THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-10-November-2019

Development of a Model to Predict Breast Cancer Recurrence Using Decision Tree based Learning Algorithms

[1]Dawngliani M S, [2]Chandrasekaran N, [3]Samuel Lalmuanawma

[1] Department of Computer Science, GZRSC, Aizawl, Mizoram, India,

[2] Visiting Professor, Martin Luther Christian University, Former Director, CDAC, India,

[3]Department of Management, Mizoram University, India

[1]dawngliani@gmail.com, [2]professor.chandra@gmail.com, [3]samuellalmuanawma@mzu.edu.in

Abstract*— Cancer has been recognized as one of the leading causes of death all over the world. Until recently, while Physicians have been trained to successfully diagnose and treat cancer patients, only the most advanced research medical centers have been able to develop support system tools and platforms that can facilitatethe specialists to perform prognosis. To achieve this objective, these research institutions have been able to put to better use the medical data that has been stored and maintained in the hospital records by linking themwith machine learning algorithms. The main objective of this paper is to build a suitable model that will employa decision tree (which is a data mining algorithmused for classification) to predict a three-year recurrence of breast cancer including chances of survival from the dreaded disease. For training purposes and to augment the learning process of the algorithm, a vast amount of medical data of breast cancer patients obtained from the records of MizoramCancer Institute, Aizawl has been utilized.

Index Terms—*breast cancer, data mining, decision tree, J48, recurrence.*

I. INTRODUCTION

Most of the recent research work carried out in the field of Data Mining has contributed immensely in the development of some advanced applications with predictive capabilities. They have found effective in a number of professional fields, such as medical care and health sciences, manufacturing industry and also when dealing with numerous other commercial issues [1]. During the last few years alone, the medical field has seen dramatic increase in attempts being made to build predictive tools and platforms using data mining algorithms based on optimization of existing predictive models. These techniques have found widespread deployment in facilitating a number of medical tasks, such as breast cancer diagnosis and prognosis, diabetes prediction, detection of certain other types of cancer and diseases, etc. The medical database that is found in the hospitals, is mostly in the form of conventional analogue files, i.e. in paper form, while some others have found entry into the computer in unstructured format. It is pertinent to note that these contain massive amounts of very useful information, which relate to personal information of patients and various other analysis data reports sought from laboratories, etc. These databases are exhibiting tremendous growth year after year, while also becoming more sophisticated. With the application of appropriate data mining methods, useful patterns of information can be observed and detected residing in this data, which can effectively be utilized to further advanced research and to prepare evaluation reports [2]. Cancer of a patient is classified as 'recurred' if the disease has been observed to occur at any other subsequent time. Otherwise is is termed as 'non-recurred'.

It is a known fact that the north-eastern part of India has the highest incidence of cancer in the country, according to the latest report of the Indian Council of Medical Research (ICMR). In men, the age-adjusted incidence rate of all types of cancer is the highest in the Aizawl district of Mizoram followed by East Khasi Hills (Meghalaya). In women, the highest incidence is in the Aizawl district followed by Kamrup urban district (Assam) [3].

According to North Eastern Regional Cancer Registries Reports [4], the leading cancer sites of Mizoram among females were: uterine cervix (15.9%), lung (15.6%), breast (13.0%), stomach (11.3%) and oesophagus (4.2%). For the year 2012-2014, the respective cumulative ration (CR) and age adjusted ratio (AAR) per 100,000 population for the sites were: uterine cervix (19.4 and 23.1), lung (19.0 and 29.3), breast (15.8 and 19.9), stomach (13.7 and 20.2) and oesophagus (5.1 and 7.6).

II. RELATED STUDY

This section highlights and reviews some of the studies that have been performed by various researchers working in the field of data mining techniques with the predictions done employing cancer dataset.

