

Evaluation of different SVM Kernels and their Comparison on Breast Cancer Detection

Sheela S¹ and Rashmi S²

¹Assistant Professor, School of Computer Science and Engineering, RV University, Bangalore.

²Professor, CSE(AI&ML), School of Computer Science and Engineering, RV University, Bangalore

Abstract— An assessment of numerous Support Vector Machine (SVM) kernel functions, which include linear, polynomial, sigmoid, and radial basis function (RBF) kernels, are used for the crucial utility of breast cancer detection. Early and correct analysis of breast cancer is of paramount importance because it immediately affects the remedy effects and survival rates of affected persons. SVMs have been demonstrated to be relatively powerful types throughout several domains. In this study, the type overall performance of those SVM kernel with DaskGridSearchcv configurations changed into carefully assessed on a breast cancer dataset, using complete assessment metrics and cross-validation techniques. The number one goal changed into discovering the kernel feature that yields the very best type accuracy for breast cancer detection, at the same time as additionally gaining insights into the respective strengths and weaknesses of every kernel type. The findings from this study have giant implications for improving sturdy and dependable breast cancer analysis structures primarily based totally on optimized SVM classifiers. The effects can manually choose the most suitable kernel feature and facilitate the fine-tuning of SVM hyperparameters to maximize the overall performance of this crucial clinical utility.

I. INTRODUCTION

Breast cancer is the most common cancer among women in India. Breast cancer is a cancer that develops from the breast tissue. Signs of breast cancer include a lump in the breast, a change in breast shape, dimpling, and others. As per research, 1 in each 28 ladies is probably going to get impacted by the illness. While breast cancer occurs almost entirely in women, around 1-2% of men are likely to get affected, too. Through machine learning techniques, we can detect breast cancer risk in its early stage and increase the chances of survival. The review of current research shows that the algorithms used for detecting breast cancer diagnosis are supervised. This paper evaluates the performance of various SVM models with its various kernel functions. This paper uses various kernel types such as linear SVM with DaskGridSearchcv, radial basis function with DaskGridSearchcv, sigmoid function with DaskGridSearchcv, and polynomial function with DaskGridSearchcv to classify whether the cancer is benign or malignant.

II. RELATED WORKS

[1] Farjana Akter Tania and Pintu Chandra Shill have presented a new method for improving SVM performance using hybrid kernel functions with a data pre-processing algorithm to reduce noisy data, successfully achieving this. To validate their proposed model, they tested it with different dataset sizes and attributes.

The simulation results shown in tables and figures, demonstrate the performance indices like accuracy, sensitivity, specificity, false discovery rate, false omission rate, and Matthews correlation coefficient for the training and testing phases on the Heart Disease Diagnosis dataset, comparing RBF, Linear, and Hybrid kernels. Their proposed method outperformed existing methods for this dataset in terms of accuracy. For the WBC dataset in the training phase, their proposed SVM classifier with the LRBF kernel achieved up to 100% sensitivity, 99.54% specificity, and 99.84% accuracy. On the WBC diagnostic dataset, it attained 99.74% accuracy, 99.58% sensitivity, and 100% specificity. On the Cleveland heart disease dataset, it achieved 99.17% accuracy, 99.25% sensitivity, and 99.06% specificity. Thus, they showed SVMs are efficient, giving the best classification accuracy by obtaining the optimum separating surface with good performance on unseen datasets. They state this research can be extended to other larger datasets.

[2] The researchers, including C.-S.L. developed a two-stage breast cancer model to improve diagnostic accuracy and reduce misdiagnosis. They first used a risk factor dataset, then an image recognition model with pre-trained ConvNets (AlexNet, ResNet101, and Inception_V3) on mammograms. Using optimization algorithms like ADAM and SGDM, they classified tumours as benign or malignant and evaluated accuracies. The stepwise approach boosted classification accuracy over single models, which could misclassify cases. Using multiple ConvNet models with soft voting allowed the correct classification of previously misclassified cases, enabling timely treatment. The research was funded by Taiwan's Ministry of Science and Technology and a university-hospital collaboration.

