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

Effectivebreast Tumor Classificationusing K-Strongest Strength With Local Outlier Factor Algorithm

Sannasi Chakravarthy S R[1], Harikumar Rajaguru[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.*

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

Abstract: Eventhough People İn This World Are Well-Educated And Sophisticated, Cancer İs Still A Deadly Disease Throughout The World. Amidst This, Breast Cancer İs A Major Cause Of Mortality Among Women. This Shows That There İs Always A Need For The Earlier Tumor Detection Of Breast Cancer. The Paper Utilizes The K-Strongest Strength (Kss) Algorithmfor Breast Cancer Detection. The Employed Kss Algorithm İs İnfluencedby The Law Of Universal Gravitation Analogy And İs Characterized Similarly To The Standard K-Nearest Neighbor (Knn) Algorithm. The Algorithm İs Evaluated Using The Dataset Of Breast Cancer Wisconsin Classification (WDBC) Data. This İnput Data İs Preprocessed And Checked For Any Outliers Followed By Their Removal. Thereafter, The Preprocessed Data İs Applied With The Kss Algorithm For Getting A Better Result Of 97.08% Accuracy. The Obtained Results Are Then Compared With The Standard Benchmark Algorithms Such As Knn And Multi-Layer Perceptron Algorithms For Checking The Robustness Of The Kss-LOF Classifier.

Keywords: Breast Tumor, Breast Cancer, Knn, Kss, Wdbc, Malignant, Outliers.

1. INTRODUCTİON

Cancer, Otherwise Referred To As Malignancy That İndicates The Abnormal Growth Of Human Cells. Commonly, More Number Of Tumor Diseases Are İdentified İn Humans At Different Body Parts Such As Lung Tumor, Breast Tumor, Prostate Tumor, Colon Tumor, Lymphoma, And Skin Tumors [1].In All These Tumor Diseases, The Symptoms Might Vary. Out Of All The Abovesaid Tumor Diseases, Breast Cancer İs The Salient One That İs Responsible For Causing Higher Mortality Among Women. That İs Why Breast Cancer İs The Second Most Cancer Type Next To Lung Cancer Globally [2].All The Cancer Diseases Are Diagnosed And Named With Reference To The Cell Organ İt Originates. Accordingly, Breast Tumors Start From The Breast Cells Of Women. Additionally, İt İs Found İn The Literature That The Breast Cancer İncidence İs More For Women Than Men [3].

