International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Effectiveclassificationframework For Breast Tumorsusing Optimized Multi-Kernel Svm With Controlled Skewness

Harikumar Rajaguru [1], Sannasi Chakravarthy S R [2]

[1] *Department Of Electronics And Communication Engineering, Bannari Amman Institute Of Technology, Sathyamangalam – 638401, India.*
[2] *Department Of Electronics And Communication Engineering, Bannari Amman Institute Of Technology, Sathyamangalam – 638401, India.*

harikumarrajaguru@Gmail.Com[1],elektroniqz@Gmail.Com[2]

Abstract: The World İs So Advanced And Sophisticated Enough Nowadays. But, Cancer Still Remains As A Deadly Disease İn Many Parts Of The World For All Living Organisms. The İncidence Rate Of Cancer Among Humans Globally İs İncreasing Steadily Day By Day.Among All Forms Of Tumors, Breast Cancer İs A Type Of İllness That Plays A Major Role İn Disturbingthesurvivalrate Of Humans Globally.Thus, There İs A Need To Predict Breast Cancer İn İts Earlier Stage. Thus, The Work İn This Paperis Aided To Design A Robust Classification Model That İnvolves The Use Of A Randomized-Parameter Optimizedmulti-Kernel Support Vector Machine (RPOMK-SVM) Classifier. Before The Stage Of Classification, The Paper Analyzes The Nature Of İnput İn Order To Obtain Promising Results. The Skewness Of The Feature Attributes İs Controlled And Reformed Using Box-Cox Transform.For This Analysis And Evaluation, The Paperemploysthe Breast Cancer Wisconsin (Diagnostic) Database,Which İs A Standard Public Dataset. The Final Results Are Then Compared Against The Existing Algorithms.

Keywords: Breast Cancer, Support Vector Machine, Malignant, Skewness, Benign, Gaussian, Wdbc.

1. INTRODUCTİON

Cancer, Otherwiseknown As Malignancy, Refers Tothe Abruptcell Growth İn The Human Body. In General, There Are More Than A Hundreddifferent Cancer Types İdentified, Some Of Them Arelung Cancer, Breast Cancer, Colon Cancer, Prostate Cancer, Skin Cancer, And Lymphoma [1]. The Symptoms May Get Vary Based On The Cancer Type. In This, Breast Tumor İs The One Type Of Cancer That Leads Tomore Mortality Among Women. Besides, Breast Cancerremainsin Second Position Among All Types Of Cancer, Next To Lung Cancer [2]. Cancers Could Be Namedand Diagnosed Based On The Type Of Cell İt Originates.In This Way, Breast Cancer Starts From The Cells Of The Breast Of The Human Body. In Addition, The Occurrence Frequency Of Breast Cancer İs More For Women Rather Than Males [3].

As Discussed Above, Breast Cancer İs The One That Heavily Affects The Survival Rate Of Women. Thus, It Is Necessary To Identify The Breast Tumor In An Earlier Way. This Will

Increase The Life Span And Will Reduce The Mortality Rate Of Breast Cancer Definitely [4].Many Researchers Are Working Towards This To Support Human Lives.In Breast Cancer, The Major Symptom Is The Lump Formation In The Cells Of The Breast [5].During The İnitial Stage Of Breast Cancer, The Primary Region Of Ducts And Lobules Are Affected. These Symptoms Are Very Hard To Feel And So The Woman Affected By Breast Cancer Cannot Able To Realize Or Recognize On İts Own [6]. The Paper Utilizes The Standard, As Well As Publicly Available Breast Cancer Wisconsin (Diagnosis) (WDBC) Database. Herein, The Breast Characteristics Were Examinedand Abstracted Through A Fine Needleprocedure At The Time Of Biopsy.Figure 1 Portrays The Workflow Followed For The Aim Of Effective Classification Of Breast Cancer.

