International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Effect Of Principal Component Analysis In Lung Cancer detection using Machine Learning Techniques

Firoz Sajad [1], Vignesh. V [2], Chanchal. C. L [3] ,Jishnu. V [4] ,Anoop. P. S.[5]

[1,2,3,4]UG Student, Department of Mechanical Engineering, MES Institute of Technology & Management, Kerala, India

[5]Asst. Professor, Department of Mechanical Engineering, MES Institute of Technology & Management, Kerala, India

Abstract - *Lung cancer is the leading cause of cancer-related mortality around the world. The lung cancer contributes over 12.3 percentage of the total number of new cases diagnosed in the recent years. Smoking is regarded as one of the main cause of lung cancer. The wide range of causes and symptoms shows the complexity of the problem. Currently there are so many tests to diagnose cancer. Imaging techniques like x-ray , CT scan are used to diagnose the disease . By x-ray the image of the lungs may reveal an abnormal mass or nodule. A CT scan can reveal small lesions in the lungs that might not be detected on an X-ray. Also, sputum cytology tests done for persons who having cough and continuously producing sputum. Biopsy tests are done for detailed diagnosis of the disease. But for accurate results it take much time to test in laboratory and to detect cancer. Here, data mining techniques of machine learning are used experimentally to detect cancer more frequently. The data of lung cancer is collected from UCI repository, It is then filtered and classified using various algorithms to find the best method. The algorithms are also used in combination with principal components and both of the results are compared. This technique will help us to detect the lung cancer earlier and can save many lives in future.*

***Key Words*: Lung cancer, Principal component Analysis(PCA), weka, Random forest, Kstar**

1.INTRODUCTION

Machine learning is the branch of artificial intelligence, which is the frequently emerging technology and that can make a huge impact to the whole humanity in future. It is the study of algorithms and statistical models by using computer systems. The algorithms used in machine learning is to make mathematical models based on the given sample data also known as training data. It is used to make decisions or predictions without being explicitly programmed to perform a task. Data mining is a field of study within machine learning, and focuses on data analysis through unsupervised learning. Data mining uses many machine learning methods, but with different motives. On the other side, machine learning also employs data mining methods as unsupervised learning or as a pre-processing step to improve learner accuracy. There are many applications for machine automating the process of automation. Here our motive is to bring machine learning applications in the field of lung cancer detection. Since, cancer is the mostly detected and dangerous in humans , of that lung cancer is the most death rates . Due to the lack of detection of the disease at earlier stages ,it is hard for people to recover. Even though there are new technologies available to detect by combining machine learning we can generate more accurate results. It is a boon that by combining mL techniques, the disease can be detected at its earlier stages.

Here, initially the data collected is made to feature selection and classification. The filtering is done by J48 algorithm and principal components and both of the results are compared. Finding the best suits for filtering and classifying. The device used here is the Waikato Environment for knowledge Analysis (WEKA). It is very helpful to use the data mining techniques. Since, the advantage of weka is that it can manipulate easily and without the need of graphs it can learn graphs. Various data mining tasks can be done in weka like feature selection, regression, classification, clustering, visualization. etc.

2. LITERATURE SURVEY

The basic idea of using machine learning tasks in cancer prognosis or fault detection has been reinvented many times. One notable work is of [Konstantina Kourou, Themis P. Exarchos , Konstantinos P. Exarchos, Michalis V. Karamouzis,Dimitrios I. Fotiadis et.al. (2015)] ,they proposed the use of machine learning applications in cancer prognosis and prediction. They presented the studies based on various ML techniques used in cancer prognosis ,like the use of ANNs,SVM,SSL based on clinical data set and SEER datasets. A Survey on Hoeffding tree stream data classification algorithms is done by [Arvind Kumar, Parminder Kaur,Pratibha Sharma et.al. (2015)] proposed various decision trees,algorithms,data mining techniques used in machine learning. One important work done by [Jyotismita Talukdar, Dr. Sanjib, Kr. KalitaInt et.al. (2015)] is about breast cancer detection using data mining tool in Weka. They collected various clinical datasets and the attributes of the data is round up to 10 attributes. Finally compared the accuracies given by two classifier algorithms mainly, ZeroR and J48.

