

Performance Evaluation Of Machine Learning Algorithm For Lung Cancer

Ms.M.Pyngkodi¹, Wilfred Blessing N.R², Dr.S.Shanthi³,
R.Mahalakshmi⁴, M.Gowthami⁵

¹Dept Of Computer Applications, Kongu Engineering College, Taimlnadu, India

²It Department, University Of Technology And Applied Science, Salalah, Oman.

³Dept Of Computer Science & Engineering, Kongu Engineering College, India,

⁴Pg Student, Dept Of Computer Applications, Kongu Engineering College, India

Email: Pyngkodikongu@Gmail.Com¹,
Wilfred.B@Sct.Edu.Om²,Shanthi.Kongumca@Gmail.Com³

Abstract. Background: Lung Cancer Is A Cancer That Is Difficult To Identify, And Deadly. Early Detection Of Cancer Can Be Effective In Curing The Disease Altogether. The Proposed Model Is An Effective Classification-Based Method, Using Machine Learning Methods To Identify Lung Cancer Diseases. The Method Can Significantly Reduce The Danger Of Disease Through Digging Out A Transparent And Understandable Model For Lung Carcinoma From A Medical Database. **Methods:** This Work Tried To Assess The Efficacy Of Machine Learning Algorithms In The Task Of Classifying Lung Cancer Based On Diagnostic Levels. In This Analysis We Analyzed And Compared Various Classifications To Identify And Predict Lung Cancer Disease. We Applied Benchmark Machine Learning Techniques Like Support Vector Machine, K-Nearest Neighbor, Random Forest, Linear Regression, And Logistic Regression. The Main Objective Is To Evaluate The Precision, Accuracy, Sensitivity And Specificity Of Data Classification Regarding The Effectiveness And Efficiency Of Algorithms. **Results:** The Outcome Of The Result Has Assesses Based On Correct And Incorrect Data Which Are Exactly Classified By A Classification Methods. The Best Performance Of These Techniques Have Been Obtained By Svm And Logistic Techniques With The Highest Accuracy. **Conclusions:** Classification Of The Detection Process Is Conducted And The Output Were Analyzed With The Comparison Of Accuracy Among Machine Learning Techniques And The Results Were Given Based On The Data, Respectively. This Technique Will Better Help Us Diagnose Lung Cancer And May Save Many Lives In The Future.

Keywords: Lung Cancer Prediction, Classification Algorithms, Knn, Svm, Random Forest, Linear Regression And Logistic Regression.

1. INTRODUCTION

Machine Learning Is A Computational Intelligence (Ci) Application Which Gives Systems The Ability To Automatically Learn Information From Experience And Other Ways. Prediction And Classification Are Two Data Analysis Approaches Which Can Be Employ To Develop Models That Describe Important Data Classes, Or Forecast Future Trends. In Several Fields Of Science Prediction Is An Important Subject Now A Day. This Paper Analyzes The Lung Cancer Prediction By Means Of A Particular Classification Algorithms. Lung Cancer Is The Most Prevalent Cause Of Cancer Mortality In The World. In Most Cases, Manifestation Of Lung Cancer In The Patient's Body Shows By Early Symptoms. The

Number Of Chain Smokers Is Directly Proportional To The Number Of Lung Cancer Sufferers. Smoking And Pollution Are Causing The Lung Cancer In Particular, And Early Detection Will Help Patients Avoid Smoking Or Other Cancer-Causing Factor. Data Mining Is A Part Of Artificial Intelligence And Uses A Variety Of Data Sets, Probabilistic And Mining Models That Use Past Outcomes To Provide A Predictive Outcome Technique [1]. This Paper Used Classification Algorithms Such As Knn, Svm, Random Forest, Linear Regression And Logistic Regression To Analyze Lung Cancer Prediction. Data Preprocessing, Feature Selection And Classification Has Been Proposed And Implemented. Pre-Processing Refers To The Changes Applied To Our Data Before Feeding It To The Algorithm. Feature Selection Can Serve As A Very Valuable Method Of Pre-Processing In Solving Classification Problems.