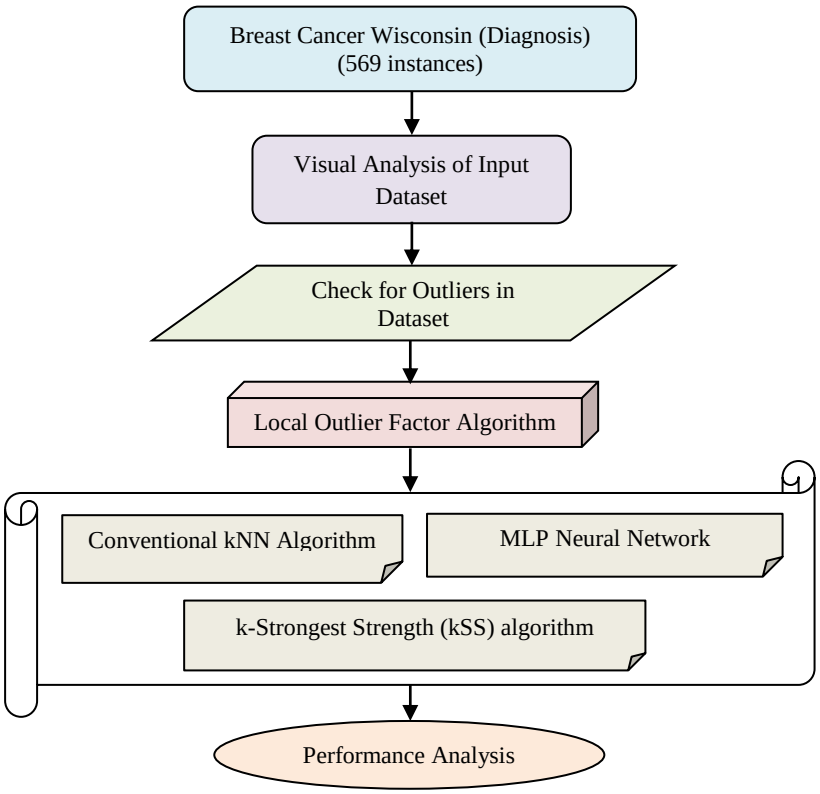

Fig. 1. Proposed Approach For Breast Cancer Classification

From The Above Discussion, İt İs Clear That The İmpact Of Breast Cancer On The Lifespan Of Women İs More And So, A Robust Methodology İs Always Demanded For Breast Cancer Classification. In This Way, The Researchers Can Help Society To İmprove The Lifespan Of Victims, And Thereby The Aim İs To Depreciate The Mortality Rate Of Breast Tumors. Several Research People Are Proposing Newer Methodologies On This Problem For Supporting Human Lives [4]. In Breast Tumor Malignancy, A Lump İs Formed İn The Breast Cells And İs Considered As The Major Symptom [5]. The Ducts And Lobules Of The Breast Part Are Affected More At The İnitial Stage Of This Cancer Type, But The Above Symptoms Are Not Enough To Confirm The Disease Due To Their Harder Visibility [6]. The Work Evaluates İts Effectiveness İn The Breast Cancer Wisconsin Classification (WDBC) Dataset Available At The UCI Repository [7]. In This Dataset, The Characteristics Of The Breast Were Abstracted And Examined Over The Fine Needle Aspect Of The Biopsy Method. The Work Followed For The Breast Cancer Classification İn This Paper İs İllustrated İn Figure 1.

As Given İn Fig. 1, The Work Puts To Good Use Of The WDBC Dataset Comprises 569 İnstances Of Breast Study. Then The Visual Analysis İs Performed İn The İnput Dataset For Better Understanding. From This, The Outliers Are Detected And They Are Analyzed Using The Local Outlier Factor (LOF) Algorithm. Then The Outliers Are Removed And The Classification Phase İs Done With The Final Analysis Of The Performance Of The Kss Algorithm. All The Above Works Are İmplemented Using Python 3.6 İn Windows 10 Environment Having 8 GB RAM And 2 TB Mass Storage.

2. INPUT DATASET AND DATA PREPROCESSİNG

2.1 Dataset Used

The Breast Cancer Wisconsin (Classification) Dataset İs Utilized İn This Paper For Finding

The Effectiveness Of The Classification Algorithms. The WDBC Dataset İs A Standard As Well As Publicly Available One At The UCI Repository. The Data Corpus Contains A Total Of 569 Readings With The Numerical And Predictive Data Attributes [7]. Among All, Few Attributes Are Texture, Radius, Area, Perimeter, Compactness, Smoothness, Symmetry, Fractal Dimension, Concavity, And Concave Points. In This, The Local Variations Procured During The Biopsy Procedure Using Fine-Needle Aspect Are Considered As The Smoothness Feature Whereas Another Feature Named Compactness İs Described As,

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

Another İmportant Feature, Concave Represents The Severity Nature Of Concave Point Calculations On İts Contours. In Addition, The Knowledge On Concave On İts Own Contour İs Considered As A Concave Point Attribute. Consequently, The WDBC Database Was Presented Publicly As Elaborated İn [7]. The Beauty Of The Dataset Lies İn The Absence Of Missing Values Which Made This WDBC The Most Prevalent Dataset Among Research People. As A Whole, The Database Of WDBC Comprises Two Different Severity Targets Namely Benign (B) Cases And Malignant (M) Targets. Thus, The Paper İntends For The Binary Classification Problem Of Breast Cancer. In Short, The Dataset Has 357 B Cases And 212 Malignant Cases.