[3] Ebrahim Edriss Ebrahim Ali and Wu Zhi Feng demonstrated the modelling of breast cancer as a classification task in their paper. They described the implementation of Neural Network (NN) and Support Vector Machine (SVM) approaches for classifying breast cancer as either benign or malignant. The researchers compared the results of both NN and SVM based on accuracy and precision. They observed that the classification implemented by the Neural Network technique in their paper is more efficient compared to SVM, as seen in the higher accuracy and precision achieved by NN. Based on their results, Ebrahim Ali and Wu Zhi Feng concluded that the NN technique is more efficient compared to the SVM technique in breast cancer detection.

[4] Tao Xie, Jun Yao, and Zhiwei Zhou proposed a novel classification algorithm, DA-CKSVM, based on SVM with a combined kernel function constructed with an RBF kernel function and a polynomial kernel function. The parameters of the combined kernel SVM classifier were optimized by DA. They compared DA-CKSVM's performance in cancer classification with four algorithms from the literature - DA-SVM, PSO-SVM, BA-SVM, and GA-SVM algorithms. Their experimental results showed that due to its excellent learning ability and generalization ability, the proposed DA-CKSVM algorithm had better classification accuracy than these algorithms with a single kernel function, reflecting its good practical value in the field of cancer classification.

[5] Primit Brata Chanda and Subir Kumar Sarkar state that breast cancer is the most common disease in women worldwide, and early detection can reduce mortality. Their work illustrated detecting and classifying different breast cancer types from mammogram images using a Support Vector Machine (SVM) classifier. The segmentation and classification approach on MIAS breast cancer images achieved over 80% linear and quadratic classification accuracy. They suggest this SVM classifier could help radiologists make quicker, more accurate diagnostic decisions, reducing unnecessary biopsies and radiologist hazards. Their future work aims to extend testing to various datasets, feature sets, and classifiers.

[6] Bouchra El Ouassif, Ali Idri, and Mohamed Hosni proposed and evaluated SVM ensembles using four kernels (LK, NP, RBF, and PUK) and MLP as a combiner rule over four historical datasets. They compared the SVM ensembles and single SVM methods using accuracy, precision, and recall criteria, with the SK test for significance tests and Borda Count to rank the best classifiers. Their findings showed that SVM ensembles outperformed single SVMs, meaning using a combination of different kernels often leads to better classifiers, and ensemble performance increases with more kernel members. While no SVM ensemble outperformed all others, using P and R kernels seemed to increase their performance. For efficiency, the single SVM-PUK can be used instead of ensembles. El Ouassif, Idri, and Hosni found that ensembles are generally time-consuming.

[7] Elsayed Badr, SultanAlmotairi, Mustafa Abdul Salam, and Hagar Ahmed made three main contributions to this work. The first was improving the performance of the support vector machine (SVM) using a recent grey wolf optimizer (GWO) for breast cancer diagnosis. The second was proposing three efficient scaling techniques against classical normalization. The third was using a parallel technique applying task distribution to improve GWO efficiency. They applied the proposed sequential model on two datasets - the Wisconsin Diagnosis Breast Cancer (WDBC) and Electronic Health Records (EHR). Experimental results on WDBC showed the proposed hybrid GWO-SVM model achieved 98.60% accuracy with normalization scaling, and using the proposed scaling techniques with GWO-SVM gave faster convergence and 99.30% accuracy. The parallel version achieved a 3.9

speedup on four CPU cores. For EHR, the hybrid GWO-SVM model achieved 93.26% accuracy with normalization against 82.05% for SVM.

[8] In the conclusion, Garima Verma and Hemraj Verma state that in this study, they made an effort to develop a model that could predict whether a patient has malignant or benign breast cancer. The model uses the Support Vector Machine algorithm based on a linear kernel. They performed experiments using the dataset from the UCI machine learning repository with 569 records. The results of the experiments are shown in the form of a confusion matrix, precision, recall, overall accuracy, and AUC-ROC curve. Garima Verma and Hemraj Verma mention that the overall accuracy of their model is 90.35%. They indicate that in the future, the study may be improved by including more machine learning algorithms and conducting a final evaluation for better performance.