Classification Performance Using Principal Component Analysis and Different Value of the Ratio R is done by [J. Novakovic, S. Rankov et.al. (2011)], which discusses data dimensionality reduction and using various methods to overcome this. Also one prominent work in other field is the work by [Rebecca Jeya Vadhanam, S. Mohan, V.V. Ramalingam and V. Sugumaran et.al. (2016)] is the performance comparison of various decision tree algorithms for the classification of advertisement and non advertisement videos. Here, the recordings are recorded in MPEG format of size 1024 x 1024 and block intensity comparison code(BICC) is applied to various block in the frames. Classification is done by tree algorithms like, J48,J48 graft,LMT,Random tree,BF tree,Rep tree and NB tree, And random tree got the most accuracy of 92.085%.

The importance of tree algorithms is shown in another work by [B.R.Manju , A.Joshuva , V. Sugumaran et.al. (2018)] .Here ,the detection of faults in wind turbine blades is done by analysing the vibration signals using adhesive mounting technique and the classification is done using the J48 algorithm and finally hoeffding tree to check the classification accuracy. One notable work in medical field by [Nagesh Shukla, Markus Hagenbuchner,Khin Than Win and Jack Yang et.al. (2018)] is to predict the breast cancer survivability. They used SOM algorithm is used for data mining process and DBSCAN to check the area of high density in the dataset and uses the dataset available in SEER program. A detailed study of PCA is done by [Liton Chandra Paul, Abdulla Al Suman and Nahid Sultan et.al. (2013)] . A methodological analysis of dimension reduction problems is performed in this paper. Also, Principal Component Analysis in ECG Signal Processing done by [Francisco Castells, Pablo Laguna, Leif Sörnmo, Andreas Bollmann, and Jose Millet Roig et.al. (2007)]. Here, the heart beat signals are extracted by a QRS detector The signal segment of a beat is represented by a column.Then,Body surface potential mapping (BSPM) to the recording and analysis of temporal and spatial distributions of ECG potentials acquired multiple sites. In their work , [Cristinel Constantin et.al. (2014)] used PCA as a powerful marketing tool. Here for PCA computation SPSS systems are used.

Another notable work in the field of breast cancer done by [Amna Ali, Kanghee Park , Dokyoon Kim, Yeolwoo An, Minkoo Kim and Hyunjung Shin et.al. (2013)] , Here the SEER dataset is used. Prediction accuracy is measured by entries in the confusion matrix. ANNs are used as the encoding and solving methods. And got 71% of classification accuracy. A work by [Dr. Prof. Neeraj,Sakshi Sharma,Renuka Purohit,Pramod Singh Rathore et.al. (2017)] uses J48 for the prediction of cancer recurrence. From the result of the experiment they concludes that patient with specific range of attribute value have more chances of recurrence cancer. A contribution to breast cancer survivability by [Rohit J. Kate and Ramya Nadig et.al. (2017)] . They collected data from SEER dataset, and used Naive bayes ,logistic regression and decision tree to predict cancer survivability. And got an overall accuracy of 92.50%. Reducing online threats and viruses by adaptive statistical compression algorithms (Dynamic Markov Compression (DMC) and Prediction by Partial Matching (PPM)) is depicted in the work of [Philip K. Chan and Richard P. Lippmann et.al. (2006)] . Here , Standard optical character recognition (OCR) software is used to extract words embedded in images and these extra words are used in addition to text in the email header and body to improve performance of a support vector machine spam classifier. The application of C4.5 algorithm to evaluate the damages and faults occur in single point cutting tool is depicted in the work of [M.Elangovan,S.Babu Devasenapati, N.R.Sakthivel and K.I.Ramachandran et.al. (2011)]. Where, the extraction of data in the form of vibration signals is done and compared the classification accuracy of PCA , C4.5 and decision tree . Finally ,concludes that decision tree with a high accuracy of 77.22%. A notable work by [Nour El Islem Karabadji, Hassina Seridi, Fouad Bousetouane, Wajdi Dhifli and Sabeur Aridhi et.al. (2017)].Here, they proposed to use good sub-training and sub-testing samples and only a subset of pertinent attributes to construct an optimal DT with respect to the input dataset. Classifiers are used in visual inspection process to examine the faults is proposed by [S. Ravikumar,K.I. Ramachandran and V. Sugumaran et.al. (2011)] ,by checking the salient features by taking images on various angles. These features have different values for the defects considered namely, sheets without scratches, sheets with minor scratches and sheets with deep scratches. Here the classifier used is C4.5 and Naïve Bayes in combination with the histogram features extracted from images.

