International Journal of Future Generation Communication and Networking
Vol. 14, No. 1, (2021), pp. 237–247

An Automatic Classification of Lung Cancer with J48 using Weka

Dr.B.Dhanalakshmi[1], Dr.Raghuraman[2], Dr.Sudha Rajesh[3]
[1]Professor, [1]Department of Information Technology,[1]Bharath Institute of Higher Education and Research
[2]Associate Professor, [2]Department of Automobile Engineering, [2]Bharath Institute of Higher Education and Research
[3]Assistant Professor, [3]Department of Computer Applications, [3]B.S.A. Crescent Institute of Science and Technology
[1]dhana.baskaran@gmail.com, [2]raguraman150807@gmail.com, [3]drsudharajesh84@gmail.com

ABSTRACT
Lung dangerous development is a sort of infection that begins in the lungs. Your lungs are two springy organs in your chest that take in oxygen when you take in and release carbon dioxide when you inhale out. Mechanized examination impacts the precise evaluation of lung malignant growth in a viable way. Lung disease influences individuals in lung parts of the body. A PC strategy ought to be inspected to analyze the lung malignant growth exactly. This is the specialist's pre-screening framework for early analysis. The related and the proposed work is looked at and analyzed. The proposed work gives the report on the order of sores from the lung malignant growth dataset with fundamental advances, for example, pre-handling and arrangement. Here J48 Decision tree examination is utilized to separate the highlights. The reenactment quantifies the exact conclusion and affirms the precision esteems up to 96% for Classification.
Keywords : J48 Decision Tree, C4.5, Weka, Pruned Tree.

INTRODUCTION
The growth of the malignant tumour in the lung that is identical to the other tumour occurs due to the abnormality in any body unit, essential for life, which is the cell. Generally, body has a balanced cell development, which mainly focuses on the partition of the cell, to generate new cells when required. Any kind of disturbance in this balance, will lead to an uncontrolled division and rapid multiplication of cells, forming a cluster termed as tumour. Doctors or health experts can easily get rid of the mild tumours, as they do not spread to the other areas of the body [7]. In the case of malignant or harmful tumour, there are likely to spread initially entering into circularity system and eventually passing to other parts of the body. This stage of spreading is known as metastases after it is developed, it turns out to be a very dangerous disease to treat.

LITERATURE REVIEW
An upgraded J48 calculation was used to improve the accuracy in discovery and implementation of novel IDs procedure. Also this helped in recognizing possible aggregation, which could improve the system secrecy. For this purpose, the examiners made use of plenty of

ISSN: 2233-7857 IJFGCN

datasets by implying different approaches like J48, Naive bayes, Random Forest and NB tree. While completing the analysis, a NSL KDD interpretation dataset was applied [1].

Then this dataset was classified into preparing dataset and testing dataset, regarding the information handling. Consequently, for assessing the fairness of the number of highlight a component assuring determination technique based on WEKA application was used. The obtained outcomes solicit that the calculations were supervisor, exact and increase effective, without the need for the highlights. Also this calculation assured that the orderliness of dataset in light yielded 99.88% of exactness and while in the usage of 10 overlay cross approval test, 90.01% of exactness, during the usage of the entire datasets beside all highlights, 76.33% of precision was recorded.

We present an immediate procedure for imperfections recognized by vortex current testing, for any defect an ECT field reaction for various EC test ways is announced or articulated to an unpredictable plane to obtain Lissajous figures. Their shapes are characterized through the usage of few geometrical parameters forming an element vector. These vectors are considered as masks of defects differentiated by the tests at various junctions and good ways from deforming [2]. The feasibility of the proposed approach is tested by predicting the implementation of AI based classifications such as Naïve Bayes, C4.5/J48 decision tree and multilayer perception neural system, through other measurements. The output agrees the value of the approach to deal with the discovery and grouping without the need for checking the flaws.

The objective of this research is to examine the link between diabetes mellitus and singular hazard elements of metabolic syndrome (Mets), in a non-traditionalist method [3]. The forecast of the upcoming of diabetics using important hazard elements of Mets and to analyze the entire execution of AI methodologies when information inspecting systems are applied to create adjusted sets. The dataset emphasized right now has 667907 records for a time period from 2003 to 2013.

To measure the responsibility of individual hazard elements of Mets in the improvement of diabetes in a less preserving method that calculates relapse examination. Outcomes say that the extended degrees of HDL (High density lipoprotein) are correlated with the initial stage of diabetes, specifically in ladies. Similarly we have suggested J48 choice tree and Naïve bayes techniques for the upcoming diabetes using hazardous elements from critical examination over adjusted and lopsided datasets. The outputs exhibited the matchless quality of naïve bayes with k-medoids under analyzing methods.

The aftermath of this examination results in the explanation of pathological essential of HDL and pathways in the improvement of diabetes. Moles are occurred on human body, because of infection provoked by human and papillomavirus. The highly affected regions of moles are hands and feet especially, which is hard to recover in crucial stages [4].

