International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Cancer Prediction Method using Effective Feature Extraction and Supervised Classification Techniques

S. Supraja, N.Vasuki, N. J. Vishwa Dhakshana, N. Kiruthiga

Abstract: ***ML, the logical investigation of calculations and measurable models that PC frameworks use to acceptably play out a specific endeavor without using unequivocal headings, subordinate upon models and deciphering. A huge bit of down to earth AI utilizes controlled learning.PC based understanding has been used in various fields of drug. Decision help instruments with presenting to AI models can be passed on for chest undermining headway figure. Such models can add to early finding and treatment and right currently results. With intensive electronic clinical records and moved AI impels, a chest compromising improvement want model can be utilized as a choice assistance mechanical get together for care intervention.Breast infection is the fifth driving clarification behind perilous advancement mortality around the world. Early affirmation of repeat chance is principal in improving desire. Chest hurtful advancement happens in chest cells, the slick tissue or the stringy connective tissue inside the chest. Chest sickness is hazardous tumors will all around become reasonably logically awful and become lively instigating end. Factors, for example, age and a family legacy of chest danger can create the risk of chest illness. Two sorts of tumors: Benign: this tumor type isn't hazardous for a human body and from time to time causes human destruction. Undermining: this tumor type is powerfully perilous and causes human passing, it is called chest unsafe advancement. The reason for this assessment is making and studying models dependent on different parameters for chest threat guess.***

Keywords: ***Breast Cancer, Machine Learning, Supervised Learning, Tumors***

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Supraja S*., Department of Computer Science & Engineering,Sri Krishna College of Technology, Coimbatore, India. Email: 16tucs235@skct.edu.in
Vasuki N., Department of Computer Science & Engineering,Sri Krishna College of Technology, Coimbatore, India. Email: 16tucs244@skct.edu.in
Vishwa Dhakshana N. J., Department of Computer Science & Engineering,Sri Krishna College of Technology, Coimbatore, India. Email: 16tucs255@skct.edu.in
Kiruthiga N., Department of Computer Science & Engineering, Sri Krishna College of Technology,Coimbatore, India. Email: n.kiruthiga@skct.edu.in

I. INTRODUCTION

Bosom malignancy (Breast Cancer) is the fifth driving reason for disease mortality around the world, as indicated by the World Health Organization (WHO) and breast cancer-related mortality is the most part brought about by metastasis and Approximately 30%–40% of patients with bosom malignancy were reported to experience the severe ill Effects and roughly 10%–15% of them were accounted for to bite the dust of malignant growth metastasis. Bosom disease results might be guessed based superficially markers of tumor cells and serum tests the estrogen receptor (ER), progesterone receptor (PR), and human epidermal development factor receptor 2 (HER2) are basic bosom malignant growth surface markers. The ER and PR have been utilized as prognostic elements for the growth of the cancer, especially in initial 5 years following determination HER2 over expression, which happens in 20%–25% of all bosom disease cases has been seen as a prognostic factor for poor in general endurance and moderately brief time frame to backslide in patients with bosom malignant growth. Breast cancer can also be detected in Mammogram. Fig. 1 shows the mass detection of breast cancer in Digital Mammogram which is based on Gestalt Psychology [1].

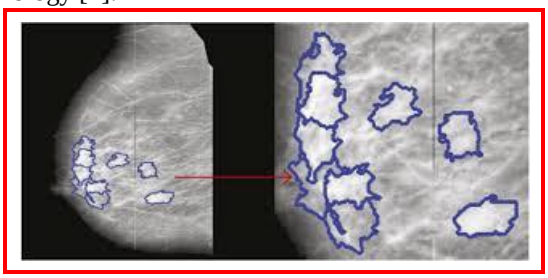

Fig.1.1.Mass Detection of Bosom Malignancy in Digital Mammogram

II. RELATED WORKS

Different change based surface investigation strategies are applied to change over the picture into another structure utilizing the spatial recurrence properties of the pixel power varieties.Habib Dhahri et al proposed the automatic breast cancer diagnosis by using some Machine Learning algorithms such as Ada Boost classifier (AB), Gaussian Naive Bayes (GNB), and linear discriminant analysis (LDA) [2].