Shukla N, et al, 2018 [5] have proposed a method to create multiple patient subgroups possessing different properties, in order to help improve the baseline accuracy of Multi Layer Perceptron (MLP) classifiers after missing data imputation. Also, MLP analysis showed that the selection of the cut-off survivability period significantly impacts MLP performances.

Mitra, et al, 2016 [6] in their study have used dataset recorded by the Cancer Registry Organization of Kerman Province, in Iran. The authors obtained better results while employing decision tree called Trees Random Forest (TRF) technique in comparison to other techniques

ISSN:0971-1260

Vol-22-Issue-10-November-2019

(NB, 1NN, AD, SVM and RBFN, MLP). The accuracy, sensitivity and the area under the Receiver Operating Characteristic (ROC) curve of TRF are 96%, 96%, 93%, respectively.

Khodary, et al, 2018 [7] in their experiment used a 60 samples dataset obtained from the National Cancer Institute (NCI) of Egypt. The authors have attempted to develop a breast cancer prediction and diagnosis system based on the Rough Set (RS) theory.

Mohebian, et.al, 2017 [8] have reported developing a Hybrid computer-aided-diagnosis system for Prediction of Breast Cancer Recurrence (HPBCR) using Optimized Ensemble Learning. Among 579 patients who participated in the study, 19.3% had a recurrence of cancer five years after diagnosis. The authors employed a hybrid of three algorithms, viz., Decision Tree, Support Vector Machine (SVM) and Multi Layer Perceptron (MLP) and this enabled them to conclude that best results were obtained when using this hybrid approach.

Bazazeh & Shubair, 2017 [9] have compared the effectiveness of the following three data mining classifiers, viz., Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). The Wisconsin original breast cancer data set was used as a training set to evaluate and compare the performance of the three ML classifiers in terms of key parameters such as accuracy, recall, precision and area under ROC curve. They were able to establish that SVM performance was the best, in terms of accuracy, specificity, and precision. However, RF had the highest probability of correctly classifying tumors.

Guo, et al., 2017 [10] by their research were able to reveal certain determinant factors for early breast cancer recurrence by a decision tree-based approach. The dataset contained information relating to 574 patients with non-metastatic invasive breast cancer, who received surgeries in the Leiden University Medical Centre with a database containing 55 attributes. They excluded 71 datasets which had some missing values. They too applied a Decision Tree algorithm based on data relating to whether the patients developed early recurrence and on similarities of their clinical and pathological diagnoses. The authors were able to establish that the classifier was able to successfully predict whether a patient developed early disease recurrence, with an estimated 70% accuracy.

Suryachandra & Reddy, 2016 [11] in their paper compared 4 machine learning algorithms for breast cancer prediction. According to their findings, Random Forest and Naïve Bayes algorithms demonstrated a better ability to make predictions more accurately than Decision Tree and other algorithms even when the available information was scanty.

III. ACRONYMS

Table I: Abbreviations used in the study

Sl. No	Attributes	Abbr	Sl. No	Attributes	Abbr
1	Age	AG	13	Pan Masala	PM
2	Sex	SX	14	Betel Nut	BN
3	Laterality	LT	15	Tuberculosis	TB
4	BMI	BMI	16	Hypertension	HP
5	Morphology	MP	17	Diabetes	DB
6	Socio economic Status	SS	18	Heart Disease	HD
7	Axillary Lymph Node	ALN	19	Asthma	AS
8	Skin Involvement	SI	20	Allergy	AL
9	Stage Grouping	SG	21	Hepatitis	HPT
10	Cigarette	CG	22	Others	OT
11	Tobacco	TBC	23	Aids	AI
12	Alcohol	ALC	24	Recurred	RC

IV. METHODOLOGY

A. *Concepts of Data Mining*

According to one definition, data mining is the extraction of implicit, previously unknown and potentially useful information from datasets. This process of discovering patterns in them involves methods that are at the intersection of machine learning, statisticsand database systems [12]. Britannica [13] defines data mining as an interdisciplinary subfieldof computer science and statistics. The overall objective of the technique is to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Eapen (2004) [14] defines the four different methods of learning approaches that can be used in data mining as follows:

- **Classification learning**: - In this method, the learning algorithms take a set of classified examples (training set) and use it for training algorithms. With the trained algorithms, classification of the test data takes place based on the patterns and rules extracted from the training set. Classification can also be termed as predicting a distinct class.
- **Numeric prediction**: - This is a variant of classification learning with the exception that instead of predicting the discrete class the outcome is a numeric value.
- **Association learning**: - The association and patterns between the various attributes are extracted and from these rules are created. The rules and patterns are used for predicting the categories or classification of the test data.
- **Clustering**:- This technique attempts to group similar instances into clusters. The challenges or drawbacks, while dealing with this type of machine learning technique is that we have to first identify clusters and assign new instances to these clusters.

B. *Tools Used*

An attempt is being made in the current study, to propose a new model and test its accuracy using WEKA. WEKA is a machine learning tool for data pre-processing, classification, regression, clustering, association rules, and visualization [15]. It is developed by the University of Waikato in New Zealand. While ease of use is one

ISSN:0971-1260

Vol-22-Issue-10-November-2019

attraction, it also provides numerous machine learning functionalities. The default input file format is an ARFF (Attribute-Relation File Format) file, which is an ASCII text file. We can easily transform our tables from Microsoft Excel into comma-delimited format (.csv) file and import it into WEKA. Once data is imported into WEKA, we can then convert this data to our desired default format, replace the missing value, remove the redundant value, etc [16]. And for the analysis, there are numerous algorithms which can effectively be used. In the current study, we have used J48 data mining classifiers to analyze the Breast Cancer dataset.

For this development work, we have chosen Eclipse Mars 1.0, which is an integrated development environment (IDE) for Java and other programming languages like C, C++, PHP, and Ruby, etc [17]. Eclipse provides full support for agile software development practices such as test-driven development and refactoring [18]. It provides easy access, while possessing an excellent user interface environment. We import weka.jar for the external reference so that all the machine learning applications can be implemented.

C. *Feature Selection*

Feature selection process finds extensive use and has become an area of active research interest pertaining to pattern recognition, statistics, and data mining in the medical domain. Feature selection invariably involves reducing the number of attributes to improve the accuracy of the outcome. The attributes are reduced by removing irrelevant and redundant attributes, which do not have much importance in determining the outcome [19]. The main goal is to find an optimal feature-subset (one that maximizes the prediction accuracy) [20]. The feature selection evaluator that we use is InfogainAttributeEval. It evaluates the worth of an attribute by measuring the information gain with respect to the class [21]. The evaluator checks each subset using the ranker method and ranks them according to their significance and relevance as shown in Table II:

Table II: Result of the rank of the attribute using Infogain evaluator

Rank	Info gain	Attribute
1	0.1454	SG
2	0.0448	ALN
3	0.0355	MP
4	0.0130	PM
5	0.0119	SI
6	0.0105	SS
7	0.0095	LT
8	0.0082	AL
9	0.0023	HD
10	0.0019	HP
11	0.0018	BN
12	0.0012	TB
13	0.0010	AS
14	0.0010	DB
15	0.0010	TBC
16	0.0009	OT
17	0.0005	CG
18	0.0003	ALC
19	0.0001	AI
20	0.0001	HPT
21	0.0000	SX
22	0.0000	BMI
23	0.0000	AG