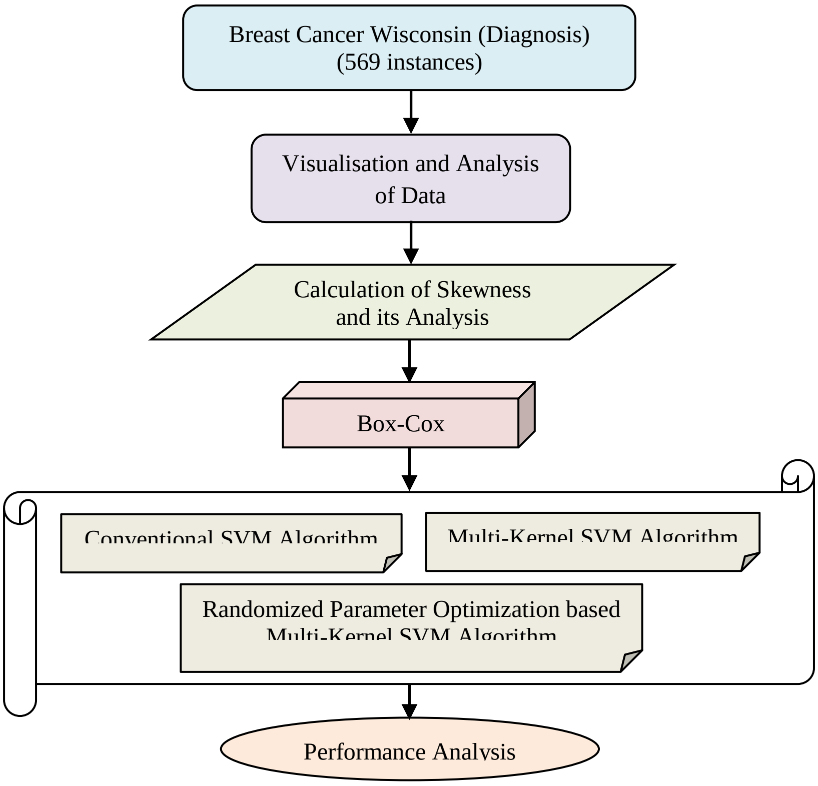

Fig. 1.The Work-Flow Of The Paper

As Portrayed İn The Above Figure 1, The Work Makes Use Of Wdbcdata Corpuscomprises Of 569 İnstances For İts Evaluation. The İnput Features Of The WDBC Dataset Are Visualized For Better Analysis. From This, The Skewness İs Calculated And Plotted For The İnput Feature Vectors. Then The Box-Cox Transform İs Used For The Correction Of Controlled Skewness Of The Dataset. The Classification Process İs Next Carried Out And Thereby The Results Are Compared And Analyzed.

2 . PREPROCESSİNG OF DATASET

2.1 Data Visualizationand Analysis

As Portrayed İn Figure 1, The Breast Cancer Wisconsin (Diagnosis) Data-Corpus İs Employed For The Breast Cancer Classification. This Publiclyavailable Data İs Widely Known İn The Name Of WDBC Dataset. And İtcomprises Of Total İnstances Of 569 Readings Together With Predictive (30) And Numerical Attributes [7]. The Sample

Attributes That Are İncluded İn Thewdbc Areradius, Texture, Area, Compactness, Perimeter, Symmetry, Smoothness, Concavity, Concave Points, And Fractal Dimension. Herein, During The Biopsy Testconducted With A Fine-Needle Procedure, The Obtained Local Variation İn Radius Length İs Said To Be Smoothness Attribute. The Compactness İs Another Attribute That İs Computed As,

$$Compactness = \frac{perimeter^2}{area-1} \quad (1)$$

The Next Concave Attributeillustratesthe Severityof The Calculation Of Concave Portions On İts Own Contour. Also, Themeasure Ofthese Concave Detailsthat Are Attained Onits Contour İs Saidto Bea Concave Point Features. In This Way, The WDBC Dataset Was Publicly İntroducedas Detailed İn [7]. No Missing Values Found İn Thedata Set Having 569 İnstances Whichmadethis WDBC Data Setmore Popular Among The Cancer Researchers. The Dataset - WDBC Consists Of Two Output Or Severity Targets That Are Denoted As Benign (B) Target And Malignant (M) Target Class. Hence, Theworkaims Toperform Binary Classification After Analyzing The Dataset. The Graphical Distribution Of Output Classes (Severity) Of The WDBC Dataset İs Plotted İn Figure 2.

