International Journal of Grid and Distributed Computing
Vol. 11, No. 2 (2018), pp.41-54
http://dx.doi.org/10.14257/ijgdc.2018.11.2.05

A Novel Approach to Perform Analysis and Prediction on Breast Cancer Dataset using R

Syed Muzamil Basha, Dharmendra Singh Rajput, N. Ch. S. N Iyengar[1] and Ronnie D. Caytiles[2]

VIT University, Vellore, T.N., India
[1]Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, India
[2]Multimedia Eng., Dept, Hannam University, Daejeon, Korea
srimannarayanach@sreenidhi.edu.in

Abstract

Screening shows impact on cancer mortality rate by decreasing the number of advanced cancers with poor diagnosis, while cancer treatment works through decreasing the case-fatality rate. The prediction of breast cancer survivability has been a challenging research problem for many researchers. The objective of this research work is to propose a Novel model that can analysis the Breast cancer data and do efficient prediction. The contributions made in this paper are as follows, we collected three different the dataset from UCI Machine Learning repositories. We propose an approach, where a detailed comparison made between feature selection algorithms. Trained the datasets using Decision Tree, Random Forest and Support vector machine (SVM) machine learning algorithms. An attempt made to understand the impact of model selection metric in predicting different classes of Brest cancer. The results indicated that the Random forest is the best predictor wit 0.98 accuracy on the holdout sample, SVM came out to be the second with 0.97 accuracy and the Decision Tree came out with 0.96 to be the worst of the four condition tree with 0.95 accuracy. Finally performed prediction using Neural Network with three hidden layers and measured the efficiency, using Root Mean Square Error (RMSE) along with its variations.

Keywords*: Correlation, Recursive Feature elimination, Genetic algorithm, Decision Tree, Random Forest and Support vector machine (SVM), RMSE.*

1. Introduction

Breast cancer is a hateful tumor that build up majority death cells without any control on it and makes difficult to separate death cells with normal cells [1]. Currently, there is a limited clinical ability to monitor breast cancer patients for signs of drug resistance at the molecular level. The Cancer Stem Cells (CSCs) hypothesis implies the existence of progenitor breast cancer cells with the novo resistance toward antioestrogen receptor (ER) therapies. In 2017, breast cancer was estimated in 1.7 million cases and 612,970 deaths worldwide. Estimation made on National Colorectal Cancer (CRC) Roundtable's goal of increasing screening prevalence to 80% by 2018 would prevent 277,000 CRC cases and 203,000 deaths by 2030 [14]. Treatment for breast cancer are categorized into two types, Local(L) and Systematic(S). Surgery and radiation are pattern of local treatments, whereas chemotherapy and hormone therapy are pattern of systematic therapies. screening made an huge contributed towards reductions in the risk of breast cancer death by reducing the fatality of advanced cancers in screening groups [10]. Data Mining Techniques are applied to analyze the data and provide the report through web for easy user access. The author in [3] made an attempt to present *ONCOMINE*, a cancer

Received (October 10, 2017), Review Result (December 19, 2017), Accepted (December 22, 2017)

ISSN: 2005-4262 IJGDC

microarray database and web-based data-mining platform to facilitate discovery of genome-wide expression analyses. In this paper, we attempt to use made a comparison on different feature selection methods (Correlation, Recursive Feature elimination, Genetic algorithms) and state the advantages of each method in detail. we have trained four classifiers on characteristics such as clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses text classification to identify breast cancer masses as benign or malignant. When run on the data, the classifiers were able to achieve up to 98% recall accuracy on a randomly sampled training set of 200 patients and test set of 400 patients. Also made a comparison on model selection metric by training the dataset using KNN algorithm and state ROC provide better result than Accuracy and Kappa metrics.

The remainder of this paper is organized as follows. Section 2 provides the reader with the understanding of breast cancer research. In Section 3, we explained in detail the data, feature selection methods, Machine Learning classification methods and Model selection metric. In Section 4, the comparative result of all the approaches are presented. The paper concludes with Section 5 where we summarize the research findings, and outline the limitations and further research directions.

2. Literature Review

In [2] the author stated that C4.5 algorithm had better performance than Artificial Neural Network (ANN) and Naive Bayes on SEER data (period of 1973-2000) with 433,272 records. In [11] the author made a study on reproducibility of Ki-67 proliferation index (KIPI) exist between three pathologists, assessed KIPI performing visual estimation and proved that KIPI-RV and DIA-2 showed no significant difference (p=0.754), and an excellent agreement (correlation coefficient =0.979; 95% correlation Index =0.975 to 0.982). The author in [12], attempts to made weather Cigarette smoking is associated with breast cancer prognosis or not and concluded that women continued to smoke after diagnosis with those who quit smoking after diagnosis will have lower mortality from breast cancer with Hazard Rate 0.67:95%, Correlation Index 0.38 to 1.19 and respiratory cancer with Hazard Rate, 0.39: 95%, Correlation Index 0.16 to 0.95. In [13] the author applied fuzzy c-means technique on co-occurrence texture features to generate discriminative features to be used in classification in which discrete wavelet transform is used to remove High-frequency information also applied Principal Component analysis in extracting the features, evaluated the performance of four classifier like Probabilistic Neural Network (PNN), SVM, Decision Tree with respect to sensitivity and specificity and concluded that PNN classifier has a good performance in both benign and malignant tumors.

3. Methodology

The overall steps performed in analyzing and predicting the different classes of Breast cancer is described in detail as sown in the Figure 1.

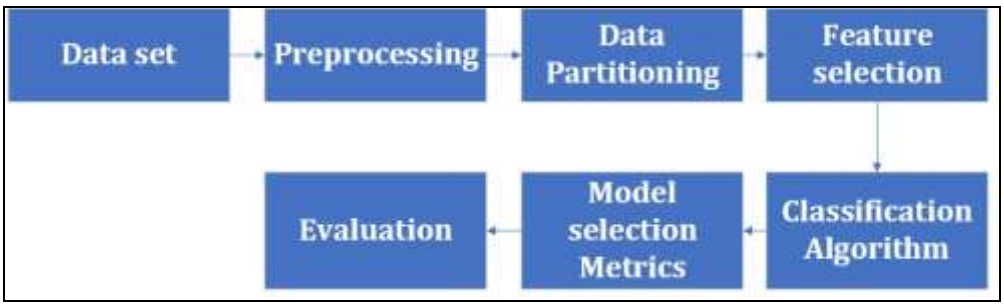


Figure 1. Process Flow of the Proposed Approach

As the First step, we aims to collect the data from (URL: https://raw.githubusercontent.com/ricardoscr/UW-Data-Science-Certificate/master/02 Methods/breast-cancer-wisconsin.data.txt) consisting of 699 observations and 10 variables as shown in Figure 2.