Non linear PCA can eliminate any type of non-linear correlation occurring in the data[Mark A. Kramer et.al. (1991)]. A notable study in 3 point neural networks is done by [A. L. Blum , R. L. Rivest et.al. (1989)]. Principal component analysis is central to the study of multivariate data [I. T. Jolliffe et.al. (1986)]. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network.[V. Krishnaiah, Dr. G. Narasimha, Dr. N. Subash chandra et.al. (2013)]. No major organization recommends screening for early detection of lung cancer, although screening has interested researchers and physicians. Smoking cessation remains the critical component of preventive primary care [Lauren G. Collins, M.D., Christopher Haines, M.D., Robert Perkel, M.D., and Robert E. Enck et.al. (2006)]. The combination of neural network

classifier along with binarization and GLCM will increase the accuracy of lung cancer detection process. This system will also decrease the cost and time required for cancer detection.[Neha panpaliya et.al. (2015)]. C4.5 and PCA-based diagnosis method has higher accuracy and needs less training time than BPNN in the fault diagnosis of rotating machinery. [Sun, W., Chen, J., & Li, J. et.al. (2007)].

3. METHODOLOGY

There are some steps to be followed in the process of lung cancer prediction using machine learning techniques. Initially the data representing the features of lung cancer of many patients are collected from UCI repository. The data must contain important features regarding lung cancer. The data is initially made to data pre processing i.e. to eliminate the unwanted features like age, sex etc. These processed data is made to training datasets. The figure given below shows the flowchart representation of the process.

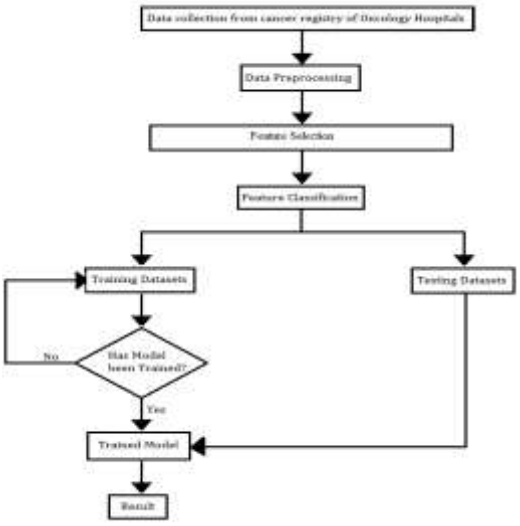

Figure-1: Methodology

The data sets check whether the model is trained or not. This process is continued until the model is trained

Initially, the prominent features of the lung cancer is to be extracted. The features maybe the details of the cells, symptoms or any other factors related to the disease. Here, the data collected is initially feature extracted data, so going to the next step.

3.2 Data pre-processing

Here, the original data is collected from the UCI repository. It contains 24 attributes, 599 instances and 3 classes. It contains certain unimportant factors in the data (like, name ,age ,sex. Etc.). So this factors are eliminated in the pre-processing stage. Now ,the data contains 21 attributes, 599 instances and 3 classes.

