2019 IEEE Conference on Information and Communication Technology (CICT)

Breast Cancer Recurrence Prediction using Machine Learning

Kaustubh Chakradeo
Information Technology Department
Savitribai Phule Pune University
Pune, India
chakradeokaustubh@gmail.com

Sanyog Vyawahare
Information Technology Department
Savitribai Phule Pune University
Pune, India
sanvyawahare@gmail.com

Dr. Pranav Pawar
Faculty of Engineering
Bar-Ilan University,
Ramat Gan, Israel
pranav21684@gmail.com

***Abstract*—The most common cancer among women is breast cancer. Around 12% of women are affected by it all over the world. Recurrent breast cancer is a term used for breast cancer which returns even after a successful treatment. This research aims to use Machine learning to detect and predict the recurrence of breast cancer; and compare all the models by using different metrics like accuracy, precision, etc. The models built can help predict the recurrence of breast cancer effectively. All the models are built using the Wisconsin Prognostic Breast Cancer Dataset(WPBC). The models built are Multiple Linear Regression, Support Vector Machine, which was build by using RBF Kernel and Leave-One-Out(K-fold Cross-Validation) and Decision Tree using metrics like Gini Index, Entropy and Information Gain. Support Vector Machine and K-fold Cross-Validation gave the best results for recurrence and non-recurrence predictions.**

***Index Terms*—Breast Cancer, Recurrence, Prognosis, SVM, Regression, Decision Tree**

I. Introduction

Breast cancer is one of the most widespread diseases in the world[22]. Recurrence of breast cancer can happen in any duration from 1 to 20 years after the original cancer is treated. Detection of breast cancer can be done using MRI, mammograms, biopsy, ultrasound, and physical checks.

Breast cancer can be diagnosed into two categories according to the tumour, benign and malignant. It is then classified into recurrent and non-recurrent. Recurrence of cancer normally depends on the number of lymph nodes which were affected by original cancer.

Breast cancer can reoccur to the same site of spread to other parts of the body(metastasize).[22] Recurrence of breast cancer can be predicted by checking various factors like the size of the original tumour, number of lymph nodes which were affected, the area of the tumour, and other similar factors. By using Machine Learning models for the prediction of breast cancer recurrence, the major advantage to be gained is an increase in accuracy and decrease in errors. By establishing a timescale for the last occurrence, and knowing the size, shape, texture and other features of the previous tumour, whether a recurrence will happen can be accurately determined. Use of machine learning saves a lot of time compared to the time taken for traditional methods like a biopsy since a model can predict the outcome within seconds.

To build the prediction models, we have selected 8-10 such factors depending on their effect on the breast cancer recurrence. The models are built individually and predict the probability of recurrence of breast cancer. The types of models to be built were selected after careful comparison between various machine learning models and their effectiveness. Using machine learning models to predict breast cancer recurrence can be effective and accurate. This model could be provided to doctors to help in lead to quicker and better diagnosis of such a serious and life-altering disease.

Organization of the Paper

The section "Literature survey and background work" describes the background work done for the research. It involves a study of existing systems for prediction of breast cancer recurrence. They were compared and thus formed the basis of motivation for the research. The next section, "Methodology" describes the dataset and the techniques used on this dataset for the machine learning models. It also gave insights to building the classifiers. Further, "Implementation" expounds on the actual models built and their results with the help of plots. The next section, "Conclusions" gives an overview of the models built along with the interpretation of the results, followed by the "References" section.

II. Literature Survey and Background Work

The authors conducted thorough reading about the work done in the chosen topic by using different datasets.

Mandeep Rana et al[1] use 4 classification techniques, Naive Bayes Algorithm, SVM, Logistic Regression and KNN to classify between benign and malignant breast cancer and in the latter case, predict recurrence of breast cancer. They used Wisconsin Prognostic and Diagnostic Datasets for their models. Selection of parameters in the data set is important. Based on this selection, it was found that KNN had the best accuracy in training and testing dataset. However, when the value of parameters was large, SVM using Gaussian Kernel gave the best results. With an increase in the number of iterations, its performance increased, but the time required to train the model increased. Also, with SVM, a maximum of two classes could be selected. Multiclass SVM, which solves

978-1-7281-5398-8/19/$31.00 ©2019 IEEE

this limitation, was not used in this paper. This limitation could seriously affect training time and accuracy, and needs to be overcome.[1]