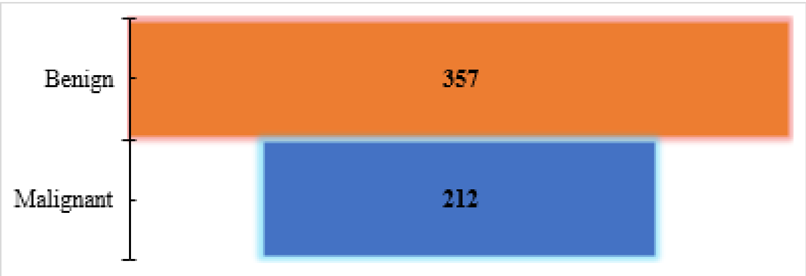

Fig. 2.The Distribution Of Output Severity Present İn WDBC Dataset

Figure 3 Portrays The Pair-Plot Visualization Of WDBC Datasetpertaining Tosample Attributes(Mean Radius, Mean Texture, Mean Perimeter, Mean Area, Mean Smoothness)Plotted Against The Output Classes. As Portrayedin Figure 3, The Scatter Plot Reveals That The İnput WDBC Features Are Highly Non-Linear. Also, The Diagonal Plot İn Figure 3 Represents The Kernel Density Estimates (KDE) Of The İnput WDBC Dataset That Reveals The Features Are Skewed At A Different Range. Thus, The Skewness İs Needed To Be Checked Before The Process Of Classification.