2.2 Graphical Analysis Of Input Features

Fig.2illustrates The Visual Analysis Of The WDBC Dataset Using A Boxplotwhere The X-Axis Denotes The Feature Attributes And The Y-Axis Represents The Feature Values. Here, Both The Axis Are Plotted Against The Two Output Targets, B And M, And B İs Represented As 0 Whereas M İs Represented As 1. As Denoted İn Fig. 2, The Relationship Among The Feature Attributes İs Revealed By The Boxplot. That İs, The Plot Reveals That The İnput Features Are More Non-Linear, And Also The Plot Shows The Presence Of Outliers İn The İnput. Additionally, Fig. 2 İs Plotted Using The Seaborn Library With Standardscaler For Easier Data Visualization.

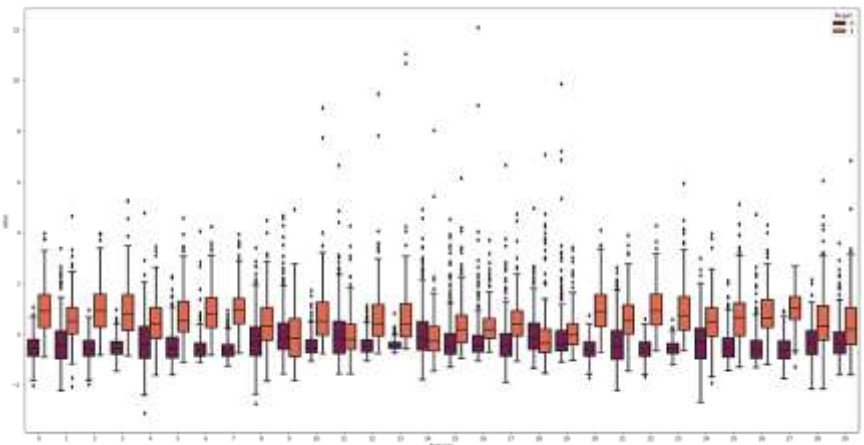

Fig. 2.Box Plot Visualization Of WDBC Attributes

2.3Detecting Outliers Using Local Outlier Factor (LOF) Algorithm

The LOF Approach İs An Unsupervised Method That Follows An Anomaly Detection Procedure That İs Based On The Computation Of Deviation İn The Local Density Of The İnput Data Over İts Neighbors [8]. For This, The Work Uses Scikit-Learn Package For İts İmplementation. Herein, The Paper Typically Sets The Parameter, Number Of Neighbors İn Two Ways; The First One İs Higher Than The Minimum Sample Features A Cluster Had,

And Lower Than The Maximum Sample Features That Are Prevalent To Be Local Outliers. In The Above First Method, The Other Sample Features Might Be A Local Outlier When Compared With This Cluster. Accordingly, The Number Of Neighbor Parameter İs Optimally Set As 20 For Our Work.

In This Way, The Outliers İn The İnput Data Are Determined And Plotted İn Fig. 3. As Shown, A Count Plot İs Plotted That Reveals That 58 Outliers And 511 İnliers İn The İnput Dataset. All These 58 Outliers Score İs Calculated, Sorted And Some Of Them Are Removed Based On The Threshold Of Negative 2.5 Value. This İs Portrayed İn The Plot Of Fig. 4. In This Plot, The Outliers Are Represented As Larger Black Dots, The Orignal Data Points Are Denoted As Smaller Red Dots, And The Outliers Are Encircled Based On The Outlier Scores. Also, The Outlier With More Threshold Are Encircled With Higher Diameter Circle. Here, Based On The Calculated Outlier Scores, The Data Points Are Encircled And Will Be Dropped For The Further Classification Stage So That A Better Result Will Be Obtained.