[9] Tina Elizabeth Mathew analyzed with an accuracy of 94.6%. The study concludes that various machine learning (ML) approaches applied to breast cancer (BC) diagnosis and prognosis have been examined utilizing the Wisconsin Breast Cancer Dataset (WBCD) as a benchmark. These ML techniques have demonstrated significant enhancements in classification and prediction accuracy. Nonetheless, given the paramount importance of accuracy in biomedical sciences, and considering the variability of methods across different datasets, these complex techniques must undergo thorough examination and rigorous assessment before their commercial utilization. The study evaluates the performance of Support Vector Machines (SVM) with different kernel functions. Additionally, it suggests further analysis of SVM performance when combined with other methods or utilized within ensembles in future research.

[10] Xueyun YE, Zhixiong ZHANG, and Yu JIANG state that this paper attempts to apply machine learning technology to detect whether a tumour is malignant or benign. By analysing and sorting out breast cancer data, they compared several different classification algorithms based on the Support Vector Machine (SVM). Using various SVM classification methods, they carried out training prediction analysis of sample values. Cross-validation and grid search were used in their experiments.

Their experiments found that adding a radial basis function into SVM greatly improved the convergence speed of the algorithm, and the accuracy and recall rate were better than the other three classifiers tested. Xueyun YE, Zhixiong ZHANG, and Yu JIANG conclude that the SVM algorithm has a high accuracy and recall rate for dichotomous problems, and can achieve good results in practical applications of detecting malignant versus benign tumours.

[11] Jiaji Wang, Muhammad Attique Khan, Shuihua Wang, and Yudong Zhang proposed a model utilizing an SVM classifier guided by SqueezeNet with complex bypass, trained on the mini MIAS breast cancer image dataset. Evaluation metrics included sensitivity, specificity, precision, accuracy, F1, MCC, and FMI. Their experimental findings showed the proposed SNSVM model surpassed existing state-of-the-art models across all performance metrics with reasonable fluctuations. To enhance SNSVM's performance, they plan to explore different algorithms, improve existing ones to better capture data patterns, conduct thorough feature engineering to extract relevant information and reduce noise, and fine-tune the model's hyperparameters to achieve improved results, aiming to optimize the model and increase its predictive power for more effective decision support.

[12] [12] Prity Vijay and Bright Keswani, investigated breast cancer, a disease predominantly affecting women, particularly those over the age of 35. With a high mortality rate if undetected early, the urgency for an effective solution is paramount. To address this, they employed Support Vector Machines (SVM) with linear, RBF, and polynomial kernels to construct a predictive model. Their findings revealed that the linear and RBF kernels exhibited exceptional performance, while the polynomial kernels yielded less favorable results. They further refined the model by adjusting the gamma and C parameters. Ultimately, they determined that the SVM model with a linear kernel ($\gamma=1$, $C=1000$) achieved the highest accuracy of 98% in predicting breast cancer, thus establishing it as the optimal model for this purpose.

III. METHODOLOGY

The Wisconsin Breast Cancer Database (WBCD) obtained from Dr. William H. Wolberg of Wisconsin University Hospitals, Madison is used. The Original data contains 699 instances and 11 attributes, the first attribute is the ID of the instances is discarded as it has no role in our predictions, and the next 9 represent the different characteristics of an instance. The cytological attributes of the breast fine needle aspiration (FNA) test are represented by values going from 1 to 10 for each instance 1 for the least problematic and 10 for the most. The last property in the dataset represents the sort of diagnosis. Each instance in the breast fine needle aspiration (FNA) test belongs to one of two possible classes: benign (value 2) or malignant (value 4). The prediction

process uses nine attributes, which are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses.

This dataset contains 699 instances, 458 instances are benign, and 241 instances are malignant (65.52% of them are benign and 34.48% of them are malignant). We excluded sixteen instances from the study due to the missing values, leaving us 683 instances (444 benign, 239 malignant) included for analysis.