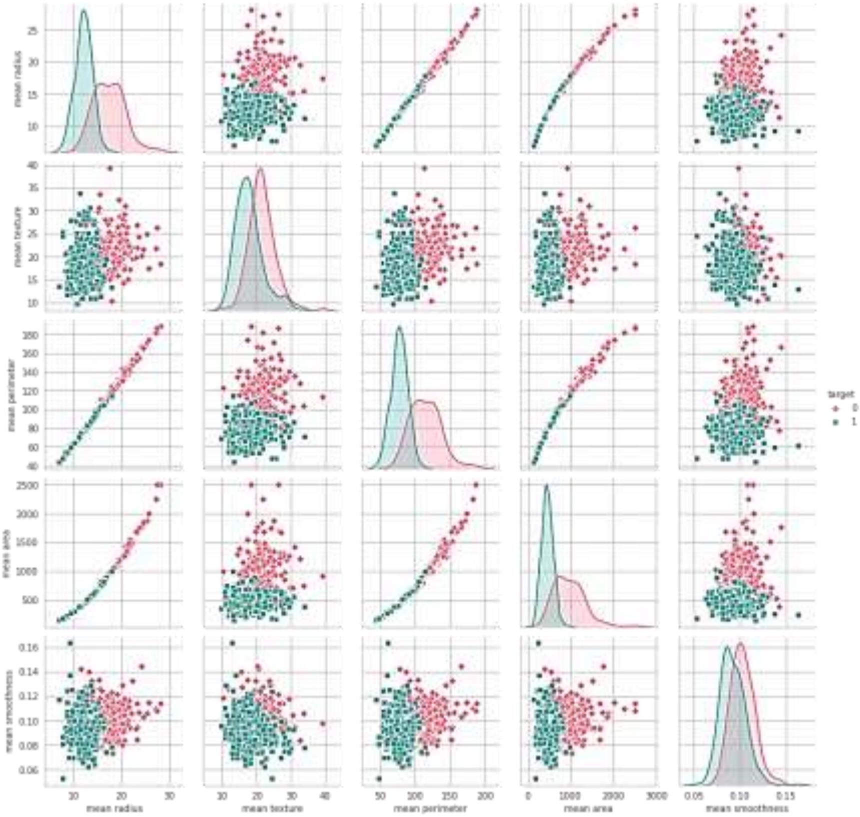

Fig. 3.Visualization Of Sample WDBC Attributes

2.2 Calculation Of Skewness And İts Analysis

In General, Skewness İs A Term That Represents How The İnput Data Distracts From The Normal Distribution. In A Normal Distribution, The Samplesare Graphically Denotedas A Bell-Shaped One, Where The Average (Mean) And The Maximum Value İn The Dataset (Mode) Are Equal [6].After Plotting The İnput Samples, İf The Right-Side Of The Curve İs Found To Be Larger Than İts Left Tail, Then İt Can Be Noted That The İnput Data Has Positive Skewness İ.E. $Mean > Median > Mode$. In Other Cases, That İs, İf The Left-Side Of The Curve İs Found To Be Larger Than İts Right Tail, Then İt Can Be Noted That The İnput Data Has Negative Skewness İ.E. $Mode > Median > Mean$.If Both The Left And Right Tails Are Normally Distributed, Then The Data İs Said To Have Symmetric Skewness. The Skewness İs Calculated For All The Attributes Of WDBC Dataset And İs Tabulated İn Table 1.

Table 1. Obtained Skewness Value For Different Attributes Of WDBC Dataset

S No	Attributes	Skewness Value	S No	Attributes	Skewness Value
1	Mean Radius	0.942380	16	Compactness Error	1.902221
2	Mean Texture	0.650450	17	Concavity Error	5.110463
3	Mean Perimeter	0.990650	18	Concave Points Error	1.444678
4	Mean Area	1.645732	19	Symmetry Error	2.195133
5	Mean Smoothness	0.456324	20	Fractal Dimension Error	3.923969
6	Mean Compactness	1.190123	21	Worst Radius	1.103115
7	Mean Concavity	1.401180	22	Worst Texture	0.498321

8	Mean Concave Points	1.171180	23	Worst Perimeter	1.128164
9	Mean Symmetry	0.725609	24	Worst Area	1.859373
10	Mean Fractal Dimension	1.304489	25	Worst Smoothness	0.415426
11	Radius Error	3.088612	26	Worst Compactness	1.473555
12	Texture Error	1.646444	27	Worst Concavity	1.150237
13	Perimeter Error	3.443615	28	Worst Concave Points	0.492616
14	Area Error	5.447186	29	Worst Symmetry	1.433928
15	Smoothness Error	2.314450	30	Worst Fractal Dimension	1.662579

The Value Of Skewness As Given İn Table 1 Tells Us How The Attributes Are Distorted From The Normal Distribution. The Highest Value Of Skewness İs Found For The Attribute 'Area Error' İ.E. 5.447186 And The Least Value Of Skewness İs Obtained For The Attribute 'Worst Smoothness' İ.E. 0.415426. In All The Machine-Learning Algorithms, Any Value Of Skewness İs Generally Undesirable, Since This Canresults İnan Excessively Large Variance İn The Estimates.Thus, For Every Classification Problem, İt İs Necessary To Decrease Skewness Value To Make The Data As Closer To A Normal Distribution Curve By Employingany Transformation Method. The Paper Adopts A Box-Cox Transform To Reduce The Skewness İn The İnput Data.