The important test in treating moles is the assorted variety of treatment on several patients. Hence it gets hard to identify accurate treatment to be carries out for treating the disease. Consequences of extra spaces are important to detect early sickness and create master

ISSN: 2233-7857 IJFGCN

platforms. This pursuing work focuses on updating exactness of J48, a twofold chance tree based classifiers by embedding characters based on hereditary programming.

For the arrangement, creators have chosen immunotherapy and cyrothreaphy datasets from UCI AI achieves, which involves the cases of patient reaction, against immunotherapy and cyrothreaphy for plants and basic moles [5]. After analysis, it is identified after the involvement of the characteristics created through hereditary programming. The outcome of grouping precision of J48 is said to be 82.22%.

METHODOLOGY

J48

Grouping a method for building a class model from plenty of records that posses class names. Choice tree calculations are used to identify the conduct of properties. The classes for recently created cases are identified based on preparation timings. This calculation attains the principle of expectation of an objective variable [8]. The basic announcement of the information is made correctly compliable with the guide of tree order classification. J48 is an expansion of ID3. The merits of J48 are representation of missing qualities, pruning of choice tress, induction of rules etc. this apparatus yields various alternatives regarding tree pruning.

During the times of potential over fitting, pruning can be used for accuracy. This classification creates limitations from which specific status of that information is produced. Basically the dynamic speculation of choice tree takes up balance, adaptability, exactness and accuracy.

Architecture

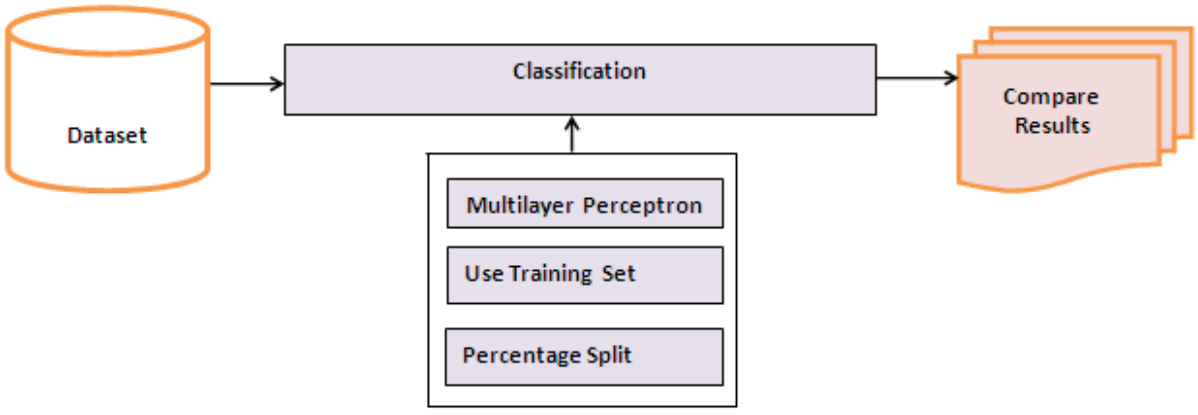

Figure1:Architecture Diagram for Proposed Method

Algorithm

make a hub N;

condition parts in D in identicalset, C onpeak

revisit N to besidecore marked with set C;

```
        conditiona_list is unfilled onpeak
revisit N to beside core with marked

use lion's divideset in D;|| dominant part casting a ballot
applya_range_scheme (D, characteristic_list)

to locate the best split_measure;
scratch hub N with split_measure;

conditions_attribute is linear-esteem and
        multiway part permitted over thepeak
        attribute_list = parting quality;

for every result j of parting paradigm
leaveDj alone the arrangement of information parts in D fulfilling result j;
wheneverDj is vacant at that peak
append a side marked over greater part

set in D to core N;
else
append corerevisit by engender
choice hierarchy(Dj, trait incline) to hub N;
end for
bring N back;.
```

RESULT AND DISCUSSION

The dataset is obtained from the specialty in identifying lung cancer. The input dataset contains plenty of attribute to prove the taken input is affected with cancer. There are nearly 41 attributes to describe the taken dataset is cancer or not. To prove the cancerous data, the class attribute is taken with 3 categories; they are completely cancer, partial cancer and no cancer. The visualization of dataset that are filtered using discretization is shown in Figure 2.