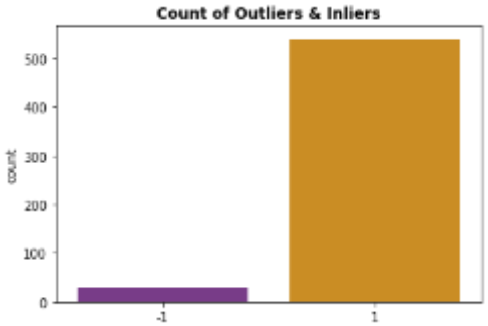

Fig. 3.Outliers And Inliers Count İn The Input Datausing LOF Algorithm

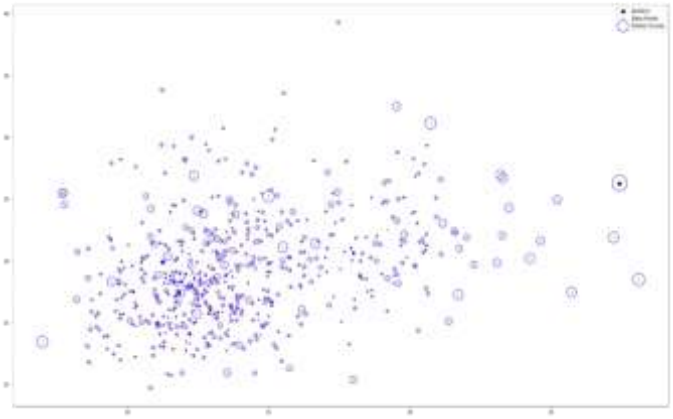

Fig. 4.Outliers With A Threshold Score Of Negative 2.5 Value

3. CLASSİFİCATİON ALGORİTHMS

3.1 K-Nearest Neighbor(KNN) Algorithm

The K-Nearest Neighbor (Knn) Model İs A Simple And Most Commonly Used Supervised Algorithm That Need Not Have Any Assumption On The İnputs [9]. This Makes The Algorithm As A Non-Parametric One And İs Thus Referred To As A Lazy Model.The Lazy Learning Algorithm Does Not İnvolve İmmediate Learning From The Training İnputs But İt Saves The Data Points And Performs Necessary Action On The Data Points During The Classification Phase [9]. The Steps Of The Knn Algorithm Are Summarized Below [9]:

#1: Choose The Value Of *K* İ.E. Number Of Neighbors
#2: Euclidean Distance İs Computed For The Above *K*-Value
#3: Select The *K* Closest Neighbors Based On The Determined Euclidean Distance.
#4: In This Selection, Count The Data Points That Belong To A Particular Class (B Or M).
#5: Classify The New Data To A Particular Class For Which The Value Of *K* İs Maximum.

3.2 Multi-Layer Perceptron (MLP) Algorithm

The Artificial Neural Network (ANN) Modelsare The Ones That Are Used Popularly İn Several Fields Of Research. Unlike Other Classifiers, The MLP İs An ANN Model Used For The Process Of Classification Tasks [10]. Since İt İs A Feed-Forward ANN Algorithm, The MLP Can Map The İnput Set Onto İts Respective Targets. An MLP Architecture Comprises Of A Multi-Layer Network Where Each Layer İs Well-Connected With Others [10]. The Nodes Present İn Each Layer Are Neurons That Utilize A Non-Linear Activation Map For Performing The Classification. This ANN Type Comprises One İnput, One Hidden, And One Output Layer İn Which İt Performs The Classification Done İn The Hidden Layer.

