A Research on Breast Cancer Prediction using Data Mining Techniques

R.Preetha, S. Vinila Jinny

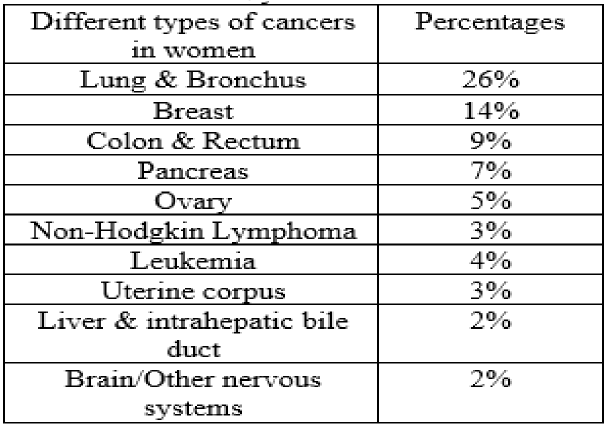
Abstract— Early detection and diagnosis of breast cancer plays a significant role in the welfare of women. The mortality rate due to breast cancer is on an all-time high. Factors such as food habits, environmental pollution, hectic lifestyle and genetics are commonly attributed to breast cancer. In order to detect and diagnose such types of cancer, intelligent systems are implemented. Automated diagnosis gets impacted by prediction accuracy when compared with surgical biopsy. Bioinformatics mining has emerged as the area of research that involves analyzing both data mining and Bioinformatics. In order to statistically find significant associations on a breast cancer data set, the result is conceivable. Using a larger data set results in discovering the correlations between a bigger set of gene. The algorithm has to be improved to perceive the interactions with low marginal. This research field affords most intelligent and reliable data mining models in breast cancer prediction and decision making. This survey reviews various data mining algorithms on large breast cancer biological datasets. The merits and demerits of various procedures and comparison of their corresponding results are presented in this work.

Keywords: Diagnosis, Breast cancer, Data mining models.

I. INTRODUCTION

Breast cancer severely affects the physical and mental health of a woman. There are many factors contribute to the cause of breast cancer. Amongst [1] those diet and genetics are considered major aspect. Research is ongoing to predict the causes and effects of this disease. According to American Cancer Society Statistical Information of 2013, Breast cancer is the second most fatal disease that causes death in women. Table 1 provides the leading type of cancer and its corresponding percentages in women.

Table 1 Summary of main cancers in women



Different types of cancers in women	Percentages
Lung & Bronchus	26%
Breast	14%
Colon & Rectum	9%
Pancreas	7%
Ovary	5%
Non-Hodgkin Lymphoma	3%
Leukemia	4%
Uterine corpus	3%
Liver & intrahepatic bile duct	2%
Brain/Other nervous systems	2%

The possibility prediction suggests that the quality of life corresponds with the bioinformatics. Presently, records mining contributes masses in medical location for the early prognosis of ailments, threat element assessment, choice making, remedy and prescription of medicine. The unseen records report of affected man or woman is used for the prediction of disorder in magnificence segment of facts mining. This disease can be classified using various morphology factors. The accuracy of this type of model provide better results than normal traditional techniques.

Data mining may be implemented in the trouble of bioinformatics in numerous packages such like [2] gene finding, [3] protein detection, [4] characteristic motif detection, [5] protein function inference, [6] disease evaluation, [7] contamination evaluation, [8] disease medicine optimization, [9] protein and gene cooperation community regeneration, data cleaning, and [10]protein sub-cellular location prediction. The facts mining method might be used for the assessment of breast maximum cancers the use of elegance algorithms. The mining process reveals several secreted and strange patterns, which might not focused before.

Two types of Tumors are diagnosticate since so far. They are Benign tumors which are generally not dangerous and Not often invade the tissues all over the place. This type of tumors may not blowout to further portions of body and this could be easily removed and not makes any further problem. The Malignant tumors seems to be hazard to life, which permeate adjacent tissues and muscles. This form of tumors spread to further portions of body and removing it is useless. Generally breast cancers are diagnosed by Mammograms, Breast Ultrasound, Breast magnetic resonance imaging Scans and Breast diagnostic assay. Typically breast cancers are identified inside the clinical workplace. But facts pushed completed math assessment on carcinoma permits to are looking forward to the survival evaluation of the infection. It is a contestable task to foresee the irruption with the employment of information mining technique. Large volume of amassed medical dataset draws the mining analyst to seek out the contamination patterns and their relationships. This carried out math mining technique ought to analyze good sized quantity of variable elements and clinical ancient facts. [11] Country wide most cancers statistics (NCDB) and additionally the [12] Surveillance, remedy, and completing (SEER) deliver dataset to the researchers. This dataset is made with relevant medical, extrapolative, demographic, mystifying, and outcome variables in real affected person citizens.

Revised Manuscript Received on August 05, 2019.