ISSN: 2233-7857 IJFGCN

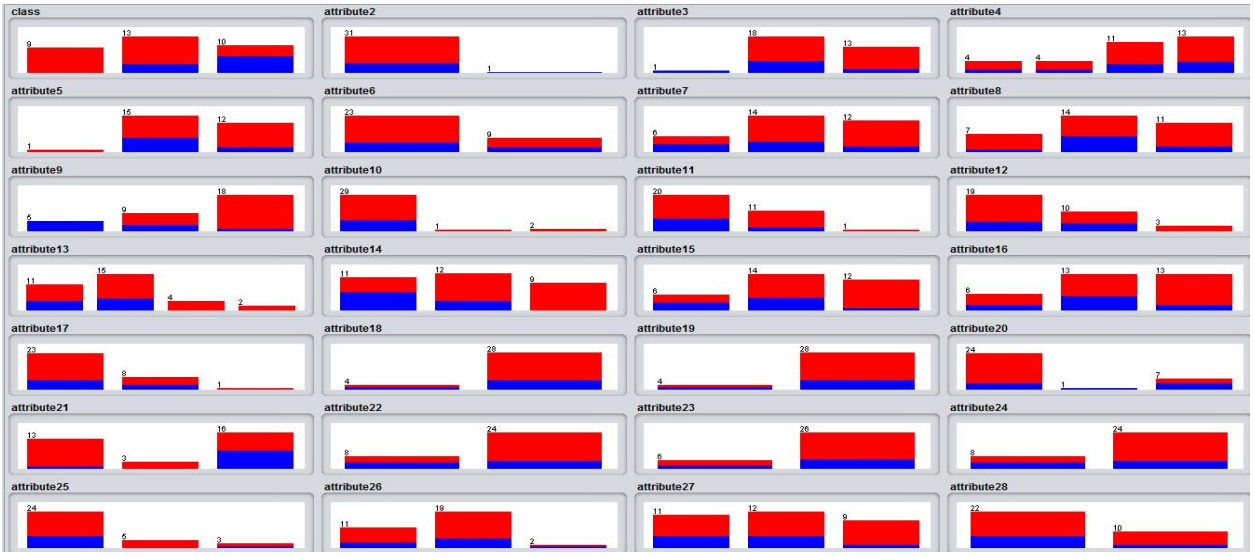

Figure2: Visualize of Inputs

After the fine filtering process, the input data is taken into classification, the J48 classification rule is used to order the data. The J48 algorithm is a sub class of C4.5 algorithm, the major usage of the rule is to divide the information into plenty of data pruned as tree. The tree with the maximum attribute is located as root node where as the remaining nodes are differentiated with respect to attribute differentiation shown in Figure 3.

```
J48 pruned tree
------------------

attribute41 = 1: 3 (3.0)
attribute41 = 2
|   attribute47 = 1: 1 (2.0)
|   attribute47 = 2
|   |   attribute24 = 1
|   |   |   attribute28 = 2: 3 (2.0)
|   |   |   attribute28 = 3: 2 (2.0)
|   |   attribute24 = 2
|   |   |   attribute34 = 1: 2 (2.0/1.0)
|   |   |   attribute34 = 2: 2 (0.0)
|   |   |   attribute34 = 3
|   |   |   |   attribute28 = 2: 2 (11.0/2.0)
|   |   |   |   attribute28 = 3: 1 (6.0/1.0)
|   attribute47 = 3: 3 (1.0)
attribute41 = 3: 3 (3.0)

Number of Leaves  :     10

Size of the tree :      16


Time taken to build model: 0.03 seconds
```

Figure3: Pruned Tree Structure

After the pruned tree, the decision tree is constructed from the maximum attribute till its pruned data the decision tree proves the maximum distribution of class is shown in Figure 4.

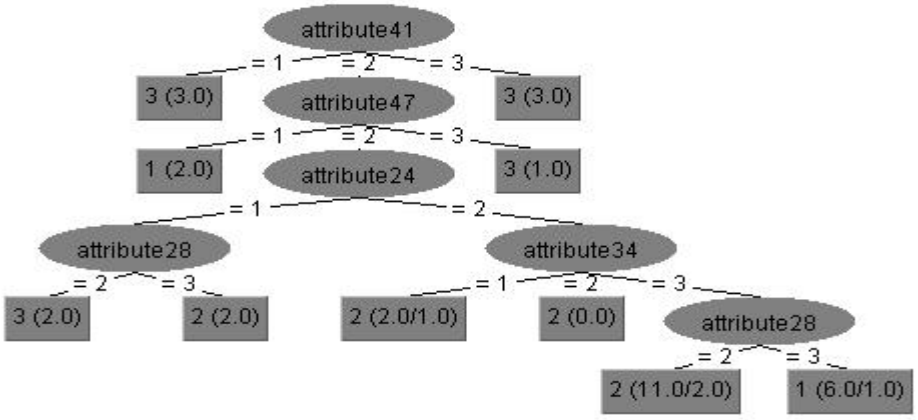

Figure4: Decision Tree

After fine distribution of decision tree, the outcomes of classification should be evaluated. The summary of classification with the classified instances and the error rate is shown in Table 1.

Table1 Summary

ProperlyCategorizedOccurrences	31	96.87 %
ImperfectlyCategorizedOccurrences	1	0.031%
Kappa measurement	0.4289	
Mean absolute fault	0.1461	
Root mean squared fault	0.1005	
Total Occurrences	32	

After clear summary report of classification and error rates, the true Positive, true negative and other components are classified is shown in Table 2 and the confusion matrix describes about the segregation of class to be classified in shown in Table 3.