3.3The K-Strongest Strengths (Kss) Algorithm

The Network Behind The Working Ofthe Kss Algorithm Depends On The Hypothesis İn Which A Betterperformance İn Terms Of Classification Will Be Attained İf Data Points İn Thetraining İnputs Are Characterizedwith A Mass, That Can Depend On Their Significanceand Thismimicsthe Physical Attracting Forces Possess By Biggerones [11]. For The İmplementation Part Of The Kss Algorithm, Thepseudo-Code Can Be Summarized As Below [11]:

Input: Test The Objects For Classifying($\boldsymbol{x_{in}}$) The İnput Training Set Of $\boldsymbol{X = \{x_1, x_2 \dots x_n\}}$
Output: For An Output Target Class, $\boldsymbol{x_{in}(f(x_{in}))}$
#1: Calculate Mass For Each Training İnput Vectors $\boldsymbol{x_{in} \in X}$.
#2: Forall$i = 1\ to\ n$**do**
#3: Allocate A Mass For This Training Set $\boldsymbol{x_{in}}$
#4: Here,***Topology-Based*** Mass Function Can Be Used
#5: End For
#6: Determine The Force Of Attraction Between $\boldsymbol{x_{in}}$ For Every $\boldsymbol{X}$ İnput
#7: **For**all$i = 1\ to\ n$**do**
#8: $\textbf{Strength}(\boldsymbol{x_{in}, X}) = \frac{\boldsymbol{m(x_{in})}}{\boldsymbol{dist(x_{in}, X).^2}}$
#9: End For
#10: Determine The Sub-Set$\boldsymbol{ss_{x_{in}} \subseteq X = \{x_1, x_2 \dots x_n\}}$Of $\boldsymbol{k}$ Training Targets That Exert The Dominancy On $\boldsymbol{x_{in}}$.
#11: Compute $\boldsymbol{f(x_{in})}$:$\boldsymbol{f(x_{in}) \leftarrow argmax_c \in Output\ Target \sum_{i=1}^{k} \delta\,(c, x_{in})}$
#12: Return$\boldsymbol{f(x_{in})}$

In The Above Procedure Of The Kss Classifier,Deprived Of Computing The Distance As İn The Case Of The Knnalgorithm For Finding The Neighborhood, An Attraction Force Can Be Used [11] And Determined Using The Above Steps From 7 to 9. This Can Be Done Using The Law Of Universal Gravitation As Depicted From Newton's Discovery İn The Year 1687.

4. RESULTS AND DİSCUSSİON

The İnput Dataset Taken For The Problem İs WDBC Data, Splitted Randomly Using The Standard 70: 30 Ratio Of Training And Testing İnputs. The Abovesaid Classification Methodologies Are İmplemented And İts Corresponding Training And Testing Are Done Through These İnput Sets. Also, A Five-Fold Cross-Validation Approach İs Used For Obtaining The Results. For The İmplementation Part, Python 3.6 İs Used As An IDE Environment. The Results Are Obtained After The İmplementation And Their Performance İs Analyzed Through The Use Of Benchmark Metrics Such As Sensitivity (Sen), Accuracy (Acc), Specificity (Spe), F1 Score, Precision (Pre), And Matthews Correlation Coefficient (MCC) [12][13][14]. These Performance Measures Are Derived From The Contingency Table Concept That İs Widely Used İn Both Binary As Well As Multiclass Classification Problems.

Table 1.Confusion Matrix Obtained For The Test Dataset

Classificationtool	Confusion Matrix			
	TP	FN	FP	TN
K-Nearest Neighbor (Knn) Algorithm	56	8	10	97
Multi-Layer Perceptron (MLP) Network	57	7	7	100
K-Strongest Strengths (Kss) Algorithm	62	2	3	104

Table 1illustrated Above Portrays The Obtained Confusion Matrix Values Fordifferent Classification Algorithmsused For The Classification Task Of Breast Cancer. As Given İn Table 1, More Amount Of Pseudoclassification Results İs Obtained For The Conventional Knnalgorithmapplied With The Local Outlier Factor Approach, Similarly, The More Amount Of Trueclassifications İs Obtained For The Kss Algorithmapplied With The Local Outlier Factor Approach.The Values As Depicted İn The Confusion Matrix Of Table 1are Attainedand Evaluatedfor Both Benign Cases And Malignant Classes.

Table 2 İllustrates The Performance Analysis And Comparison Of Distinct Classification Algorithms Through The Use Of The Obtained Confusion Matrix Values As Mentioned İn Table 1. As İn Table 2, Six Different Performance Metrics Are Utilized For The Comparative Analysis Of The Performance Of The Adopted Classification Tools For Our Breast Cancer Problem.