3.2 Feature selection

3.2.1 Using J48 decision tree*:-* The data collected is now made to feature selection. J48 is a decision tree algorithm which can be used for feature selection. I.e. when one attribute is selected and removed. The accuracy is checked after classifying using J48. And compares it with initial accuracy. If the accuracy got is greater than the previous accuracy. Then that attribute is eliminated permanently. Or if it is lesser then that attribute is valuable.

3.2.2 Using principal component analysis*:-* The principal components is used as a filter. So that it will randomly select the prominent features and eliminates the unimportant features. PCA is used in combination with various algorithms to classify. Thus, the time and load is reduced. PCA can be regarded as an unsupervised learning method which uses statistical methods in machine learning. It uses the ranker search method. Initially, the mean is subtracted and forms covariance matrix. Then generates the eigenvalues and eigenvectors of the covariance matrix. Then choosing components form a feature vector. This is constructed by taking the eigenvectors that you want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.

$$\text{Feature vector} = (\text{eig}_1\ \text{eig}_2\ \text{eig}_3\ \ldots.\ \text{eig}_n)$$

Finally new data is derived.[Lindsay I Smith(2002)]

3.3 Feature classification

After the feature selection feature classification is done. Here, the classification is done by using four algorithms.

3.3.1 Classification using random forest *:-* Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. A random forest is a classifier consisting of a collection of tree structured classifiers {h(x,Θk), k=1, ...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.[Leo Breiman(2001)]

3.3.2 Classification using Kstar *:-* The Kstar algorithm can be defined as a clustering method that divides *n* data into *k* clusters, where each data entry in a particular cluster with an average viewing distance nearby. The Kstar algorithm is an instance-based learner algorithm

that uses entropy to measure the distance [Wiharto W., MCom, Hari Kusnanto, DrPH and Herianto H., DrEng(2016)]

4. RESULTS AND DISCUSSION

4.1 Classification using Kstar

The table 1 shows the stratified cross validation details of the classifier, Table 2 gives the detailed accuracy by class, Table 3 shows the confusion matrix and Table 4 gives values for objects of the trained Kstar algorithm. Tables 5,6,7,8 shows corresponding values using pca filter.

Classification via Kstar without using PCA filter:-

Table-1: Stratified cross validation

Summary	
Correctly Classified Instances	589
Incorrectly Classified Instances	10
Kappa statistic	0.9749
Mean absolute error	0.0313
Root mean squared error	0.1134
Relative absolute error	7.0641%
Root relative squared error	24.0647%
Total number of instances	599

Table-2 : Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.005	0.990	1.000	0.995	0.992	0.997	0.979	Low
0.979	0.015	0.969	0.979	0.974	0.962	0.980	0.972	Medium
0.972	0.005	0.991	0.972	0.981	0.971	0.987	0.979	High
0.983	0.008	0.983	0.983	0.983	0.975	0.988	0.977	

Table-3: Confusion matrix

Classified as	Low	Medium	High
Low	190	0	0
Medium	2	190	2
High	0	6	209

Table-4: Value of objects trained by Kstar

Attribute	Values
Global blend(B)	80

The confusion matrix table(table 3) indicates that 190/190 samples were correctly classified as low, then 190/194 samples were correctly classified as medium and 209/215 instances were correctly classified as high. The classifier got a maximum accuracy of 98.3306% after training without principal components.

Classification via Kstar with using PCA filter:-

Table-5 : Stratified cross validation

Summary	
Correctly Classified Instances	589
Incorrectly Classified Instances	10
Kappa statistic	0.9749
Mean absolute error	0.0391
Root mean squared error	0.1149
Relative absolute error	8.8153%
Root relative squared error	24.3864%
Total number of instances	599

Table-6 :Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.005	0.990	1.000	0.995	0.992	0.996	0.975	Low
0.979	0.015	0.969	0.979	0.974	0.962	0.978	0.968	Medium
0.982	0.005	0.991	0.982	0.981	0.971	0.985	0.979	High
0.983	0.008	0.983	0.983	0.983	0.975	0.986	0.974	

Table-7: Confusion matrix

Classified as	Low	Medium	High
Low	190	0	0
Medium	2	190	2
High	0	6	209

Table-8: Value of objects trained by Kstar

Attribute	Values
Global blend(B)	80

The confusion matrix table(table 7) indicates that 190/190 samples were correctly classified as low, then 190/194 samples were correctly classified as medium and 209/215 instances were correctly classified as high. The classifier got a maximum accuracy of 98.3306% after training with principal components.