R.Preetha, Research Scholar, Department of Computer science and Engineering,Noorul Islam Center for Higher Education,, Kumaracoil , Tamil Nadu – 629180, India

S. Vinila Jinny, Associate Professor, Department of Computer science and Engineering, Noorul Islam Center for Higher Education,, Kumaracoil , Tamil Nadu – 629180, India

The effective automated pc Aided analysis device need to classify the breast most cancers dataset that modified into gathered from numerous oncologist from various worldwide places. The accuracy of the beauty strategies suffers with the first-rate of information. The outliers and imbalanced bioinformatics dataset affects the stableness and popular usual overall performance of the prediction models. This prediction facts mining algorithms has to face the following challenges.

- Selecting appropriate attributes to gather the extrapolation model.
- Identity and removal of outliers with the bioinformatics records.
- Handling imbalanced information in elegance.
- Differentiating regular genetic skills and infection morphological factors.

the primary cause of this survey is to summarize numerous statistics mining methodologies on evaluation and assessment of breast most cancers.It provide all recent research carried out related to this corresponding field with various breast cancer dataset.

In this process, the various issues in preprocessing and classification are analyzed. The different existing approaches and their merits and demerits were analyzed. It briefly explore the risk factors which causes breast cancer to the humankind. The main aim of this survey works like a compass to direct a best CADS to protect the women from breast cancer disease.

This survey is well ordered as follows. The section II exhibits the risk factors that increase the possibilities of breast cancer in women. Section III describes the major appliances of data mining classification like neural networks. The section IV studies the Problems of breast cancer data for effective mining. *S*ection V and VI delivers the data mining steps and description of various datasets and tools respectively. Finally, remarks of the data mining classifications in breast cancer are presented in the section VI.

II. RISK FACTORS

Several risk factors are determined for the cause of breast cancer. Though all women has the chance to be affected by the breast cancer, these factors influence more. Some of these risk factors can be reduced and some cannot be changed. Parker and Folsom[13] discovered that Intentional weight loss of women increase the chance of getting breast cancer. Wasserman et al.[14] stated that stoutness and nourishment habits influence the menace of breast cancer in postmenopausal females[15]. Kullberg et al[16] studied both white color and blue color job women dataset and concluded that white color job women are at high risk. It is because of life style and reproduction style. McPherson et al.[17] surveyed all risk factors which cause breast cancer and concluded that developed countries are more inconvenient than developing country. Kamdar et al.[18] Akerstedt et al [19]focused the women who are working in night shift and concluded that they are in very risk zone for breast cancer. Cybulski et al.[20] studied the influence of genetic disorders and genetical issues accompanying with breast cancer. This work discovered many genetically related proteins which was reason for causing breat cancer.

Majoor et al.[21] analyzed the age risk factor in every countries. The breast cancer is directly proportional to the age risk factor. The color of the women also participates in causing breast cancer in women. Evans et al.[22] proposed the influence of Family history, and stated that the women has high risk chance of getting breast cancer when anyone of her close blood relation having this disease. Kamińska [23]proposed various risk factors which causes breast cancer. This research discovered that the women has high risk of breast cancer when they are with condensed breast tissue. If there is sudden change in the size of breast tissue, it may lead to breast cancer. Reed et al.[24]Lobular carcinoma situation arise when the cells that appear similar to cancer cells are in the Breast tissue, on the other hand they may not mature over the wall of the lobules. This condition is not a hunky-dory cancer or pre-cancer. Eventually the probabilities of possibility increases with having LCIS. Harris[25] statistically studied that Women who instigated periods earlier to teen age or attaining menopause later the age of 55 have a marginally increased risk of breast cancer[26]. [27] proposed that the risk increases when teen-age women concedes radiation treatment because of immature breasts. Troisi et al.[28]proposed that the women without having children has higher risk. This work also revealed that the women who postpone the first child still to age 30 has high risk. This research also forced the pregnancy women not to consume the diethylstilbestrol medicine because of risk of breast cancer. This paper also suggested early pregnancy and attaining many pregnancies. Need et al. [29] suggested not to consume estrogen and progesterone later menopause surges the possibility of attaining breast cancer.

III. DATA MINING IN BREAST CANCER

The analysis can be done with breast cancer dataset for the disease diagnosis, treatment prediction, and prognosis of diseases, the risk awareness prediction and survival of breast cancer. The advantage of using data mining in the disease diagnosis is that it can mine the large medical dataset within a short time. The prediction accuracy is high and reliable. The result will be based on the various attributes in multidimensional view.

Neural networks is a promising approach in many classification systems and business forecasting. It is a bio inspiration approach, which works like a brain in dealing with information processing. This method works based on three stages that are1.architecture or model 2. Learning algorithm 3. Activation functions.

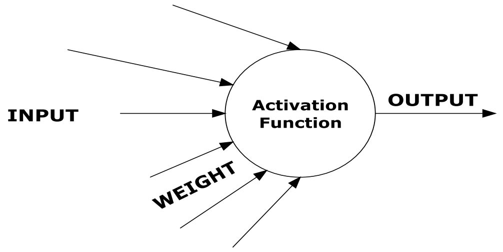


Figure. 2 Artificial Neuran

This method implies supervised learning using Multi-Layer Perceptron (MLP) with backprogration by applying a set of _weights to unite between input and output units. This method seems to be complex and understandable structure. However, the prediction accuracy is very high. It can tolerate high noisy data and consumes very long running time. There are many neural network models to classify the breast cancer medical data.