Zheng et al joined K-implies calculation and a help support vector machine (SVM) for bosom malignant growth finding [3]. Praful Agarwal et al proposed the screening mammograms which results in major detection of breast cancer and also the
Mass detection is based on saliency.
Normally, some model parameters must be tuned to accomplish the ideal execution from a calculation. For example, the learning rate for preparing a neural system and the parameter C and sigma parameter of SVMs [5] are indicated physically in light of the fact that there is no logical recipe to process the correct worth.
A. P. Charate et al proposed the preprocessing methods of mammogram images for breast cancer detection [4]. Hence, picking the last tuning parameters of any proposed model has not yet been settled. The master in ML picks the fitting strategy for the present issue space [6]. In any case, the non-experts in ML invest a great deal of energy to advance their proposed models and to accomplish the objective execution [7].

III. EXISTING SYSTEM

In existing framework, informational collection contains just scarcely any records and size is very low. So that precision rate is less in ML models. Early detection of repeat hazard is basic in improving prognosis. Features are not well-known for classification and also the combination of hybrid algorithms were not used [8].Though few risk reduction can be obtained with certain measures and these ML techniques cannot remove completely that have growth in low and middle –income areas.

Early detection methods that include early finding or awarness of early signs and manifestations in Symptomatic populaces so as to encourage determination and early treatment and screening that is the efficient utilization of a screening test in a probably asymptomatic population.It means to recognize people with a variation from the norm reminiscent of malignant growth [9].

Fig. 3.1,3.2 shows the Logistic Regression with PCA training set and Logistic Regression with PCA testing set. Ch. Shravya et al proposed the Prediction of Breast Cancer Using Supervised Machine Learning Techniques [10].