2.3 Box-Cox Transform

In General, Box-Cox Transform İs Usedfor The Transformation Of Non-Normal (Skewed) Dependent Values İn The İnput Data İnto A Normal (Bell-Shaped) Shape.This Will İnfluence Or Enhance The Performance Of Any Classification Framework. The Advantage Ofbox-Cox Transform İs That İt İs A Configurable Data Transformationprocedurethat Also Supports The Square-Root And Log-Transformation Methods [8]. Moreover, This Transform Can Be Configurableforautomatic Evaluation Of A Suite Of Mathematical Transforms And Thus Provides A Best-Fit For The İnput Data.The Box-Cox Transform Can Be Defined As [8],

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ \mathrm{Log}(y_i) & if\ \lambda = 0 \end{cases} \quad (2)$$

Where The Transform Reduces The Skewness Of İnput Data Based On The Box-Cox Parameter (λ).Figures 4 And 5 Show The Skewness Of The 'Area Error' Attribute Before And After Applying The Box-Cox Transform.

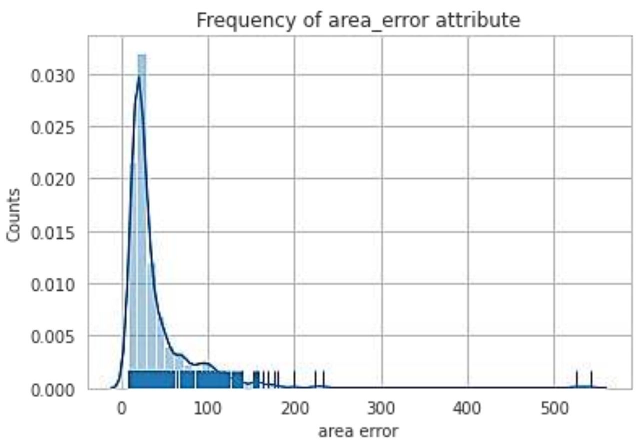

Fig. 4.Skewness Plot Of 'Area Error' Attribute Of WDBC Dataset Before Box-Cox Transform

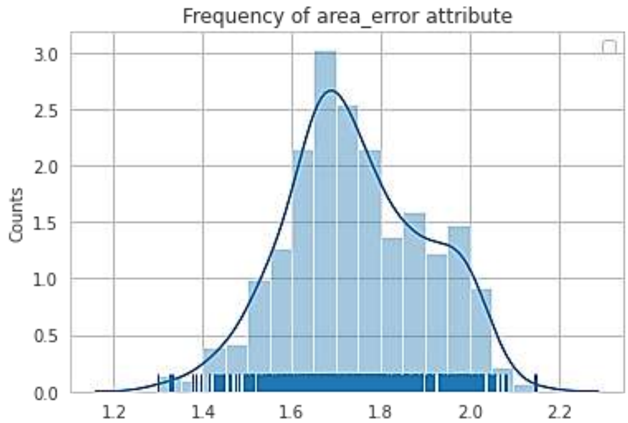

Fig. 5.Skewness Plot Of 'Area Error' Attribute Of WDBC Dataset After Box-Cox Transform

As From Table 1, The Attribute Having A Larger Skewness İs The 'Area Error' Attribute. The Skewness Plot Of The 'Area Error' Plot İs Shown İn Figure 4, Where The Skewness İs Not Symmetry. By Using Box-Cox Transform, The Skewness İs Reduced From 5.447186 To 0.058528, And Thus İt Resembles The Curve Of Normal Distribution As Shown İn Figure 5. In This Way, The Skewness Of All The Attributes İs Reduced To Obtain Better Classification Performance.

3. CLASSİFİCATİON ALGORİTHMS

3.1 Multi-Kernelsupport Vector Machine (MK-SVM) Classifier

The Support Vector Machine (SVM) Model İs A Type Of Supervised Learning Algorithm, Used Popularly For The Task Of Classification Problems [9]. The SVM Algorithm Follows A Discriminative Classification Strategythat İs Otherwise Done By Using Separating Hyperplanes. This İmplies That The SVM Classifier Provides An Optimal Hyperplane Used For Categorizing The Newer İnput Samples. By Using These Hyperplanes, The SVM Can Classify Well Enough For Linear İnputs [10]. But For The Application Of Non-Linear Classification Problems, The SVM Model Makes Use Of Kernel Functions For Categorizing The Newer İnput Samples. The MK-SVM Classifier İncludes The Combination Of Two Different Kernels And İs Given By,

$$MK - SVM = \frac{1}{2}[(lin(a,b) + rbf(a,b)] \quad (3)$$

Where The lin Represents The Use Of Linear Kernel And The rbf Represents The Use Of Radial Basis Function İn The Multi-Kernel SVM Algorithm.