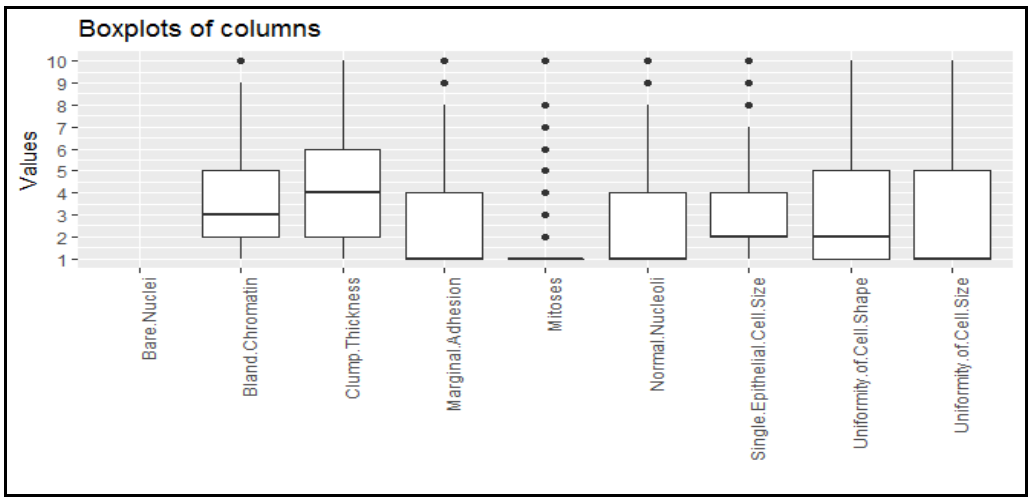


Figure 2. Box Plot of Attributes

After analyzing the columnar values of the dataset considered we found that, limiting/adjusting the threshold value improving the correlation the same is plotted in Figure 3 as before adjustment and after adjustment and achieved 0.762 of correlation value after adjustment from 0.718.

```
Correlation with Class BEFORE: 0.7182754
----------------------------------------
- AFTER adjustment:
COLUMN: Normal.Nucleoli
    Benign Malignant
2      342        36
3       12        26
4       12       131
----------------------------------------
Correlation with Class AFTER: 0.762556
```

Figure 3. Correlation Value

To validate the above adjustment made in improving the correlation between the class variables in the dataset hypothesis test is performed using t-test and obtained 16.492 as t value with 162.29 degree of freedom and P value as $< 2.2e^{-16}$. In [6] the author, identified Micro RNAs whose pattern was correlated with specific breast cancer biopathologic features, such as estrogen and progesterone receptor expression, tumor stage, vascular invasion, or proliferation index and demonstrated the existence of a breast cancer–specific Micro RNA signature. Where as in [7], the author stated that Micro-RNAs are extensively involved in cancer pathogenesis of solid tumors and support their function as either dominant or recessive cancer genes.

As the next step we aims to use machine learning techniques in selecting the best feature in the dataset collected from UGI Data repositories (Three different datasets: Dataset 1, 2, 3) using Random Forest algorithm, in which first we use determine to plot a cluster Dendrogram in understanding the importance of attributes as shown in Figure 4.

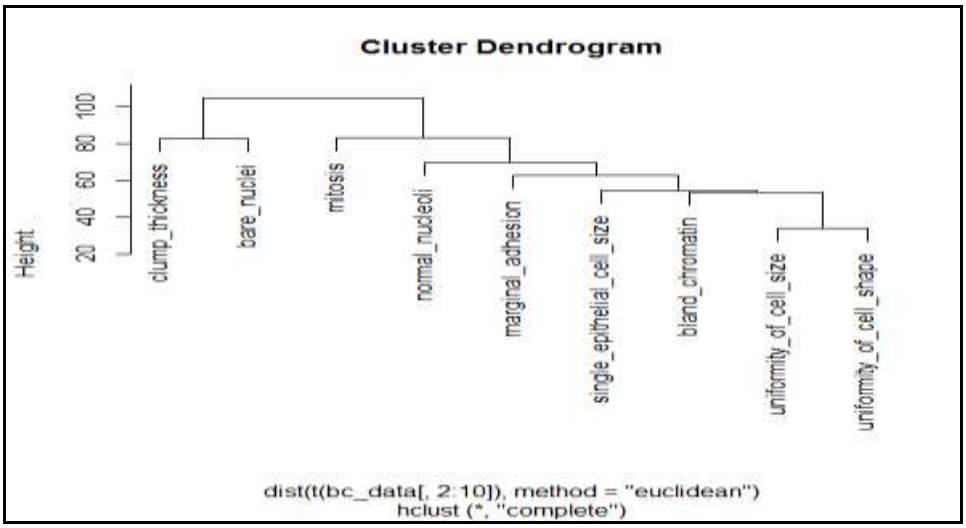


Figure 4. Cluster Plot using R Packages on Dataset1

The distribution of variables are taken in to consideration in understanding the behavior of each attribute in prediction as shown in Figures 5, 6 and 7.