This Paper Introduces A New Feature Selection Strategy Which Is Demonstrated On A Real Data Set. Namely, The Proposed Approach Solution Establishes Endorsed Subsets Based On Two Parameters: (1) Individual Attributes Have High Power To Discriminate (Classify); And (2) The Sub-Set Attributes Are Complementary- That Is, They Misclassify Various Classes. The Method Uses Confusion Matrix Information, And Evaluates One Attribute At A Time. The Uncertainty Matrix Is Thus A Strategy For Summing Up The Results Of A Classification Algorithm. That Shows How Confused The Classification Model Is When Predictions Are Made. Comparison Of Two ML Models, Based On A Statistical Test. They Used Different Datasets To Test These Two Models, And Then Used A Statistical Test To Compare Their Accuracy Results To See If There Was Any Statistical Significance. The Statistical Tests Discussed Below Tend To Provide A P-Value And A Test Statistic.

This Paper Mainly Focuses On The Disease Of Lung Cancer Using Various Classification Algorithms With The Help Of Python-Based Datamining Tools. Only If You Recognize The Annual Stage We Will Cure Lung Cancer. And Here We Use Machine Learning Algorithms To Detect Lung Cancer. That Can Be Made More Accurate And Quicker. Even Though New Technologies Are Available To Detect Through The Combination Of Machine Learning, We Can Produce More Accurate Results. It Is A Boon That The Disease Can Be Detected In Its Earlier Stages By Combining Machine Learning Techniques.

2. LITERATURE SURVEY

The Most Significant Source Of Death For Both Women And Men Is Lung Cancer, A Disease Of Uncontrolled Cell Growth In The Lung Tissues[1]. Data Analysis Is A Crucial Role In The Growth Of The Discovery Of Data In Datasets. It Has Many Potential Uses. The Performance Of Classifiers Is Highly Dependent On The Data Set Used For Learning. This Results In Improved Effective Classification Models In Terms Of Predictive Or Descriptive Accuracy, Reduced Computational Time Needed To Construct Models As They Learn More Quickly, And A Greater Understanding Of The Models. This Offers A Comparative Study Of Data Classification Precision, Using Lung Cancer Data In Various Scenarios. This Compares Predictive Performance Of Rising Classifiers In Quantitative Terms. Prashantnaresh[2] Applied A Pattern Prediction Software To A Lung Cancer Prediction System, Lung Cancer Risk Prediction System Could Help Diagnose A Person's Predisposition To Lung Cancer. Early Detection Of Lung Cancer May Play A Key Role In The Diagnosis Process And For A Effective Plan For Prevention [2]. Svm Implementation Uses Two Basic Steps As The Training Of Instances And Checking. The Initial Phase Allowed The Svm To Feed New Data Across From What Was Previously Known. Van Belle Et Al[3] Are Proposing To Categorize Unknown Data In The Training Set An Svm Gets Its Understanding Runxuan Zhang Et Al[4] Suggest Pneumoconiosis Detection Through The Use Of Various Sub-Sets Of Lung

Disorders Based On Support Vector Machines Shubpreet Kaur[5] Future Developments In Data Mining In Medical Health System Forecasting Complex Diseases. The Application Of Data Mining In The Medical Sector In Their Paper Poses An Extremely Daunting Challenge For The Medical Profession. These Are Characteristic Of Widespread Processes That Involve A Detailed Understanding Of Health Needs. A Work On Critical Classification Algorithms For Recognizing And Diagnosing Lung Cancer Diseases N.V. Ramana Murty[6]. In Their Paper They Experimented An Analytical Process Using Weka Tool With Various Data Mining Classification Techniques And Confirmed That The Naive Bayesian Method Gives Better Performance In All Respects Over The Other Classification Algorithms. Comparative Analysis Of Recent Trends In Cancer Prediction Using Data Mining Techniques, Satyam Shukla Et[7] They Use Data Mining Techniques In Their Paper, Such As The Rank-Based Method Where Reversal Pairs Arerea And Can Be Easily Independent Of Samples By Helping In The Diagnosis Of Cancer. A Research On Mining Lung Cancer Data Using Optimization Techniques Called Ant Colony Optimization For Predictive Quality Increasing Or Decreasing Disease, J.Jamara Banu [8]. With Multiple Data Mining Classification Techniques, They Performed Successfully In Their Paper And Concluded That Data Mining Could Contribute Significantly To Research Into Lung Cancer And Eventually Improve The Quality Of Healthcare For Lung Cancer Patients. Neha Panpaliya, Neha Tada [9] Thesis On Early Detection And Lung Cancer Prediction. In Their Paper[10][11][12][13] We Conclude That Combining Neural Network Classifier Withr With Binarization And Glcm Would Increase The Accuracy Of The Lung Cancer Detection Techniques. By Using This Method, The Cost And Time Required For The Diagnosis Of Cancer Will Also Be Reduced, And Also If The Patient Is Not Diagnosed With Lung Cancer, The System Will Continue The Predictive Process. Risk Detection Of Early Lung Cancer, Using Data Mining. They Show Experimental Findings In Their Paper Are Split Into Two Parts. The Important Recurring Patterns Are Discovered, And Another Is Predictive Tools For Predicting Lung Cancer. We Use Data From The Data Warehouse To Identify The Relevant Trends For Predicting Lung Cancer [14].