Table 2 Complete Accuracy by Class

Period	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
1	0.444	0.174	0.500	0.444	0.471	0.281	0.630	0.391
2	0.538	0.474	0.438	0.538	0.483	0.064	0.551	0.450
3	0.500	0.136	0.625	0.500	0.556	0.389	0.834	0.657
Weighted Mean	0.500	0.284	0.514	0.500	0.502	0.226	0.662	0.498

ISSN: 2233-7857 IJFGCN

Table 3 Confusion Matrix

Class a	Class b	Class c	Classified
4	5	6	Class a is 1
3	7	3	Class b is 2
1	4	5	Class c is 3

After perfect classification, plenty of reports to be generated, after classification there could be some error rate which is shown in Figure 5 and the margin curve to improve the efficiency is shown in Figure 6.

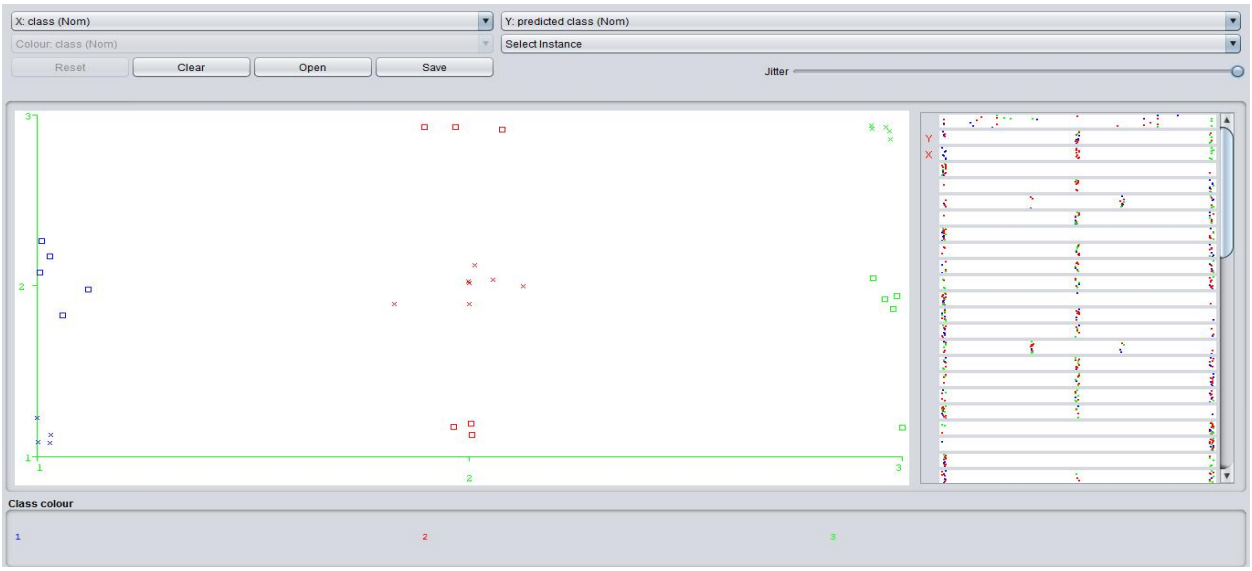

Figure5: Classify Error of Classification

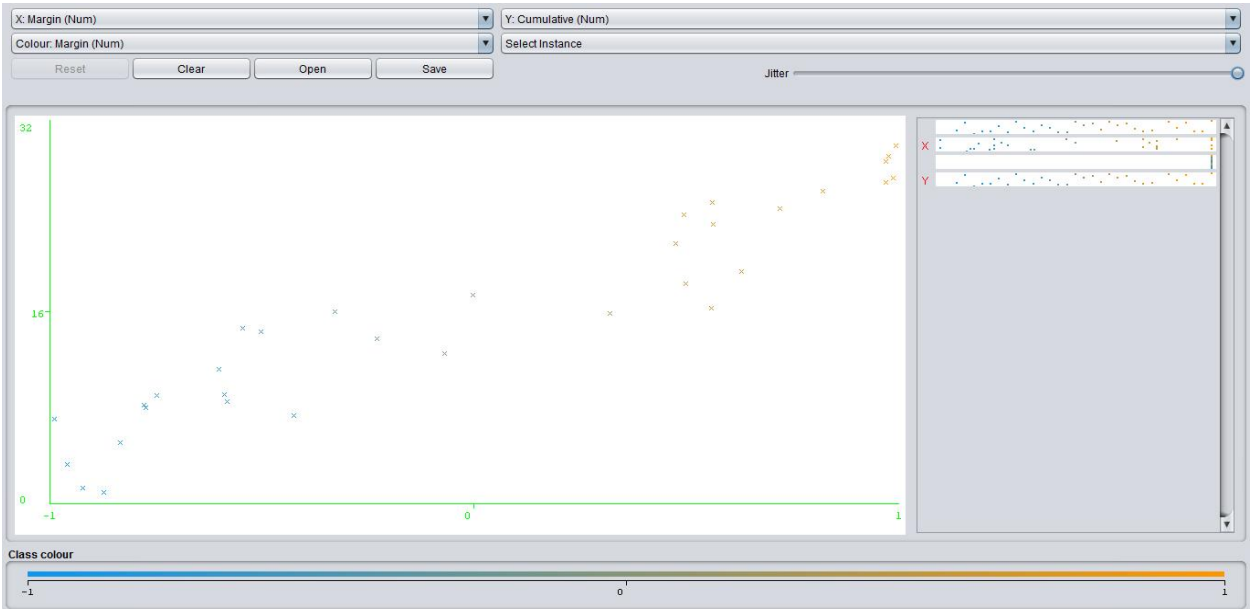

Figure6: Margin Curve of classification

The threshold curve is constructed to prove the threshold value of each class. The class is generated behalf of class values. The class value improves the entire solution; threshold curve is shown in Figure 7 for class1, Figure 8 for class 2 and Figure9 for class 3.