Decision tree induction approach is one of fundamental approach in knowledge discovery process. The tree structure consist of root node, intermediate nodes and leaves. The intermediate nodes symbolize the decision test and the results are denoted by the edges. The leaf node is associated with a class label. This method works based on the rules established. This technique may be very effective method for the magnificence of breast most cancers. The gain of this model is that it is very fast and there's no need of know-how area information.

Rule based really is one of the classical smooth comprehensible method in statistics mining. Each characteristic involves in condition check analysis and consequences are generated. The situation checking may be both from choice tree or from separate schooling dataset. Many studies studies have hired in rule-based definitely algorithms to assemble prediction regulations the use of breast most cancers facts gadgets.

Manual Vector device uses neural networks to discover a linear finest hyper aircraft to get the out of the margin close by the extrication hyper aircraft. the choice feature is certainly quantified with the aid of a subgroup of schooling samples. The SVM model produce higher accuracy and flexible consequences. Widespread have a observe has been partaken with help Vector machine (SVM) to supply the prediction fashions

IV. PROBLEMS OF BREAST CANCER DATA IN DATA MINING

In order to accomplish accuracy in classification results, purity of breast cancer dataset must be high. The noisy data is the biggest problem of data mining. Selection of appropriate purified dataset can have great impact on quality of extracted knowledge. Generally, the cancer data is retrieved manually from the database, which is not the exact data-mining format. The dataset suffers from missing data, outliers and imbalanced data. These type of impurities directly affects the accuracy of the prediction models.

Missing data:

Missing data mentions that some of the data will not be available or applicable. This type of missing values severely degrade the classification algorithms. Breast cancer dataset suffers with three forms of missing data. Secondly, the unrecorded events happened in the lab. Finally, the intermediate results, which was neglected at the time of data aggregation.

Outliers:

Outliers are the invariant data that may follow the general category. This type of data severely affects the prediction results. These outlier data can be detected using various approaches like including, outlier filtering, outlier correction and robust algorithms.

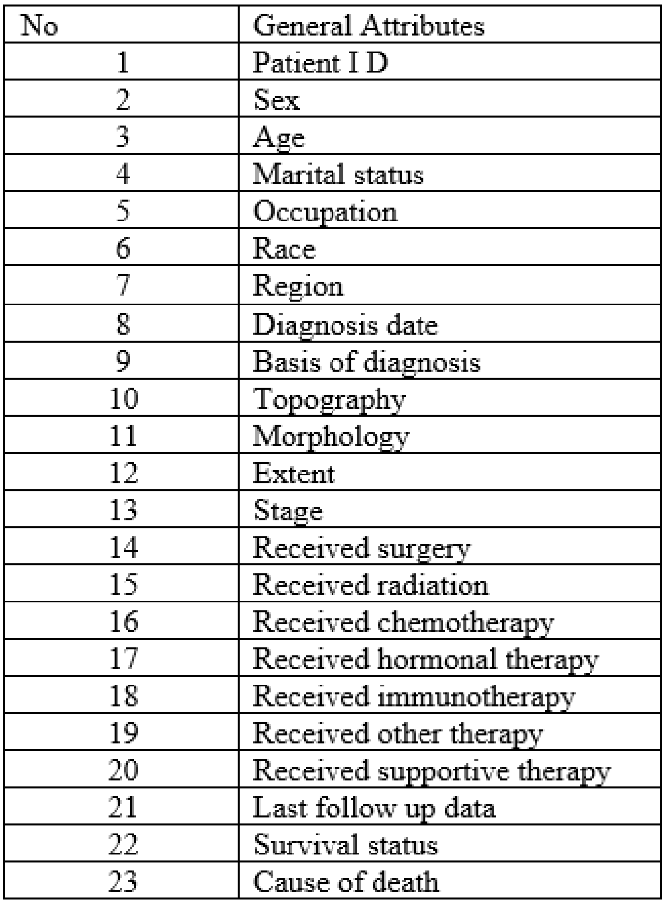
Imbalanced data:

The various dataset suffers with unbalanced attributes for matching. The set of attributes of one particular class may not exist same in other class of dataset. Generally sampling methods are used to solve this problem in the phase of preprocessing.

Breast cancer attributes:

The accuracy of the data mining classification increases with the increasing of attributes in the dataset. Every researcher use the attributes corresponding to the dataset they choose. Table 2 indicates various attributes used for the classification.

Table. 2 Common List of attributes for breast cancer data



No	General Attributes
1	Patient I D
2	Sex
3	Age
4	Marital status
5	Occupation
6	Race
7	Region
8	Diagnosis date
9	Basis of diagnosis
10	Topography
11	Morphology
12	Extent
13	Stage
14	Received surgery
15	Received radiation
16	Received chemotherapy
17	Received hormonal therapy
18	Received immunotherapy
19	Received other therapy
20	Received supportive therapy
21	Last follow up data
22	Survival status
23	Cause of death

Every general attributes has sub attributes related to the breast cancer patient.

