Prognosis on Stratification of Breast Cancer using Data Mining Models

Sreelakshmi S Pai, AnnMary Simon, G S Anisha

Abstract: **Breast cancer classification can be useful for discovering the genetic behavior of tumors and envision the outcome of some diseases. Through this paper we are predicting the noxious behavior of a tumor. The prediction models used are Random Forest, Naïve Bayes, IBK (Instance Based Learner), SMO (Sequential minimal optimization), and Multi Class Classifier. This prediction model which can potentially be used as a biomarker of breast cancer is based on physical attributes of a breast mass and which is gathered from digitized image of Fine Needle Aspirate (FNA). These can be helpful in prediction and reduction of invasive tumors**

Keywords **: Breast Cancer, Benign, Data Mining, Malignant.**

I. INTRODUCTION

Bosom Cancer is one of the cancer-causing illness most usually found in ladies. Cancer is a kind of illness which makes the body cells differ its attributes and cause anomalous development of cells. This paper explains the breast cancer identification by using various characteristics. Cells which grow out of control start the formation of cancer. The manifestations of carcinoma incorporate bump inside the bosom, blood release from the areola and changes inside the shape or surface of the areola or bosom. The breast cancer which affects milk vessel and does not spread is called non-invasive also called as Benign, invasive breast cancer which spreads to other organs is called malignant. Therefore, it is necessary to detect the affected cells before the spreading to other nearby organs. Early identification can forestall the demise pace of bosom carcinogenic patients, which requires early finding and dependable analysis strategy that permits doctors to recognize benign tumors from malignant ones. Programmed conclusion of bosom malignant growth is a significant, certifiable restorative issue thus, finding an exact and successful determination strategy is significant.

As of late Information Mining strategies have been broadly utilized in expectation, especially in therapeutic field. The characterization methods on bosom malignancy information can be valuable to anticipate the result of certain maladies or find the hereditary conduct of tumors. Foreseeing result of an infection is one among the first intriguing and testing errands in creating information handling applications. Utilization of PCs with robotized device, enormous volumes of the restorative information are being gathered and made accessible to the medicinal research gatherings. Accordingly, information preparing strategies turned into a popular research instrument for therapeutic scientists to spot and adventure designs, ailment expectation, and connections among sizable measure of factors. Right now, propose a prescient model to recognize a bosom malignant growth at beginning time utilizing five Data Mining systems like random Forest, naïve Bayes, IBK, SMO, multi Class classifier, and also measure the accuracy.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author

Sreelakshmi S Pai*,PG Student, Department of Computer Science &IT Amrita School of Arts and Sciences, Kochi Amrita Vishwa Vidyapeetham, India **sree.achu97@gmail.com**

AnnMary Simon, PG Student, Department of Computer Science &IT Amrita School of Arts and Sciences, Kochi Amrita Vishwa Vidyapeetham, India **annmary019@gmail.com**

G S Anisha, Assistant Professor, Department of Computer Science &IT Amrita School of Arts and Sciences, Kochi Amrita Vishwa Vidyapeetham, India **gs.anisha21@gmail.com**

I. LITERATURE REVIEW

From different reviewed papers it shows us that there have been several studies on the early discovery and avoidance of Mammary Carcinoma using data mining Approaches.

In [1], they used three data mining techniques Naïve Bayes, Neural Networks and C4.5 for mammary cancer examination and prognosis dilemma. They have likewise sketched out and settled the issues, calculations, and procedures for the issue of bosom malignancy survivability forecast in SEER database. According to their experimental result they found that C4.5 has more accuracy. The objective of [2] is to present a note on breast cancer to create predictive models for breast tumor endurance. They utilized three mainstream information mining Models (Naïve Bayes, RBF Network, J48), additionally utilized 10-crease cross-approval strategies to figure the unprejudiced gauge of the three forecast models for execution examination purposes. The result (in view of normal precision bosom malignancy dataset) confirmed that the Naïve Bayes is the best prophet with 97.36% exactness.

In [3], the proposed research work is to predict mammary tumor using classification algorithm and decision tree. The parameters like effectively classified instances, erroneously grouped cases, time taken, kappa measurement, relative absolute error, and root relative squared mistake are error for contrasting outcomes. From the empirical results it is observed that the functioning of the Decision tree better than K Nearest algorithm.[4], Plans to appraise and perceive a precise model to foresee the frequency of bosom tumor dependent on discrete number of patients investigative records.

The data mining models applied are Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes Classifier, AdaBoost tree. Principle Component analysis method is applied to data mining models as a preprocessing technique, to analyze the test lapse of each model; a 10-crease cross-approval technique has been executed. Result of the analysis demonstrates an exhaustive trade-off between procedures.

II. METHODOLOGY

For analyzing the data we have applied classification and prediction methods. This model is developed to anticipate whether a tumor is Lethal or Benevolent. Here, we have used five algorithms to develop this model.

A. Data Mining

The process of collection of pragmatic information from a massive set of data is called as Data mining. The raw data is transformed to an understandable structure through a process called pre-processing. There are different algorithms like Classification, Clustering, Regression and prediction in Data mining and these algorithms can be used to predict the noxious behavior of tumors when applied on the data which consist of the physical features of breast mass.

