© July 2020 | IJIRT | Volume 7 Issue 2 | ISSN: 2349-6002

Performance Comparison Of Classification Alogrithms For Diagnosis Of Breast Cancer

Parimala.S[1], Dr.Senthil Vadivu[2]
[1]*Research Scholar, Assistant Professor, Department of Computer Science, SRM Institute of Science and Technology, Chennai, India*
[2]*Head, Department of Computer Applications, Hindustan Arts and Science College, Coimbatore, India*

***Abstract*— Breast cancer is the one of the maximum analyzed cancer amongst women everywhere in the world. The growth and the expansion diagnosis tools is indispensable assist the pathologists to correctly infer and classify between malignant and benign tumors. Breast Cancer which took away 6, 25,000 lives alone. Because of the less attention of the breast cancer by not detecting the early stage and if it is done preventive measures can be taken to reduce the death rate. There are massive amount of technologies and methods are available in data mining area in predicting Brest cancer. This simulation paper work emphases on dissimilar classification techniques and implementation for data mining in prophesying malignant and benign breast cancer. The experimental dataset is taken from the Breast Cancer Wisconsin data set UCI repository while parameter clump thickness being used as assessment class.**

***Index Terms*- Breast cancer, Classification algorithms, UCI repository, data mining**

I. INTRODUCTION

Data mining is the technology of examining the possible stimulating, valuable, and previously unidentified outlines of an enormous cluster of data. Data mining is a field, which consists of areas like, pattern recognition, artificial intelligence, statistics information retrieval, machine learning, and knowledge-based systems data visualization. Database technology, neural networks, and high-performance computing. Of the 184 major countries in the world, breast cancer is the most common cancer diagnosis in women in 140 countries (76%) and the most frequent cause of cancer mortality in 101 countries (55%) [1].

Cancer is one of the extreme shared diseases in the world that the common consequences is death. Cancer is originated by the unrestricted growth of cells in any of the tissues or parts of the body. Table 1 shows the ranking of top twenty-five countries most affected by breast cancer [2]. Only when we find a person's health condition at the very beginning stage can avoid spreading the cancer to malignant stage to save the patient's life .Only timely finding of cancer at the beginning stage and avoidance from distribution to other parts in malignant stage could save a person's life. Cancer is a possibly noxious disease created typically by ecological issues that modify the genes encrypting dangerous cell-regulatory proteins.

The subsequent aberrant cell behavior leads to extensive masses of abnormal cells that abolish nearby normal tissue and can spread to energetic organs, resulting in scattered disease [3], usually a harbinger of imminent patient death. The poor regions can be made aware by familiarizing with early diagnosis program, as state by World Health Organization. It includes early diagnosis, screening, mammography and Clinical Breast Exam (CBE) [4].

II. RELATED WORKS

A. Farah Sardouk et al, [5] analyzed the performance of six different algorithms from each classification and found out which algorithm performs best in prediction using the dataset. They have concluded that RBFS - Recursive Best-First Search Algorithm and Multilayer Perceptron are the top two best algorithms in neural network algorithms.

B. Amr hassan et al, [6] Experimented in his paper that SVM parameter finds the near optimal solution for the evaluating the efficiency of the proposed model. The dataset are being taken from Breast Cancer Coimbra,UCI Library. The results of the experiments illustrates that this results yield the Experimental results demonstrated that our proposed approach can yield hopeful results for breast cancer diagnosis in evaluation to some preceding works and conservative classification methods

C. .Chaurasia V et al, [7], discusses the use of available technological advancements to develop prediction models for breast cancer. The manuscript used Naïve Bayes, RBF Network and J48 to develop prediction model by 10 fold cross-validation method for measuring the unbiased estimate of these models for performance comparison [7]. Verma D et al,.[8], used five classification algorithms Naïve bayes, SMO, REP Tree, J48 and MLP upon two data sets which are breast cancer and diabetes respectively, from the UCI ma-chine learning repository [8]

D. .Rodrigues. B.L, associates two machine learning techniques to create classifier that can distinguish benign from malignant breast lumps [9]. Wisconsin Breast Cancer Diagnosis data set is used for this resolution. The script also deliberates the vision of statistics and how to contract with the lost values and avoid overfitting or underfitting of the applied classifiers [10]. Saxena S, executed neural network for classification of breast cancer data. The paper studies various techniques used for the analysis of breast cancer using ANN and discusses its accuracy [11].