3. MATERIALS AND METHODS

There Are Certain Steps To Be Taken When Predicting Lung Cancer Using Machine Learning Techniques. Data Representing The Features Of Many Patients ' Lung Cancer Are Initially Collected From The Uci Repository. The Data Must Include Important Features Relating To Lung Cancer. The Approach Used Here For The Prognosis Technique Is Based On A Thorough Review Of The Signs And Risk Factors Associated With Lung Cancer.

3.1 Dataset Description

The Data Sets Used In This Study Are More Reliable And Precise To Improve The Predictive Accuracy Of The Data Mining Algorithms. Symptom Attributes Are Used To Identify Disease To Be Effectively Managed In Order To Obtain The Optimum Outcome From The Data Mining Process. Then Split The Dataset Into Two Parts , I.E., Data Training And Testing. Here Test Data Is 20%, And Train Data Is 80%. These Test And Train Data Were Taken, And Were Applied To Classification Techniques. This Defines The 1 And 0 Which Is 1 In The Lung Cancer Area, Shows That The Patient Has A Lung Cancer. The Results Shows 0 Then The Patient Is Identified As Free From Lung Cancer Diseases.

Table 1. Dataset Description Of Lung Cancer

Dataset	Genes	Samples	Classes
Gse6044	9870	76	2

3.2. Data Pre-Processing

The First And Formost Stage Of The This Work Is The Preprocessing Of Data By Which Raw Data Can Be Translated Into The Precise Format For Which The Data Set Is Given For Testing And Training And Further Study. Pre-Processing The Data To Enhance The Data Quality. There Are Many Missing Values To The Raw Data. Missing Value For Big Data Sets Is A Common Property. Some Algorithms Do Not Like The Missing Values, So We Can Remove Missing Values From Rows. Here, We Use Handling The Dataset, Splitting The Data, Finding The Missing Value And Encoding The Categorical Data Are Done.

3.2.1 Feature Selection

It Is The Method Used To Select The Prevailing Features From The Data Set And To Delete The Features That Are Not Related To The Function To Be Performed. Feature Selection Can Be Extremely Helpful In Reducing Data Dimensionality To Be Processed By The Classifier, Minimizing Execution Time And Increasing Predictive Accuracy[15]. Selection Of Features When The Initial Units Are In Is Superior To The Transition Function[18]. Reducing Data Dimensionality Reduces Computational Complexity For Larger Data Sets Such As Data, Resulting In Faster Execution Times. Statistical Testing And Normalization Methods Are Used In This Feature Selection Technique. There Are Five Statistical Tests That Are Used For Comparing The Data Samples, Namely Mann-Whitney U Test, T-Test, Friedman Tests , Kruskalwallis And Wilcoxon Rank Test. The Kruskalwallis H And Friedman Tests For Comparing More Than Two Data Samples[19]. The Mann-Whitney U Test To Compare Independent Data Samples. The Wilcoxon Rank Test For Comparing Paired Samples Of Data.