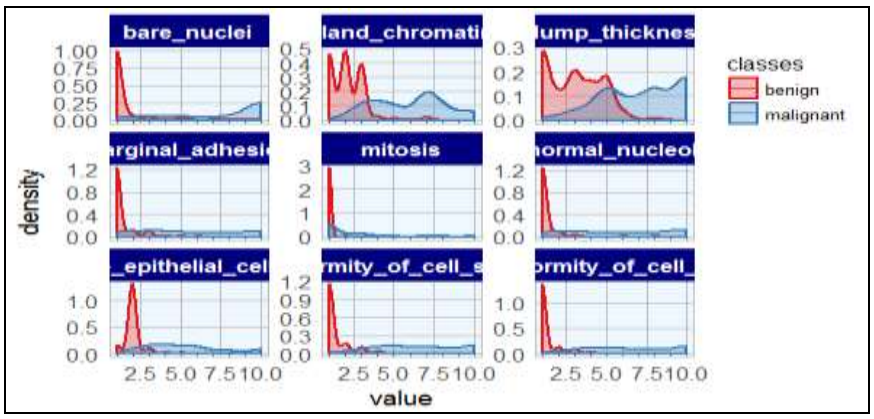


Figure 5. Density Plot on Dataset1

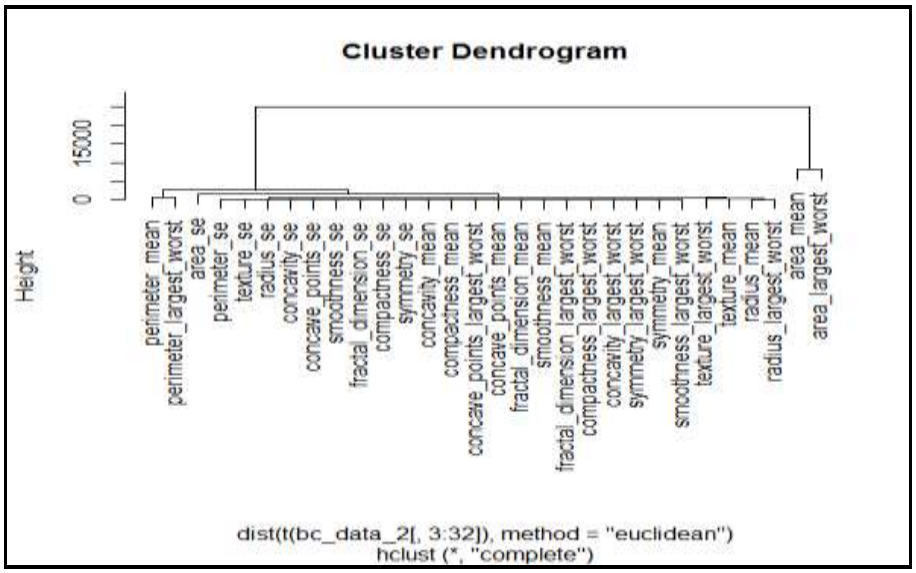


Figure 6. Cluster Plot using R Packages on Dataset2

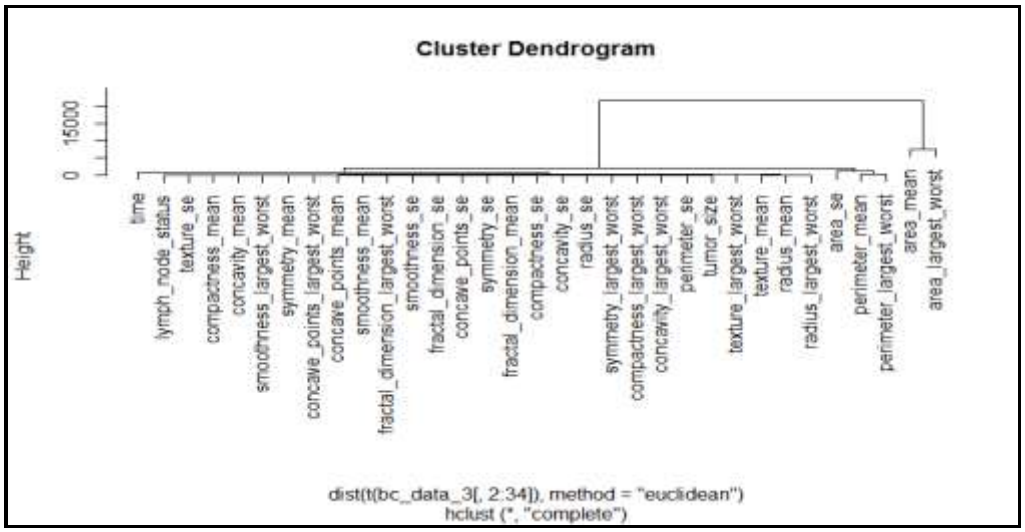


Figure 7. Cluster Plot using R Packages on Dataset3

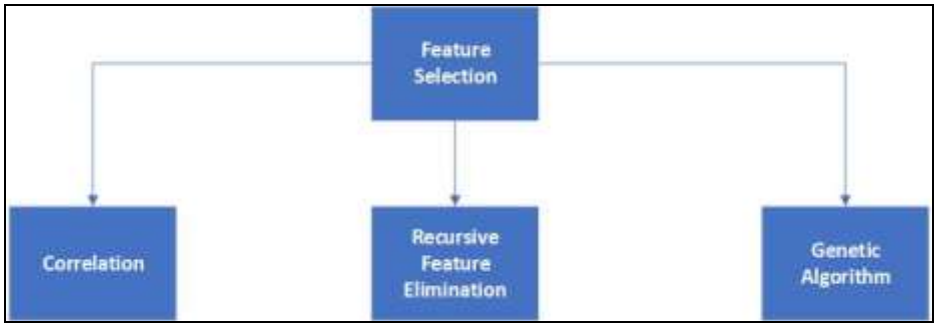


Figure 8. Techniques used in Feature Selection

In the Next Step, we are likely to perform feature selection with the Methods as described in Figure 8. Considering all the three dataset considered in our experiment, we have listed out the important features and plotted as shown in the Figures 9, 10 and 11 using Random Forest and Repeated Cross Validation with ten repeats to understand the impact of each feature on overall prediction.