The classifier depends on the variable global blend. The variation of this parameter with & without pca vs. the classification accuracy of algorithm is shown in chart 1.

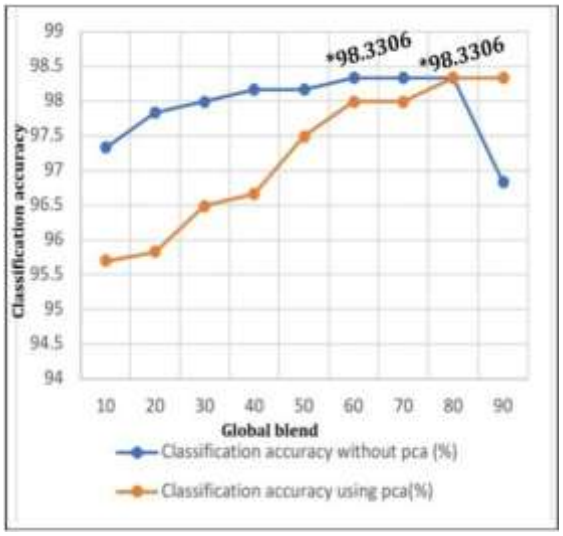

Chart-1: Global blend vs classification accuracy

The chart-1 shows the global blend vs classification accuracy. Here, x-axis denotes the global blend and y-axis denotes the classification accuracy.

The parameter 'Global blend' is varied from 10 to 100 , without PCA in steps of '20' it gradually increases and became constant at '60', after '80' it drops down. When using PCA the value initially start increasing and gradually increases to '80' and became constant.

4.2 Classification using random forest

The table 9 shows the stratified cross validation details of the classifier, Table 10 gives the detailed accuracy by class, Table 11 shows the confusion matrix and Table 12 gives values for objects of the trained random forest and tables 13,14,15,16 shows corresponding values using pca filter.

Classification via random forest without using PCA filter:-

Table-9: Stratified cross validation

Summary	
Correctly Classified Instances	589
Incorrectly Classified Instances	10
Kappa statistic	0.9749
Mean absolute error	0.0662
Root mean squared error	0.137

Relative absolute error	14.9129%
Root relative squared error	29.0811%
Total number of instances	599

Table-10: Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.005	0.990	1.000	0.995	0.992	0.997	0.979	Low
0.979	0.015	0.969	0.979	0.974	0.962	0.981	0.966	Medium
0.972	0.005	0.991	0.972	0.981	0.971	0.985	0.975	High
0.983	0.008	0.983	0.983	0.983	0.975	0.988	0.973	

Table-11: Confusion matrix

Classified as	Low	Medium	High
Low	190	0	0
Medium	2	190	2
High	0	6	209

Table-12: Value of objects trained by random forest

Attribute	Values
Number of iterations(I)	100
Number of features(K)	1
Seed(S)	1

The confusion matrix table(table 11) indicates that 190/190 samples were correctly classified as low, then 190/194 samples were correctly classified as medium and 209/215 instances were correctly classified as high. The classifier got a maximum accuracy of 98.3306% after training without principal components.