Uma Ojha and Dr. Savita Goel[2] evaluate the performance of clustering and classification algorithms by using some important attributes which perform a major role in predicting the possibility of breast cancer recurrence in advance using C5.0 algorithm. They used 4 clustering algorithms including KMeans, EM, PAM, and Fuzzy CMeans. KNN, Naive Bayes, SVM and C 5.0 were the classification algorithms used. The metrics for performance used for measurement were accuracy, specificity, and sensitivity; these included results having True Positive, False Positive, True Negative and False Negative values. These models were built on using Wisconsin Diagnostic Breast Cancer dataset. On careful analysis, it was found that classification algorithms give better accuracy than clustering algorithms. SVM was found to have the best accuracy out of all the methods studied. It was found that Perimeter is the most critical factor for predicting the recurrence of the disease. However, a larger database is needed to verify the identified critical parameters to predict the recurrence of the disease. The results obtained may change when working on a larger dataset.[2]

Ahmed Pritom et al[3] look at the data mining techniques to find out recurrence of breast cancer. Their paper used Wisconsin Diagnostic Dataset for the model. By selecting an efficient feature selection algorithm the accuracy of each model was improved significantly. The methods used were Naive Bayes, C 4.5 Decision Tree, and SVM classification algorithm. This paper mainly deals with finding the best attributes from the dataset, and selection of those which contribute most significantly to improve the efficiency of the model. Ranker algorithm was used for the same. The authors found that Time, Lymph Node Status and Perimeter are the best contributors to the accuracy of the model; while Concave points and Concavity std error do not contribute to the model at all. In the results, it was found that SVM gives the best accuracy when proper parameters are selected. However, the paper only gives the validation for recurrence, it does not give a time frame for the same. Also, the accuracy obtained by the best model was still 75.75 percent, which is on the lower side. This could be improved by changing the feature selection methods or using a hybrid model for prediction.[3]

Faezeh Roshani et al proposed a new data mining method called CDF tree by combining the Decision tree (DT) with Frequent-Pattern Growth tree (FP-growth) which extracted rules for predicting the class of unknown cases. They used online Wisconsin Prognostic Breast Cancer (WPBC) dataset available on the UC Irvine Machine Learning Repository[21] for building the model. After pre-processing, 80% of data was used for training in which there was an equal number of both cases to study (recur and non-recur). This data training was done using CDF tree and the prediction was made using the Mamdani interference. This was done by centroid defuzzification method and various S/T-norms and with an acceptable threshold value of 1.5. They first converted the quantitative dataset into a transactional dataset, then split the dataset into different classes with several cases in the path between They further split the sub-data by sorting with the lowest entropy attribute repeating this step until all attributes are used. At last pruning and rule extraction was performed. Their model gave the accuracy of around 93%. Results can be further improved with good splitting criteria and dataset with a high number of recurrence cases to handle class imbalance issue which is the main concern in building the recurrence model. The accuracy and sensitivity of the model can be improved significantly by considering more features to train the model. [4]

Daad Abdullah Almuhaidib et al[5] built three models namely NaïveBayes, C 4.5 decision tree and random forest to predict recurrence of breast cancer recurrence, for this, they used the dataset online using UC Irvine Machine Learning Repository called Wisconsin Prognostic Breast Cancer (WPBC). After eliminating the rows with missing data values they approach the class imbalance issue with an under-sampling problem. They also used normalization and examined model before and after normalizing. Other pre-processing steps they used on dataset were data extraction using principal component analysis, feature selection using recursive feature elimination and feature ranking using the random forest. The best 10 features in their research include lymph node status, mean fractal dimension, mean symmetry, tumor size, worst concave points, and standard error in radius. After pre-processing K-fold cross-validation was used to split the dataset upon which the 3 models as mentioned were built. Later, using multiple approaches like normalization, under-sampling, and feature selection and building model before and after applying these methods. The best models were selected and then the process of ensemble learning was applied. Ensemble learning is a technique in which multiple models are built based on different or similar conditions. The best models out of those are combined to build an improved model with high accuracy. Then the simple average mean was calculated. They tested efficiency by generating a confusion matrix. The measures of performance were accuracy, specificity and sensitivity area under ROC curve. [5]

From the background survey, it was clear that more features needed to be used for getting better accuracy.Some papers achieved low accuracy, which could be increased by changing the features or selection methods or by using hybrid models. The features selected also need to be normalized in some cases. By using a different method for feature selection, accuracy can be increased.