B. Random Forests

Random forests or random decision forests are a method for plan, backslide and other endeavors that controls by raising a swarm of decision trees at preparing time and results the class that is the style of the arrangement (order) or standard expectation (relapse) of the singular trees.

C. Multiclass Classifier

Multiclass classifier is used to perform classification tasks with more than two classes. Multiclass classification assumes that each sample is assigned to one label.

D. SMO (Sequential minimal optimization)

SMO is a calculation that is utilized for taking care of the Quadratic programming Problem that happens during the preparation of Support Vector Machine (SVM). SMO Breaks the entire problem into sub-problems and are solved analytically.

E. IBK (Instance Based Learner)

IBK is Nearest Neighbor also known as Collaborative Filtering or Instance Learning. It is a very fruitful Information Mining Technique that permits you to prophet an uncharted yield estimation of another information by using your prior information cases with realized yield esteems.

F. Naïve Bayes

Naïve Bayes is an Information Mining Classification procedure that is based on Bayes Theorem. It is not a single algorithm but a collection of algorithm where they have a common principle. Where, the principle states that every pair of countenance of being classified is nonpartisan of each other.

III. PROPOSED SYSTEM

Here, we are developing a model to predict the breast cancer diagnosis. The Fig. 1, shows overall procedures through which we have developed the model.

The data collected is undergone through resampling method in preprocessing stage, we are separating entire information into training set and test set in 70 and 30 rates individually. In the next stage train the dataset with Random Forest, Naïve Bayes, IBK, SMO and Multi Class Classifier algorithms then save and load each model and re evaluate each model with supplied test set.

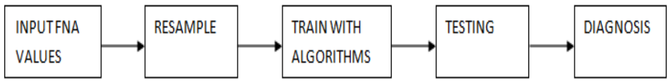

Fig. 1. Diagnosis Classifier

IV. EXPERIMENTAL RESULTS

The dataset considered for this study which shows the physical features of breast mass consist of 569 instances, in which 398 instances are taken as train data and the remaining 171 are taken as test data.

A. SMO (Sequential minimal optimization)

The SMO obtained the highest accuracy of 97.188%. But for 16 records it was unable to classify with this classifier and shows an Error Rate of 2.811%.

Table- I. Confusion Matrix of SMO

a	b	Classified as

198
2
	14
	355
		a=Malignant
		b=benign

B. Random Forest

The Random Forest obtained the accuracy of 96.6608 %.But shows an Error Rate of 3.3339%.

Table- II. Confusion Matrix of Random Forest

a	b	Classified as

205
12
	7
	345
		a=Malignant
		b=benign

C. Multi Class Classifier

The Multi Class Classifier obtained the accuracy of 96.4851 %.But shows an Error Rate of 3.5149%.

Table- III. Confusion Matrix of Multi Class Classifier

a	b	Classified as

201
9
	11
	348
		a=Malignant
		b=benign

D. IBK (Instance Based Learner)

The IBK obtained the accuracy of 95.7821 %.But shows an Error Rate of 4.2179%.

Table- IV. Confusion Matrix of IBK

a	b	Classified as

197
9
	15
	348
		a=Malignant
		b=benign

E. Naïve Bayes

The naïve Bayes obtained the accuracy of 92.7944 %.But shows an Error Rate Of 7.2056%.

TABLE V. Confusion Matrix of naïve Bayes

a	*b*	*Classified as*
190	22	a=Malignan t
19	338	b=benign

From the confusion matrixes of all the above algorithms we can observe that benign or b class is more classified. Based upon which we come to the conclusion that the results of the dataset obtained from cancer patient exhibits, that they are having benign cells which is non spreading in nature

V. PERFORMANCE COMPARISON

After training the data, the performance of the models are compared by calculating the Accuracy. The exactness of the grouping model can be calculated by determining the level of accurately arranged examples. Fig 2. Shows the accuracy.

Accuracy =correctly classified instances/total number of instances.

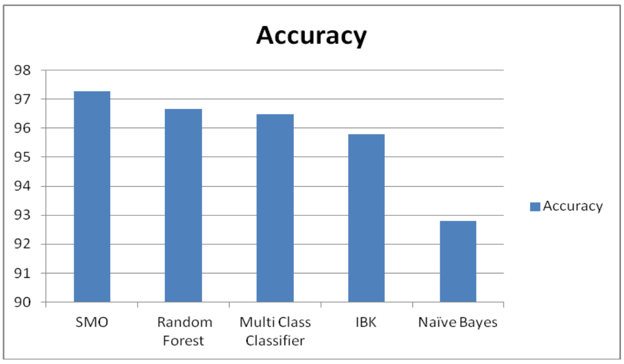

Fig. 2.Comparison of Accuracy of Algorithms

The dataset is evaluated for the Multiclass classifier and Decision Tree J48 algorithms using TP rate, FP rate, precision and recall. The resultant graph is shown in Fig. 3. And the corresponding values are as shown in Table VI.