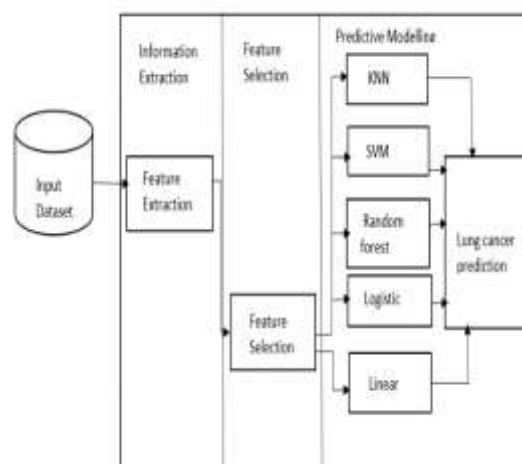


Fig 1: Workflow Diagram For Lung Cancer Prediction

4. CLASSIFICATION

This Work Generates Various Predictive Models Using The Algorithm-Based Train Data Collection, Such As Knn, Svm, Random Forest, Linear Regression, And Logistic Regression. Such Models Are Also Provided With The Test Dataset, And The Confusion Matrix Is Generated For Each Of The Following Models, And The Accuracy Of These Models Is Evaluated. **K-Nearest Neighbor:** One Of The Simplest Classification Strategies In Machine Learning Methods Is The K-Nearest Neighbor (Knn) Algorithm. Knn Calculation Is Among The Least Difficult Of All Ai Calculations. It Is Calculated By Distance And Given Number Of K, Therefore The Method Is Called K-Nearest Algorithm Of Neighbor. If The K Value Is Four Then The Dataset Is Simply Searched By Taking The Most Four Nearest Neighbors And Verifying Them. **Support Vector Machine:** Support Vector Machine Is A Simple Yet Powerful Supervised Machine Learning Algorithm Capable Of Performing With Data Sets That Are Both Linearly Separable And Nonlinear. In The Model Building Process Hyperplane Is Determined To Separate These Classes. Svm Constructs Hyperplanes Linearly And Has To Find The Ideal Hyperplane By Reducing The Distance Between Support Vector Points And Reducing The Chance Of Test Dataset Misclassification. Svms Are Used In Categorizing Text, Classifying Images, Recognizing Handwriting And Science. **Random Forest:** Random Forest Is A Classifier That Comprises A Collections Of Decision Trees On Various Subsets Of The Specified Dataset And Takes The Average To Improve The Data Set's Predictive Accuracy. Rather Than Relying On One Decision Tree, The Random Forest Takes Prediction From Each Tree And Predicts The Final Output Based On The Majority Vote Of Predictions. It Predicts Performance With High Accuracy, And It Runs Efficiently Even For The Large Dataset. The Greater Number Of Trees In The Forest Contributes To Greater Accuracy And Avoids The Problem Of Overfitting. **Logistic Regression:** Logistic Regression Is A Statistical Model That Uses A Logistic Function To Model A Binary Dependent Variable In Its Basic Form Although There Are Many More Complex Extensions. Logistic Regression Is A Regression Model In Which The Response Variable (Dependent Variable) Includes Categorical Values Such As True / False Or 0/1. This Is A Kind Of Statistical Analysis Used To Estimate A Dependent Variable Outcome Based On Prior Observations. **Linear Regression:** Linear Regression Is A Statistical Method Of Modeling The Relationship Between A Variable Dependent And A Set Of Independent Variables. Analysis Of Regression Helps In Evaluating The Relation Of Cause And Effect Between Variables. Other Variables (Called Dependent Variable) May Be Predicted If The Values Of Independent Variables Can Be Predicted Using Either A Graphical Method Or The Algebraic Method.

5. EXPERIMENTAL RESULTS

In This Study, Mainly Classification Algorithms Such As Knn, Svm, Random Forest, Linear Regression And Logistic Regression Are Used To Predict The Lung Cancer Disease From Data Set Instances And The Proposed Algorithms Are Applied And The Output Is Calculated On The Dataset Form Lung Cancer Disease. In This Several Range Of Experiments Are Carried Out Using Confusion Matrix And Statistical Test Methods Which Are Described In These Sections. From Table Two To Six Performance Of Various Algorithms Are Given.

Table 2. Accuracy Level For Different Classifications Algorithms

Classification Algorithms	Accuracy
Knn	85%

Svm(Linear Classifier)	100%
Svm(Rbf Classifier)	98%
Random Forest	96%
Logistic Regression	100%

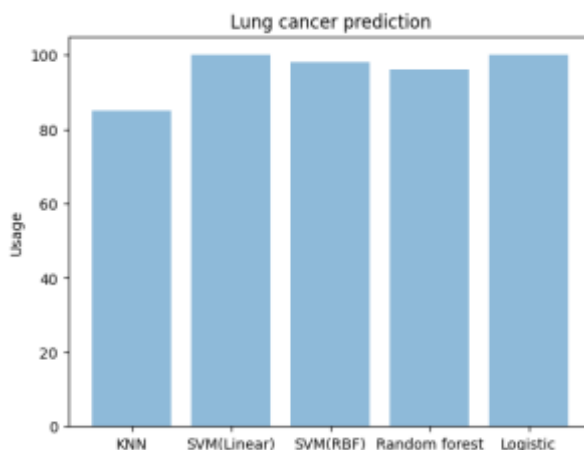


Fig 2. Lung Cancer Prediction Analysis

Figure 2 Shows That The High Precision Is A Part Of Various Classification Algorithms. Based On This Accuracy Most Algorithms Display A Better Accuracy For A Prediction. Depending On This Accuracy Many Algorithms For A Prediction Show A Higher Accuracy. The Simplest Method Employed Is Linear Regression, One Of The Oldest And Most Commonly Used Correlation Methods. The Goal Of The Methods Is To Fit A Straight Line Into A Set Of Data Points Using A Series Of Multiplied Coefficients Such As A Weighting Function And An Intercept. The Linear Regression Is Used To Determine The Value Of An Outcome Variable Y, Based On One Or More Input Predictor Variables X.