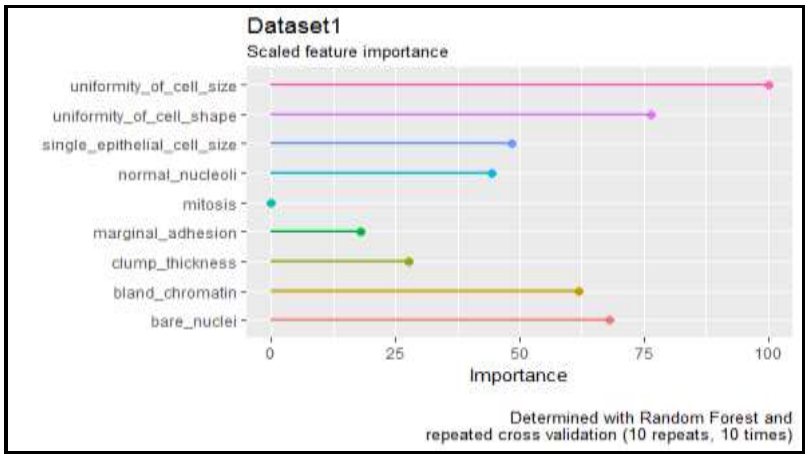


Figure 8. Impact of each Feature on Overall Prediction

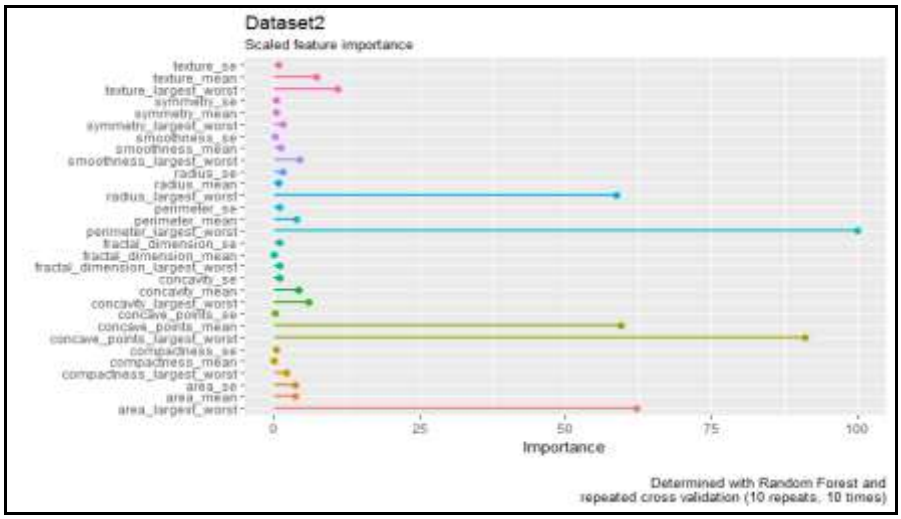


Figure 9. Impact of each Feature on Overall Prediction

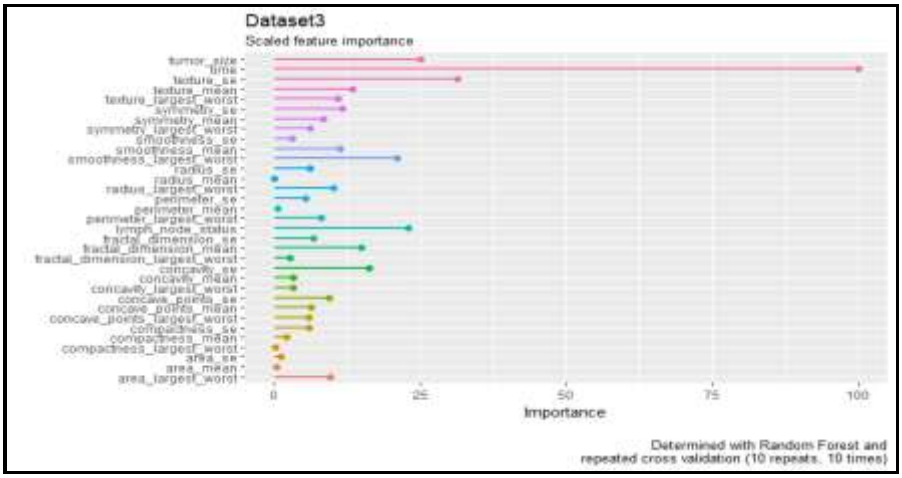


Figure 10. Impact of each Feature on Overall Prediction

In applying the correlation method is to measure the Correlation Index between the features is plotted in Figure 11. The interpretation of the same is plotted in Figure 12.

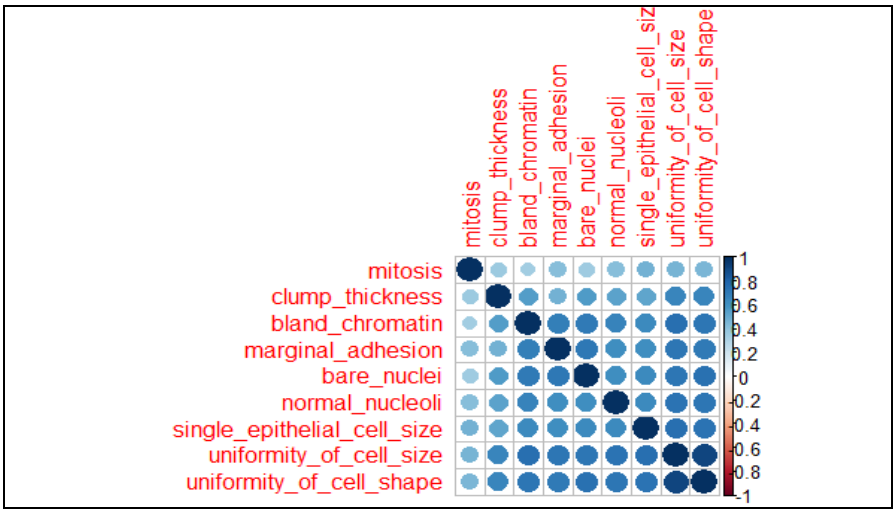


Figure 11. Correlation Matrix of Dataset1

```
Compare row 2  and column  3 with corr  0.913
  Means:  0.716 vs 0.601 so flagging column 2
Compare row 3  and column  7 with corr  0.725
  Means:  0.677 vs 0.579 so flagging column 3
Compare row 7  and column  6 with corr  0.703
  Means:  0.6 vs 0.544 so flagging column 7
Compare row 6  and column  4 with corr  0.719
  Means:  0.58 vs 0.526 so flagging column 6
All correlations <= 0.7
```

Figure 12. Interpretation of Correlation Matrix on Dataset1

Later Genetic algorithm is used to train the all the dataset with random population size as 5 as shown in Figure 13.

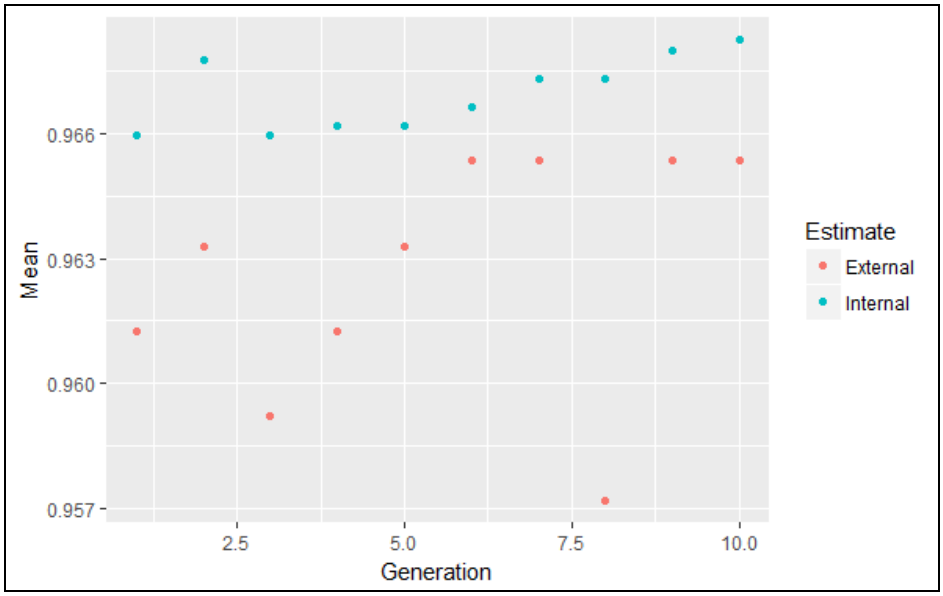


Figure 13. Plot of Mean Fitness Accuracy by Generation

Similarly, the performance of Correlation, Regressive feature elimination and Genetic algorithm in selecting the best feature from the dataset considered are plotted using Venn diagram in Figure 14.

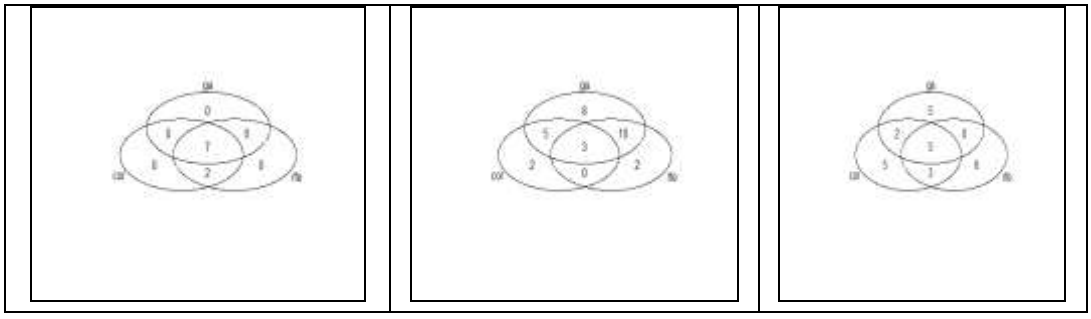


Figure 14. Plot of Venn Diagram

The efficiency of feature selection methods are evaluated in terms of sensitivity and specificity and the details are plotted in Figure 15. In which Correlation with 96.65%,

Regressive feature elimination with 95.88% and Genetic Algorithm with 83.05% of classification accuracy on Dataset 3.