Classification via random forest with using PCA filter:-

Table-13: Stratified cross validation

Summary	
Correctly Classified Instances	589
Incorrectly Classified Instances	10
Kappa statistic	0.9749
Mean absolute error	0.061
Root mean squared error	0.1344
Relative absolute error	13.7507%
Root relative squared error	28.5252%
Total number of instances	599

Table-14: Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.005	0.990	1.000	0.995	0.992	0.997	0.988	Low
0.979	0.015	0.969	0.979	0.974	0.962	0.982	0.962	Medium
0.972	0.005	0.991	0.972	0.981	0.971	0.986	0.976	High

0.983	0.008	0.983	0.983	0.983	0.975	0.988	0.975	

Table-15: Confusion matrix

Classified as	Low	Medium	High
Low	190	0	0
Medium	2	190	2
High	0	6	209

Table-16: Value of objects trained by random forest

Attribute	Values
Number of iterations(I)	100
Number of features(K)	0
Seed(S)	1

The confusion matrix table(table 15) indicates that 190/190 samples were correctly classified as low, then 190/194 samples were correctly classified as medium and 209/215 instances were correctly classified as high. The classifier got a maximum accuracy of 98.3306% after training with principal components.

The classifier depends on three variables which are number of iterations, number of features and seeds. The variation of these parameters with & without PCA vs. the algorithms classification accuracy is plotted in charts 2-4.

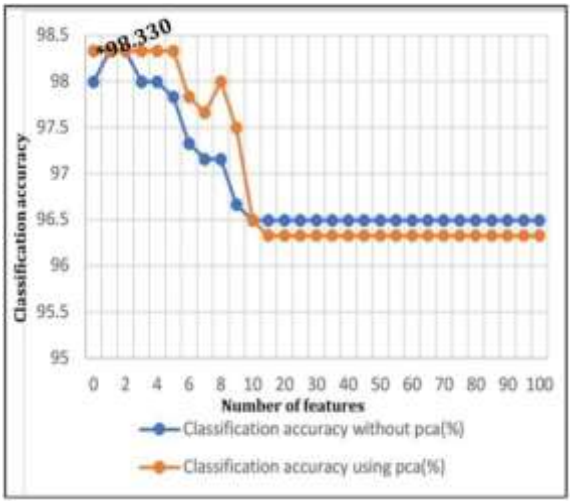

Chart-2: Number of features vs classification accuracy

The chart-2 shows the variation of number of features vs classification accuracy. The x-axis shows number of features and y-axis shows the classification accuracy.

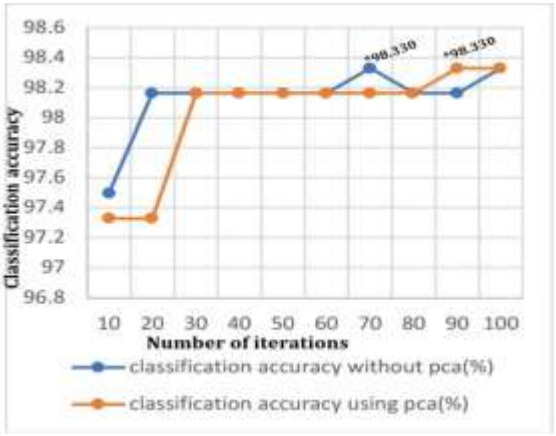

Chart-3: Seed vs classification accuracy

The chart -3 shows the variation of seed vs classification accuracy. The x-axis shows the seed and y-axis shows classification accuracy.

By varying the parameter 'Number of features' from 0 to 100 without PCA in steps of '1' gives the maximum accuracy then gradually decreases and reaches a constant value at '10'. When combined with PCA up to '6' the value is constant , then gradually decreases and reaches a constant value at '20'. Also, the parameter 'seed' is varied from 1 to 100 when classified without PCA the value is collected up to '15' and then drops down at '20' , then regains the value at '25' and became constant. When combined PCA, the value is initially constant at sometime and drops at some point then

regains the constant value. This process goes on up to '100'.