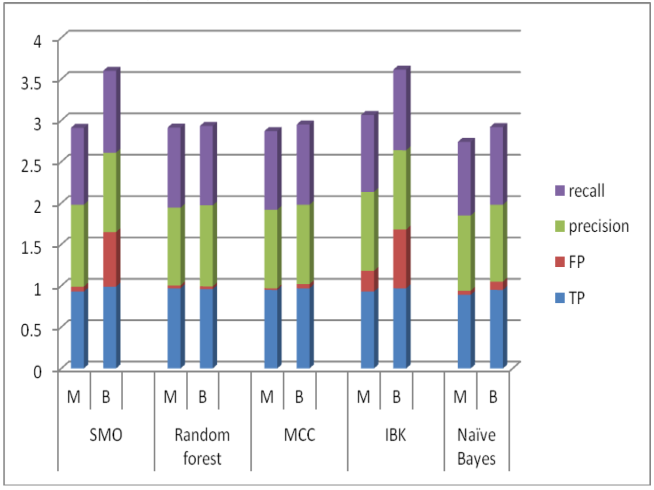

Fig. 3.Performance Comparison of different models

.

Table-VI : Performance analysis of Algorithms

Algorithm	Class Rate	TP Rate	FP Rate	PRECISION	RECALL
SMO	M B	0.93 0.99	0.06 0.66	0.99 0.96	0.93 0.99
Random Forest	M B	0.97 0.96	0.034 0.033	0.94 0.98	0.97 0.96
MCC	M B	0.95 0.97	0.02 0.05	0.95 0.96	0.95 0.97
IBK	M B	0.93 0.97	0.25 0.71	0.956 0.959	0.929 0.975
Naïve Bayes	M B	0.89 0.95	0.05 0.1	0.91 0.93	0.89 0.94

VI. CONCLUSION

The records of the datasets collected from various physical features of patients' can be helpful in prediction and reduction of invasive tumors. In this dissertation we have applied five data mining methods on FNA dataset of 569 records, we focused to predict whether a cancerous cell is noxious or not by labeling Benign (B) or Malignant (M) based upon models characteristics and behaviors obtained from the experimental results, we conclude that SMO data mining method shows more accuracy. Pragmatic experiments with weka tool endow with considerable amount of accuracy which we have already noted in the result part. This can be very advantageous to doctor for early analysis of bosom tumor and early treatment for patients. The future scope of this paper pertains to exercising of various algorithms on similar dataset.

REFERENCES

1. H. K. K. Zand, "A Comparative Survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction," Indian Journal of Fundamental and Applied life Sciences, vol. 5, pp. 4330-4339, 2015.
2. V. Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer survivability," International Journal of computer science and Mobile Computing, vol. 3, no. 1, 2014.

3. D. C. Nalini and T. Poovozhi, "Data Mining Classification Technique Applied For Breast Cancer," vol. 119, pp. 10935-10945.
4. H. Wang and S. W. Yoon, "Breast Cancer Prediction Using Data Mining Methods," Proceeding of the 2015 Industrial and System Engineering Research Conference, 2015.
5. B. Padmapriya and T. Velmurugan, "A survey on Breast Cancer analysis Using Data Mining Techniques," IEEE International Conference on Computational Intelligence and Computing Research, 2014.
6. D. Delen, G. Walker and A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Techniques," ELSEVIER, vol. 34, no. 2, 2005.
7. R. Aavula and R. Brahmaramba, "XBPF: An Extensible Breast Cancer Prognosis Framework For Predicting Susceptibility, Recurrence and Survivability," International Journal of Engineering And Advanced Technology, vol. 8, no. 5, 2019.
8. C. Shah and A. G. Jivani, "Comparison of Data Mining Classification Algorithm for Breast Cancer Prediction," 2014.
9. N. A. Farooqui and Ritika, "A Study on Early Prevention and Detection of Breast Cancer using Three-Machine Learning Techniques," International Journal of Advanced Research in Computer Science, vol. 9, no. 2, 2018.
10. K. P. Ch. Shravya and S. Shubani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques," International Journal Of Innovative Technology and Exploring Engineering, vol. 8, no. 6, 2019.
11. J. Talukdar and D. S. Kr.Kalita, "Detection of Breast Cancer Using Data Mining Tool(WEKA)," International Journal of Scientific and Engineering Research, vol. 6, no. 11, 2015.
12. J.-Y. Wang, "Data Mining Analysis(breast-cancer data)," AI term Project, 2003.

AUTHORS PROFILE

Sreelakshmi S Pai Integrated MCA Final Year student in Amrita school of Arts and Sciences, Amrita Vishwa Vidyapeetham University**.**

AnnMary Simon Integrated MCA Final Year student in Amrita school of Arts and Sciences, Amrita Vishwa Vidyapeetham University. .

G S Anisha Master Degree in Computer Applications, Master philosophy in computer science, almost 6 Years of teaching experiences. Currently working as Asst.Professor inAmrita Vishwa Vidyapeetham.