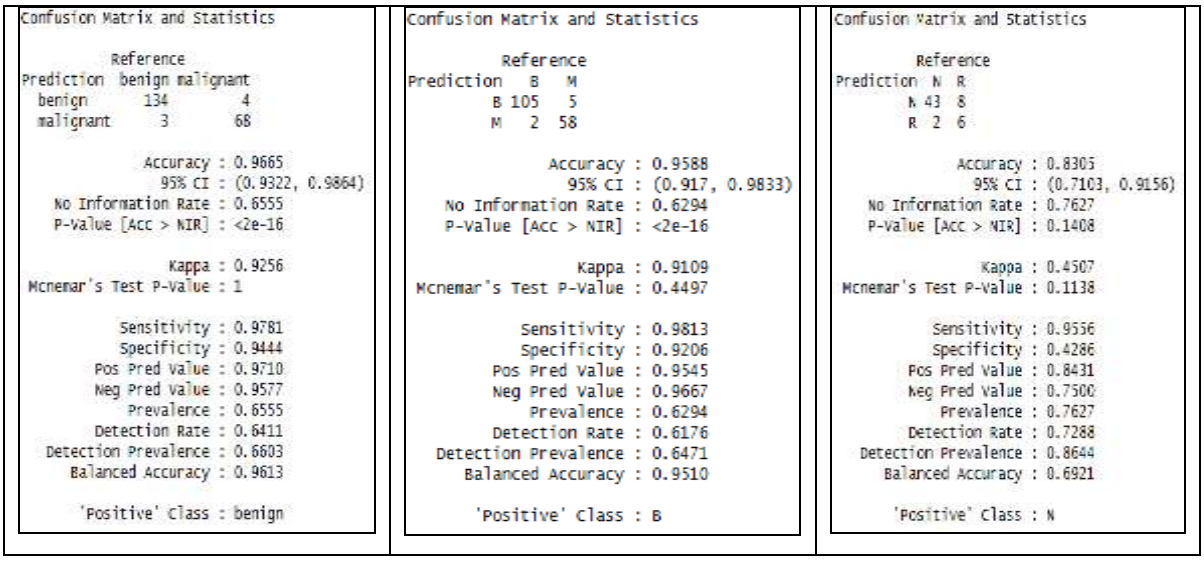


Figure 15. Confusion Matrix and Statistics of all the Feature Selection Methods

In classification, we use to map the data in to predefined targets, called supervised learning where targets are predefined. The goal of the classification is to build a classifier based on one attribute to describe the group of the objects used in predicting the group of attributes[5]. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules. Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules. The constructed Condition Tree on Dataset is plotted in Figure 16.