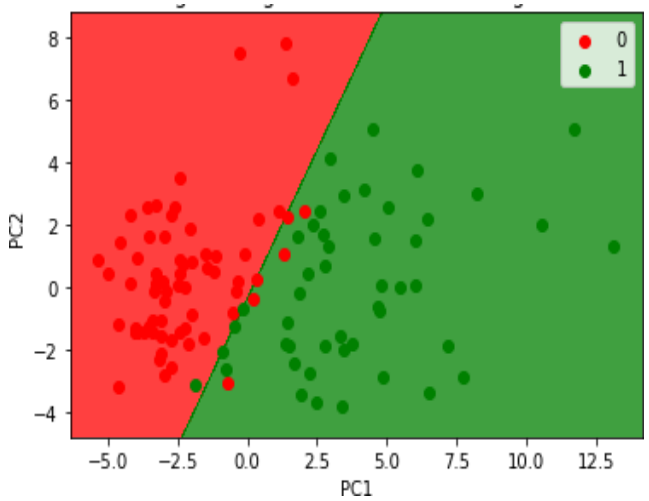

Fig.3.1.Logistic Regression with PCA training set

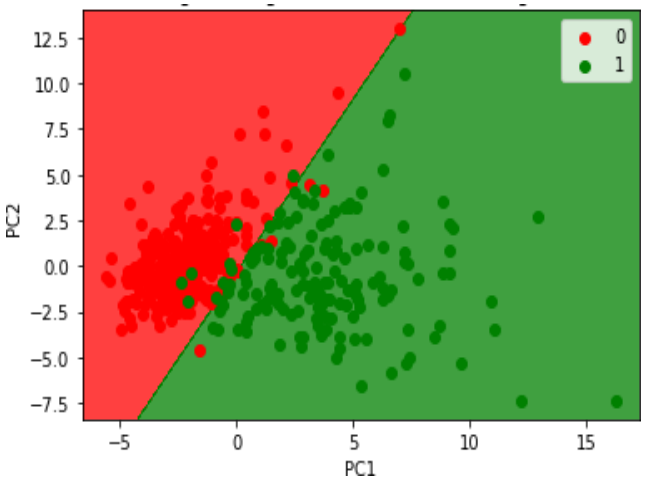

Fig.3.2Logistic Regression with PCA testing set

IV. PROPOSED METHODOLOGY

The intention of the proposed system is to make sure that the decision support tools based on machine learning model which can be developed for breast cancer prognosis. To ensure such models, that can be treated for breast cancer. Feature selection methods are used to select the best features for classification algorithms and these algorithms are used to detect whether it is B(Benign) or M(Malignant).

V. MODULES

The proposed methodology consists of four phases namely 1) Preprocessing, 2) Classification Algorithm, 3) Analysis and Insights and 4) Final Prediction.
Fig. 4.1. shows the overflow chart of proposed methodology to process. Whereas, Preprocessing consists of two algorithms such as LDA and PCA, Classification Algorithm represents Naive Bayes Classification and it analysis to predict the final prediction.

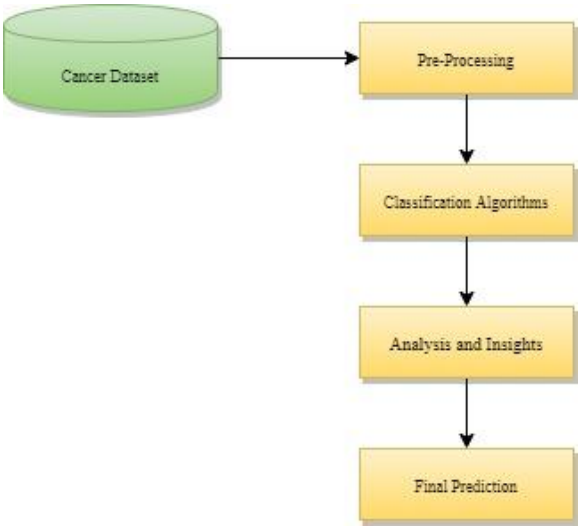

Fig.4.1.Overflow Chart of Proposed Methodology

VI. IMPLEMENTATION

ML methods involved in proposed frameworks are PCA, LDA, Naïve Bayes, Random Forest and Neural Networks. These are represented in fig.5.1.