TABLE I: BREAST CANCER DATASET ATTRIBUTE DETAILS

Sl No	Attribute	Range of the values	Comparison of benign and malignant cells	
			Malignant	Benign
1	clump thickness	1-10	Seen in multilayers	Seen in monolayer
2	size uniformity	1-10	Size differs	Uniform size
3	shape uniformity	1-10	Shape differs	Uniform shape
4	marginal adhesion	1-10	Cells do not stick together	Cells stick together
5	epithelial size	1-10	enlarged	small
6	bare nucleoli	1-10	Have bare nucleoli	No bare nucleoli
7	bland chromatin	1-10	Coarse in texture	Uniform texture
8	normal nucleoli	1-10	Nucleus is bigger	Nucleus is small
9	mitoses	1-10	Mitosis is more	Not many are there
	CLASS	2 for benign 4 for malignant		

Table I defines the instances and its range

A. Support Vector Machine

The Support Vector Machine (SVM) is a regulated machine learning calculation utilized for both characterization and relapse tasks. It can handle continuous and categorical variables and is more reasonable for working with a small dataset rather than a larger dataset. SVM was developed by Vladimir Vapnik, seeking to find an ideal hyperplane that isolates unique classes of information by maximizing the margin between the closest training examples and the hyperplane. SVM transforms the data into a higher-dimensional space using kernel functions, taking into account more mind boggling decision boundaries this method is used for nonlinear issues with various features.

Some normally utilized piece capabilities are linear, nonlinear, polynomial kernels, used in image processing, Gaussian kernel, Gaussian Radial basis function (RBF), and Laplace RBF kernels, used when there is no prior information regarding the data, Hyperbolic tangent Kernel, used in Neural Networks, Sigmoid, as proxy in Neural Networks, Linear splines kernel in one-dimension, used in large sparse data sets, Anova Radial basis Kernels, used in regression problems and Bessel function of the first kind Kernel, used to eliminate cross terms in functions. The most ordinarily involved kernel for disease prediction is the radial basis function. Five kernel functions are taken and studied. The work was done in Python utilizing the sci-kit-learn implementation of SVC.

B. SVM with Radial basis Function Kernel and DaskGridsearchcv

The radial basis function kernel (RBF) is defined as $k(x,y) = \exp(-\gamma \|x-y\|^2)$ where x and y are two input vectors, γ , is variance. The dataset is normalized and partitioned into 70-30 training-testing sets. DaskGridSearchcv is a tool that extends sci-kit-learn's GridSearchCV to work with Dask which is an adaptable library for equal registering in Python and distributed hyperparameter tuning across large datasets. It works similarly to GridSearchCv but influences the force of Dask to distribute computations across multiple workers or machines for efficient processing. From this, we can get the best arrangement of model hyperparameters by specifying the estimator, parameter grid, and cross-validation strategy Dask Grid search cross-validation was used to find the best penalty parameter values- 'C' and 'gamma' parameter values. The model is built in light of the best parameter values obtained. The data is normalized before use as normalization enhances the accuracy rates. The presentation of the classifier is shown in table II

TABLE II: CLASSIFICATION REPORT

Class	Precision	Re-call	F1-scor-e	Supp-ort	Confusion matrix	Parameter used
2-Benig-n	1.00	0.95	0.97	82	[[85, 4], [0, 55]]	C:5, gamma: 0.1
4-Malig-nant	0.93	1.00	0.96	55		
Accuracy			0.97	137		
macro avg	0.97	0.98	0.97	137		
weighted avg	0.97	0.97	0.97	137		

Table II represents the Classifications Report of RBF-DGSCV kernel based on Confusion Metrics

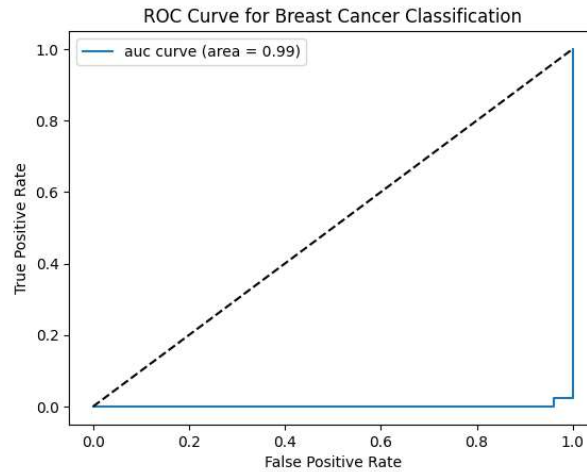


Fig 1. ROC Curve for RBF - DGSCV Kernel

In the fig 1 AUC (Area Under the Curve) implies a solitary measurement addressing the general exhibition of a twofold gathering model considering the region under its ROC bend. ROC Curve (Receiver Operating Characteristic Curve) means a graphical plot illustrating the compromise between Evident Positive Rate and False Positive Rate at various classification thresholds.

