Detection of Breast Cancer Using Machine Learning Techniques

RafeeK . A[1], Abhijith. B[2], Anandhu. P[3], Anoop. P.S[4]

[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 -*Breast cancer is one of the most common cancer among the women. Breast cancer can be classified into two benign and malignant. The etiological reason behind the cancer is due to the generation of superoxides or free radicles. X-Ray, MRI, Biopsy test are used to detect diseases in now a days. It take long time for diagnosis .In our study Data mining techniques of machine learning are used to detect disease. In this experiment, we uses different classifiers such as J48, LMT can be used. The breast cancer data set can be taken from UCI repository and it can be filtered with the help of random projection and find the appropriate method. The classifiers which can be used without filters gives the greater output at an accuracy of 97.36%.*

Key Words**: Breast cancer, Random Projection, LMT, weka, Random forest**

1. INTRODUCTION

Machine learning is the theory based on principle of computational statistics which focuses on making statement using computer. Data mining is a field of study within machine learning and focuses on research data analysis through unauthorized learning. Data mining uses many machine learning methods but with different goals, on the other side machine learning also work on data mining methods. Un authorized learning or as pre processing step to enhance learners accuracy. There are many applications for machine learning including agricultural anatomy, adaptive websites etc

Breast cancer is the second dangerous cancer after the lung cancer. According to data provided by World health organisation (WHO) two million new cases are reported and which of 626,679 were died in year 2018. At this situation the importance of machine learning can be realised.

For the early detection of disease machine learning is very helpful. Here for mostly data collected is made to feature selection and classification the separation is done by J48 algorithm and random projection and together the result are cross checked and finding the best filter and classifier.

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 analyzing 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 binarizationand 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. METHEDOLOGY

There are some steps to be followed in the process of breast cancer prediction using machine learning techniques. Initially the data representing the features of breast cancer of many patients are collected from UCI repository. The data must contain important features regarding breast 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 Flow chart representation of the process.

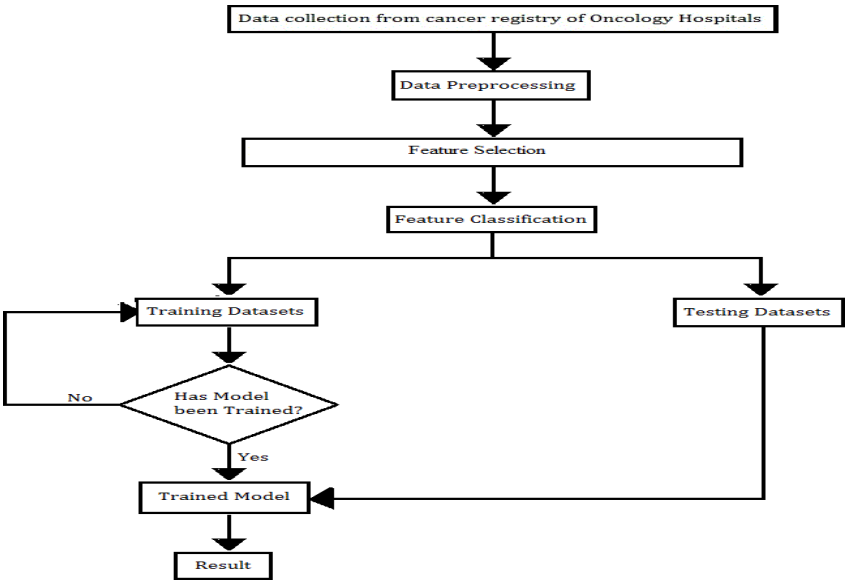

Figure-1: Methodology

3.1 Data pre-processing

The breast cancer data can be collected from the UCI repository. It contains 32 attributes, 569 instances, 2 classes. It contains the unwanted factors in the data (like name, age, sex etc) this can be eliminates at the pre processing stage now the data contain 8 attributes, 569 instances and 2 classes.

3.2 Feature selection

3.2.1 Using J48 Decision tree:- In this stage the collected data can be made to feature selection. J48 algorithm can be used for the feature selection process. The unwanted attributes can be eliminated with the help of J48 classifier.