3.2 Randomized-Parameter Optimized MK-SVM (RPOMK-SVM) Classifier

For Solving Any Type Of Problem, The Exhaustive Grid Search İs Widely Used For Optimizing The Parameters Of SVM. The Parameter Tuning Using An Exhaustive Grid Search Technique İncreases The Performance But Decrease The Overall Efficiency Of The System. The Paper Makes Use Of A Randomized Search-Over Parameter Method [11], Where Every Setting İsobtainedas A Distribution Againstall The Possible Value Of Parameters. In This Technique, How Parameters Are Needed To Be Sampled İs Carried Out Using A Dictionary As Followed İn All Optimization Techniques. For Every Parameter Values, Either A List Of Discrete Choices Or A Distribution Over Possible Values Are Versatile To Optimize The SVM Parameters [12].

Let's Discussabout The Resultsofusing The RPOMK-Svmclassifier Along With Box-Cox

Transform To Obtaina Better Classification Of Breast Cancer İn The Next Section.

4. RESULTS AND DİSCUSSİON

The WDBC Data Setis Randomly Divided İnto Training (80%) And Testing (20%) Sets. The Above-Mentioned Classification Strategieswill Be İmplemented And Their Respectiveanalysis İs Madeusingthese Training Set And Testing Set. All The Works Discussed Are Done Using Intel (Vpro) Core-İ5 Processor, 4 TB Hard-Drive Memory, 8 GB RAM, İnstalled With Python 3.6 İn Windows 7 Operating System.After Obtaining The Classification Results, They Are Analyzed Using Standard Evaluation Metrics - Accuracy (Acc), Sensitivity (Se), Precision (Pr), Specificity (Sp), F1 Score, And Matthews Correlation Coefficient (MCC). The Above Evaluationmetricswill Be Derived From The Concept Of Confusion Matrix,Which İs Used Popularly For Binary And Multiclass Classification Tasks.

Table 2. Obtained Confusion Matrix For Different Algorithms

Classificationframework	Confusion Matrix			
	TP	FN	FP	TN
SVM With Box-Cox Transform	168	44	49	308
MK-SVM With Box-Cox Transform	186	26	34	323
RPOMK-SVM With Box-Cox Transform	201	11	16	341

Table 2shown Above Shows The Values Ofthe Confusion Matrix Obtained Fordistinct Classification Frameworks That Are Adopted For Breast Cancer Classification Tasks. As İn Table 2, The More Number Ofpseudopredictions İs Attained For The Conventional Svmmodel With Box-Cox Transform And The More Number Of Correct Predictions İs Attained For The RPOMK-SVM With Box-Cox Transform.The Confusion Matrix Values As Given İn Table 2 Are Obtained And Assessedfor Both Severities - Benign And Malignant Outputs.

Table-3.Comparativeanalysis Of Different Classificationframeworks

Classification Framework	Performance Measures (%)					
	Se	Sp	Acc	Pr	F1 Score	MCC
SVM With Box-Cox Transform	79.25	86.27	83.66	77.42	78.32	65.22
MK-SVM With Box-Cox Transform	87.74	90.48	89.46	84.55	86.11	77.65
RPOMK-SVM With Box-Cox Transform	94.81	95.52	95.25	92.63	93.71	89.91

Table 3portrays The Comparison Of Performance Analysis Of Different Classification Frameworks Using The Confusion Matrix Elements As Shown İn Table 2. From Table 3, Six Distinctevaluationmeasures Are Adopted For The Performance Analysis Of The Different Classification Frameworksfor Our Classification Problem.

