Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776830

IMPROVED TECHNIQUE TO IMPROVE BREAST CANCER PREDICTION

Alpna Sharma[1] , Nisheeth Joshi[2] , Vinay Kumar[3]
Alpna@vips.edu[1], jnisheeth@banasthali.in[2], vinay5861@gmail.com[3]
[1]Department of Computer Science, Apaji Institute, Banasthali University, India
[2]Department of Computer Science, Apaji Institute, Banasthali University, India
[3]Ex Scientist GOI, Ex Professor, VIPS – Vivekananda Institute of Professional Studies
Corresponding author: alpna@vips.edu

ABSTRACT
Breast cancer is a critical disease witnessed among women. Owing to the high number of features, predicting the breast cancer using the digital method is increasingly challenging. To predict the breast cancer accurately and efficiently, this work has proposed the feature selection based technique coupled with the dimension reduction technique for the efficient prediction of breast cancer. Results collected and analyzed before applying the feature selection technique and after feature selection technique. Experiment highlighted that the KNN based machine learning algorithm with Extra Tree based feature selection method proved to be extremely effective in terms of performance. Computation time recording during the experimentation is also presented to measure the time gain received before and after the proposed approach. Proposed technique will be able to predict the breast cancer with the higher degree of accuracy, at the same time, prediction time needed would decline subsequently. Therefore, proposed approach will greatly aid the medical staff in confirming the breast cancer.

Keywords
Breast cancer, machine learning in breast cancer, feature selection in breast cancer, recall score in breast cancer.

INTRODUCTION

Breast cancer is a critical disease caused to women. Women across the world are suffering from breast cancer. There is sharp increase in the cases of breast cancer. Earlier it used to occur in elderly women; however, its presence in women under 40 is not rare. In the year 2012 itself around 1.7 million new cases were surfaced.In recent years, Breast cancer has overtaken lung cancer in terms of new cases (approximately 2.3 million) and has been the most often diagnosed cancer. Belgium, Denmark, and Netherland are some of the worst hit countries. Mortality rate also varies from one country to another. On the other hand survival rate is highest in Belgium where many women are surviving even after five years of diagnosis of cancer. Breast cancer develops when healthy breast cells become aberrant, proliferate uncontrollably, and form tumours. One in every eight women will be diagnosed with breast cancer over their lifetime while in men the risk is one in one thousand. Breast cancer risk factors include advanced age, a parent or sibling with breast cancer, dense breast, obesity and specific hereditary genemutations known as BRCA1 and BRCA2. Breast cancer can behave in invasive, non-invasive, recurring, or metastatic ways. To address the growing breast disease risk, it is necessary to

enhance the availability of prompt detection, appropriate treatment, care and extensive data gathering via robust cancer registries.

BACKGROUND

Data analysis is transforming with rapid pace. Begin with an ordinary analysis of data, now it has evolved and begin using intelligence. Machine learning, Fuzzy logic, and artificial intelligence are some of the other popular methods utilized to gain insight of a dataset. Machinelearning based method involves supervisory and non-supervisory approaches in order to learn the data. Once learned the data, proposed approach can predict the data with a high degree of accuracy. Machine learning based methods can be used to predict the data according to the label of the dataset. If the label of data is categorical in that case classification methods are used. On the other hand, in case of continuous data, regression method of prediction can be used. Clustering methods can be used to cluster the data. Based on the method employed in clustering, new instances can be predicted.

Based on the spread of a data within the dataset, classification methods are categorized into linear and non-linear. Data scattered in a linear manner can be predicted with the help of linear model.On the other hand, for the non-linear data, decision tree, neural network, SVM, KNNcan be employed. Decision tree, neural network, SVM and KNN are the widely used algorithms employed during the classification phase. Neural network based method have high potential of usage. Perceiving its benefit, it is also been employed on advance neural network based method termed as deep learning method.

Dimensionality curse is one of the widely seen in gene expression data. In such dataset, number of dimensions spread from few hundreds to thousands. Under such circumstances, employing any model turns challenging at the same time it turns time consuming. The major functionality to be drawn with the help of visualization lacks. None of the model can neatly visualize such a large number of dimensions. Under this scenario, there is a need to reduce the number of dimensions, without compromising the basis characteristics of the dataset. This can be done with the help of feature selection and dimension reduction techniques.Once feature selectionand dimension reduction techniques are employed thereafter, characteristics of dataset is not lost.