3.2.2 Using random projection:- This is used as a filter. It will randomly selects the appropriate features and eliminate the unwanted features. It is generic and simple approach can be taken for extracting features from the data. It is very easy to implement and compute. It can be used for any conventional machine learning algoritm for clearing the task.

3.3 Feature classification

3.3.1 Classification using LMT:- It is a supervised learning algoritm which combines the logistic regression and decision tree. Its leaf load is logistic regression function

3.3.2 Classification using random committee:- It is a base classifier each base classifier is built using a different random number seed the final prediction is the straight averages of individual base classifiers.

RESULTS AND DISCUSSIONS

Classification using Logistic Model Tree (LMT)

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 logistic model tree Table 5,6,7,8 shows the corresponding values using random projection for filter.

Without using random projection filter

Table-1: Stratified cross validation

Summary	
Correctly Classified Instances	554
Incorrectly Classified Instances	15
Kappa statistic	0.9433
Mean absolute error	0.0581
Root mean squared error	0.1498
Relative absolute error	12.4308%
Root relative squared error	30.9857%
Total number of instances	569

Table-2 : Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F- Measure	MCC	ROC Area	PRC Area	Class
0.953	0.014	0.976	0.953	0.964	0.944	0.993	0.992	Low
0.986	0.047	0.972	0.986	0.979	0.944	0.993	0.994	Medium
0.974	0.035	0.974	0.974	0.974	0.944	0.993	0.993	High

Table-3: Confusion matrix

Malignant	Benign	Classified as
202	10	malignant
2	352	benign

Table-4: Value of objects trained by LMT

Attribute	Values
Number of boosting iterations	9
Minimum number instances(M)	1
Weight trim beta	0.2

The confusion matrix table (table 3) indicates that correctly classified as Malignant and Benign. The classifiers got maximum accuracy of 97.3638%after training without using random projection. *With using random projection filter*

Table -5: Stratified cross validation

Summary	
Correctly Classified Instances	541
Incorrectly Classified Instances	28
Kappa statistic	0.8938
Mean absolute error	0.0731
Root mean squared error	0.2
Relative absolute error	15.6322%
Root relative squared error	41.3718%
Total number of instances	599

Table-6: Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
	0.915	0.028	0.951	0.915	0.933	0.894	0.969	**M**
	0.972	0.085	0.951	0.972	0.961	0.884	0.969	**B**
Wgt Avg	0.951	0.064	0.951	0.951	0.951	0.894	0.969	

Table- 7: Confusion matrix

Malignant	Benign	Classified as
194	18	**malignant**
10	347	**benign**

Table- 8: Value for objects of the trained LMT

Attribute	Values
Number of Boosting interactions(I)	30
Minimum number instances(M)	1
Weight trim Beta(W)	0.1

The confusion matrix table(table 7) indicates that 190/190 samples were correctly classified as Malignant and Benign. The classifier got a maximum accuracy of 95.0791% after training with Random Projection.

The confusion matrix table(table7)indicates that correctly classified as Malignantand Benign**.** The classifier got maximum accuracy of 95.0791 % after training with Random Projection.