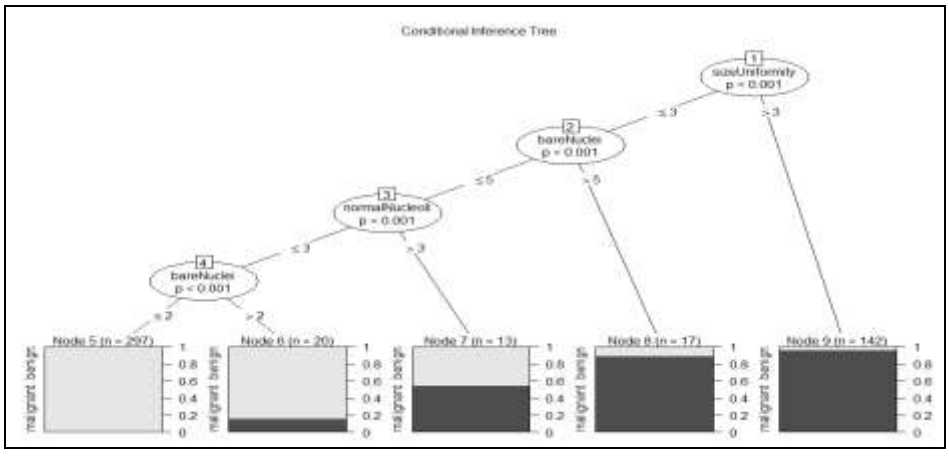


Figure 15. View of Condition Tree on Dataset1

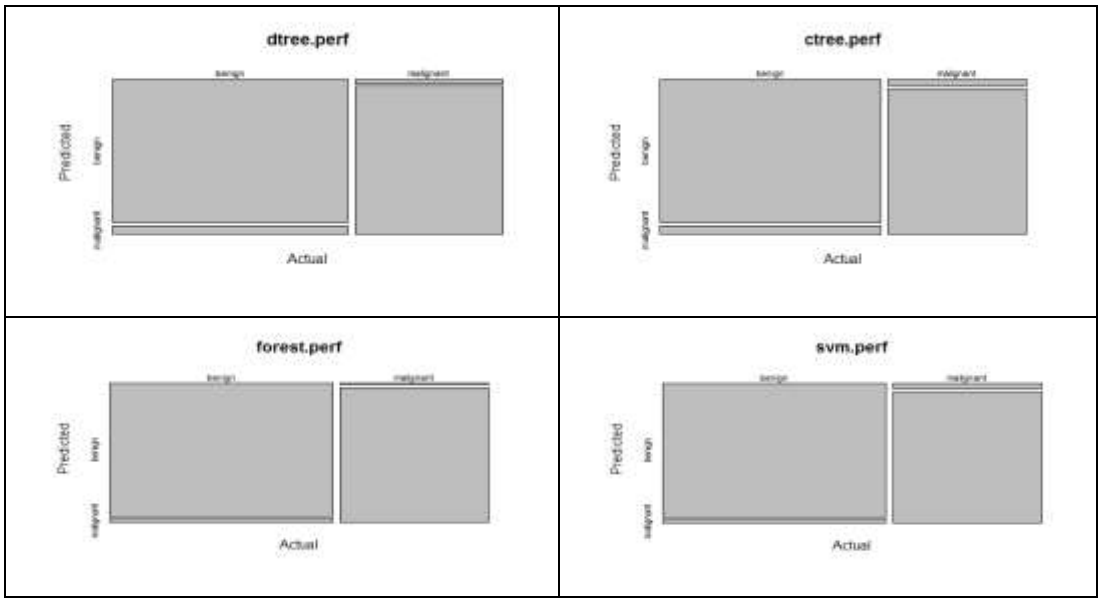
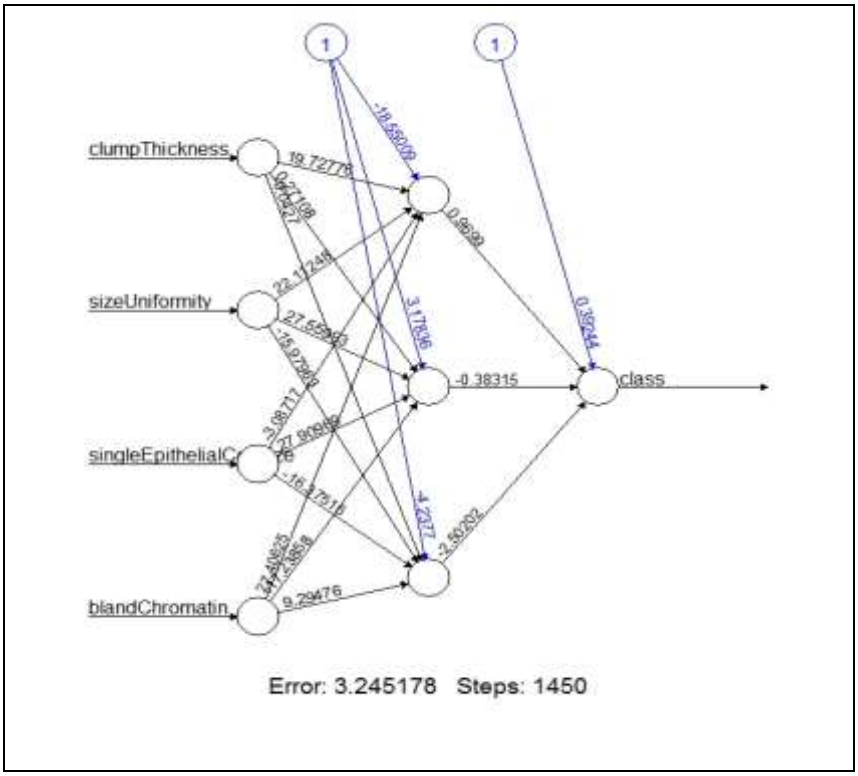
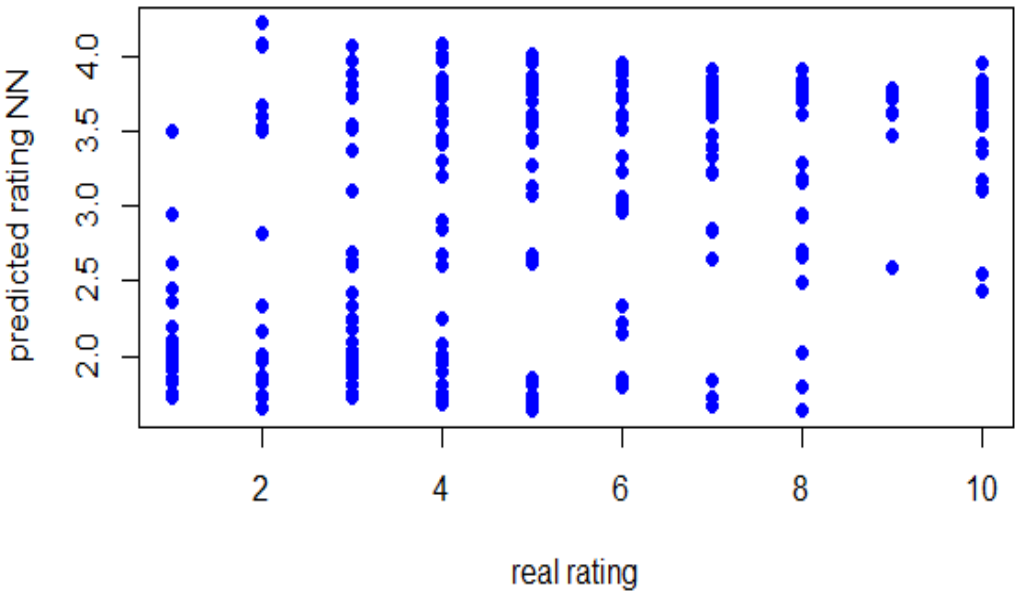


Figure 16. Plotting the Performance of Classification Algorithm on Dataset1

In [8] the author, used seven algorithms (Logist Regression model, ANN, NB, Bayes Net, DT, DT with NB, DT-ID3, DT-j48 and achieved classification accuracy as 85.8%, 84.5%, 83.9%, ,83.9%, 84.2%, 82.3%, 85.6%. Where as in our experiments we use four algorithms and achieved 98.0% using Random forest. In [9] the author, stated that early changes in tumor total hemoglobin (tHb) concentration can predict a pathologic complete response (pCR) to neoadjuvant chemotherapy in patients with operable breast cancer and achieved 81.2% sensitivity/47.0% specificity and 93.7% sensitivity/47.7% specificity.





Our aim is to perform prediction on Breast cancer data sets collected from Machine Learning Repositories (URL: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data) consisting of 569 observations and 31 variables. This dataset is trained using K nearest neighbor method and evaluated the accuracy by varying the number of neighbors ranging [10-40] and the results are plotted as shown in the Figure 17. The confusion matrix is generated using Accuracy as model selection metric as plotted in Figure 18.