D. *Data Pre-processing*

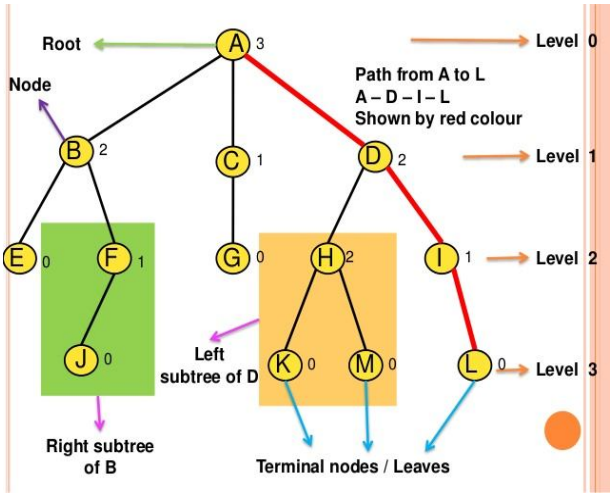
Completeness, accuracy, and consistency of the data are factors that define data quality. Data preprocessing is an important step in the data mining process to satisfy data quality requirements, as it ensures transforming raw data into an understandable format [22]. It then becomes necessary to codify the collected attributes having numeric value into groups ranging from 1 to 10. Since machine learning models are based on mathematical equations, we need to encode the categorical value into numbers [23]. This necessitates the allocation of numbers (nominal values) when the data is entered in text form, based on the order it follows.

E. *Decision Tree*

A decision tree is a decision support system tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [24]. It is one way to display an algorithm that only contains conditional control statements. They are inexpensive to construct, easy to interpret and integrate with database systems and more interestingly, they are capable of producing better performance in many data mining applications [25].

The decision tree starts with a root node at the top, shown as 'A' in Fig. 1 below. B to L are other nodes.

Fig. 1:



Decision tree nomenclature [26]

The core algorithm for building decision trees called ID3 was formulated by J. R. Quinlan which employs a top-down approach and a greedy search through the space of possible branches without any backtracking. This technique is commonly applied for gaining

ISSN:0971-1260

Vol-22-Issue-10-November-2019

information for the purpose of decision-making when (if…else) type of conditional rule becomes applicable and it usually starts with the root node [27]. The most common decision tree algorithms are J48, ID3, C4.5, and CART. In the current study, the J48 algorithm has been used to analyze the performance both with and without feature selection techniques. Figure 2 is a typical example of Decision Tree generated by WEKA.

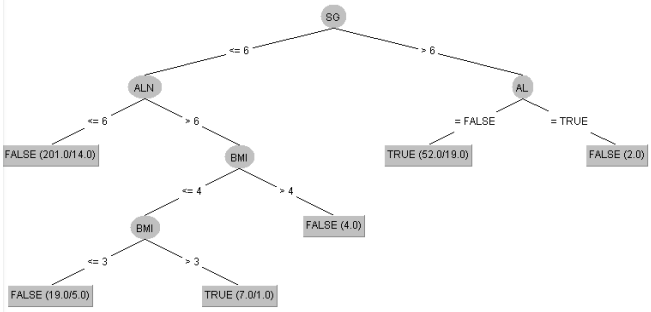


Fig 2: Example of Decision tree generated by WEKA

V. EXPERIMENTS AND RESULTS

This section describes the results obtained from the computational work carried out using the data mining algorithm namely J48 decision tree built inside WEKAand Eclipse Mars 1.0. We use a breast cancer dataset collected from Mizoram Cancer Institute, Aizawl, Mizoram, India. The generated J48 pruned tree is presented as follows:

J48 pruned tree

```
------------------
SG <= 6
|   ALN <= 6: FALSE (201.0/14.0)
|   ALN > 6
|   |   HP = FALSE
|   |   |   SG <= 2: FALSE (5.0)
|   |   |   SG > 2
|   |   |   |   BMI <= 4
|   |   |   |   |   BMI <= 3
|   |   |   |   |   |   LT = Right
|   |   |   |   |   |   |   DB = FALSE
|   |   |   |   |   |   |   |   AG <= 4: TRUE (3.0/1.0)
|   |   |   |   |   |   |   |   AG > 4: FALSE (2.0)
|   |   |   |   |   |   |   DB = TRUE: TRUE (2.0)
|   |   |   |   |   |   LT = Left: FALSE (6.0)
|   |   |   |   |   |   LT = Both: TRUE (1.0)
|   |   |   |   |   BMI > 3: TRUE (6.0)
|   |   |   |   BMI > 4: FALSE (3.0)
|   |   HP = TRUE: FALSE (2.0)
SG > 6
|   AL = FALSE
|   |   AI = FALSE
|   |   |   ALC = FALSE
|   |   |   |   AS = FALSE: TRUE (44.0/12.0)
|   |   |   |   AS = TRUE: FALSE (4.0/1.0)
|   |   |   ALC = TRUE: FALSE (2.0)
|   |   AI = TRUE: FALSE (2.0)
|   AL = TRUE: FALSE (2.0)
Number of Leaves  :     15
Size of the tree :  28
```

Our current analysis of the trained dataset using J48 gives 90.1754% accuracy. The dataset is given a class 'True' and 'False', which determine whether a patient has disease recurrence (True) or not (False) (no recurrence at all). The statistical result of J48 decision tree is depicted in Table III.

Table III: Results of J48 analysis

Class	TRUE	FALSE	Weighted Avg.
TP Rate	0.741	0.943	0.902
FP Rate	0.057	0.259	0.218
Precision	0.768	0.934	0.901
Recall	0.741	0.943	0.902
F-Measure	0.754	0.939	0.901
ROC Area	0.866	0.866	0.866
PRC Area	0.675	0.938	0.885

Other Statistical Results:

Kappa statistics : 0.693
Mean absolute error : 0.1626
Root mean squared error : 0.2851
Relative absolute error : 49.9639 %
Root relative squared error : 70.8177 %

The prediction system is implemented using a Windows 10 OS. Both WEKA and Eclipse require to operate under Java Runtime Environment. The main page consists of five menus such as analysis using data mining algorithm, prediction application, classifier comparison table, ROC value and mean error table and ROC curves. Figure 3 captures the main menu.

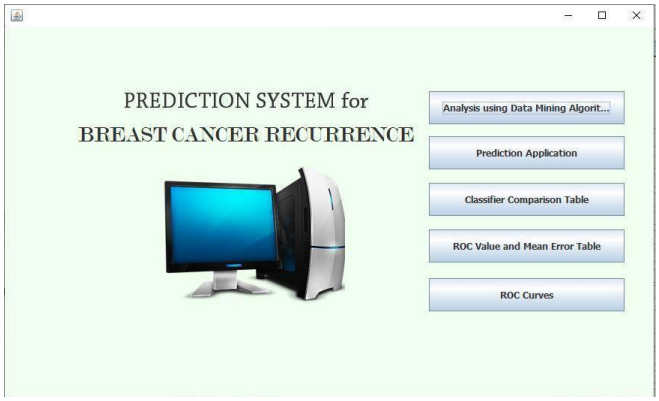


Fig. 3: Main menu of breast cancer prediction system

The second menu, which is a prediction application is the main application for the prediction of breast cancer recurrence implemented using the J48 decision tree. Figures 4 and 5 show the results obtained from the application for prediction of negative and positive recurrences once the analysis of the patient data has been completed. The patient's identity is not disclosed in any part of our research.