III. METHODOLOGY

A.THE DATASET

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset [12] was used to implement the machine learning algorithms (in Section 2.4) for breast cancer diagnosis. K. Sivakami et al., [13] recommended a for breast cancer patients, a hybrid classification algorithm which integrate DT and SVM algorithms. The proposed algorithm was self-possessed of two main phases. The first stage is Evidence treatment and selection abstraction followed by DT-SVM hybrid model forecasts. The Classification had two principal stages, Training and testing stages. The input parameters for SVM were enhanced using DT algorithm. It's estimated more than 2 million new cases of breast cancer occurred worldwide among women and men in 2018 [14].

The below table depicts the various attributes of the dataset. Wisconsin Breast Cancer Diagnosis dataset from UCI repository and other public domain available data set are used to train the model [16-21].

Table 1: List of Attributes of the dataset

Sl. No	Attributes	Scales
1	Sample code number	ID number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 - 10
6	Single Epithelial Cell Size	1 - 10
7	Bare Nuclei	1 - 10
8	Bland Chromatin	1 - 10
9	Normal Nucleoli	1 - 10
10	Mitoses	1 - 10
11	Class	2 (Benign) , 4 (Malignant)

IV MACHINE LEARNING TOOLS AND BREAST CANCER DATA SET & RESULTS

The collected real world databases from different experts of oncologists are extremely supported with vulnerable and noise threats. More missing values were recognized from the collected data and for mining purpose it will not be in proper format. Hence the crucial step of preprocessing must be used to such a type of databases for the purpose of classification. There are different datasets available for the researchers to estimate their algorithms. Some of the common breast cancer datasets are exposed in table.2.

Dataset Name	Description
Breast Cancer Wisconsin (Original)	Original Wisconsin Breast Cancer Database
Breast Cancer Wisconsin(Prognostic)	Prognostic Wisconsin Breast Cancer Database
Breast Cancer Wisconsin (Diagnostic)	Diagnostic Wisconsin Breast Cancer Database
Haberman's Survival	Breast cancer survival data
Breast Tissue	Electrical impedance quantities of freshly cut out tissue sections from the breast.

V. COMPARATIVE STUDY OF VARIOUS PROPOSED WORKS

A. Novel Multi Layered Method:

P. Ramachandran *et al.*, [2] suggested a novel model effort in which the self-possessed data are pre-processed and put in the knowledge base to build the archetypal. Seventy five percent of the total data are taken as the training set to construct the cataloguing and clustering model the residual of which is taken for stimulating purpose. The model is then tested for correctness, indulgent the model is being tested and then finally integrated it to the knowledge base. To conclude the prototype is being calculated using SVM.

B. Cancer Prediction System

K. Arutchelvan *et al.*, [3] suggested a planning, in which data mining technique placed cancer prediction system relating the prediction system with mining technology was used. In this model, the authors have used one of the classification algorithms called decision tree.

C. Adaptive Neuro Fuzzy Inference System (ANFIS)

A new Adaptive Neuro Fuzzy Inference System (ANFIS) projected by C. Kalaiselvi *et al.*, [10]. The chronic and long-lasting, severe diseases like diabetes and cancer have difficult connection. Adaptive neuro-fuzzy inference system (ANFIS) is one such model, and the findings point to better results of machine training techniques (such as artificial neural networks) than advanced statistical methods based on cox-regression [24].

Author	Method	Algorithm	Evaluation
P. Ramachandran *et al.,*	Novel Multi Layered Method	Decision tree k-means clustering algorithm	Weka
K.Sivakami *et al.,*	Hybrid Classification Algorithm	DT and SVM algorithms	Weka
K. Arutchelvan *et al.,*	Cancer Prediction	Decision Tree	Weka
C. Kalaiselvi *et al.,*	Adaptive Neuro Fuzzy Inference System	k-means Algorithm	Weka