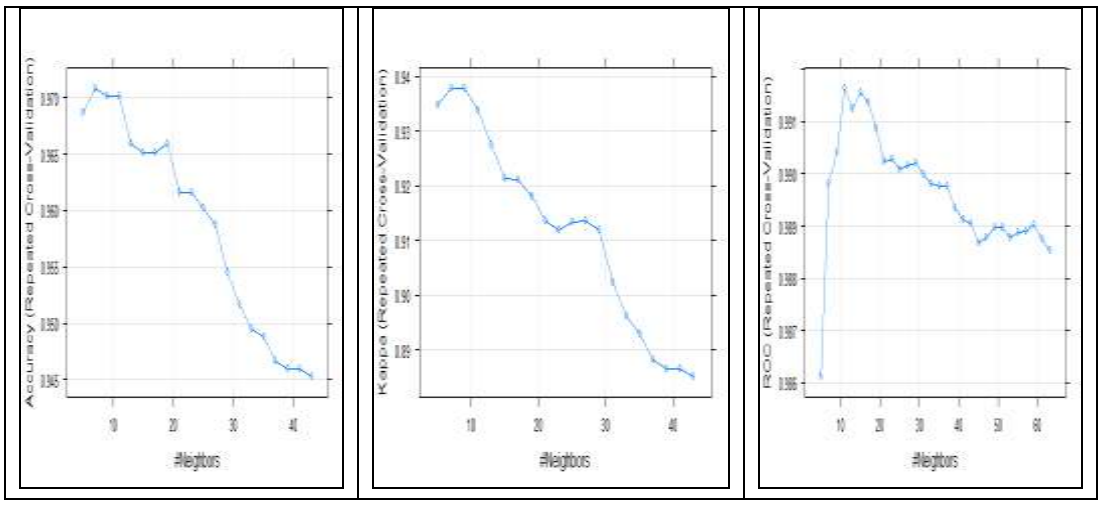


Figure 17. Performance of Model Selection Metric

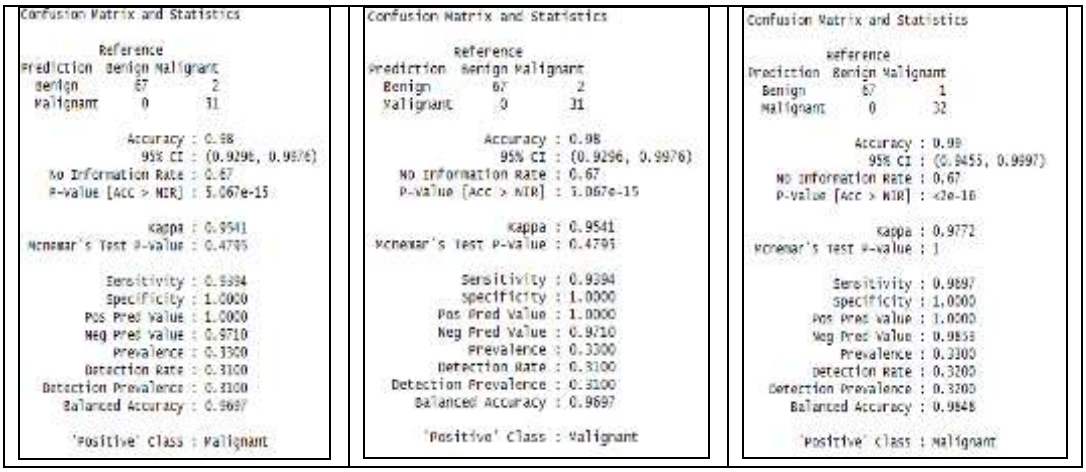


Figure 18. Confusion Matrix and Statistics of Model Selection Metric

Next is to construct confusion matrix for the same dataset. But the metric for model selection is Kappa and the value reaches 0.80 as the number of neighbor is 40 as shown in Figure 3. and the confusion matrix in Figure 4. Similarly, in the next step we consider ROC as metric in selecting a model and the confusion matrix is selected as shown in the Figure 5 and Figure 6. while Comparing the above three constructed model. we found that ROC gives as 0.99 of accuracy with k as 11 as shown in Figure 7.

4. Results & Discussion

In our Experiments, we have considered three Breast cancer datasets for performing analyses and prediction. In our approach we have compared the performance of feature selection methods and finds that all the feature selection are performing same with 96%±0.02 of Accuracy on Dataset1 and Dataset2. Whereas GA on Dataset2 have 100% sensitivity as the other methods have 98% of sensitivity as in Table 1. Similarly, the performance of classification algorithms on Dataset1 is plotted in Table 2. In which the Random forest the best with a classification accuracy of 98.0% which is better than SVM came with 97.0%, Decision Tree came out with 96% to be the worst of the four, condition tree with 95% accuracy. Finally, one of the dataset is trained using K nearest neighbor (KNN) method and evaluated the accuracy by varying the number of neighbors ranging from 10 to 40 to estimate the impact of model selection metrics and found that ROC is giving 99.0% of accuracy as in Figure 19.

Table 1. Comparison of Feature Selection Methods

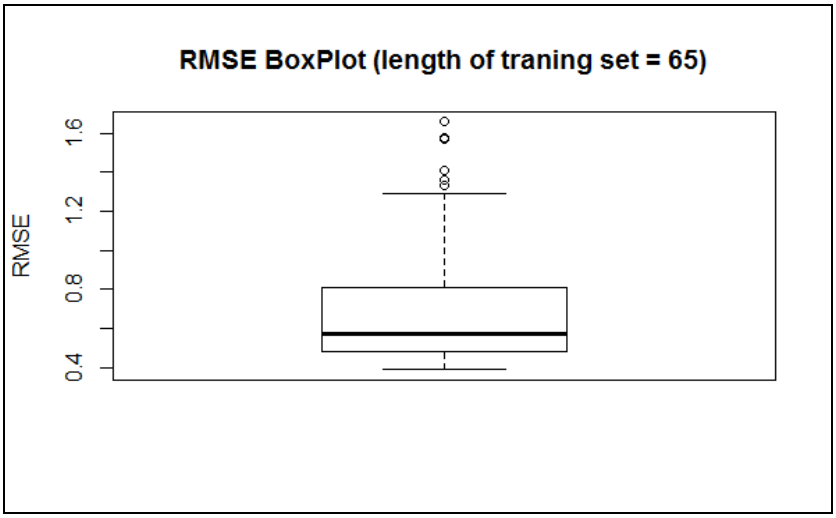
	Methods	Accuracy	Kappa	Precision	Sensitivity	Specificity
Dataset 1	**COR**	0.96	0.93	0.97	0.98	0.95
	RFE	0.96	0.93	0.97	0.98	0.95
	GA	0.96	0.92	0.97	0.95	0.95
Dataset 2	**COR**	0.95	0.91	0.93	0.97	0.90
	RFE	0.96	0.92	0.95	0.98	0.91
	GA	0.97	0.95	0.96	**1.00**	0.91
Dataset 3	**COR**	0.80	0.40	0.80	0.95	0.40
	RFE	0.80	0.30	0.83	**1.00**	0.20
	GA	0.80	0.40	0.90	0.90	0.55

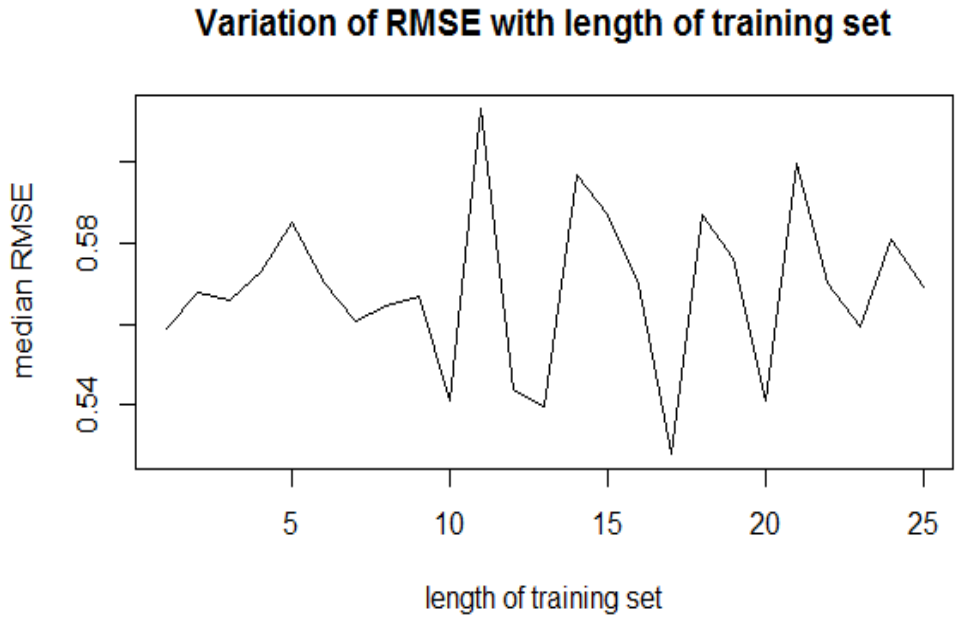
Table 2. Comparison of Classification Methods