LITERATURE REVIEW

Breast cancer cases in women under the age of 40 have increased dramatically. Tumor-intrinsic genetic changes, however, affect tumour growth, progression, and metastatic potential even in the early stages, thus limiting the usefulness of routinely used prognostic signals. In many circumstances, the present standard of care does not accurately assess a patient's prognosis[1]. Breast cancer cells carry either estrogen receptor (ER), progesterone receptor (PR), or both receptors.For the treatment of such cancers, medicines specifically hormone therapy medicines can be used that will inhibit the estrogen level or either lower it. This type of treatment is beneficial for hormone receptor-positive breast cancers, but it does not effective on hormone receptor-negative tumours (both ER- and PR-negative). Triple-negative breast cancer cells lack oestrogen and progesterone receptors and produce no or excessive amounts of the protein HER2. Cancers that are triple-positive are ER-positive, PR-positive, and HER2-positive.The Differential Gene Expression(DGE) analysis aids in understanding inherent biologicalmechanisms and gene expression stochasticity[2]. Identification of genes, that are widely

expressed in a group of cells without any or little prior knowledge about primary cell subtypes is a challenge in DGE analysis. [3] Prospective clinical trials are being conducted for several gene expression assessments, including MammaPrint, Prosigna, and EndoPredict. [4][5].
Trend hasbeen analyzed for around 10 years in a single hospital[6], According to the findings, while there has been an increase in invasive and later stage breast cancers, the overall risk of breast cancer has not changed.Computation based method yields promising outcome in analyzing the genetic pattern predicting breast cancer. Proposed approach has used several classification based approaches that include Lasso logistic regression[7]. A review for gene expression signature was carried out by Latha et al.[8], authors have reviewed the strategies prevalent that can help in diagnosis of the cancer during early stages. In another work, a logic based analysis based on association between TNF, TGFBI and EGF was proposed[9]. Authors claimed their model is efficient enough in providing the novel insight on aggressive breast cancer subtypes.Kajala et al[10]had proposed the machine learning based algorithm for early diagnosis of breast cancer. A comparative study employing the classification approaches was proposed that include random forest classifier, logistic regression, naïve bayes etc. Authorshave touched the accuracy level of 96.5 percent. An approach to combine variety of classification approach was proposed byTahmooresi et al. for early detection of breast cancer[11]. To diagnose the breast cancer, supervised machine learning algorithm were proposed by Mohana & Sahaaya Arul Mary, [12].

RESEARCH METHODOLOGY

Breast cancer dataset is a popular dataset used to learn the features impacting the breast cancer. This work has used the popular machine learning algorithm namely, a) support vector machine linear and radial basis function[13] b) Classification Tree [14] c) KNN [15]d) Neural network[16]. Sample comprised of the null values, since the data is sensitive hence missing values were not dropped instead, this work relied on mean values and same is substituted forthe missing value. In order to draw the sample cross validation method has been used. Drawn sample was divided into 2 folds and repeated 5 times. Outcome thus received is averaged out for the higher reliability.
In order to bring all the dimensions on same scale, min max method was used. During this work, python on Anaconda has been used as a language for coding. Since the numbers of dimensions present in the dataset were 16384, hence, important features were dawn with the help of following feature selection methods.

- L1 penalty
- ExtraTree Classifier
- Chi square

Outcome is generated by applying the machine learning algorithm with sample drawing techniques. Outcome thus generated are placed under the before feature selection method. After selecting the feature selection, outcomes were generated, and same are discussed on the two measurements that are:

- Average recall score before and after applying the feature selection
- Time needed to fit the model

RESULT ANALYSIS AND DISCUSSION

Description about the dataset

Dataset is comprising of 16384 features and positive and negative reports of breast cancer. Out of the complete dataset 529 instances were denoting the positive cases and 61 cases were related to the negative instances. Visualizations of the entire dataset have been illustrated in the figure using two principal components. Scattering of the data shows that the majority of the components can be well separated and only very small amount of overlapping has been featuredin the plot, refer figure 1.

Accuracy of the dataset is measured with the help of the variance in the category of the predicted model and the one bearing by the dataset. Hence, true positive and false positive, true negative and false negative are termed. Recall and accuracy score is treated as one of the robust method of categorization. Although, confusion matrix is one of the largest used method to explore the accuracy.

Table 1: Predicting accuracy

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True negative

Hence Recall and precision can be computed as

Recall=(TP)/(TP+FN)

Precision=(TP)/(TP+FP)