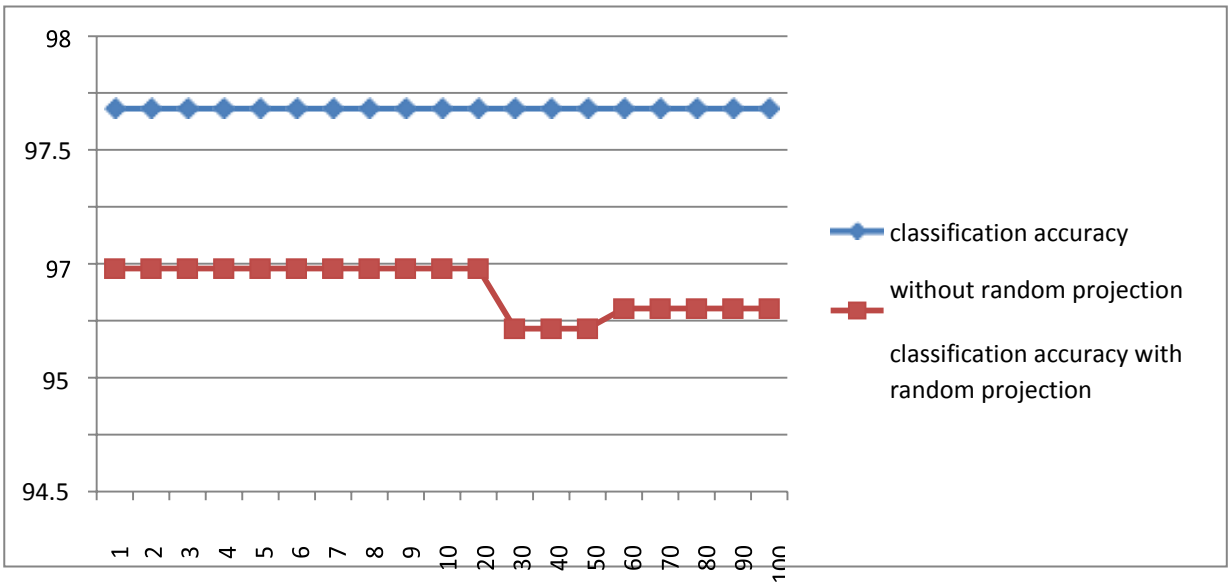

chart-1: Minimum no of instances V/S Classification Accuracy

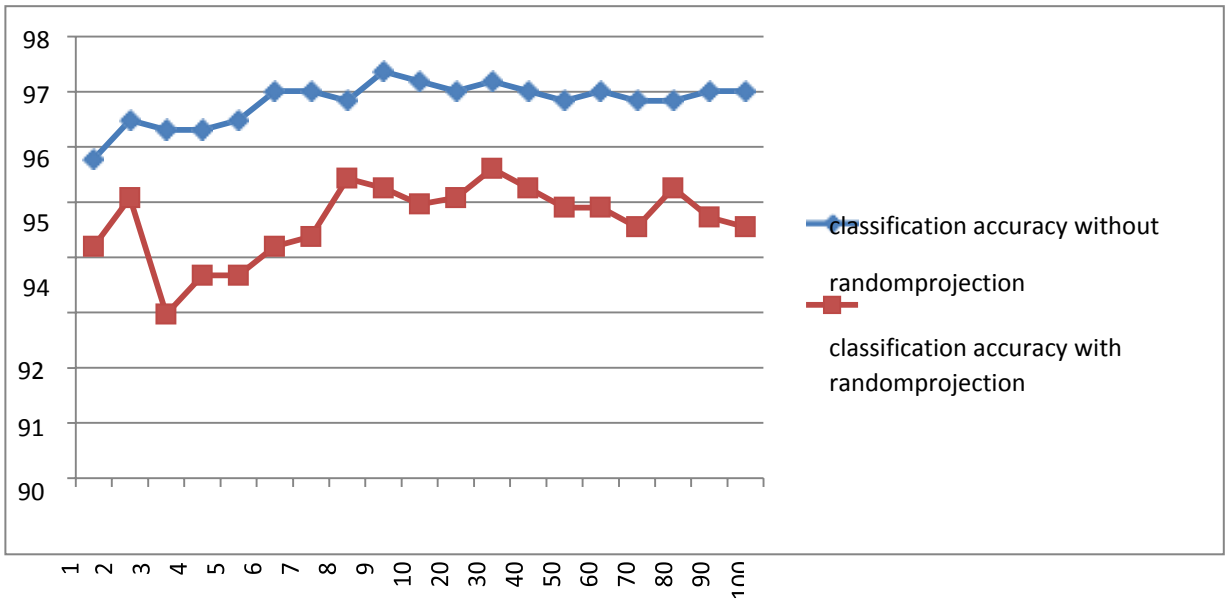

chart-2: Number of Boosting iterations v/s classification accuracy

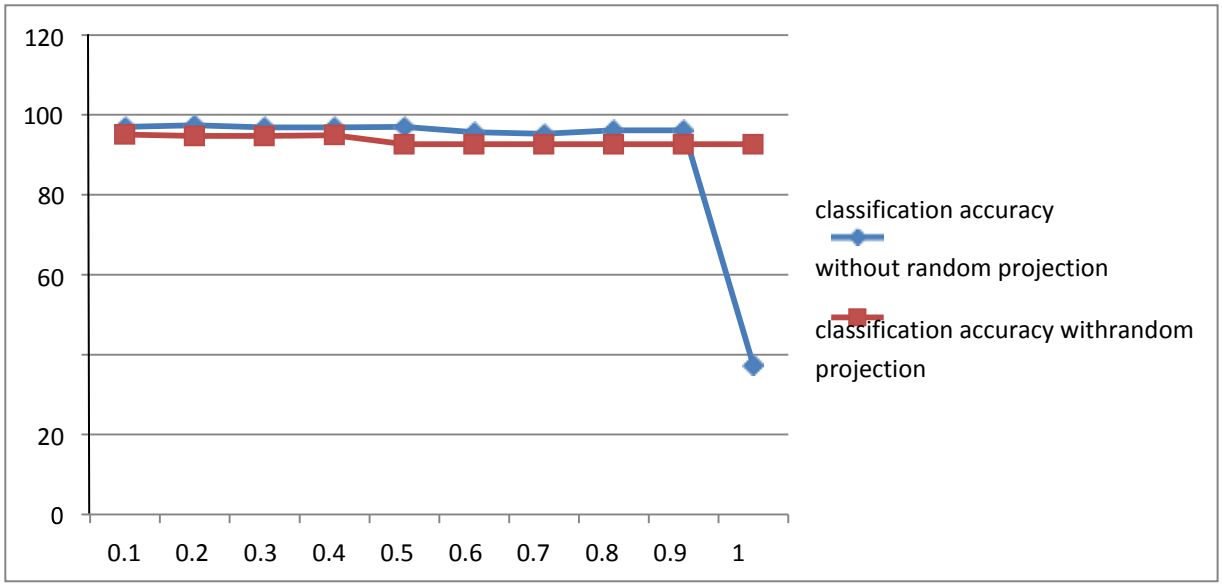

chart-3: Weight trim Beta v/s classification accuracy

Classification using J48 decision tree

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 Decision tree J48. Tables 13, 14, 15 & 16 shows corresponding data using random projection Classification Using J48

Without using random projection filter

Table-9:stratified cross validation

Summary	
Correctly Classified Instances	545
Incorrectly Classified Instances	24
Kappa statistic	0.9096
Mean absolute error	0.0553
Root mean squared error	0.2003
Relative absolute error	11.8282%
Root relative squared error	41.4225%
Table-11: Confusion matrix	
Total number of instances	569

Table-10: Confusion matrix

Malignant	Benign	Classified as
199	13	**malignant**
11	346	**benign**

Table-11: value for objects of the trained j48

Attribute	Values
Confidence Factor (C)	0.1
Minimum number of objects(M)	3

Table-12: Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
	0.939	0.031	0.948	0.939	0.943	0.910	0.943	**M**
	0.969	0.061	0.964	0.964	0.966	0.910	0.943	**B**
Wgt Avg	0.958	0.050	0.958	0.958	0.958	0.910	0.943	

The confusion matrix table (table11) indicates that correctly classified as Malignant and Benign**.** The classifier got a maximum accuracy of 95.7821 % after training without Random Projection. *With using random projection filter*

Table-13: stratified cross validation

Summary	
Correctly Classified Instances	534
Incorrectly Classified Instances	35
Kappa statistic	0.8678
Mean absolute error	0.0808
Root mean squared error	0.2412
Relative absolute error	17.2694%
Root relative squared error	49.893%
Total number of instances	569

Table-14: Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
	0.906	0.042	0.928	0.906	0.916	0.868	0.920	**M**
	0.958	0.094	0.945	0.958	0.951	0.868	0.920	**B**
Wgt Avg	0.938	0.075	0.938	0.938	0.938	0.868	0.920	

Table-15: Confusion matrix.

Malignant	Benign	Classified as
192	20	**malignant**
15	342	**benign**

Table-16: value for objects of the trained j48.