From Fig 1 we can say that the auc score is 0.99

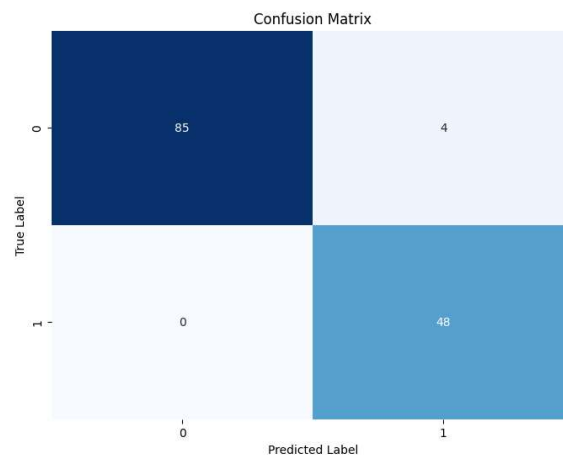


Fig 2. Confusion matrix for RBF - DGSCV Kernel

A confusion matrix is a matrix that sums up the execution of an AI model on a lot of test information. True positives (TP): occur when the model precisely predicts a positive data point.

True negatives (TN): occur when the model accurately predicts a negative data point.

False positives (FP): occur when the model predicts a positive data point incorrectly.

False negatives (FN): occur when the model mispredicts a negative data point.

From the fig 2, we can say $TN = 48$, $TP = 85$, $FP = 4$, $FN=0$

Linear SVM with DASKGRIDSEARCHCV

By utilizing this type of kernel we can linearly separate the data, it is the most regularly utilized kernel. It can be characterized as $(k(x, y) = x^T y + c_0)$ where x and y are the input vectors

DaskGridSearchcv evaluated various feature subsets and SVM configurations to recognize the optimal combination that maximized model performance. Cross-validation will partition our dataset into 10 subsets for thorough evaluation using 10-fold cross-validation.

TABLE III: CLASSIFICATION REPORT

Class	Precision	Re-call	F1-scor-e	Supp-ort	Confusion matrix	Parameter used
2-Benig-n	1.00	1.00	1.00	82	[[82, 0], [0,55]]	C: 0.1, gamma: 0.5,
4-Malig-nant	1.00	1.00	1.00	55		
Accuracy			1.00	137		
macro avg	1.00	1.00	1.00	137		
weighted avg	1.00	1.00	1.00	137		

Table III represents the Classifications Report of Linear-DGSCV kernel based on Confusion Metrics

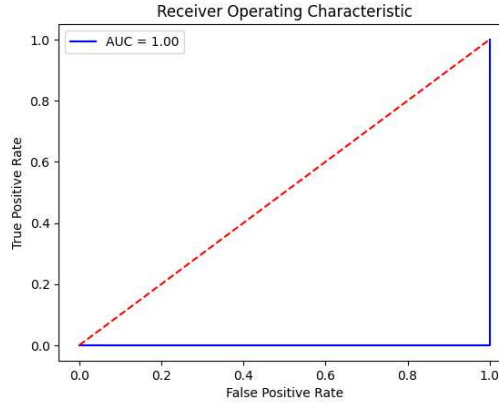


Fig 3. ROC Curve for Linear SVM with DGSCV Kernel

From the fig 3 we can say that linear kernel is performing better as the auc is 1.00