VI. CONCLUSION AND FUTURE WORK

Data mining is the technique of significant patterns in massive data sets encompassing approaches at the connotation of machine learning, database systems and statistics. The main objective of the data mining process is to mine statistics from the dataset and modify it into a comprehensible format for additional usage. Worldwide the cancer has come to be the principal cause of death. The finest operational method to diminish cancer deaths is to notice it formerly. Numerous persons avoid cancer screening due to the value involved in taking many tests for examination. The prediction system may offer easy and a cost effective way for showing cancer and may play a key role in the previous analysis process for dissimilar types of cancer. By using the data mining classification algorithms because classification algorithms the future work can be focused on to analyze the data set of breast cancer. Classification algorithms can illustrate the improved performance in predicting Breast Cancer emphasis based the correctness and exactness of the datasets.

REFERENCES

[1]The Cancer Atlas- http://canceratlas.cancer.org/the-burden/breast-cancer/

[2]American Institute of Cancer Research Statistics - https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics

[3]K. Arutchelvan and R. Periyasamy, “Cancer Prediction System Using Data Mining Techniques”, International Research Journal of Engineering and Technology, Vol.2, No. 8, pp. 1179-1183, Nov. 2015.

[4] Kumar, V., Tiwari. P, Mishra. B.K., Kumar. S.: Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 7(1), pp. 30–41. DOI: https://doi.org/10.4018/ijkdb.2017010103 (2017).

[5] Farah Sardouka*, Dr. Adil Deniz Durub, Dr. Oğuz Bayat “Classification of Breast cancer using Data mining “,American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), Volume 51, No 1, pp 38-46(2019)

[6] Hamouda, Saeed and Hassan, Amr and Wahed, Mohammed E. and Ail, Mohammed and Farouk, Osama, Tuning to Optimize SVM Approach for Breast Cancer

Diagnosis with Blood Analysis Data (February 12, 2020). Available at SSRN: https://ssrn.com/abstract=3537067

[7]Chaurasia V., Pal., S, Tiwari., BB.: Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, Vol. 12(2), pp. 119–126. DOI: http://dx.doi.org/10.1177/1748301818756225. (2018).

[8]. Verma, D., Mishra. and N.:Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques. In Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS), pp 533-538, (2017).

[9]Charles Edeki and Shardul Pandya, "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability", Mediterranean journal of Social Sciences., Vol. 3, No. 14, pp. 49-56, Nov. 2012.

[10]Rodrigues., B.L.: Analysis of the Wisconsin Breast Cancer Dataset and Machine Learn-in for Breast Cancer Detection. In: Proceedings of XI Workshop de Visão Computa-tional, pp 15-19, (2015).

[11] Saxena., S, Burse., K.: A Survey on Neural Network Techniques for Classification of Breast Cancer Data. In; International Journal of Engineering and Advanced Technology, Volume-2, Issue-1, pp 234-237, (2012).

[12] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/] (1992).

[13] K. Sivakami, "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model", International Journal of Scientific Engineering and Applied Science, Vol. 1, No. 5, pp. 418-429, Aug. 2015.

[14] International Agency for Research on Cancer (IARC) and World Health Organization (WHO). GLOBOCAN 2018: Breast. http://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf, 2018.

[15] Cohen J (1960) A coefficient of agreement for nominal scales. Education and Psychological Measurement 20: 37–46., Google Scholar

[16]Saxena., S, Burse., K.: A Survey on Neural Network Techniques for Classification of Breast Cancer Data. In; International Journal of Engineering and Advanced Technology, Volume-2, Issue-1, pp 234-237, (2012).

[17]UCI Machine Learning Repository: Breast Cancer Wisconsin Dataset, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29

[18]. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[19]. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

[20]. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear pro-gramming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadel-phia 1990, pp 22-30.

[21]. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Method and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

[22] P. Ramachandran, N. Girija, and T. Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications, Vol. 97, No. 13, pg. 48-53, July 2014.

[23] C. Kalaiselvi and G.M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer using ANFIS", World Congress on Computing and Communication Technologies, pp. 188-190, 2014.

[24] Akl A, Ismail AM, Ghoneim M. Prediction of graft survival living-donor kidney transplantation: nomograms or artificial neural networks? Transplantation. 2008; 86(10):1401-6. DOI: 10.1097/TP.0b013e31818b221f PMID: 19034010