Attribute	Values
Confidence Factor (C)	0.1
Minimum number of objects(M)	1

Table-16: value for objects of the trained j48

The confusion matrix table(table15)indicates that correctly classified as MalignantandB enign**.** The classifier got a maximum accuracy of 93.8489 % after training with RandomProjectio

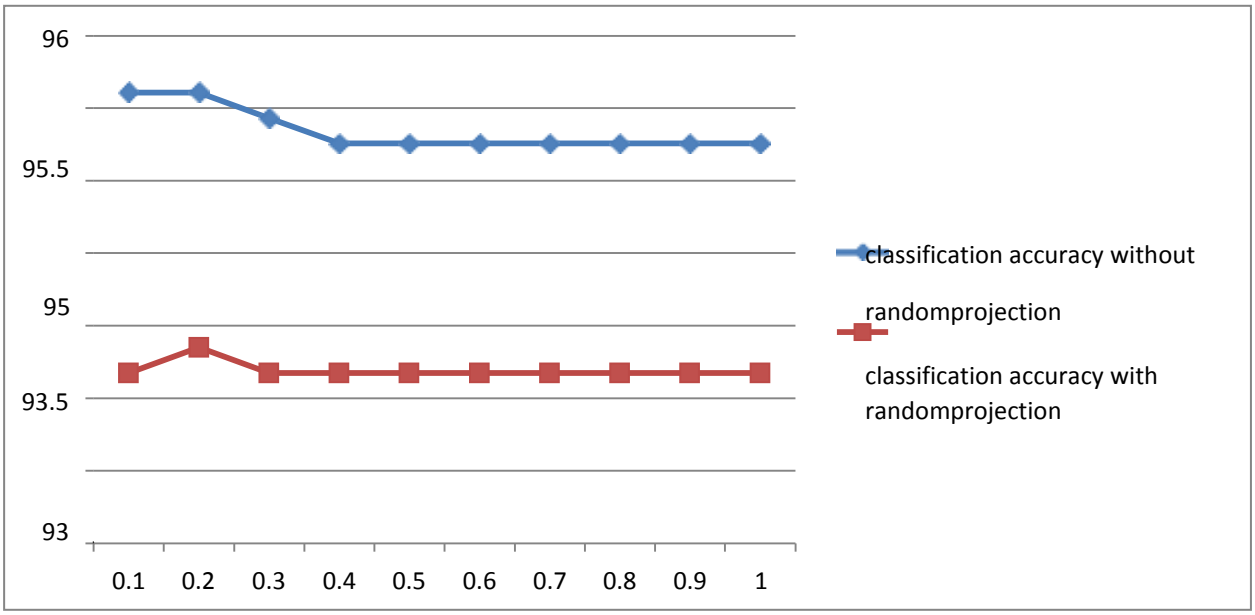

chart-4: confidence factor v/s classification accuracy

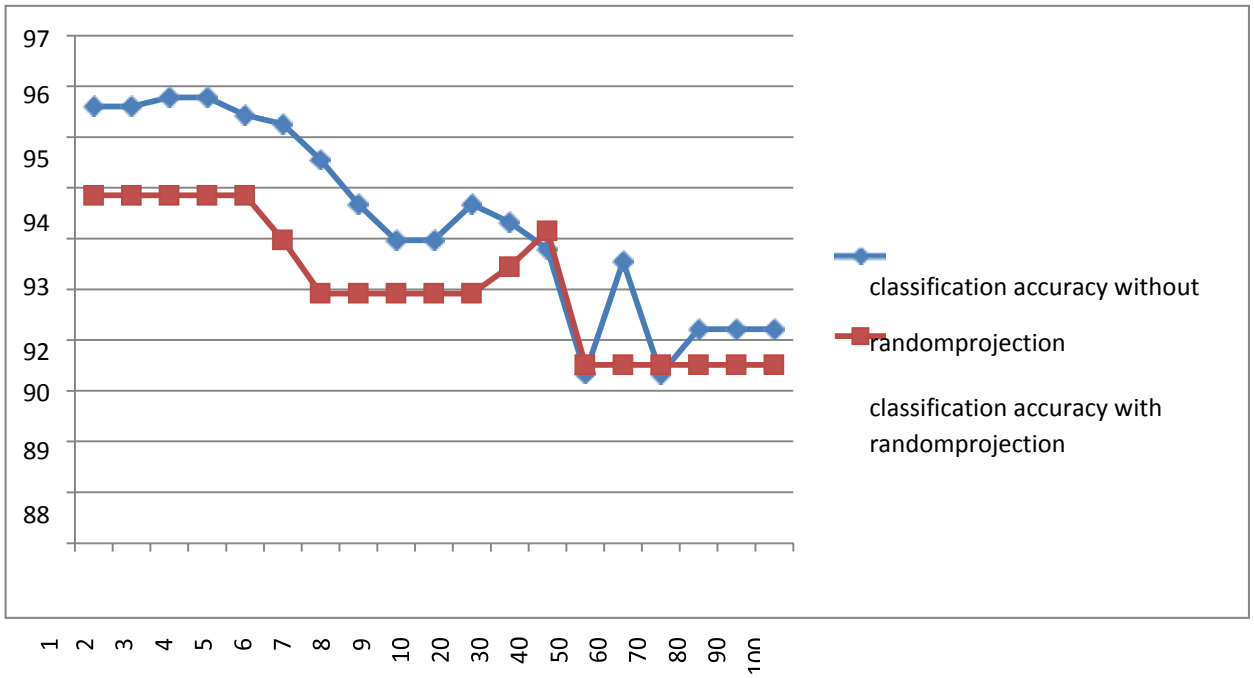

chart-5: Minimum number of objects vs. Classification accuracy

CONCLUSION

Here we focus about the advancement of predictive models by using supervised machine learning method to achieve better accuracy. The classifier such as J48 and LMT can be compared with or without random projection filter. LMT got the better accuracy of 97.368%. The accuracy of using filter is secondary. The classifiers without filter are efficient for the breast cancer detection.

REFERENCES

- Konstantina Kourou, Themis P. Exarchos , Konstantinos P. Exarchos, Michalis V. Karamouzisand Dimitrios I. Fotiadis(2015)Machine learning applications in cancer prognosis and prediction, Computational and Structural BiotechnologyJournal,vol.13,pp.8–17
- Arvind Kumar, Parminder Kaur and Pratibha Sharma(2015),A Survey on Hoeffding Tree Stream Data Classification Algorithms,CPUH-Research Journal, vol.1,Issue.2, pp.28-32,ISSN :2455-6076
- Jyotismita Talukdar, Dr. Sanjib and Kr. KalitaInt(2015)Detection of Breast Cancer using Data Mining Tool (WEKA),, International Journal of Scientific & Engineering Research, Vol.6, Issue 11,ISSN 2229-5518
- J. Novakovic, S. Rankov (2011)Classification Performance Using Principal Component Analysis and Different Value of the Ratio R,Int. J. of Computers, Communications & Control,Vol.6, pp. 317-327,ISSN 1841-9836
- B. Rebecca Jeya Vadhanam, S. Mohan, V.V. Ramalingam and V. Sugumaran(2016)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,ISSN :0974-5645
- B.R. Manju,A. Joshuvaand V. Sugumaran(2018)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, pp. 1061– 1079,ISSN: 0976-6359
- Nagesh Shukla, Markus Hagen Buchner, Khin Than Win and Jack Yang (2018) Breast cancer data analysis for survivability studies and prediction,, Computer Methods and Programs in Biomedicine 155,pp.199–208