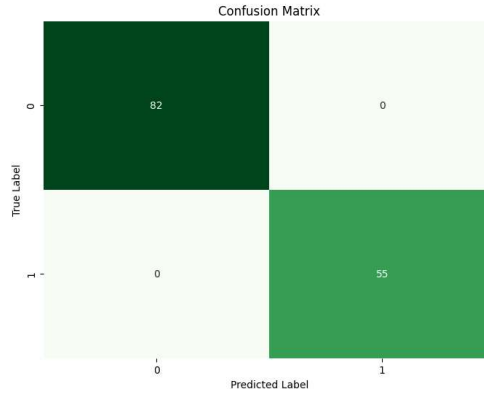


Fig 4. Confusion matrix for linear - DGSCV Kernel

From the fig 4, we can say $TN=55, TP=82, FP=0, FN=0$

D. Polynomial Kernel with DASKGRIDSEARCHCV

The polynomial kernel is defined as $k(x, y) = (\gamma x^T y + c_0)^d$ for the input vectors x and y , d is the degree of the kernel. To find the best C and γ values DaskGridSearchcv is used.

TABLE IV: CLASSIFICATION REPORT

Class	Precision	Re-call	F1-scor-e	Supp-ort	Confusion matrix	Parameter used
2-Benig-n	0.99	1.00	0.99	82	[[82, 0], [1,54]]	C: 0.1, gamma: 0.5,
4-Malig-nant	1.00	0.98	0.99	55		
Accuracy			0.99	137		
macro avg	0.99	0.99	0.99	137		
weighted avg	0.99	0.99	0.99	137		

Table V represents the Classifications Report of sigmoid-DGSCV kernel based on Confusion Metrics

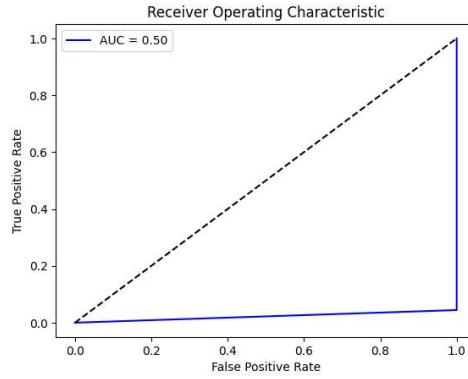


Fig 7. ROC Curve for Sigmoid SVM-DGSCV Kernel

From Fig 7 we can say that the auc score is 0.50

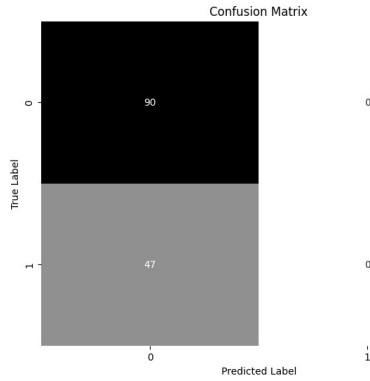


Fig 8. Confusion matrix for Sigmoid -DGSCV Kernel

From the fig 8, we can say $TN=0, TP=90, FP=0, FN=47$

IV. COMPARISON AND RESULTS

The performance study of all the SVM kernels can be done through using the evaluation metrics that is Precision, Recall, Sensitivity, F1-score and Area under the ROC curve can be calculated using the equation (1) to (5) which is shown below:

- $Recall = TP / (TP + FN)$
- $Precision = TP / (TP + FP)$
- $Sensitivity = TP / (TP + FN)$
- $F1\ score = 2 * ((precision * recall) / (precision + recall))$
- $accuracy = (TP + TN) / (TP + TN + FP + FN)$

Where TP = True positive, FP = False positive, TN = True Negative, FN = False Negative

TABLE VI: COMPARISON RESULTS OF DIFFERENT TYPES OF KERNEL

SL.NO	SVM KERNELS	PRECISION	RECALL	F1-SCORE	ACCURACY	AUC	CONSFUSION MATRIX
1	RBF-DGSCV	0.97	0.97	0.97	0.97	0.99	[[78,4] [0.55]]
2	LINEAR-DGSCV	1.00	1.00	1.00	1.00	1.00	[[82,0] [0.55]]
3	POLYNOMIAL-DGSCV	0.99	0.99	0.99	0.99	0.99	[[82,0] [1.55]]
4	SIGMOID-DGSCV	0.43	0.66	0.52	0.52	0.66	[[90,0] [47,0]]

From Table VI, we can analyze the performance of each type of kernel with the optimal values of c and gamma which are obtained by the use of the DaskGridSearchcv. AUC is a better tool than measures such as accuracy and visualizes the performance. The higher the value of AUC the better the classifier performance. Recall is the

ability of the model to identify all relevant instances Linear-DGSCV has the best f1-score, recall, AUC score, and accuracy. Sigmoid-DGSCV had the lowest accuracy, auc score, f1-score, and recall.