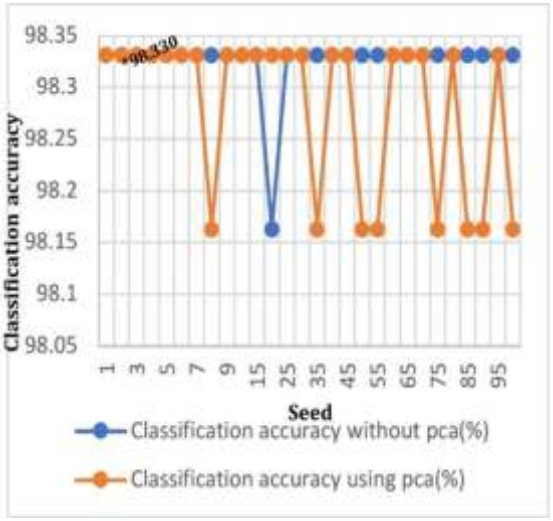

Chart-4: Number of iterations vs classification accuracy

The chart-4 shows the variation of number of iterations vs classification accuracy. The x-axis shows the number of iterations and y-axis shows the classification accuracy.

When the parameter 'Number of iterations' is varied from 10 to 100. Initially when used without PCA the value initially increases from 10 to 20 and became constant up to '60' and then fluctuates. When combined with PCA, initially the value is constant up to '20' and then increases to a value at '30' and became constant up to '80'. And then fluctuates.

5. CONCLUSION

Data mining techniques helps to eradicate information from a large dataset. In the health care field a huge amount of data on diseases are available. Here the filter methods PCA and J48 is compared. For random forest and Kstar got a higher accuracy of 98.3306% either filtered with or without PCA. So here we can conclude that random forest and kstar can be either combined with or without PCA for better accuracy in the prediction of lung cancer. And PCA filter is giving exact same accuracy as that got when J48 is used as a filter. Thus principal components can give no effect to the data as a filter.

REFERENCES

[1] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis and Dimitrios I. Fotiadis, "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal,vol.13 ,Nov. 2014,pp.8–17

[2] Arvind Kumar, Parminder Kaur and Pratibha Sharma, "A Survey on Hoeffding Tree Stream Data Classification Algorithms" ,CPUH-Research Journal, vol.1,Issue.2, Jan.2015, pp.28-32,ISSN : 2455-6076

[3] Jyotismita Talukdar, Dr. Sanjib and Kr. KalitaInt ,"Detection of Breast Cancer using Data Mining Tool (WEKA)" , International Journal of Scientific & Engineering Research, Vol.6, Issue.11, Nov. 2015., ISSN : 2229-5518

[4] J. Novakovic, S. Rankov , "Classification Performance Using Principal Component Analysis and Different Value of the Ratio R" ,Int. J. of Computers, Communications & Control,Vol.6 , June. 2011., pp. 317-327,ISSN 1841-9836

[5] B. Rebecca Jeya Vadhanam, S. Mohan,V.V. Ramalingam and V. Sugumaran, "Performance Comparison of Various Decision Tree Algorithms for Classification of Advertisement and Non Advertisement Videos", Indian Journal of Science and Technology, Vol.9, Issue.48 , Dec.2016.,ISSN : 0974-5645

[6] B.R. Manju,A. Joshuva and V. Sugumaran , "A data mining study for condition monitoring on wind turbine blades hoeffding tree algorithm through statistical and histogram features", International Journal of Mechanical Engineering and Technology,Volume.9, Issue.1, Jan.2018., pp. 1061–1079,ISSN: 0976-6359

[7] Nagesh Shukla, Markus Hagen Buchner, Khin Than Win and Jack Yang, "Breast cancer data analysis for survivability studies and prediction", Computer Methods and Programs in Biomedicine ,vol.155,Jan.2018,pp.199–208

[8] Liton Chandra Paul, Abdulla Al Suman and Nahid Sultan, "Methodological Analysis of Principal Component Analysis (PCA) Method", IJCEM International Journal of Computational Engineering & Management, Vol. 16 Issue 2,Mar.2013, ISSN: 2230-7893