V. DATA MINING CLASSIFICATION IN BREAST CANCER DATASET

Data mining is a sub task of discovery way. discovery is described as extracting worthwhile from the given dataset. It has the following iterative device.

1. Data cleansing: noisy and beside the point records are purified and detached from the assembled information.

2. Data integration: more than one heterogeneous information assets, mixed to a commonplace supply.

3. Data desire: the preferred statistics is chosen from the database for in addition research.

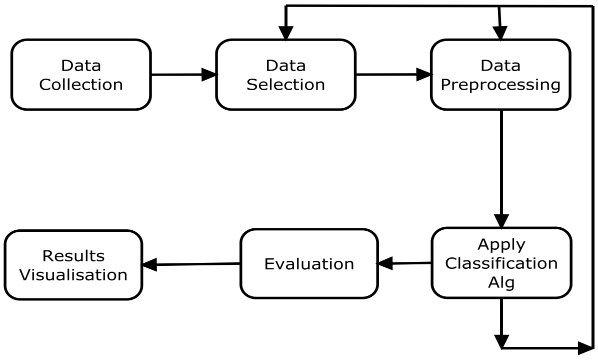


Figure. 1 Research Method Overview

4. Data transformation: the chosen suitable information is transformed to the perfect form if you want to perform statistics mining.

5. Facts mining: techniques are employed to extract effective styles

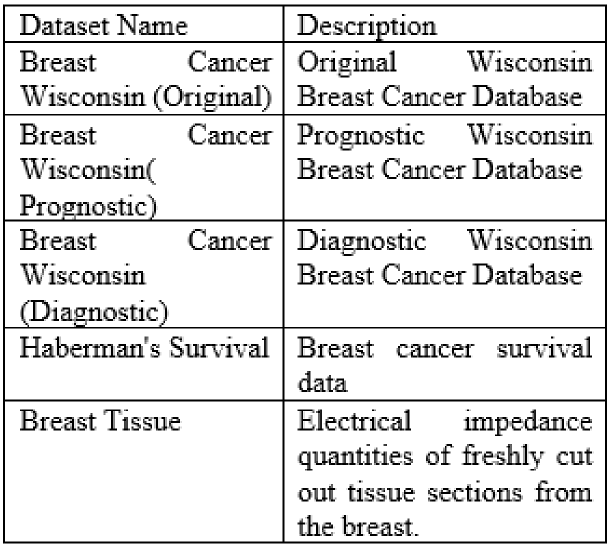
6. Pattern assessment: thrilling patterns are completed which is the image of information are diagnosed subjected on given measures.

7. Expertise example: uncovered information are visually exemplified to the consumer.

VI. MACHINE LEARNING TOOLS AND BREAST-CANCER DATA SET & RESULTS

The real world databases that is collected from various oncologist are highly assisted with noise and vulnerable threats. These collected data will have more missing values and the data will not be in correct format for mining. Hence preprocessing is a essential step to use such kind of databases for the classification purposes. There are many datasets available for the researchers to evaluate their algorithms. Some of the common breast cancer datasets are exposed in table.2

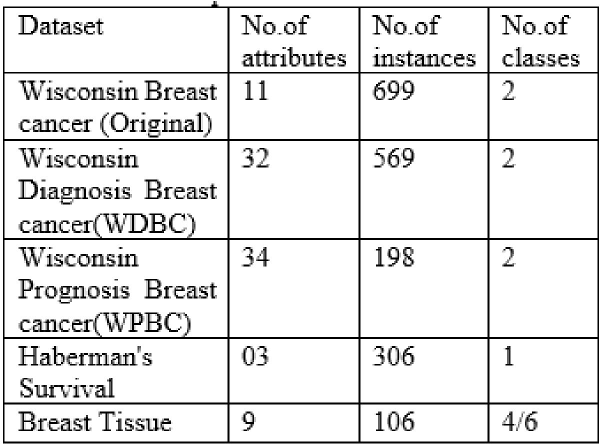
Table 2. List of Breast Cancer Dataset



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.

Every dataset has different attributes and different instances. Table. 3 mention the number of attributes for common breast cancer dataset.

Table.3 Description of Breast Cancer Dataset



Dataset	No.of attributes	No.of instances	No.of classes
Wisconsin Breast cancer (Original)	11	699	2
Wisconsin Diagnosis Breast cancer(WDBC)	32	569	2
Wisconsin Prognosis Breast cancer(WPBC)	34	198	2
Haberman's Survival	03	306	1
Breast Tissue	9	106	4/6

At hand several open source data mining tools are obtainable to classify our dataset. We can list out some of the data mining tools which is very effective for our breast cancer prediction. Rapid Miner, **Sponsor Note,** WEKA**, R-Programming,** Orange, KNIME and NLTK. Among this tools, the WEKA is very useful tool which need not to know more about the tool mining techniques. It supports GUI based learning mechanism. It is a data mining tool has been accustomed to find out attention-grabbing Patterns in the designated dataset. This open source tool has all the functions and machine learning algorithms for the classification of dataset directly or calling code from Java.