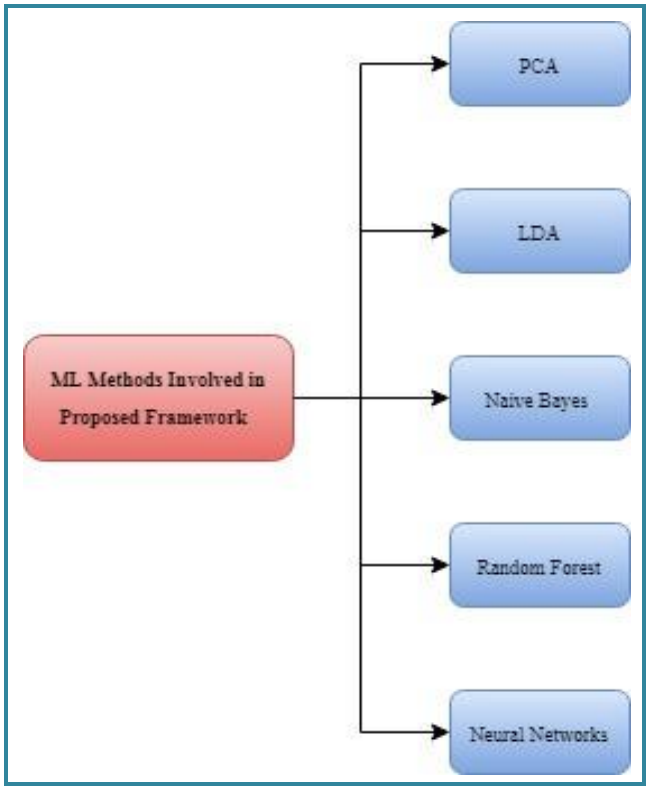

Fig.5.1.ML methods

A. PCA

The primary thought of head part examination (PCA) is to diminish the dimensionality of an informational index comprising of numerous factors associated with one another, either intensely or gently, while holding the assortment present in the dataset, up to the most extraordinary degree.. The equivalent is finished by changing the factors to another arrangement of factors, which are known as the vital segments (or basically, the PCs) and are symmetrical, requested with the end goal that the maintenance of variety present in the first factors diminishes as we descend in the request. Along these lines, right now, first head part holds most extreme variety that was available in the first segments. The key parts are the eigen vectors of a covariance grid, and consequently they are symmetrical.

B. LDA

LDA is a calculation generally restricted to just two-class arrangement issues. This post is expected for designers keen on applied AI, how the models work and how to utilize them well. All things considered no foundation in measurements or direct variable based math is required, despite the fact that it helps on the off chance that you think about the mean and variance of an appropriation. LDA is a straightforward model in both planning and application.

There is some intriguing measurements behind how the model is arrangement and how the expectation condition is determined. Consider expelling anomalies from your information. These can slant the fundamental measurements used to isolate classes in LDA such the mean and the standard deviation.

C. Naive Bayes Classifier

A Naive Bayes classifier is a probabilistic AI model that is utilized for order task. The core of the classifier depends on the Bayes hypothesis. Utilizing Bayes hypothesis, we can discover the likelihood of A happening, given that B has happened. Here, B is the proof and A is the speculation. The doubt made here is that the markers/features are free.That is nearness of one specific element doesn't influence the other. Subsequently it is called naive.The Naive Bayes Classifier system depends on the alleged Bayesian hypothesis and is especially fit when the dimensionality of the data sources is high. In spite of its straightforwardness, Naive Bayes can frequently outflank progressively refined grouping techniques.

D. Random Forest

Random forest is a supervised learning calculation. The "forest" it constructs, is a troupe of choice trees, normally prepared with the "stowing" strategy. The general thought of the packing strategy is that a blend of learning models builds the general outcome. Irregular timberland has about a similar hyper parameters as a choice tree or a packing classifier. Luckily, there's no need to join a choice tree with a sacking classifier since you can easily utilize the classifier-class of arbitrary timberland. With arbitrary timberland, you can likewise manage relapse assignments by utilizing the calculation's regressor. Arbitrary woodland adds extra arbitrariness to the model, while developing the trees. Rather than looking for the most significant element while parting a hub, it scans for the best component among an irregular subset of highlights. This outcomes in a wide decent variety that for the most part brings about a superior model.

E. Neural Network

A neural system could be a progression of calculations that makes an attempt to understand basic connections in an exceedingly heap of data through a procedure that imitates the way within which the human mind works. Right now, systems advert to frameworks of neurons, either natural or counterfeit in nature. Neural systems will comply with evolving input; during this method, the system produces the foremost ideal outcome while not expecting to update the yield criteria. the concept of neural systems, that has its underlying foundations in artificial insight, is quickly studying ubiquitousness within the advancement of trading frameworks. Neural systems, within the realm of finance, assist within the improvement of such procedure as time-arrangement prognostication, algorithmic exchanging, protections order, credit hazard demonstrating and developing exclusive pointers and price derivatives.