The Performance Analysis Of The Classification Frameworks İs Calculated And Graphically Portrayedin Figure 6. As Shown İn Table 3, The Classification Accuracy İs Obtainedhigh For The RPOMK-SVM With The Box-Cox Transformframework.In This, The Highervalue Of 95.25% Classification Accuracy İs Yielded For This RPOMK-SVM Algorithm Together With Box-Cox Transform. Despite The Fact Thatthe Conventional Svmuses The Trick Of Simple Hyperplanes, İt Provides A Classification Accuracy Of 83.66%. This İs Possible Because Of The Use Of Box-Cox Transform Employed For The Reduction Of Skewness Of

İnput Data. The MK-SVM Algorithm With Box-Cox Transform Provides A Better Classification Of 89.46% Of Accuracy. And These Classification Performances Arecompared And Analyzed İn Table 3 Are Graphically Shown İn Figure 6.

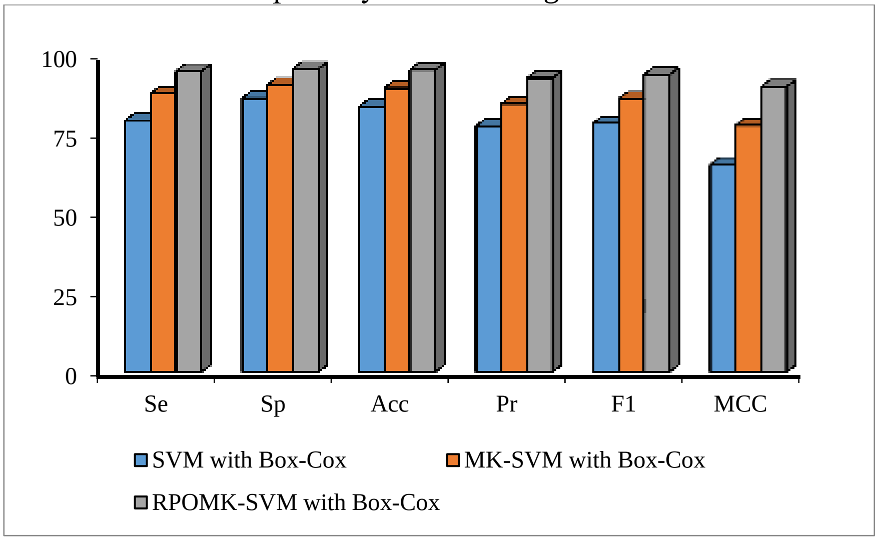

Fig. 6.Graphical Comparison Of Different Classification Frameworks

As Portrayed İntable 3 And Figure 6, The Performance Of The RPOMK-SVM Algorithm Together With Box-Cox Transform İs Significantly Very High Than The Conventional And Multi-Kernel SVM Techniques. As Portrayed İn Figure 6, Thisclassification Framework Provides Thesuperior Performance Of Classification Thatwell Differentiate The Benign (B) And Malignant (M)İnputs. That İs, Thatthe RPOMK-SVM Algorithm Together With Box-Cox Transformprovides Usthesuperior Values Of Sensitivity, F1 Score, Specificity, Precision, MCC, And Accuracy As Compared With Others.