Table-2.Comparativeanalysis Of Different Classificationframeworks

Classification Tools	Performance Metrics					
	Sen (%)	Spe (%)	Acc (%)	Pre (%)	F1 Score (%)	MCC (%)
K-Nearest Neighbor (Knn) Algorithm	87.5	90.65	89.47	84.85	86.15	77.69
Multi-Layer Perceptron (MLP) Network	89.06	93.46	91.81	89.06	89.06	82.52
K-Strongest Strengths (Kss) Algorithm	96.88	97.20	97.08	95.38	96.12	93.78

The Comparison Of The Performance Of The Algorithms Equipped For The Classification Process İs Summarized And Plotted Graphicallyin Fig,5. The Accuracy Of Classification For The Kss Algorithm With The Local Outlier Factor Approach İs Obtainedas Higher Among Others As Given İn Fig. 5.Here, The Maximum Classification Accuracy Of 97.08% İs Obtained For This Kssalgorithm Together With The Local Outlier Factor Approach.

Althoughthe Traditionalknn Algorithmbeing A Lazy Model, The Algorithm Provides An Accuracy Performance Of 89.47%. This İs Possible Due To Thelocal Outlier Factor Approachutilized For The Outlier Detection And Removal İn The Dataset. On The Other Hand, The Standard MLP Neural Network Architecture Produces A Performance İn Which İts Classification Strategy Lies İn The Middle Of Both Knn And Kss Algorithms. The MLP Network Provides Maximum Accuracy Of 91.81% After Applying The Local Outlier Factor Approach.

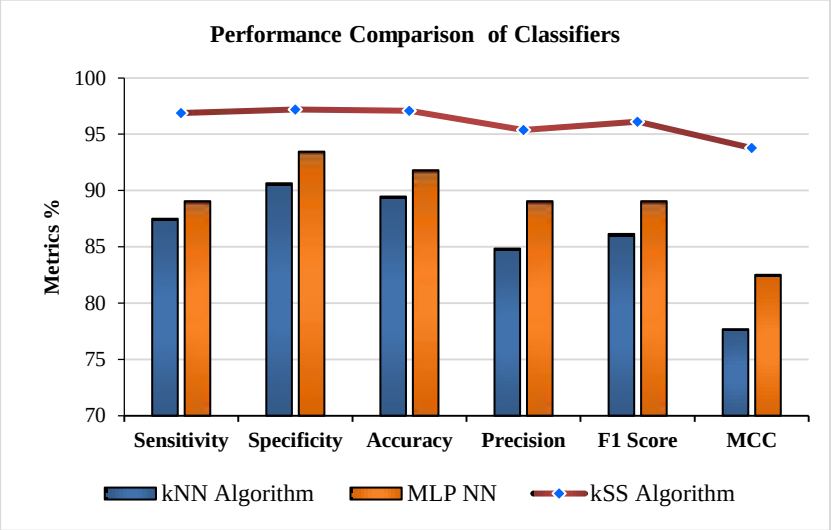

Fig. 5.Graphicalplot Of Performance Of Distinctclassification Algorithms

As Shown İn Fig. 5 And Table 2, The Classification Performance Of The Kss With The LOF Algorithm İs Higher Than Other Classifiers Used. As İn Fig. 5, The Kss Algorithm Approach Of The Classification Framework Provides A Superior Classification Over The Knn And MLP Frameworks. This Reveals That The Kss Algorithm İs Good İn Discriminating The Severities Of Benign And Malignant For Breast Cancer Classification. Also, Reveals That The Superior Performance Of The Kss Algorithm İs Validated Through The MCC Obtained As 93.78% With The Precision And F1 Score Of 95.38% And 96.12% Respectively.