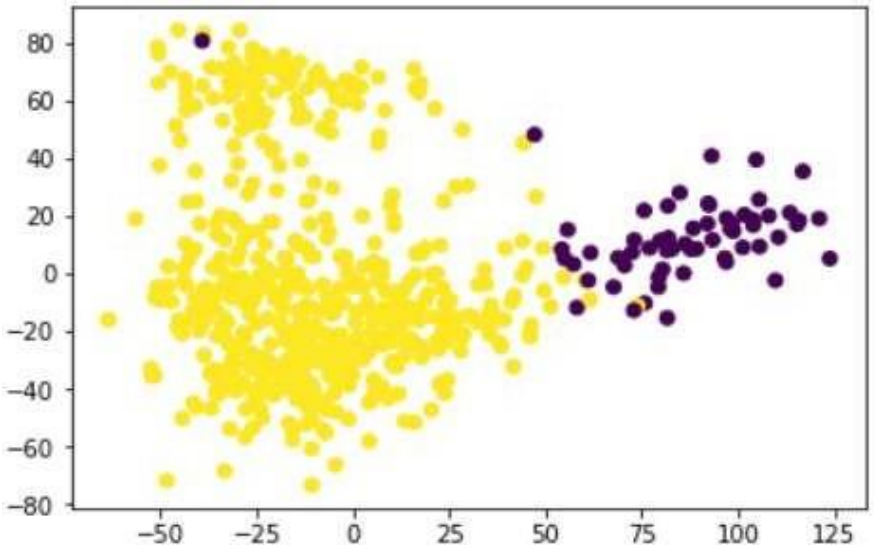

Figure 1: Dataset in two major components

The entire dataset has been categorized into two major components and the resultant visualization has been presented in the figure 1.

Table 2: Average fit time(Sec)

Time	Measure	Algorithm	Value
Before	Average Fit time	SVM Linear	0.653455
Before	Average Fit time	SVM rbf	1.01877
Before	Average Fit time	Classification Tree	1.240168
Before	Average Fit time	KNN Tree	1.286076
Before	Average Fit time	Neural Network	0.885252
L1 Penalty	Average Fit time	SVM Linear	0.637838
L1 Penalty	Average Fit time	SVM rbf	1.076283
L1 Penalty	Average Fit time	Classification Tree	1.174557
L1 Penalty	Average Fit time	KNN Tree	1.324622
L1 Penalty	Average Fit time	Neural Network	0.637837
ExtraTree	Average Fit time	SVM Linear	0.0066
ExtraTree	Average Fit time	SVM rbf	0.035201
ExtraTree	Average Fit time	Classification Tree	0.077004
ExtraTree	Average Fit time	KNN Tree	0.016003
ExtraTree	Average Fit time	Neural Network	0.085005
chi2	Average Fit time	SVM Linear	0.0014
chi2	Average Fit time	SVM rbf	0.0016
chi2	Average Fit time	Classification Tree	0.002
chi2	Average Fit time	KNN Tree	0.002
chi2	Average Fit time	Neural Network	0.004

Table 2: Recall Score

Feature Selection Method	Measure	Algorithm	Value
Before	Average Recall	SVM Linear	0.9875
Before	Average Recall	SVM rbf	0.5
Before	Average Recall	Classification Tree	0.967058405
Before	Average Recall	KNN Tree	0.996261356
Before	Average Recall	Neural Network	0.5
L1 Penalty	Average Recall	SVM Linear	0.9875
L1 Penalty	Average Recall	SVM rbf	0.5
L1 Penalty	Average Recall	Classification Tree	0.933725071
L1 Penalty	Average Recall	KNN Tree	0.996261356
L1 Penalty	Average Recall	Neural Network	0.5

ExtraTree	Average Recall	SVM Linear	0.9875
ExtraTree	Average Recall	SVM rbf	0.970833333
ExtraTree	Average Recall	Classification Tree	0.985576923
ExtraTree	Average Recall	KNN Tree	1
ExtraTree	Average Recall	Neural Network	0.7875
chi2	Average Recall	SVM Linear	0.779301994
chi2	Average Recall	SVM rbf	0.762635328
chi2	Average Recall	Classification Tree	0.880414449
chi2	Average Recall	KNN Tree	0.813778288
chi2	Average Recall	Neural Network	0.5

In order to gain deep insight of the dataset, both the linear and non-linear method of classification has been applied. Once classification based approach is applied, KNN proved to be a promising algorithm and neural network demonstrated the poor performance. However, once the feature selection methods applied and the results were collected it was learned that KNN has touched the magical value of 100 percent accurate prediction. High accuracy demonstrated by the KNN remained prevalent in all the feature selection method.

On time front, all the algorithms have generated encouraging outcome after the implementation of the chi square method. In all the experiment, outcome proved to be more efficient than any other techniques.

CONCLUSION

Breast cancer is the major concern in women and no country is untouched with this menace. Loss in terms of life is really alarming. Number of features indicating the breast cancer is enormous hence poses immense challenge in visual analytics and accurate prediction. Machine leaning algorithms once applied along with the preprocessing based method can select the features bearing high impact on cancer determination. Feature selection techniques proved tobe critical and relating the features in determining the cancer. Outcome generated predicted that KNN based machine learning based method proved to be highly efficient in predicting the breast cancer method. Once applied with the extraTree based feature selection techniques, outcome yielded encourages introducing the proposed method on real time environment after rigorous testing.