A Confusion Matrix Is A Table That Will Be Useful In Describing A Selected Classifier 'S Functionality Over A Cluster Of Test Data That Already Recognizes The Actual Values For. It Is Greatly Valuable For Obtaining Precision, Recall, Specificity, Sensitivity And Accuracy[16-17].

Table 3. Performance Of Knn

Class	Precision	Recall	F1-Score	Support
0(Low)	0.84	1.00	0.91	16
1(High)	0.00	0.00	0.00	3

Table 4. Performance Of Svm -Linear

Class	Precision	Recall	F1-Score	Support
0(Low)	0.84	1.00	0.91	16
1(High)	0.00	0.00	0.00	3

Table 5. Performance Of Svm -Rbf

Class	Precision	Recall	F1-Score	Support
0(Low)	0.87	0.81	0.84	16
1(High)	0.25	0.33	0.29	3

Table 6. Performance Of Random Forest

Class	Precision	Recall	F1-Score	Support
0(Low)	0.84	1.00	0.91	16
1(High)	0.00	0.00	0.00	3

In Statistical Significance Checks For The Comparison Of Machine Learning Based Algorithms. It Is Shown In Table 7. In General, The Probability Of Two Data Samples Being Observed Is Quantified By A Statistical Hypothesis Check To Fit Samples, Despite The Assumption That The Samples Have An Equal Distribution.

Table 7. Performance Of Statistical Tests

Statistical Hypothesis Tests	Statistical Value	P-Value
T-Test	1.370	0.174
Mann-Whitney U-Test	1142.0	1.244
Wilcoxon Signod Rank Test	1182.0	0.268
Kruskal-Wallis H Test	0.148	0.099

The Statistical Test Provide A P-Value And A Test Statistic. These Test Shows The Null Hypothesis That There's No Difference If We Should Reject Or Fail To Reject. When The P-Value Is Below A Given Threshold (Often 0.05), It Will Reject The Null Hypothesis And The Difference May Be Statistically Important.

6. CONCLUSION

Using Data Mining Classification Techniques, A Research Method For Predicting The Lung Cancer Disease Is Developed. These Program Draws A Hidden Knowledge From A Database Of Historic Lung Cancer Disease. Here, We Can Conclude That Support Vector Machine And Logistic Regression Gives Better Accuracy In Lung Cancer Prediction. In Certain Cases, The Signs Associated With Lung Cancer Are Not Present In The Advanced Stage Of Lung Cancer Patients. Because Of The Lack Of Awareness, There Are Many Patients Is Needed To

Reduce Life Losses. The Method Used To Predict Lung Cancer May Be Further Enhanced And Extended. Other Data Mining Techniques Can Be Introduced As Well. For Starters, The Time Series, The Clustering And The Rules Of Association. We Would Like To Build Web-Based Software In Future Work To Evaluate The Output Of Various Classifiers Where Users Can Simply Upload Their Data Set And Evaluate The Results On The Fly