VII. SUMMARY

We surveyed the automated breast cancer classification approaches from the manual biopsy test to the recent research work going in this field. First stage concern with dataset collection which can be real world dataset or breast cancer dataset. The second includes the development of procedures in preprocessing to enhance facts high-quality in case you want to decorate the overall performance, stability and effectiveness prediction fashions. The final stage involve with .Knowledge discovery which can use various classification algorithms for the prediction. We surveyed classification process with major models like neural networks, choice tree technique, rule primarily based mining strategies and help Vector device. The various classification approaches tries to find a better optimal discovery mechanism. But the resultant accuracy is not still up to the mark. A complete assessment of the self-regulated analysis of breast cancer detection approaches are characterized in Table 4.

Table 4 Comprehensive Review On Breast Cancer Prediction Exhausting Data Mining Techniques

Technique	Author & Ref No.	Year	Description	Advantages	Disadvantages
Performance comparison of Naïve bayes and J48 decision trees process.	Williams et al.[32]	2015	J48 choice bushes exhibited a better accuracy with lesser mistakes expenses related to that of the naïve bayes' method through the usage of the usage of 17 attributes with LASUTH breast most cancers facts.	1. Great outcomes in assessments. 2. J48 selects most discriminatory features. 3. High tolerant of noisy data.	1. Computational cost of J48 is high. 2.J48 reduces unambiguousness to the related model.
Mining the microarray structure of genes related information	Jauhari and Rizvi[33]	2014	SVM classification based on Gene importance and enrichment ranking techniques.	1. High accuracy because of mining gene expression data. 2. Diagnosis of complex proteins in gene data.	1. It is difficult to collect gene expression data. 2. 458 genes are identified yet which cause breast cancer.
Classification based on Taxonomy of patients history (Taxofinder)	Kang and Haghigh [34]	2016	Taxonomy learning approach collects domain related concepts and represent it via graphs. Then analytic algorithm learn the taxonomy related classification.	1.This method has no certainty predefined lexico syntactic designs that dependably require additional examination	1. The dependency of events and attributed are uncategorized. 2. Very complicated.
probability of breast cancer Recurrence used for classification algorithms.	Pritom and Sabab et al[35]	2016	Ranker algorithm is used to foretell the recurrence probability of breast cancer. The best feature selection approach is used to pinpoint the recurrence possibility.	1. Extracts meaningful technique for efficient classification	1. This is not adoptable for real time database implementation.
DNN and multi criteria decision making combined with Analytical Hierarchy Process	Soliman et al.[36]	2016	The output of DNN Multilayer Perceptron layer is integrated with AHP which works based on multiple criteria decision.	1. Accuracy, efficiency, less time for classification tasks. 2. Strong against noisy data. 3. useful for large and small datasets	1. DNN over fit with attributes and mislead and can cause health complication. 2. need of huge memory. .
K-means algorithm is integrated with PSO	Salma and Doreswamy [37]	2016	Select top N strong features from high dimensional breast cancer dataset for better classification of breast cancer disease.	1. Very helpful approach for attribute selection. 2. Robust and balanced than conventional techniques.	1. Not applicable for small data.

NN used for classification using genome variability.	Boutorh and Guessoum [38]	2015	The dimensionality feature reduction of SNP was achieved through hybrid mechanism of Association Rule Mining (ARM) and Neural Networks	1. Association rule method is suitable for extracting hidden information from dataset. 2. This hybrid approach is useful to discover relationship patterns in complex SNP data.	1. Very easy to understand. But it is time consuming. 2. It is not suitable for large dataset.
Biclustering and BP neural network algorithm	Chen et al.[39]	2016	Ultrasound imaging features are biclustered and BP NN algorithm was applied to classify the breast tumors	1. Effective Treatment can be achieved. 2. A robust approximator function classifies effective prediction. 3. adaptable to new attributes of dataset.	1. Time complexity of algorithm is more.
multivariate statistical approach combined with AI	Jhajharia et al.[40]	2016	The features are extracted using PCA algorithm and the resustant features and relevant data is trained using ANN approach.	1. PCA involves in dimensionality reduction when there is need. 2. When PCA works along with ANN, the resultant results will be effective.	
Random tree classifier and ANN	Kaushik and Kaur [41]	2016	Features extracted from the mammogram dataset are used to predict the biopsy results by using random forest classifier and NN.	1. Better performance and reduced dimension. 2. Highly accurate classification.	1. It's more complex. **2.** It's more difficult to implement. **3. H**igh computation cost
Gene expression data are used for classification using Elitism PSO	Nagpa and Shrivas [42]	2015	EPSO is feature selection based classifier based on lazy IBK.	1. Exact selection of gene makes easy prediction of breast cancer. 2. High dimensional data can be mined effectively.	1.Not applicable for small data
Sequential rule and pattern matching	Cheng et al. [43]	2017	Wealthy set of sequential guidelines are hired for the type set of regulations.	1. Minimal help recollect is used to healthy the sequential patterns.	1.Inconsistent and Incomplete records 2. bizarre visits of patients influences sequential approach 3.Imbalance issue 4.Post-analysis issue
Ensemble approach for feature selection mechanism	Dhakate et al.[44]	2015	The ensemble approach is used for feature selection and the accuracy of results are compared without using the technique with same dataset.	1.It provides better accuracy when compare with other techniques	1.Relatively higher computation cost
Comparison of J48, NB, SMO, IBK with three dataset.	Salama et al.[45]	2012	The combined classification of MLP and J48 along with PCA is greater than other classifiers employing WBC data set	1. This work examined various combination of classifiers.	1. Some amalgamation of classifier over fit with data set and some will be under fit with data set.