- Liton Chandra Paul, Abdulla Al Suman and Nahid Sultan(2013)Methodological Analysis of Principal Component Analysis (PCA) Method,IJCEM International Journal of Computational Engineering & Management, Vol. 16 Issue 2,ISSN:2230-7893
- Francisco Castells, Pablo Laguna, Leif Sornmo, Andreas Bollmannand Jose Millet Roig(2007)Principal Component Analysis in ECG Signal Processing, Hindawi Publishing Corporation EURASIP Journal on Advances in Signal ProcessingVolume
- Kanghee Park, Amna Ali, Dokyoon Kim, Yeolwoo An, Minkoo Kim and Hyunjung Shin(2013)Robust predictive model for evaluating breast cancer survivability, Engineering Applications of Artificial Intelligence,vol.26,pp.2194–2205
- Dr Prof. Neeraj,Sakshi Sharma,Renuka PurohitandPramod Singh Rathore(2017)Prediction of Recurrence Cancer using J48 Algorithm, Proceedings of the 2nd International Conference on Communication and Electronics Systems(ICCES)
- Philip K. Chan and Richard P. Lippmann(2006)Machine Learning for Computer Security, Journal of Machine Learning Research,vol. 7,pp.2669-2672
- M. Elangovan,S. Babu Devasenapati, N.R. Sakthivel and K.I. Ramachandran(2011)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. 38pp.4450–4459
- Nour El Islem Karabadji, Hassina Seridi, Fouad Bousetouane, Wajdi Dhifli and Sabeur Aridhi(2017) An evolutionary scheme for decision tree construction, Knowledge-Based Systems vol.119,pp.166–177
- S. Ravi Kumar, K.I. Ramachandran and V. Sugumaran (2011), Machine learning approach for automated visual inspection of machine components, Expert Systems with Applications, vol.38,pp.3260–3266
- Rohit J. Kate and Ramya Nadig(2017), Stage-specific predictive models for breast cancer survivability,

  International Journal of Medical Informatics,vol.97,pp.304–311
- Cristinel Constantin(2014)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
- P.HamsagayatriandP.Sampath(2017),Decisiontreeclassifiersforclassificationofbreastcancer, International Journal of Current Pharmaceutical Research,Vol.9, Issue. 2,ISSN :0975-7066
- R. Jegadeeshwaran and V. Sugumaran(2013),Comparative study of decision tree classifier and best first tree classifier for fault diagnosis of automobile hydraulic brake system using statistical features, Measurement ,vol.46,pp.3247–3260
- Ajith Abraham(2005), Artificial neural networks, Nature & scope of AItechniques,vol.2,pp.901-908
- Jennifer Listgarten, Sambasivarao Damaraju, Brett Poulin,Lillian Cook, Jennifer DuFour, Adrian Driga, John Mackey, David Wishart,Russ Greiner andBrentZanke(2004), Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms, Clinical CancerResearch,vol.10,pp.2725- 2737
- Jaree Thongkam,Guandong Xu andYanchun Sang(2008), Breast cancer survivability via AdaBoost algorithms, Health data and knowledge management,vol.80
- V. Sugumaran, V. Muralidharan andK.I.Ramachandran(2007),Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing, Mechanical Systems and SignalProcessing,vol.21,pp.930-942
- Hui-Ling Chen,Bo Yang, Jie Liuand Da-You Liu(2011),A support vector machine classifier with rough set- based feature selection for breast cancer diagnosis, Expert Systems with Applications,vol.38,pp.9014-9022
- Tüba Kiyanand Tülay Yildirim(2004), Breast cancer diagnosis using statistical neuralnetworks, Journal of electrical & electronics engineering,vol.4,pp.1149-1153

- Andrej Bratko,Gordon V. Cormack, Bogdan Filipic, Thomas R. Lynamand BlazˇZupan(2006), Spam Filtering Using Statistical Data Compression Models, Journal of Machine LearningResearch,vol.7,pp.2673-2698
- José M. Jerez-Aragonés, José A. Gómez-Ruiz,Gonzalo Ramos-Jiménez, José Muñoz-Pérez and Emilio Alba- Conejo(2003), A combined neural network and decision trees model for prognosis of breast cancer relapse, Artificial Intelligence in Medicine,vol.27,pp.45–63
- B.Rebecca Jeya Vadhanam,S.MohanV.Sugumaran(2016), Application of Artificial Immune Recognition System for Identification of Advertisement Video Frames using BICC Features, Indian Journal of Science and Technology,Vol.9,Issue.14,ISSN:0974-6846
- V. Sugumaran and K.I. Ramachandran(2011), Effect of number of features on classification of roller bearing faults using SVM and PSVM, Expert Systems with Applications,vol.38,pp.4088–4096
- Levent Civicik,Burak Yilmaz,Yüksel Özbay and Ganimee Dilek Emlik(2015), Detection of micro calcification in digitized mammograms with multistable cellular neural networks using a new image enhancement method: automated lesion intensity enhancer (ALIE), Turkish Journal of Electrical Engineering & Computer Sciences,vol.23,pp.853 –872