7. REFERENCES

- [1] Tapas Ranjanbhitharu , Subhendu Kumar Pani A, “ Comparative Study Of Data Mining Classification Techniques Using Lung Cancer Data”, International Journal Of Computer Trends And Technology, Vol. 22, 2015.
- [2] Prashan Naresh, “Early Dectection Of Lung Cancer Using Neural Network Techniques”, Journal Of Engineering Research And Application, Vol.4, No.8, 2014.
- [3] Van Belle V, Pelckmans K, Van Huffel S And Suykens, “Comparision Between Ranking And Regression”, Doi: **10.1016/J.Artmed.2011.06.006, Aug. 2011.**
- [4] Runxuan Zhang, Guang-Bin Huang, N.Sundararajanand P.Saratchandran, “Multicaegory Classification Usingan Exreme Learning Machine For Microarray”, Transactions On Computational Biology And Bioinformaics, Vol. 4, No.3, 2007.
- [5] Shupreet Karu And Dr.R.K.Bawa, “ Future Trends Of Data Mining In Predicting The Various Diseases In Medical Healthcare System”, International Journal Of Energy, Information And Communications , Vol. 6, No.4, 2015.
- [6] N.V.Ramana Murty, “A Critical Study Of Classification Algorithms For Lung Cancer Disease Detection And Diagnosis”, International Journal Of Computational Intelligence Research, Pp. 1041-1048.
- [7] Satyam Shukla,” Comparitive Study Of Recent Trends On Cancer Diseases Prediction Using Data Mining Techniques”, International Journal Of Database Theory And Application, Vol.9, No.9, 2016.
- [8] J.Jamera Banu, “A Study On Mining Lung Cancer Data For Increasing Or Decreasing Disease Prediction Value By Using Ant Colony Optimization Techniques”, International Journal Of Advanced Networking And Applications, 2015.
- [9] Neha Panpaliya, Neha Tada Surabhi Bobade, Rewti Aglawe, Akshay Gudadhe, “A Survey On Early Dectection And Prediction Of Lung Cancer “ , International Journal Of Computer Science And Mobile Computing, Vol.4 Issue.1, January- 2015, Pg. 175-184.
- [10] M.Pyingkodi, .S.Shanthi, T.M.Saravanan, K Thenmozhi, K. Nanthini, D.Hemalatha, M. Muthukumaran, M. Dhivya, 2020, ‘Performance Study Of Classification Algorithms Using The Microarray Breast Cancer Dataset’, International Journal Of Future Generation Communication And Networking, Vol. 13, No. 2, Pp. 1238-1245.

- [11] M.Pyngkodi, S.Shanthi, "Composition Of Feature Relevancy Based Biomarker Gene Selection In Gene Expression Dataset", International Journal Of Recent Technology And Engineering (Ijrte) Vol.8, No.4, Pp. November 2019
- [12] Pyngkodi, M., Shanthi, S., Muthukumar, M., Nanthini, K., Thenmozhi, K., "Hybrid Bee Colony And Weighted Ranking Firefly Optimization For Cancer Detection From Gene Regulatory Sequences", International Journal Of Scientific And Technology Research, 2020
- [13] K Thenmozhi, N Karthikeyani Visalakshi, S Shanthi, M Pyngkodi, "Distributed Icsa Clustering Approach For Large Scale Protein Sequences And Cancer Diagnosis", Asian Pacific Journal Of Cancer Prevention: Apjcp 19 (11), 3105, 2018.
- [14] Kawsar Ahmed, Abdullah-Ai-Emran, Tasnuba Jesmi, Roushney Fatima Mukti, "Early Detection Of Lung Cancer Risk Using Data Mining" , Asian Pacific Journal Of Cancer Prevention , **Vol 14, 2013.**
- [15] M Pyngkodi And R.Thangarajan, "Informative Gene Selection For Cancer Classification With Microarray Data Using A Metaheuristic Framework", Asian Pacific Journal Of Cancer Prevention Vol 19, 2018.
- [16] Animesh Hazra, Nanigopal Bera, Avijit Mandal, " Predicting Lung Cancer Survivability Using Svm And Logistic Regression Algorithms", International Journal Of Computer Applications, Vol. 174, No.2, 2017.
- K. Santra, And C. Josephine Christy, "Genetic Algorithm And Confusion Matrix For Document", Science Issues, Vol. 9, Issue 1, No. 2, 2012.
- [17] Sujatha, K & Shalini Punithavathani, D, 'Optimized Ensemble Decision-Based Multi-Focus Image Fusion Using Binary Genetic Grey-Wolf Optimizer In Camera Sensor Networks', Spinger, Doi: 10.1007/S11042-016-4312-3, Multimedia Tools And Applications
- [18] V.R. Balaji, Maheswaran S, M. Rajesh Babu, M. Kowsigan, Prabhu E., Venkatachalam K, Combining Statistical Models Using Modified Spectral Subtraction Method For Embedded System, Microprocessors And Microsystems, Volume 73, 2020
- [19] M. Kowsigan, J. Rajeshkumar, B. Baranidharan, N. Prasath, S. Nalini, And K. Venkatachalam, "A Novel Intrusion Detection System To Alleviate The Black Hole Attacks To Improve The Security And Performance Of The Manet," *Wireless Personal Communications*, Pp. 1-21, 2021. Lustering", Ijcsi International Journal Of Computer