Kernel based model

It is essential to understand this behavior. As we know that, the AUC-ROC score represents FPR (False positive Rate) and TPR (True Positive Rate 0 for different threshold values. Looking at the confusion matrix obtained for sigmoid-based kernel model, it is evident that True Negative value is zero and hence, specificity value zero. Now, FPR (False positive Rate) is one minus specificity. Hence, obtained value for FPR is one. Since value for FPR is one, the AUC-ROC will be almost 100% and it leads to false interpretation on matter of selection of kernel to fit the model. This particular case leads to the conclusion against traditional assumption of considering AUC value to determine the performance evaluation measure. This result also depicts that AUC or Accuracy cannot be used to determine as sole indicator of model selection. It is necessary to cross verify in cases where AUC is very high or low using the Accuracy measure along with other measures like F1-Score, Jaccard-similarity index and of course the Sensitivity and Specificity. It is also important to note from the observations that Jaccard-similarity score yield almost similar score that we observed for Accuracy. This is because the classification is of binary type. It is more appropriate when the classifier classifies more than one class.

VI. CONCLUSION

Two important and results are obtained which are very much unusual and different from expectations. One is related the choice of kernel and second is related to the performance evaluation measurement using AUC. Both these results are concluded at end part of this section. Classification problem is pertaining to human cell and to classify them based on their feature sets as malignant or benign using Support vector machine based model. The model is trained using the train datasets n-folds approach (n=4). Each fold is used as test-dataset whereas balance all three folds collectively used as training set to train the model. Dataset classification is not linear, hence to achieve higher dimensionality; kernel is used. Kernel transforms the dataset to higher dimensionality, which is separable using hyper plane. Objective of the work is to identify the kernel that is most appropriate for given problem and yield optimum performance. Four kernels are used to analyze the performance evolution of model. These four kernels are linear kernel, Polynomial kernel, Radial Basis Function (RBF) kernel and sigmoid kernel. Classifier model performance analysis is measured by obtaining Accuracy, specificity, sensitivity and precision. Jaccard-similarity score is also used to analyze the similarity score among actual and predicted. F1-score is also used for performance evaluation of models. It is observed that sigmoid kernel performance based on statistical performance analysis is very low. Accuracy obtained for model fitted using sigmoid kernel is 32.75% that is very poor. Precision, sensitivity and specificity observed are 50.51%, 47.86% and 0.00% that shows drastically low performance. Looking at the f1-score and Jaccard-score that are 0.3173 and 0.3274, we can conclude that sigmoid kernel is failed to accept to implement for given problem of classification. Considering the performance of Linear, Polynomial and RBF kernel accuracy maximum accuracy and precision are obtained in case of Polynomial kernel based model that is 95.32% and 92.73%. Linear kernel accuracy and Precision are 94.74% and 92.73% that is very much close to the performance of Polynomial kernel performance. It is observed that precision is observed to be similar in case of Polynomial and linear kernel based models. However, accuracy is observed higher in case of Polynomial kernel based model. RBF kernel based model's accuracy, precision is observed as 93.57% and 90.91% that is lower than the performance of Linear, and Polynomial kernel based models. Considering the sensitivity and specificity; 100% sensitivity is obtained for Polynomial kernel based model which is significantly higher compared to linear and RBF kernel. Specificity observation in case of Polynomial kernel is also highest among rest all kernels that is observed as 88.41%. However, it is also important to note that sensitivity and specificity of model based on linear kernel are 99.03% and 88.23% respectively. It is also significantly high and very close to the model based on Polynomial kernel. F1-score and Jaccard-score obtained for Polynomial kernel based model are also higher than rest models. Based on these observations, it can be concluded that Polynomial Kernel based model performance for given classification problem is more appropriate compared to the linear, RBF or sigmoid kernel based model. It is also important to note that RBF kernel is widely used for Support Vector machine based classification models, but use of kernel is problem specific and it is essential to note that Polynomial Kernel yield better performance to model the classifier in this particular scenario.

(i) Among the Models based on four Kernels, 95.32% accuracy is obtained while Polynomial kernel is used. This is highest accuracy among all models based on rest three kernels.

(ii) Accuracy is crossed verified using F1-score and Jaccard similarity score that results 95.37% and 95.32% respectively.

(iii) AUC score obtained for model based on Polynomial kernel yield 99.92% that is highly significant. Considering Accuracy, F1-score and AUC, we can conclude that Polynomial kernel based model is most significant and adoptable for particular problem.

(iv) One important observation obtained for this study is about sigmoid-based model performance observation. Accuracy and AUC score obtained for sigmoid kernel based model are highly contradictory. This result and observations lead us to conclude that only Accuracy or AUC score cannot be determined to predict the fitness and adoptability of model.

References

Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

[1] Boujelbene, Siwar and Ben Ayed, Dorra and Ellouze, Noureddine. 2010. Improving SVM by Modifying Kernel Functions for Speaker Identification Task, JDCTA. 4(6.12): 100-105.

[2] S.,Amari , S.,Wu. 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, Elsvier, 12: 783-789.

[3] S.,Sathiya Keerthi,Oliver Chapelle, Dennis Decoste. 2006. Building Support Vector Machines with Reduced Classifier complexity. Journal of Machine Learning Research, 7:1493-1515.

[4] Asuncion and Newman. 2007. UCI Machine learning repository, http://mlearn.ics.uci.edu/MLRepository.html.

[5] T. Ayer, O. Alagoz, J. Chhatwal, J.W. Shavlik, C.E. Kahn, E.S. Burns"ide. 2010. Breast cancer risk estimation with artificial neural networks revisited Cancer. 116 :3310-3321.

[6] M. Waddell, D. Page, J. Shaughnessy Jr. 2005. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma, ACM Journal. 1: 21-28.

[7] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, et al. 2004. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res. 10:2725-2737.

[8] A. Stojadinovic, A. Nissan, J. Eberhardt, T.C. Chua, J.O.W Pelz, J. Esquivel. 2011. Development of a Bayesian belief network model for personalized prognostic risk assessment in colon carcinomatosis. Am Surg, 77:221-230.

[9] K.P. Exarchos, Y. Goletsis, D.I. Fotiadis. 2012. Multi-parametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans Inf Technol Biomed, 16:1127-1134.

[10] C. Park, J. Ahn, H. Kim, S. ParkI. 2014. Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. PLoS One, 9:e86309.

[11] A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A.R. Razavi, L.G. Ahmad. 2013. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4:124.

[12] Y.-C. Chen, W.-C. Ke, H.-W. 2014.ChiuRisk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med, 48:1-7.

[13] S.-W. Chang, S. Abdul-Kareem, A.F. Merican, R.B. Zain. 2013. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinforma, 14:170.

[14] X. Xu, Y. Zhang, L. Zou, M. Wang, A. LiA. 2012. Gene signature for breast cancer prognosis using support vector machine. IEEE , 928-931.

[15] P. Rosado, P. Lequerica-Fernández, L. Villallaín, I. Peña, F. Sanchez-Lasheras, J.C. de Vicente. 2013. Survival model in oral squamous cell carcinoma based on clinic pathological parameters, molecular markers and support vector machines. Expert Syst Appl, 40:4770-477.