VII. RESULT

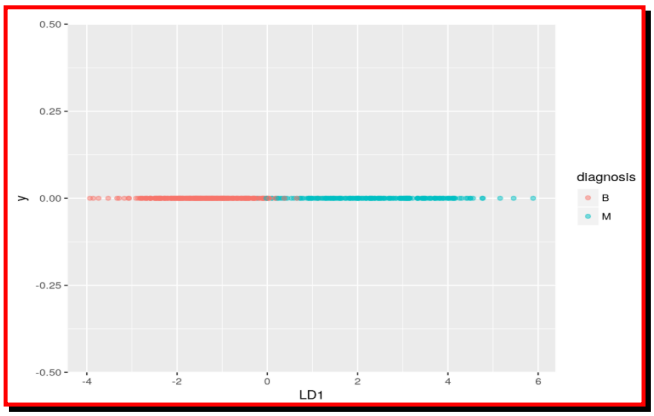

Fig.6.1.LDA pre-processing

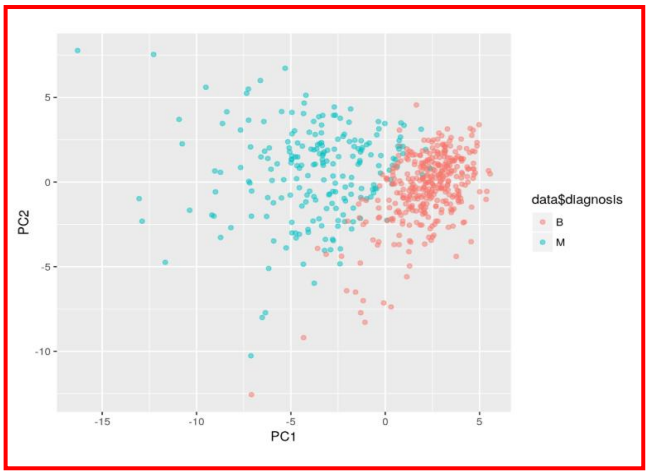

Fig.6.2.ROC Comparision Curve

Fig.6.3. Shows the ROC (Receiver Operating characteristic Curve) curve which is the another metric to measure the performance of a classifier model. The models LDA_NNET and LDA_NB achieve a great ROC.

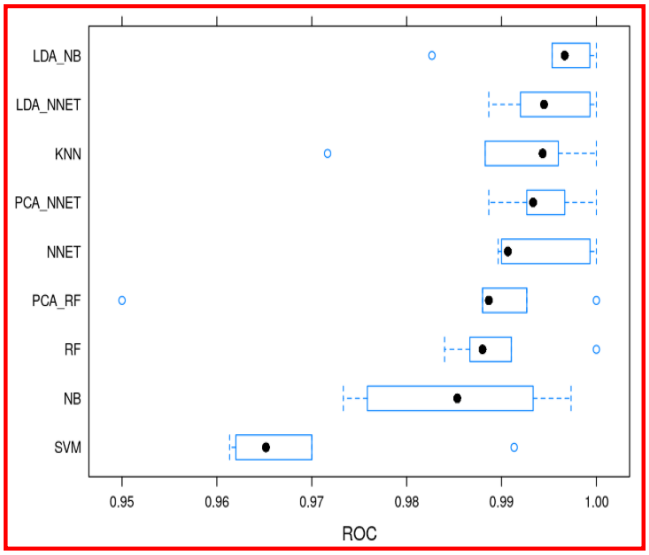

Fig.6.3.ROC Comparision Curve

Table-I represents the comparitive analysis of various machine learning algorithms such as SVM (Support Virtual Machine), Naive Bayes, LDA, PCA and Neural Networks.

Table-I: Various Machine Learning: Comparative Analysis

Classifier	*Accuracy*	*Sensitivity*	*Specificity*	*Kappa*
Support Vector Machine	*76%*	*0.968*	*0.57*	*0.47*
Naive Bayes	*90%*	*0.888*	*0.925*	*0.8115*
Naive Bayes + LDA	*97%*	*0.968*	*0.99*	*0.962*
Random Forest	*94%*	*0.9*	*0.98*	*0.8978*
Random Forest + PCA	*92%*	*0.8413*	*0.9907*	*0.8571*
Neural Networks	*96%*	*0.9524*	*0.9813*	*0.9367*
Neural Networks + PCA	*94%*	*0.9524*	*0.9439*	*0.8876*
Neural Networks + LDA	*99%*	*0.9841*	*0.9907*	*0.9748*

VIII. FUTURE ENHANCEMENT

ML has been utilized in different fields of medication. Choice help devices dependent on ML models can be produced for bosom malignant growth guess. Such models can add to early analysis and treatment and along these lines improved results. With exhaustive electronic clinical records and propelled ML advancements, a bosom malignancy metastasis forecast model can be utilized as a choice help apparatus for care intercession.