III. METHODOLOGY

A. Data Set Description

For the study, the authors utilized "Breast Cancer Wisconsin (Prognostic) Data Set" from UC Irvine machine learning repository.[21] The dataset is built from a real-life study of patients by Dr. Wolberg since 1984. The dataset has 34 attributes and 198 instances. The Outcome attribute specifies if cancer has recurred (R) or not (N). "The dataset has the following features

- ID No.
- Time (if R then recurrence exists, if N then disease free)
- Outcome (R = recurrence, N = nonrecurrence)
- Tumor size
- Lymph node status

while the remaining 30 features are real-valued features. They are computed from cell nucleus. The features under this computation are as follows:

- Radius of Tumor
- Area of Tumor
- Perimeter of Tumor
- Fractal Dimension ("coastline approximation" - 1)
- Smoothness
- Texture (standard deviation of gray-scale values)
- Compactness (perimeter$\hat{2}$ / area - 1.0)
- Symmetry
- Concave Points (number of concave portions of the contour)
- Concavity (severity of concave portions of the contour)"[21]

Each of the 10 above features has three forms Mean, Standard Error and Worst (mean of three largest values).

B. Data Pre-processing

Data preprocessing is the first step in building a machine learning model. The data may have values which need to be converted, predicted or removed in order to feed the model with data to maximize the prediction.

NA Values: NA values are those whose values are missing. The dataset used in this paper had some instances with NA values. To tackle NA values there are 2 options either remove the instance completely or fill the NA value with either the mean, mode or EM. After comparing the result by predicting NA values and removing the instances the later gave better result so all the instances with NA values were removed (Count 4).[6]

Feature Selection: This process creates a subset of the features which contribute the most towards prediction variable. In this paper, we extracted features using a correlation matrix with heatmap. The matrix helps understand how the features are related to each other as well as the target variable. To create a heatmap a Seaborn library was used.[7]

Label Encoding: Label Encoding is done because to train a model we need numerical data. The labels need to be converted into numerical values for machine to understand. The 'Outcome' variable in the dataset has N and R values which need to be converted into 0 and 1 respectively.

Conversion to Categorical Data: For decision tree categorical data was needed. In order to train DT model dataset was converted to categorical data.

One Hot Encoding: This process converts the categorical data into a form which can be later fed to the model for training and prediction purpose. One hot encoding solves the problem created by label encoding where the model tries to correlate the feature values with each other.

C. Building Classifiers (Processing)

SVM: Support vector machines (SVM) are a set of supervised learning methods used for classification, regression and outliers detection[23]. It is a discriminative model defined by a hyperplane. The hyperplane classifies the new instances into either of the n-classes. For example, a hyperplane is a line in 2-Dimensions. To define a good hyperplane we try to maximize the margin, which is calculated by finding the distance between the two closest points from the hyperplane of either class. The hyperplane with the highest margin has fewer chances of miss classifying the instances. Hyperplane can be defined with a set of points $\vec{x}$ satisfying the equation below

$$\vec{w} \cdot \vec{x} - b = 0 \tag{1}$$

where the normal vector of the hyper plane is given by $\vec{w}$[24].The offset of hyperplane is calculated as $\frac{b}{||\vec{w}||}$.

RBF kernel was used in constructing the SVM model in this paper.The RBF kernel is given by equation

$$K(x, x') = \frac{exp(||x - x'||^2}{2\sigma^2)} \tag{2}$$

$||x - x'||$ is the squared euclidean function distance between the two vector applied on 2 sample x and x' represented as feature vector in input space[24]. σ represents a free parameter and which also have a equivalent definition as $\gamma = \frac{1}{2\sigma^2}$.

$$K(x, x') = exp(-\gamma||x - x'||^2) \tag{3}$$

In equation above $\gamma > 0$ and parametrized as above[24].

Decision Tree: Decision trees are supervised learning algorithms capable of handling non-linear relationships. They are tree-like structures, each branch exploring a possible solution space. A decision tree calculates the probability of a categorical target variable belonging to each class. A decision tree categorizes a instance by labeling it to the class which resembles it most closely . The population is split into two or more homogenous sets. The root of the decision tree is the desired outcome of the target variable.

There are three terms used to build a decision tree algorithm. 1) Gini Index: A Gini index tells us about the distribution of two classes which are created by the split by showing their heterogeneity. It measures how often a randomly chosen

element would be incorrectly identified. Gini Index can be obtained by using the formula

$$GiniIndex = p^2 + q^2 \tag{4}$$

Where, p is the probability of success of a node, and q is the probability of failure of a node.

2) Entropy: Entropy is the amount of heterogeneity or uncertainty in a dataset. In other words, it is a degree of disorganization. Entropy is given by the formula

$$Entropy = -plog_2p - qlog_2q \tag{5}$$

Where, p is the probability of success of a node, and q is the probability of failure of a node.

3) Information Gain- Information gain calculates the drop or rise in entropy value once the split occurs. When a node partitions the training instances into smaller subsets, the entropy of the set changes. This change in entropy, or reduction in the randomness in the dataset, is called information gain. It can be calculated by using the equation below

$$InformationGain(T, X) = Entropy(T) - Entropy(T, X) \tag{6}$$

Where T is the target variable, Entropy(T) is the entropy of target variable before splitting of the attributes, X is chosen the attribute. E(T, X) is the entropy of the target variable after splitting into different attributes.