En based clustering and fuzzy logic implementation combined with PCA	Nilashi et al.[46]	2017	Fuzzy rules are applied with the EM based clustering Mammographic mass datasets. The PCA supports to avoid multi-collinearity in dataset.	1. Prediction accuracy was high.	1. Practical issues at the time of implementation.
fuzzy-rough nearest neighbor method.	Onan et al. [47]	2015	Fuzzy-tough instance choice take away the noise of records. Then consistency-based totally totally sincerely feature choice is carried out. Fuzzy-hard nearest neighbor set of guidelines is used for the elegance.	1.A combined approach 2.Dealt the uncertainty data very well 3. Assisted with noisy data. 4. works fast for small training sets	1. It suffers with large training data.
GA is used for FS and multiple classification approach is employed.	Alicovic et al.[48]	2017	Genetic algorithms are used for putting off of useful and amazing abilities. Discrete person and multiple classifier coordination have been used to make specific technique for breast most cancers cataloguing.	1. Individual accuracy and diversity. 2. Very sensitive and high accuracy. 3. Subset relationship is maintained.	1. More time complexity. 2. individual observation might be a member of no classes or too many classes
Foraging lung cancer using Ant colony optimization	Christopher et al.[49]	2015	The various attributes are taken to compute the multi objective function and determines the increased risk of disease.	1. Changes of attributes can be considered. 2. Multi objective function involved in prediction	1. Attributes cannot be increased much. 2.Computational complexity.
prognosticating breast cancer using random forest classifier	Nguyen et al.[50]	2013	Hybrid technique of random forest classifier and function desire method classifies the breast most cancers records.	1. Perform good with large dataset. 2. Good attribute selection approach. 3. It measure importance of each feature.	1. smaller groups are Favored over larger groups.
FS approach based on Filter and Wrapper	Tomar et al.[51]	2015	The filter approach filters the features based on the ranking and wrapper method validate it with SVM and cross validation approach	1. Non-regular data can be handled easily.	1. High algorithmic complexity. 2. More memory requirement.
K-means clustering combined with SVM	Zheng at al.[52]	2014	The similarities of breast tumors are calculated using K-SVM approach. K-means detects the hidden information of tumors and SVM classifies the new data with the existing pattern.	1. This method applied pattern matching system using SVM. 2. The required attributes only can be selected from large attribute set.	1. Accuracy is less.

VIII. CONCLUSION AND FUTURE WORK

From the aforementioned survey, it is evident that a suitable breast cancer diagnosis model assist scientific practitioners with accurate and dependable prediction effects. Early detection of breast cancer makes it treatable which may lead to save lives. The implied mining technique involves explicit attributes also with implicit gene expression data. With the help of various data mining algorithms, we can detect hidden cancer associated patterns for the classification of risk analysis of breast cancer. Positive and negative consequences of various data mining approaches are discussed in table 4. Various techniques such as neural networks, choice tree technique, rule primarily based definitely absolutely mining techniques and resource vector system have been comprehensively used in prognostication of breast most cancers associated styles. Among the ones techniques, fuzzy logic seems to be useful for the effective diagnosis. We conclude that the automated timely prediction of breast cancer ensures preventive measures for the disease. We need efficient applications for detection and prevention of breast cancer.

REFERENCES

1 T. Srinivasan, A. Chandrasekhar, J. Seshadri, and J. Siddharth, "Knowledge discovery in clinical databases with neural network evidence combination," in *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*, 2005, pp. 512-517.

2 T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, *et al.*, "antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters," *Nucleic acids research,* vol. 43, pp. W237-W243, 2015.

3 J. Ji, A. Zhang, C. Liu, X. Quan, and Z. Liu, "Survey: Functional module detection from protein-protein interaction networks," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, pp. 261-277, 2014.

4 J. D. Watson, R. A. Laskowski, and J. M. Thornton, "Predicting protein function from sequence and structural data," *Current opinion in structural biology,* vol. 15, pp. 275-284, 2005.

5 M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein–protein interaction data," *Journal of computational biology,* vol. 10, pp. 947-960, 2003.

6 [6] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications,* vol. 17, pp. 43-48, 2011.

7 M.-J. Huang, M.-Y. Chen, and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis," *Expert systems with applications,* vol. 32, pp. 856-867, 2007.