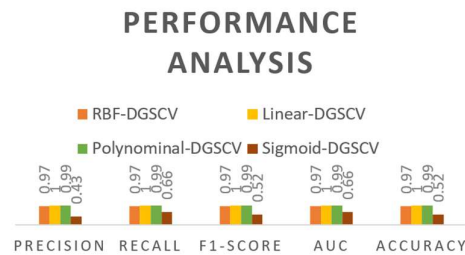


Fig 9 represents the performance analysis of each kernel type

V. CONCLUSION

In our analysis, we evaluated the performance of Support Vector Machines (SVM) with various kernel functions. While different kernels of SVM have demonstrated their considerable efficacy and accuracy, there remains an opportunity to explore its performance when integrated with other methods. From this study, we can say that there is still a way for ML techniques to improve the ability of classification and prediction accuracy.

However, in the field of medical science accuracy plays a vital role and since the model varies for different datasets, the models that are difficult to comprehend must be examined and assessed thoroughly before using this model regularly. The performance of SVM with different kernel functions was assessed.

ACKNOWLEDGMENT

The authors would like to thank School of Computer Science and Engineering , RV, University, Bangalore, Karnataka, India.

REFERENCES

- [1] Farjana Akter Tania and Pintu Chandra Shill, "An Improved Support Vector Machine for Disease Diagnosis", 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1-7, 18-20 December 2019.
- [2] Rong-Ho Lin, Benjamin Kofi Kujabi, Chun-Ling Chuang, Ching-Shun Lin, and Chun-Jen Chiu, "Application of Deep Learning to Construct Breast Cancer Diagnosis Model", Applied Sciences, vol. 12, no. 4, p. 1957, 2022.
- [3] Ebrahim Edriss Ebrahim Ali and Wu Zhi Feng, "Breast Cancer Classification using Support Vector Machine and Neural Network", International Journal of Science and Research (IJSR), vol. 5, no. 3, pp. 1-6, March 2016.
- [4] Tao Xie, Jun Yao, and Zhiwei Zhou, "DA-Based Parameter Optimization of Combined Kernel Support Vector Machine for Cancer Diagnosis", Article, Received: 31 March 2019; Accepted: 30 April 2019; Published: 6 May 2019.
- [5] Pramit Brata Chanda and Subir Kumar Sarkar, "Detection And Classification Technique Of Breast Cancer Using Multi Kernel SVM Classifier Approach", Article.
- [6] Bouchra El Ouassif, Ali Idri, and Mohamed Hosni, "Homogeneous Ensemble based Support Vector Machine in Breast Cancer Diagnosis", Article.
- [7] Elsayed Badr, Sultan Almotairi, Mustafa AbdulSalam, and Hagar Ahmed, "New Sequential and Parallel Support Vector Machine with Grey Wolf Optimizer for Breast Cancer Diagnosis", Article, Received 8 December 2020, revised 31 May 2021, accepted 9 July 2021.
- [8] Garima Verma and Hemraj Verma, "Predicting Breast Cancer using Linear Kernel Support Vector Machine", Article.
- [9] Tina Elizabeth Mathew "A Comparative Study of the Performance of Different Support Vector Machine Kernels in Breast Cancer Diagnosis", Article.
- [10] Xueyun Ye, Zhixiong Zhang, and Yu Jiang, "Prediction of Breast Cancer of Women Based on Support Vector Machines", International Conference/Journal Name (if available), Article.
- [11] Jiaji Wang, Muhammad Attique Khan, Shuihua Wang, and Yudong Zhang, "SNSVM: SqueezeNet-Guided SVM for Breast Cancer Diagnosis", Computers, Materials & Continua, vol. 78, no. 2, pp. 3445-3463, August 2023, DOI: 10.32604/cmc.2023.041191.
- [12] Prity Vijay and Bright Keswani, "Support Vector Machine (SVM) Kernels based approach for detection of Breast Cancer", HEB Cass