[9] Francisco Castells, Pablo Laguna, Leif Sornmo, Andreas Bollmann and Jose Millet Roig , "Principal Component Analysis in ECG Signal Processing", Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing Volume, Jan.2007

[10] Cristinel Constantin, "Principal component analysis - A powerful tool in computing marketing information", Bulletin of the Transylvania University of Braşov Series V: Economic Sciences, Vol. 7 (56), No.2, Jan.2014

[11] Kanghee Park, Amna Ali, Dokyoon Kim, Yeolwoo An, Minkoo Kim and Hyunjung Shin, "Robust predictive model for evaluating breast cancer survivability", Engineering Applications of Artificial Intelligence,vol.26, Oct.2013., pp.2194–2205

[12] Dr Prof. Neeraj,Sakshi Sharma,Renuka Purohit and Pramod Singh Rathore , "Prediction of Recurrence Cancer using J48 Algorithm", Proceedings of the 2nd International Conference on Communication and Electronics Systems (ICCES), Jan.2017

[13] Rohit J. Kate and Ramya Nadig, "Stage-specific predictive models for breast cancer survivability",

International Journal of Medical Informatics, vol.97, Nov.2016.,pp.304–311

[14] Philip K. Chan and Richard P. Lippmann, "Machine Learning for Computer Security", Journal of Machine Learning Research,vol. 7, Dec.2006.,pp. 2669-2672

[15] M. Elangovan,S. Babu Devasenapati, N.R. Sakthivel and K.I. Ramachandran, "Evaluation of expert system for condition monitoring of a single point cutting tool using principle component analysis and decision tree algorithm", Expert Systems with Applications,vol. 38,Apr.2011, pp.4450–4459

[16] Nour El Islem Karabadji, Hassina Seridi, Fouad Bousetouane, Wajdi Dhifli and Sabeur Aridhi , "An evolutionary scheme for decision tree construction", Knowledge-Based Systems , Vol.119,Aug.2017,pp.166–177

[17] S. Ravi Kumar, K.I. Ramachandran and V. Sugumaran, "Machine learning approach for automated visual inspection of machine components", Expert Systems with Applications, Vol.38, Apr.2011,pp.3260–3266

[18] Mark A. Kramer, "Nonlinear Principal Component Analysis Using auto associative Neural Networks", AIChE Journal, Vol.37, No.2, Feb.1991, pp.233-243

[19] Avrim L. Blum & R. L. Rivest, "Training a 3-Node Neural Network is NP-Complete", Neural networks, Feb.1989., pp.494-501

[20] I.T. Jolliffe, " Principal Component Analysis", Springer-Verlag, New York, 1986.

[21] V. Krishnaiah , Dr. G. Narasimha, Dr. N. Subash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, Vol. 4 ,Issue.1 , Sep.2013, pp.39 – 45

[22] Lauren G. Collins, M.D., Christopher Haines, M.D., Robert Perkel, M.D. and Robert Enck E. M.D. , "Lung Cancer: Diagnosis and Management" ., Article in American family physician.

[23] Neha Panpaliya, Neha Tadas, Surabhi Bobade, Rewti Aglawe, Akshay Gudadhe, "A Survey On Early Detection and Prediction of lung cancer" , International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 1, Jan.2015,pp.175 – 184

[24] Sun. W., Chen. J., & Li. J, "Decision tree and PCA-based fault diagnosis of rotating machinery", Mechanical Systems and Signal Processing, Vol.21, Issue.3, Apr.2007,pp.1300–1317.

[25] Lindsay. I. Smith, "A tutorial on principal component analysis", Feb.2002

[26] Leo Breiman, "Random forests", Statistics Department, University of California, Berkeley, Jan.2001

[27] Wiharto W., MCom, Hari Kusnanto, DrPH and Herianto H., DrEng, "Intelligence System For Diagnosis Level of Coronary Heart disease with Kstar Algorithm" , Korean society of Informatics Research, Vol.22, Issue.1, Jan.2016, pp.30-38