8 M. G. Tsipouras, T. P. Exarchos, D. I. Fotiadis, A. P. Kotsia, K. V. Vakalis, K. K. Naka, *et al.*, "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling," *IEEE Transactions on Information Technology in Biomedicine,* vol. 12, pp. 447-458, 2008.

9 J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein–protein interaction network," *Bioinformatics,* vol. 22, pp. 2800-2805, 2006.

10 B. Stapley, L. A. Kelley, and M. J. Sternberg, "Predicting the sub-cellular location of proteins from text using support vector machines," in *Biocomputing 2002,* ed: World Scientific, 2001, pp. 374-385.

11 L. Kujtan, K. F. Kennedy, S. Manthravadi, J. R. Davis, and J. Subramanian, "MINI01. 09: Outcomes of Early Stage Large Cell Neuroendocrine Lung Carcinoma (LCNELC): A National Cancer Database (NCDB) Analysis: Topic: Medical Oncology," *Journal of Thoracic Oncology,* vol. 11, pp. S261-S262, 2016.

12 N. Howlader, A. Noone, and M. Krapcho, "Surveillance, Epidemiology, and End Results (SEER) Program. SEER Cancer Statistics Review, 1975-2012 (updated August 20, 2015). National Cancer Institute, Bethesda, MD, based on November 2014 SEER data submission, posted to the SEER web site, April 2015," ed, 2016.

13 E. Parker and A. Folsom, "Intentional weight loss and incidence of obesity-related cancers: the Iowa Women's Health Study," *International journal of obesity,* vol. 27, p. 1447, 2003.

14 L. Wasserman, S. Flatt, L. Natarajan, G. Laughlin, M. Matusalem, S. Faerber, *et al.*, "Correlates of obesity in postmenopausal women with breast cancer: comparison of genetic, demographic, disease-related, life history and dietary factors," *International journal of obesity,* vol. 28, p. 49, 2004.

15 C. Friedenreich, "Review of anthropometric factors and breast cancer risk," *European Journal of Cancer Prevention,* vol. 10, pp. 15-32, 2001.

16 C. Kullberg, J. Selander, M. Albin, S. Borgquist, J. Manjer, and P. Gustavsson, "Female white-collar workers remain at higher risk of breast cancer after adjustments for individual risk factors related to reproduction and lifestyle," *Occup Environ Med,* pp. oemed-2016-104043, 2017.

17 K. McPherson, C. Steel, and J. Dixon, "ABC of breast diseases: breast cancer—epidemiology, risk factors, and genetics," *BMJ: British Medical Journal,* vol. 321, p. 624, 2000.

18 B. B. Kamdar, A. I. Tergas, F. J. Mateen, N. H. Bhayani, and J. Oh, "Night-shift work and risk of breast cancer: a systematic review and meta-analysis," *Breast cancer research and treatment,* vol. 138, pp. 291-301, 2013.

19 T. Åkerstedt, A. Knutsson, J. Narusyte, P. Svedberg, G. Kecklund, and K. Alexanderson, "Night work and breast cancer in women: a Swedish cohort study," *BMJ open,* vol. 5, p. e008127, 2015.

20 C. Cybulski, J. Carrot-Zhang, W. Kluźniak, B. Rivera, A. Kashyap, D. Wokołorczyk, *et al.*, "Germline RECQL mutations are associated with breast cancer susceptibility," *Nature genetics,* vol. 47, p. 643, 2015.

21 B. C. Majoor, A. M. Boyce, J. V. Bovée, V. T. Smit, M. T. Collins, A. M. Cleton-Jansen, *et al.*, "Increased risk of breast cancer at a young age in women with fibrous dysplasia," *Journal of Bone and Mineral Research,* vol. 33, pp. 84-90, 2018.

22 D. G. Evans, S. Astley, P. Stavrinos, E. Harkness, L. S. Donnelly, S. Dawe, *et al.*, "Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study," 2016.

23 M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, "Breast cancer risk factors," *Przeglad menopauzalny= Menopause review,* vol. 14, p. 196, 2015.

24 A. E. M. Reed, J. R. Kutasovic, S. R. Lakhani, and P. T. Simpson, "Invasive lobular carcinoma of the breast: morphology, biomarkers and'omics," *Breast cancer research,* vol. 17, p. 12, 2015.

25 H. Harris, L. Titus, D. Cramer, and K. Terry, "Long and irregular menstrual cycles, polycystic ovary syndrome, and ovarian cancer risk in a population-based case-control study," *International journal of cancer,* vol. 140, pp. 285-291, 2017.

26 R. Laforest, S. E. Lapi, R. Oyama, R. Bose, A. Tabchy, B. V. Marquez-Nostra, *et al.*, "[89 Zr] Trastuzumab: Evaluation of Radiation Dosimetry, Safety, and Optimal Imaging Parameters in Women with HER2-Positive Breast Cancer," *Molecular Imaging and Biology,* vol. 18, pp. 952-959, 2016.