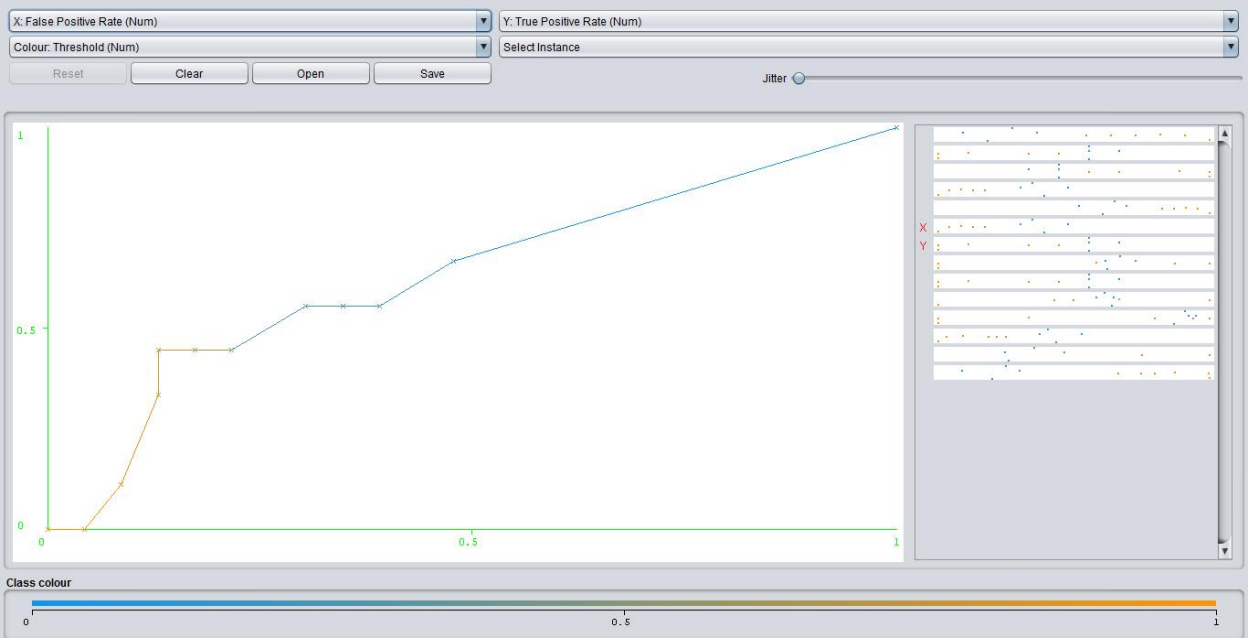

Figure7:Threshold Curve for class 1

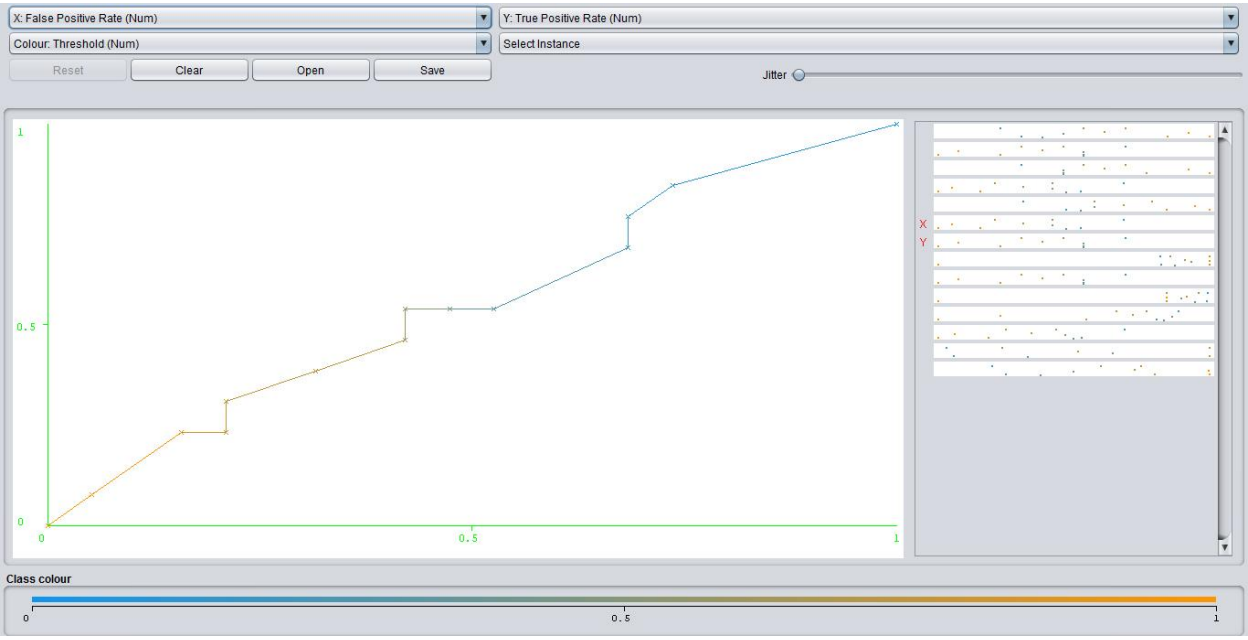

Figure8:Threshold Curve for class 2

ISSN: 2233-7857 IJFGCN

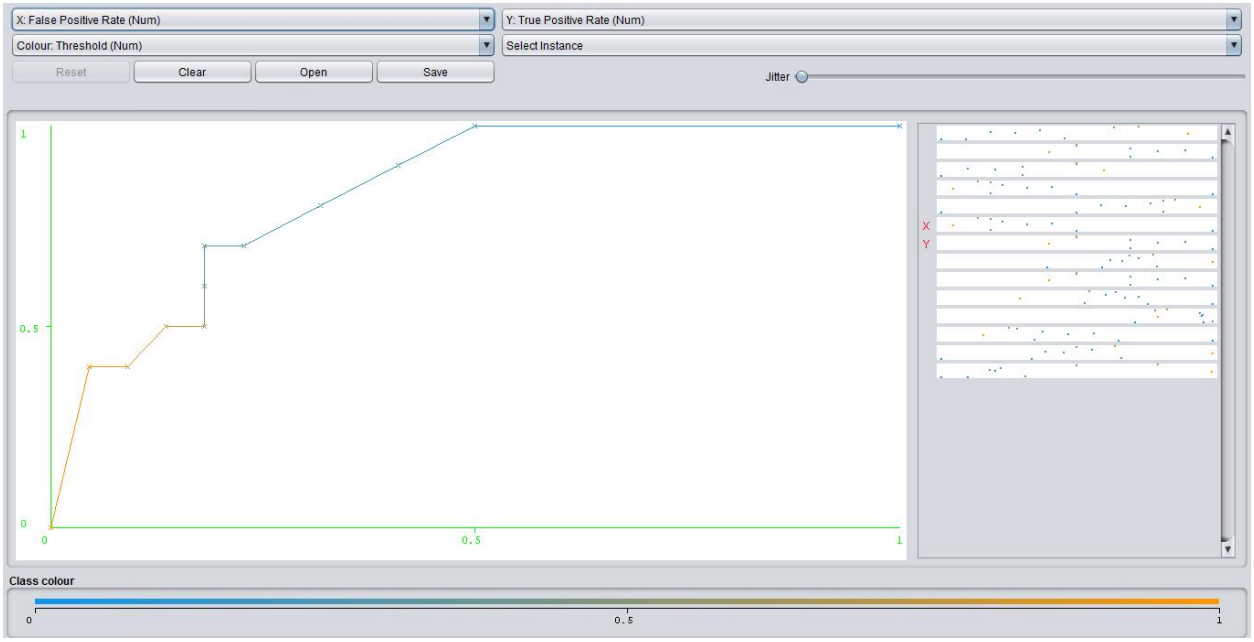

Figure9:Threshold Curve for class 3

To improve the accuracy of the classification Cost/Benefit curve is drawn to visualize the improvement of cost/benefit, if there is any updation of class file shown in Figure 10 for class1, Figure 11 for class 2 and Figure 12 for class3.