IX. CONCLUSION

ML has been utilized in different fields of medication. Choice help devices dependent on AI models can be produced for bosom malignant growth guess. Such models can add to early analysis and treatment and along these lines improved results. With exhaustive electronic clinical records and propelled AI advancements, a bosom malignancy metastasis forecast model can be utilized as a choice help apparatus for care intercession.

REFERENCES

1. Wang, J. Feng, Q. Bu et al., “Breast mass detection in digital mammogram based on gestalt psychology,” Journal of Healthcare Engineering, vol. 2018, Article ID 4015613, 13 pages, 2018
2. Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi “Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms”, 2019.
3. Bichen Zhengh, Sang Won Yoon, Sarah S,Lam “Breast Cancer Diagnosis based on feature extraction using a hybrid of K-means and svm algorithms”, 2014
4. Praful Agarwal, Mayank Vatsa, Richa Singh “Saliency based mass detection from screening mammograms”. 2013.
5. N. Mao, P. Yin, Q. Wang et al., “Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study,” Journal of the American College of Radiology, vol. 16, no. 4, pp. 485–491, 2019

6. J. Ferlay, I. Soerjomataram, R. Dikshit et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," International Journal of Cancer, vol. 136, no. 5, pp. 359–389, 2014.
7. H. Wang, J. Feng, Q. Bu et al., "Breast mass detection in digital mammogram based on gestalt psychology," Journal of Healthcare Engineering, vol. 2018, Article ID 4015613, 13 pages, 2018
8. A. P. Charate and S. B. Jamge, "The preprocessing methods of mammogram images for breast cancer detection," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 1, pp. 261–264, 2017.
9. Ch.Shravya, K.Pravika, Shaik Subhani, "Prediction of Breast Cancer using Supervised Mzchine Learning Techniques", vol. 8,no. 6, 2019.
10. Sreeja N.K., Sankar A, "Pattern matching based classification using Ant Colony Optimization based feature selection", Applied Soft Computing Journal, vol.31, Issue 2818, 2015
11. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012.
12. L. Breiman. Random forest.Machine Learning, 45:5–32, 2010.
13. Selvy P.T., Palanisamy V., Purusothaman T, "Performance analysis of clustering algorithms in brain tumor detection of MR images", European Journal of Scientific Research, vol 62, Issue 3, 2011.
14. Cancer Genome Atlas Research Network. Nature, 513(7517):202–209, 2014.
15. J. Wu, X. Zhao, Z. Lin, and Z. Shao. A system level analysis of gastric cancer across tumor stages with RNA-seq data. Molecular BioSystems, 11(7):1925–1932, 2015.

AUTHORS PROFILE

S. Supraja doing her Bachelor of Engineering in Sri Krishna College of Technology. Her area of interests are Neural Networks and Machine Learning. She published her papers in various journals.

N. Vasuki doing her Bachelor of Engineering in Sri Krishna College of Technology. Her area of interests are Neural Networks and Machine Learning. She published her papers in various journals.

N. J. Vishwa Dhakshana doing her Bachelor of Engineering in Sri Krishna College of Technology. Her area of interests are Neural Networks and Machine Learning. She published her papers in various journals.

Kiruthiga N. is currently working as Assisant Professor in the Department of Computer at **Sri Krishna College of Technology**, Coimbatore,Tamil Nadu,India. She received her Bachelor's degree from Anna University, Chennai in 2013 and her Master's as a gold medalist in the year 2015. er research interests include IoT & Wireless Networks.She is a life member of ISRD, member in ISTE and IAENG.