	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Accuracy
DT	0.98	0.95	0.92	0.98	0.96
CT	0.96	0.95	0.92	0.98	0.95
RF	0.99	0.98	0.96	0.99	0.98
SVM	0.96	0.98	0.96	0.98	0.97

	k	metric	TN	TP	FN	FP	acc	sens	spec	PPV	NPV
Traditional Approach	21	NA	67	32	2	0	0.98	0.94	1	1	0.97
Approach 1	7	Accuracy	67	31	2	0	0.98	0.94	1	1	0.97
Approach 2	7	Kappa	67	31	2	0	0.98	0.94	1	1	0.97
Approach 3	11	ROC	67	32	1	0	0.99	0.97	1	1	0.99

Figure 19. Comparing of Model Selection Methods





5. Conclusion

In this paper, we used three different Breast cancer Dataset, three popular data feature selection methods, four Machine learning classifiers are trained to develop the prediction models and impact of three model selection metrics on prediction. In this research, we defined that correlation method after proper adjustments improve the correlation factor and helps in achieving better accuracy compared to other two feature selection methods. The aggregated results indicated that the Random forest the best with a classification accuracy of 98.0% which is better than SVM came with 97.0%, Decision Tree came out with 96% to be the worst of the four, condition tree with 95% accuracy. Finally, one of the dataset is trained using K nearest neighbor method and evaluated the accuracy by varying the number of neighbors ranging from 10 to 40 to estimate the impact of model selection metrics and found that ROC is giving 99.0% of accuracy. Our ongoing research efforts are towards incorporating new capabilities into our approach along the lines of extended research direction. we would like to propose most accurate prediction models and their hybrids for all possible cancer types.

References

[1] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial intelligence in medicine, vol. 34, no. 2, **(2005)**, pp. 113-127.
[2] A. Soltani Sarvestani, A. A. Safavi, N. M. Parandeh and M. Salehi, "Predicting breast cancer survivability using data mining techniques", In Software technology and Engineering (ICSTE), 2010 2nd international Conference on IEEE, vol. 2, **(2010)**, pp. V2-227.
[3] D. R. Rhodes, R. Daniel, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pander and A. M. Chinnaiyan, "ONCOMINE: a cancer microarray database and integrated data-mining platform", Neoplasia, vol. 6, no. 1, **(2004)**, pp, 1-6.
[4] S. Gupta, D. Kumar and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis", Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 2, **(2011)**, pp. 188-195.
[5] M. Karabatak and M. Cevdet Ince, "An expert system for detection of breast cancer based on association rules and neural network", Expert systems with Applications, vol. 36, no. 2, **(2009)**, pp. 3465-3469.
[6] M. V. Iorio, M. Ferracin, C.-G. Liu, A. Veronese, R. Spizzo, S. Sabbioni and E. Magri, "MicroRNA gene expression deregulation in human breast cancer", Cancer research, vol. 65, no. 16, **(2005)**, pp. 7065-7070.
[7] S. Volinia, G. A. Calin, C.-G. Liu, S. Ambs, A. Cimmino, F. Petrocca and R. Visone, "A microRNA expression signature of human solid tumors defines cancer gene targets", Proceedings of the National academy of Sciences of the United States of America, vol. 103, no. 7, **(2006)**, pp. 2257-2261.
[8] A. Endo, T. Shibata and H. Tanaka, "Comparison of Seven Algorithms to Predict Breast Cancer Survival (< Special Issue> Contribution to 21 Century Intelligent Technologies and Bioinformatics)", International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association, vol. 13, no. 2, **(2008)**, pp. 11-16.
[9] S. Ueda, N. Yoshizawa, T. Shigekawa, H. Takeuchi, H. Ogura, A. Osaki and T. Saeki, "Near-Infrared Diffuse Optical Imaging for Early Prediction of Breast Cancer Response to Neoadjuvant Chemotherapy: A Comparative Study Using 18F-FDG PET/CT", Journal of Nuclear Medicine, vol. 57, no. 8, **(2016)**, pp. 1189-1195.
[10] P. Autier, M. Boniol, M. Smans, R. Sullivan and P. Boyle, "Observed and predicted risk of breast cancer death in randomized trials on breast cancer screening", PloS one, vol. 11, no. 4, **(2016)**: e0154113.
[11] B. Acs, L. Madaras, K. Attila Kovacs, Anna-Maria Tokes, Janina Kulka and A. Marcell Szasz, "Abstract P1-01-02: Ki-67 proliferation index supported by digital quantitation in breast cancer: A comparative study", **(2016)**, P1-01.
[12] M. N. Passarelli, P. A. Newcomb, J. M. Hampton, A. Trentham-Dietz, L. J. Titus, K. M. Egan, J. A. Baron and W. C. Willett, "Cigarette smoking before and after breast cancer diagnosis: mortality from breast cancer and smoking-related diseases", Journal of Clinical Oncology, vol. 34, no. 12, **(2016)**, pp. 1315-1322.
[13] K. Nie, S. Glaßer, U. Niemann, G. Mistelbauer and B. Preim, "Classification of DCE-MRI Data for Breast Cancer Diagnosis Combining Contrast Agent Dynamics and Texture Features", In Bildverarbeitung für die Medizin 2017, Springer Vieweg, Berlin, Heidelberg, **(2017)**, pp. 325-330.
[14] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi and A. Jemal, "Colorectal cancer statistics, 2017", CA: a cancer journal for clinicians, vol. 67, no. 3, **(2017)**, pp. 177-193.

Authors

Syed Muzamil Basha, he had his Bachelor of Science in Information Technology at SITAMS, MTech in Information Technology (Networking) at VIT University and currently doing his research at VIT University. His research area are Wireless Sensor Networks, Text Mining and Big Data Predictive Analytics.

Dharmendra Singh Rajput working as Associate Professor in the Department of Software and Systems Engineering, School of Information Technology and Engineering, VIT University. His research area are Data Mining and Big Data Predictive Analytics.

N. Ch. S. N. Iyengar (b 1961), he currently Professor at the Sreenidhi Institute of Science and Technology (SNIST) Ghatkesr, Hyderabad, Telengana, India. His research interests include Intelligent Computing, Network Security, Cloud Computing, Data Science and Fluid Mechanics. He had 32+ years of experience in teaching and research, guided many scholars, has authored several textbooks and had nearly 200+ research publications in reputed peer reviewed international journals. He served as PCM/reviewer/keynote speaker/ Invited speaker. He received 2017 achievement award for his contributions in teaching and research by IOSRD.

Ronnie D. Caytiles, he had his Bachelor of Science in Computer Engineering- Western Institute of Technology, Iloilo City, Philippines, and Master of Science in Computer Science– Central Philippine University, Iloilo City, Philippines. He finished his Ph.D. in Multimedia Engineering, Hannam University, Daejeon, Korea. Currently, he serves as an Assistant Professor at Multimedia Engineering department, Hannam University, Daejeon, Korea. His research interests include Mobile Computing, Multimedia Communication, Information Technology Security, Ubiquitous Computing, Control and Automation