27 C. S. Moskowitz, J. F. Chou, S. L. Wolden, J. L. Bernstein, J. Malhotra, D. Novetsky Friedman, *et al.*, "Breast cancer after chest radiation therapy for childhood cancer," *Journal of Clinical oncology,* vol. 32, pp. 2217-2223, 2014.

28 R. Troisi, E. E. Hatch, L. Titus, W. Strohsnitter, M. H. Gail, D. Huo, *et al.*, "Prenatal diethylstilbestrol exposure and cancer risk in women," *Environmental and molecular mutagenesis,* 2017.

29 E. F. Need, L. A. Selth, A. P. Trotta, D. A. Leach, L. Giorgio, M. A. O'Loughlin, *et al.*, "The unique transcriptional response produced by concurrent estrogen and progesterone treatment in breast cancer cells results in upregulation of growth factor pathways and switching from a Luminal A to a Basal-like subtype," *BMC cancer,* vol. 15, p. 791, 2015.

30 R. M. Tamimi, D. Spiegelman, S. A. Smith-Warner, M. Wang, M. Pazaris, W. C. Willett, *et al.*, "Population attributable risk of modifiable and nonmodifiable breast cancer risk factors in postmenopausal breast cancer," *American journal of epidemiology,* vol. 184, pp. 884-893, 2016.

31 K. A. Hirko, W. Y. Chen, W. C. Willett, B. A. Rosner, S. E. Hankinson, A. H. Beck, *et al.*, "Alcohol consumption and risk of breast cancer by molecular subtype: Prospective analysis of the nurses' health study after 26 years of follow-up," *International journal of cancer,* vol. 138, pp. 1094-1101, 2016.

32 K. Williams, P. A. Idowu, J. A. Balogun, and A. I. Oluwaranti, "Breast cancer risk prediction using data mining classification techniques," *Transactions on Networks and Communications,* vol. 3, p. 01, 2015.

33 S. Jauhari and S. Rizvi, "Mining gene expression data focusing cancer therapeutics: a digest," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* vol. 11, pp. 533-547, 2014.

34 Y.-B. Kang, P. D. Haghigh, and F. Burstein, "TaxoFinder: a graph-based approach for taxonomy learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, pp. 524-536, 2016.

35 A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in *Computer and Information Technology (ICCIT), 2016 19th International Conference on*, 2016, pp. 310-314.

36 T. H. A. Soliman, R. Mohamed, and A. A. Sewissy, "A hybrid Analytical Hierarchical Process and Deep Neural Networks approach for classifying breast cancer," in *Computer Engineering & Systems (ICCES), 2016 11th International Conference on*, 2016, pp. 212-219.

37 M. U. Salma, "PSO based fast K-means algorithm for feature selection from high dimensional medical data set," in *Intelligent Systems and Control (ISCO), 2016 10th International Conference on*, 2016, pp. 1-6.

38 A. Boutorh and A. Guessoum, "Classication of SNPs for breast cancer diagnosis using neural-network-based association rules," in *Programming and Systems (ISPS), 2015 12th International Symposium on*, 2015, pp. 1-9.

39 Y. Chen, L. Ling, and Q. Huang, "Classification of breast tumors in ultrasound using biclustering mining and neural network," in *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on,* 2016, pp. 1787-1791.

40 S. Jhajharia, H. K. Varshney, S. Verma, and R. Kumar, "A neural network based breast cancer prognosis model with PCA processed features," in *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, 2016, pp. 1896-1901.

41 D. Kaushik and K. Kaur, "Application of Data Mining for high accuracy prediction of breast tissue biopsy results," in *Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016 Third International Conference on,* 2016, pp. 40-45.

42 R. Nagpal and R. Shrivas, "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data," *Journal of Scientific and Technical Advancements,* vol. 1, pp. 19-23, 2015.

43 Y.-T. Cheng, Y.-F. Lin, K.-H. Chiang, and V. S. Tseng, "Mining Sequential Risk Patterns From Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease," *IEEE journal of biomedical and health informatics,* vol. 21, pp. 303-311, 2017.

44 P. P. Dhakate, K. Rajeswari, and D. Abin, "An ensemble approach for cancerious dataset analysis using feature selection," in *Communication Technologies (GCCT), 2015 Global Conference on*, 2015, pp. 479-482.

45 G. I. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis," in *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on,* 2012, pp. 180-185.

46 M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telematics and Informatics,* vol. 34, pp. 133-144, 2017.

47 A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Systems with Applications,* vol. 42, pp. 6844-6852, 2015.

48 E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," *Neural Computing and Applications,* vol. 28, pp. 753-763, 2017.

49 T. Christopher and B. Jamera, "A study on mining lung cancer data for increasing or decreasing disease prediction value by using ant colony optimization techniques," in *Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications*, 2015.

50 [50] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *Journal of Biomedical Science and Engineering,* vol. 6, p. 551, 2013.

51 D. Tomar and S. Agarwal, "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes," *Advances in Artificial Neural Systems,* vol. 2015, p. 1, 2015.

52 B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications,* vol. 41, pp. 1476-1482, 2014.