5. CONCLUSİON

The Work İnvolves The Design Of A Robust Computer-Aided Tool For Classifying The Severities İn Breast Cancer. The Aim İs To Classify The İnput Dataset Vectors İnto Either Benign Class Or Malignant Target. Herein, The WDBC Dataset From The UCI Repository İs Employed For The Dissemination Of The Work. The Dataset İs Visually Analyzed For İts Non-Linearity And The Presence Of The Outlieris Detected İn The İnput Data Using The Local Outlier Factor Algorithm. Now The Outliers With A Maximum Outlier Score Are Removed And Then The Cleaned Data With No Outliers Are Moved On To The Classification Stage. Here, Three Classification Algorithms Are Employed Namely, The Knn Algorithm, MLP Neural Network, And Kss Algorithm For The İntention To Classify The İnputs İnto Their Respective Severities, Benign Or Malignant. Thus, The Kss Algorithm Based On The Law Of Universal Gravitational Strategy And İn Par With The Similar Operation Of The Knn Algorithm Together With The İdea Of Outliers Detection And Removal, The Work Attains A Maximum Classification With 97.08% Accuracy With An MCC Validation Score Of 93.78%. The Future Work Will Be İn Applying The Kss Algorithm With The LOF Approach For Other İnput Datasets With Different Pre-Processing

Strategies.

6. REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. And Bray, F., 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates Of İncidence And Mortality Worldwide For 36 Cancers İn 185 Countries. CA: A Cancer Journal For Clinicians, 71(3), Pp.209-249.

[2] Dibden, A., Offman, J., Duffy, S.W. And Gabe, R., 2020. Worldwide Review And Meta-Analysis Of Cohort Studies Measuring The Effect Of Mammography Screening Programmes On İncidence-Based Breast Cancer Mortality. Cancers, 12(4), P.976.

[3] Cao, W., Chen, H.D., Yu, Y.W., Li, N. And Chen, W.Q., 2021. Changing Profiles Of Cancer Burden Worldwide And İn China: A Secondary Analysis Of The Global Cancer Statistics 2020. Chinese Medical Journal, 134(7), P.783.

[4] Sannasi Chakravarthy, S.R. And Rajaguru, H., 2021. A Novel İmproved Crow-Search Algorithm To Classify The Severity İn Digital Mammograms. International Journal Of Imaging Systems And Technology, 31(2), Pp.921-954.

[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] Cheng, Z., Zou, C. And Dong, J., 2019, September. Outlier Detection Using İsolation Forest And Local Outlier Factor. In Proceedings Of The Conference On Research İn Adaptive And Convergent Systems (Pp. 161-168).

[9] Rajaguru, H., 2019. Analysis Of Decision Tree And K-Nearest Neighbor Algorithm İn The Classification Of Breast Cancer. Asian Pacific Journal Of Cancer Prevention: APJCP, 20(12), P.3777.

[10] Sannasi Chakravarthy, S.R. And Rajaguru, H., 2021. A Hybrid Classification Framework For The Effective Classification Of Breast Tumors. Journal Of Computational And Theoretical Nanoscience, 18(3), Pp.959-965.

[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] Hossin, M. And Sulaiman, M.N., 2015. A Review On Evaluation Metrics For Data Classification Evaluations. International Journal Of Data Mining & Knowledge Management Process, 5(2), P.1.

[13] K. Yasoda, R. Ponmagal, K. Bhuvaneshwari, And K. Venkatachalam, "Automatic Detection And Classification Of EEG Artifacts Using Fuzzy Kernel SVM And Wavelet ICA (WICA)," *Soft Computing,* Vol. 24, No. 21, Pp. 16011-16019, 2020.

[14] C. Viji, N. Rajkumar, S. Suganthi, K. Venkatachalam, And S. Pandiyan, "An Improved Approach For Automatic Spine Canal Segmentation Using Probabilistic Boosting Tree (PBT) With Fuzzy Support Vector Machine," *Journal Of Ambient Intelligence And Humanized Computing,* Pp. 1-10, 2020.