REFERENCES

1. E. Ellsworth Rachel, J. Decewicz David, D. Shriver Craig and L. Ellsworth Darrell, Breast Cancer in the Personal Genomics Era, Current Genomics 2010; 11(3) . https://dx.doi.org/10.2174/138920210791110951

2. Andrew McDavid, Greg Finak, Pratip K. Chattopadyay, Maria Dominguez, Laurie Lamoreaux, Steven S. Ma, Mario Roederer, Raphael Gottardo, Data exploration, quality

control and testing in single-cell qPCR-based gene expression experiments, *Bioinformatics*, Volume 29, Issue 4, February 2013, Pages 461–467, https://doi.org/10.1093/bioinformatics/bts714

3. Stegle, O., Teichmann, S. &Marioni, J. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145 (2015). https://doi.org/10.1038/nrg3833

4. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol. 2017 Oct;14(10):595-610. doi: 10.1038/nrclinonc.2017.74. Epub 2017 May 31. PMID: 28561071.

5. Senkus E, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rutgers E, Zackrisson S, Cardoso F; ESMO Guidelines Committee. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2015 Sep;26 Suppl 5:v8-30. doi: 10.1093/annonc/mdv298. PMID: 26314782.

6. Dodelzon, K., Starikov, A., Reichman, M., Cheng, E., Lu, C. M., Blackburn, A., Reznik, E., Kim, J., Bosec, A., Thomas, C., Askin, G., & Arleo, E. K. (2021). Breast cancer in women under age 40: A decade of trend analysis at a single institution. *Clinical Imaging*. https://doi.org/10.1016/j.clinimag.2021.03.031

7. Nandagopal, V., Geeitha, S., Kumar, K. V., & Anbarasi, J. (2019). Feasible analysis of gene expression –a computational based classification for breast cancer. *Measurement: Journal of the International Measurement Confederation*, *140*, 120–125. https://doi.org/10.1016/j.measurement.2019.03.015

8. Latha, N. R., Rajan, A., Nadhan, R., Achyutuni, S., Sengodan, S. K., Hemalatha, S. K., Varghese, G. R., Thankappan, R., Krishnan, N., Patra, D., Warrier, A., & Srinivas, P. (2020). Gene expression signatures: A tool for analysis of breast cancer prognosis and therapy. In *Critical Reviews in Oncology/Hematology* (Vol. 151, p. 102964). Elsevier Ireland Ltd. https://doi.org/10.1016/j.critrevonc.2020.102964

9. Jo, K., Santos-Buitrago, B., Kim, M., Rhee, S., Talcott, C., & Kim, S. (2020). Logic-based analysis of gene expression data predicts association between TNF, TGFB1 and EGF pathways in basal-like breast cancer. *Methods*, *179*, 89–100. https://doi.org/10.1016/j.ymeth.2020.05.008

10. Kajala, A., & Jain, V. K. (2020). Diagnosis of Breast Cancer using Machine Learning Algorithms-A Review. *Proceedings - 2020 International Conference on Emerging Trends in Communication, Control and Computing, ICONC3 2020*, *29*(3 Special Issue). https://doi.org/10.1109/ICONC345789.2020.9117320

11. Tahmooresi, M., Afshar, A., Bashari Rad, B., Nowshath, K. B., & Bamiah, M. A. (2018). Early detection of breast cancer using machine learning techniques. *Journal of Telecommunication, Electronic and Computer Engineering, 10*(3–2), 21–27.

12. Mohana, S., & Sahaaya Arul Mary, S. A. (2020). Diagnosis of breast cancer using supervised machine learning techniques. *Test Engineering and Management, 83*(6), 9293– 9300.

13. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS One. 2017 Jan 6;12(1):e0161501. doi: 10.1371/journal.pone.0161501. PMID: 28060807; PMCID: PMC5217832.

14. Ponnuraja C, Lakshmanan B. C, Srinivasan V, Prasanth B. K. Decision tree Classification and Model Evaluation for Breast Cancer Survivability: A Data Mining Approach. Biomed Pharmacol J 2017;10(1).

15. R. MurtiRawat, S. Panchal, V. K. Singh and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 534-540, doi: 10.1109/ICESC48915.2020.9155783.

16. Saad Awadh Alanazi, M. M. Kamruzzaman, Md Nazirul Islam Sarker, Madallah Alruwaili, Yousef Alhwaiti, Nasser Alshammari, Muhammad Hameed Siddiqi, "Boosting Breast Cancer Detection Using Convolutional Neural Network", Journal of Healthcare Engineering, vol. 2021, Article ID 5528622, 11 pages, 2021. https://doi.org/10.1155/2021/5528622