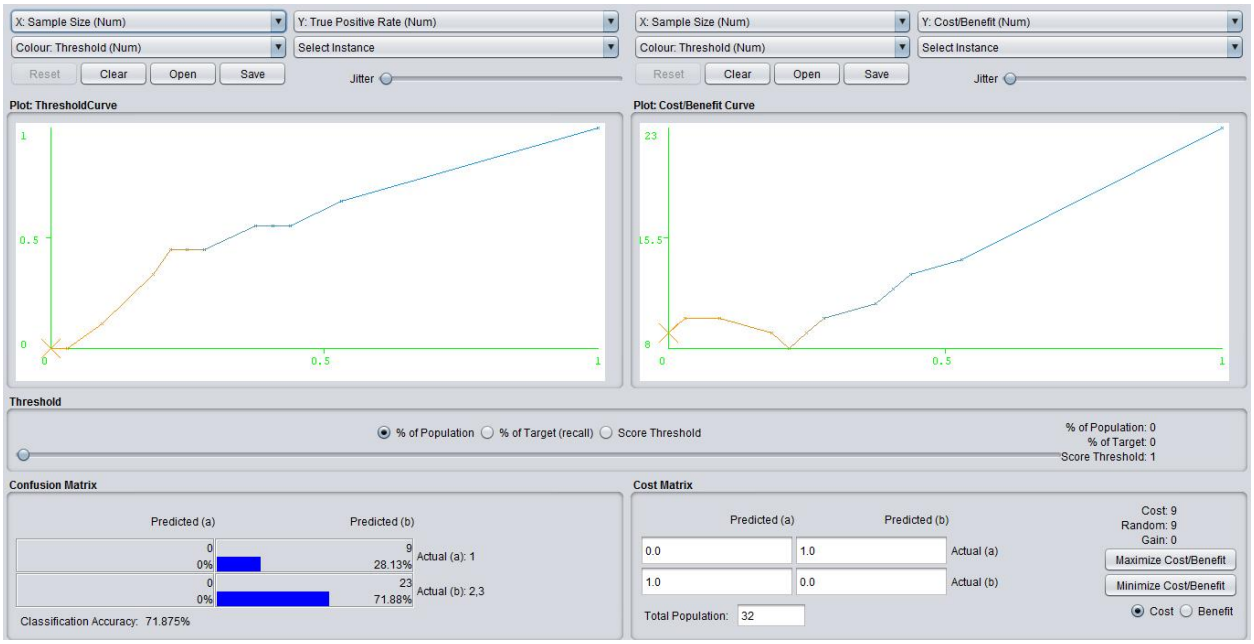

Figure 10:Cost/Benefit Curve for class 1

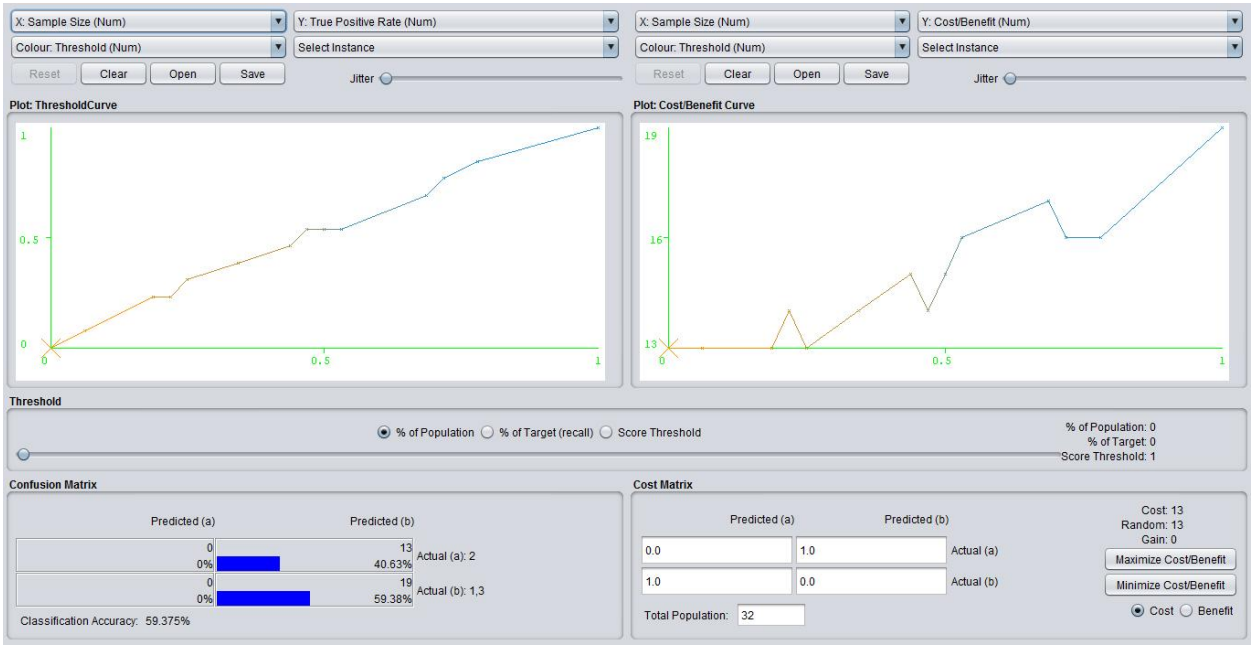

Figure11:Cost/Benefit Curve for class 2

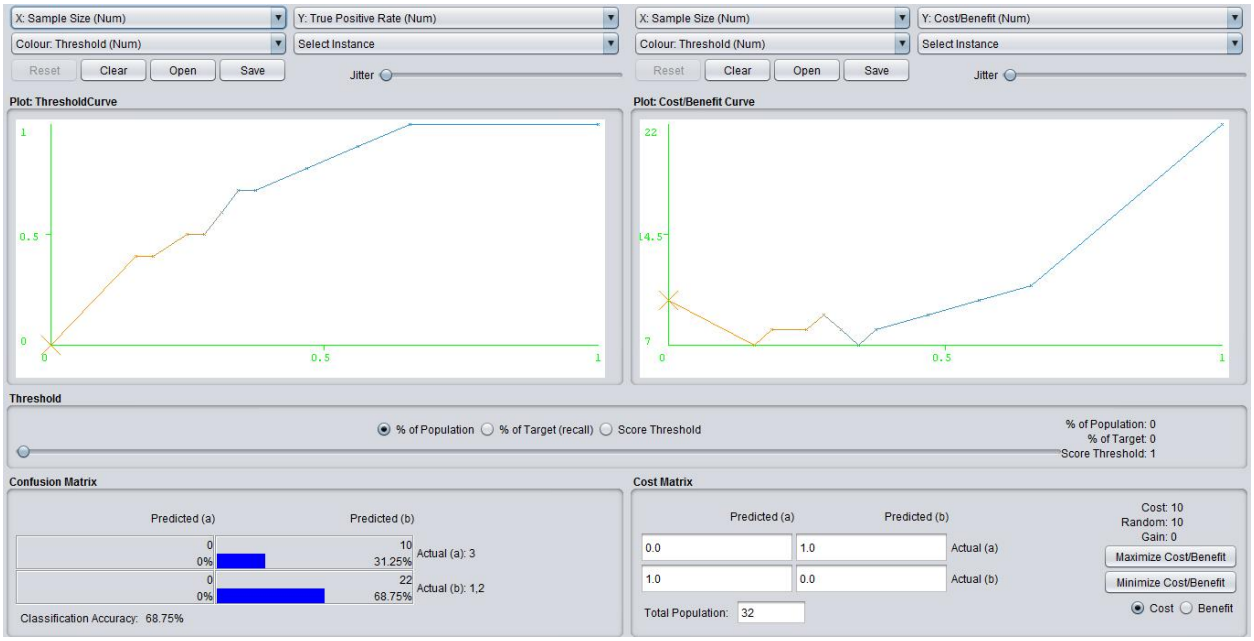

Figure 12:Cost/Benefit Curve for class 3

CONCLUSION

After a fine extraction of highlights for arrangement, GLCM ends by producing parameters to make a class for weka. The Discretize work changes over the GLCM highlights to ostensible information. The ostensible information is imagined to show the isolation proportion of qualities. J48 percept's its given ostensible information into choice tree and thinks about all tree leaves to one another to give the best results. The ostensible information is differentiate

ISSN: 2233-7857 IJFGCN

property of dataset, contains on correlation of every datum. At last adjusted information is taken into percept with other characteristic. The edge bend gives the ideal perception of edge and examples. The edge bend pictures the scope of limit of each estimation of ostensible property. At long last the money saving advantage examination portrays the ideal efficiency up to 96% of characterization.

REFERENCES

1. Aljawarneh, S., Yassein, M.B. &Aljundi, M.An enhanced J48 classification algorithm for the anomaly intrusion detection systems. Cluster Comput 22, 10549–10565 (2019). https://doi.org/10.1007/s10586-017-1109-8
2. G. D'Angelo, M. Laracca, S. Rampone and G. Betta, "Fast Eddy Current Testing Defect Classification Using Lissajous Figures," in IEEE Transactions on Instrumentation and Measurement, vol. 67, no. 4, pp. 821-830, April 2018.