5. CONCLUSİON

The Design Of Computer-Aided As Well As Automatic Classification Approach İs Discussed İn This Paper For Classifying Thebenign And Malignant Type Of Severities Pertains Tobreast Cancer. The Performance Analysis Of The RPOMK-SVM Algorithm Together With The Box-Cox Transform İs Compared And Analyzed With İts Variants Such As Conventional And Multi-Kernel Support Vector Machine Algorithmsfor The Purpose Ofseverity Classification Of Benign And Malignant Ones. Herein, The Dataset Used İn WDBC Data, And İt İs Found That Some Of The Attributes Of The Dataset Are Highly Skewed Up To The Value Of 5.447186. This Skewness Will Affect The Performance Of Any Classification Models And Thus The Box-Cox Transform İs Used To Remove The Skewness Of İnput Before Proceeding To The Stage Of Classification. Hence, Therandomized-Parameter Optimized Multi-Kernel Support Vector Machine Algorithm Along With The Box-Cox Transform Yields Superior Performance Over Other Techniques. The Future Work Will Be The Use Of Clinical Data For The Designed Classification Framework And For Classifying Other Severities.

6. REFERENCES

[1] Desantis, C.E., Ma, J., Gaudet, M.M., Newman, L.A., Miller, K.D., Goding Sauer, A., Jemal, A. And Siegel, R.L., 2019. Breast Cancer Statistics, 2019. CA: A Cancer Journal For Clinicians, 69(6), Pp.438-451.

[2] Northouse, L.L., Templin, T., Mood, D. And Oberst, M., 1998. Couples' Adjustment To Breast Cancer And Benign Breast Disease: A Longitudinal Analysis. Psycho-Oncology: Journal Of The Psychological, Social And Behavioral Dimensions Of Cancer, 7(1), Pp.37-48.

[3] Evans, A., Whelehan, P., Thomson, K., Mclean, D., Brauer, K., Purdie, C., Jordan, L., Baker, L. And Thompson, A., 2010. Quantitative Shear Wave Ultrasound Elastography: İnitial Experience İn Solid Breast Masses. Breast Cancer Research, 12(6), Pp104.

[4] Chaurasia, V., Pal, S. And Tiwari, B.B., 2018. Prediction Of Benign And Malignant Breast Cancer Using Data Mining Techniques. Journal Of Algorithms & Computational Technology, 12(2), Pp.119-126.

[5] Abirami, C., Harikumar, R. And Chakravarthy, S.S., 2016, March. Performance Analysis And Detection Of Micro Calcification İn Digital Mammograms Using Wavelet Features. In 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet) (Pp. 2327-2331). IEEE.

[6] Sannasi Chakravarthy, S.R. And Rajaguru, H., 2020. Detection And Classification Of Microcalcification From Digital Mammograms With Firefly Algorithm, Extreme Learning Machine And Non-Linear Regression Models: A Comparison. International Journal Of Imaging Systems And Technology, 30(1), Pp.126-146.

[7] Blake, C.L. And Merz, C.J., 1998. UCI Repository Of Machine Learning Databases. University Of California, Irvine, Dept. Of Information And Computer Sciences.

[8] Zhang, Y., Xiong, R., He, H. And Pecht, M.G., 2018. Lithium-İon Battery Remaining Useful Life Prediction With Box–Cox Transformation And Monte Carlo Simulation. IEEE Transactions On Industrial Electronics, 66(2), Pp.1585-1597.

[9] Dadi, H.S. And Pillutla, G.M., 2016. Improved Face Recognition Rate Using HOG Features And SVM Classifier. IOSR Journal Of Electronics And Communication Engineering, 11(4), Pp.34-44.

[10] Padierna, L.C., Carpio, M., Rojas-Domínguez, A., Puga, H. And Fraire, H., 2018. A Novel Formulation Of Orthogonal Polynomial Kernel Functions For SVM Classifiers: The Gegenbauer Family. Pattern Recognition, 84, Pp.211-225..

[11] Demidova, L., Nikulchev, E. And Sokolova, Y., 2016. The Svm Classifier Based On The Modified Particle Swarm Optimization. Arxiv Preprint Arxiv:1603.08296.

[12] Bamakan, S.M.H., Wang, H., Yingjie, T. And Shi, Y., 2016. An Effective İntrusion Detection Framework Based On MCLP/SVM Optimized By Time-Varying Chaos Particle Swarm Optimization. Neurocomputing, 199, Pp.90-102.