ISSN:0971-1260

Vol-22-Issue-10-November-2019

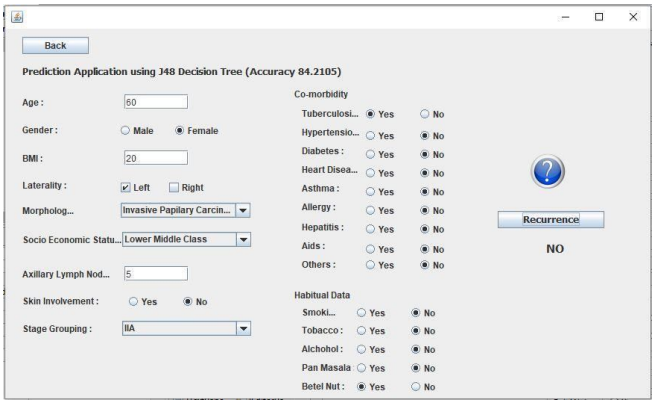


Fig. 4: Prediction system results for negative recurrence

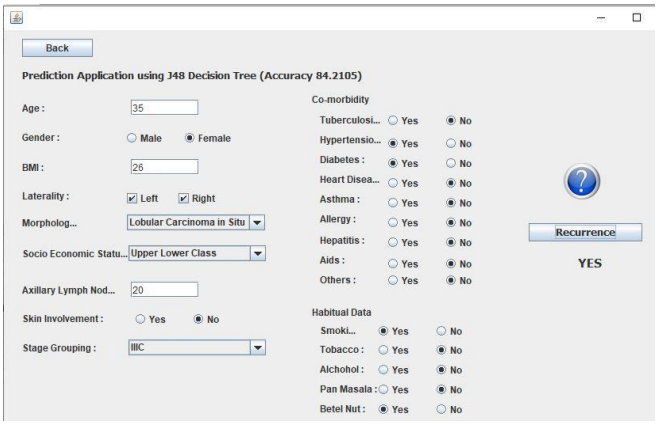


Fig. 5: Prediction system results for positive recurrence

VI. EVALUATION AND TESTING

In order to evaluate our results, we are using a 10 fold cross-validation and ROC curve. Cross-validation is a technique, which is normally employed to evaluate predictive models by partitioning the dataset into a training set to train the model, and a test set to evaluate it [28]. After 10 fold cross-validation, the accuracy of J48 is 84.2105%. The receiver operating characteristic (ROC) curve is used to check or visualize the performance of classifiers at various threshold settings. It shows how much a classifier is capable of distinguishing between classes [29]. Higher the AUC, the better the model is at predicting true as true and false as false. We have obtained an ROC value in our current study of 0.866, which is considered to be good in determining the class [30].

The ROC curve is plotted with false positive rate along X axis against the true positive rate along the Y axis, see Figure 6. The sensitivity, specificity and accuracy are calculated as follows:

Sensitivity (true positive rate) = TP/ (TP+FN)
= 0.7679

Specificity (true negative rate) = TN/ (FP+TN)
= 0.9345

Where TP is true positive, TN is true negative, FP is false positive and FN is a false negative.

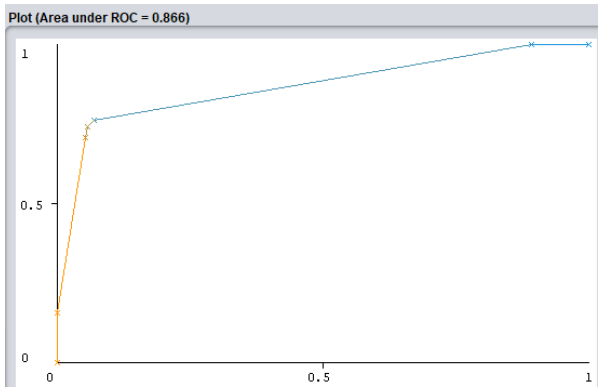


Fig. 6: ROC curve of the J48 decision tree

CONCLUSION

Data mining plays a vital role in the prediction of diagnosis and prognosis in the domain of health care. Our proposed prediction model for breast cancer recurrence has an accuracy of 84.2105% which is validated using follow up data of the patients. The proposed prediction system can be used as a powerful tool by medical practitioners to identify their patients, who are likely to have a high chance of recurrence of breast cancer.

We would like to thank Dr. Jerry Lalrinsanga, Medical Oncologist of Mizoram Cancer Institute (MCI), Aizawl, for supporting this research work by giving us permission to collect breast cancer datasets from MCI. The authors express their sincere gratitude to MLCU for facilitating and extending all possible help to complete this particular research work.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second edi. Morgan Kaufmann Publications, 2006.