3. S. Perveen, M. Shahbaz, K. Keshavjee and A. Guergachi, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques," in IEEE Access, vol. 7, pp. 1365-1375, 2019.
4. SabitaKhatria, Deepak Aroraa, Anil Kumara, "Enhancing Decision Tree Classification Accuracy through Genetically Programmed Attributes for Wart Treatment Method Identification", Procedia Computer Science, Volume 132, 2018, Pages 1685-1694
5. S. Sameen, M. Sharjeel, R. M. A. Nawab, P. Rayson and I. Muneer, "Measuring Short Text Reuse for the Urdu Language," in IEEE Access, vol. 6, pp. 7412-7421, 2018.
6. N. Emaminejad et al., "Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 5, pp. 1034-1043, May 2016.
7. Y. Yin et al., "Tumor Cell Load and Heterogeneity Estimation From Diffusion-Weighted MRI Calibrated With Histological Data: an Example From Lung Cancer," in IEEE Transactions on Medical Imaging, vol. 37, no. 1, pp. 35-46, Jan. 2018.
8. J. Jiang et al., "Multiple Resolution Residually Connected Feature Streams for Automatic Lung Tumor Segmentation From CT Images," in IEEE Transactions on Medical Imaging, vol. 38, no. 1, pp. 134-144, Jan. 2019.
9. N. Kureshi, S. S. R. Abidi and C. Blouin, "A Predictive Model for Personalized Therapeutic Interventions in Non-small Cell Lung Cancer," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, pp. 424-431, Jan. 2016.
10. L. Hussain, W. Aziz, A. A. Alshdadi, M. S. Ahmed Nadeem, I. R. Khan and Q. Chaudhry, "Analyzing the Dynamics of Lung Cancer Imaging Data Using Refined Fuzzy Entropy Methods by Extracting Different Features," in IEEE Access, vol. 7, pp. 64704-64721, 2019.

ISSN: 2233-7857 IJFGCN