[2] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Third Edit. Morgan Kaufmann, San Francisco, 2011.

[3] DowntoEarth, "cancer-rises-in-the-northeast." 2019.

[4] NCRP, "North Eastern Regional Cancer Registries Reports," 2006.

[5] N. Shukla, M. Hagenbuchner, K. T. Win, and J. Yang, "Breast cancer data analysis for survivability studies and prediction," *Comput. Methods Programs Biomed.*, vol. 155, pp. 199–208, 2018.

[6] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," *Technol. Heal. Care*, vol. 24, no. 1, pp. 31–42, 2016.

[7] S. Khodary, M. Hamouda, M. E. Wahed, R. H. Abo, and K. Riad, "Computer Methods and Programs in Biomedicine Robust breast cancer prediction system based on rough set theory at National Cancer Institute of Egypt," *Comput. Methods Programs Biomed.*, vol. 153, pp. 259–268, 2018.

ISSN:0971-1260

Vol-22-Issue-10-November-2019

[8] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. Angel, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 75–85, 2017.

[9] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," *Int. Conf. Electron. Devices, Syst. Appl.*, pp. 2–5, 2017.

[10] J. Guo *et al.*, "Revealing determinant factors for early breast cancer recurrence by decision tree," *Inf. Syst. Front.*, vol. 19, no. 6, pp. 1233–1241, 2017.

[11] P. Suryachandra and P. V. S. Reddy, "Comparison of machine learning algorithms for breast cancer," *Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016*, vol. 2016, 2016.

[12] KDD.org, "Curriculum Design : Introduction," *KDD.org*. 2018.

[13] C. Clifton, "Data Mining _ Britannica." 2010.

[14] A. G. Eapen, "Application of Data mining in Medical Applications," University of Waterloo, 2004.

[15] M. Hall, E. Frank, G. Holmes, I. H. Witten, and S. J. Cunningham, "Weka: Practical Machine Learning Tools and techniques," in *Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 2007.

[16] S. S. Aksenova, "Machine Learning with WEKA," *Mach. Learn.*, vol. 11, no. 1, pp. 1–37, 2006.

[17] T. Point, "Eclipse Tutorial." 2019.

[18] B. M. Dexter, "Eclipse And Java For Total Beginners Companion Tutorial Document," pp. 1–45, 2007.

[19] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," *Proc. 2014 Sci. Inf. Conf. SAI 2014*, pp. 372–378, 2014.

[20] Mehul Ved, "Feature Selection and Feature Extraction in Machine Learning: An Overview," *Medium*. 2018.

[21] J. Brownlee, "How to Perform Feature Selection With Machine Learning Data in Weka," *Machine learning mastery*. 2013.

[22] Mohit Sharma, "What Steps should one take while doing Data Preprocessing?," *Hackernoon*. 2018.

[23] M. Rajaratne, "Machine Learning - Part 1 - Data Preprocessing." 2018.

[24] R. Brid, "Decision Trees — A simple way to visualize a decision," *October, 26th*. 2018.

[25] M. S. Dawngliani, N. Chandrasekaran, and S. Lalmuanawma, "A Comparative Study between Data Mining Classification and Ensemble Techniques for Predicting Survivability of Breast Cancer Patients," vol. 8, no. 9, 2019.

[26] N. Bichwani, "https://www.slideshare.net/NiteshBichwani/data-structures-73261516" .

[27] A. K. Pujari, *Data Mining Techniques*, Third edit. Universities Press Private Limited, 2013.

[28] F. Wikipedia, "Cross-validation (statistics) - Wikipedia." 2019.

[29] S. Narkhede, "Understanding AUC - ROC Curve – Towards Data Science." pp. 1–9, 2018.

[30] STARD, "The Area Under an ROC Curve," *Area*. pp. 2–3, 2012.