Linear Regression: Linear regression is a statistical approach to modeling the relationships between a scalar or dependent variable and one or more explanatory variables. A regression model looks for statistical relationships against a deterministic relationship. It looks for a line that best fits the data points. The best fit line is the line for which the total prediction error is the least. Regression model is represented by the linear equation between x and y, where x is a set of input feature and y represents the target variable. For a simple regression, the model will be represented by using the formula

$$Y = \beta 0 + \beta 1 * X \tag{7}$$

Where β is the slope of the regression line

When there are multiple inputs, the Ordinary Least Squares method is used to estimate the value of coefficients. It minimize the squared sum of distance between the regression line and data points.

D. Results

To check how good the model is at predicting the instances into the correct class it's important to calculate the accuracy of the model. To calculate the same the below methods were used.

Confusion Matrix: It is a metric which is used to measure how well a classification model performed. It is also called as an error matrix because it can help figure out where the model is getting the prediction wrong. After the prediction, the confusion matrix calculates the number of a correct and incorrect prediction. The table is arranged as

- Predicted Class Column-wise
- Expected Class Row-wise

In a binary classification we can assign once class as positive and one as negative row-wise and True or False column-wise this gives us

- **true positive**: Correct recurrence predicted
- **true negative**: Correct non-recurrence predicted
- **false positive**: Incorrect recurrence predicted
- **false negative**: Incorrect non-recurrence predicted

.

Classification Report: A classification report is used to measure the quality of the classification model. It uses the below metrics:

- **Precision**: "Precision is defined as the number of true positives over the number of true positives plus the number of false positives.

  $$Precision = \frac{TP}{TP + FP} \tag{8}$$

  where TP is true positive and FP is false positive
- **Recall**: Recall is defined as the number of true positives over the number of true positives plus the number of false negatives.

  $$Recall = \frac{TP}{TP + FN} \tag{9}$$

  where TP is true positive and FN is false negative
- **F1-Score**:It is defined as the harmonic mean of precision and recall"[23].

  $$F1 - Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{10}$$

  where TP is true positive and FP is false positive

Accuracy Score: It computes the accuracy of the model. The default value it returns is a fraction of correctly made predictions of instances to total instances while if normalize is set to false then it returns the count of correctly predicted instance.

- MAE: Mean Absolute Error, given by

  $$\frac{1}{n}\sum_{1}^{n}|yi - \hat{y}_i| \tag{11}$$

  here the $|yi - \hat{y}_i|$ is the absolute difference between predicted and actual observation.
- MSE: Mean Squared Error, given by

  $$\frac{1}{n}\sum_{i}^{n}(yi - \hat{y}_i)^2 \tag{12}$$

- RMSE: Root Mean Squared Error. Taking the square root of MSE gives the value of RMSE.

IV. Implementation

A. Multiple Linear Regression

For the creation of the Regression Model, the features from the dataset needed to be reduced. This was done to get the maximum accuracy of the model. The dimentionality reduction was possible with the help of Heatmap of correlation of features with the target feature. [8]This was done by using Seaborn Library. Using this, it was found out that the correlation between MeanRadius, Time, MeanPerimeter, TumorSize, ConcavitySE, MeanArea, LymphNodeStatus was the highest with the target variable that is Outcome. Thus, these eight features were selected for the regression model. Label Encoding was done on the Outcome feature. The dataset was split into testing and training data, taking various random splits on each different run. The regression model was then run on this data frame and metrics like Mean Absolute Error, Root Mean Squared Error and Mean Squared Error were used to predict the results. A scatter plot was produced on 3 different runs. Along with this, a distance plot for prediction was also plotted.

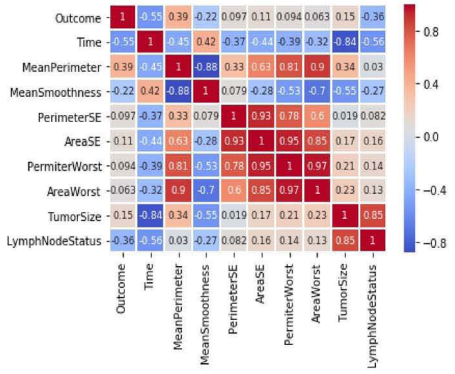

Figure 1. **Heat Map for Correlation**

B. Support Vector Machine with RBF Kernel

For building the SVM model for breast cancer recurrence prediction, it was important to find the features best suited to get an accurate result. Using the subset selection algorithm the below features were found - MeanRadius, MeanPerimeter, MeanArea, SmoothnessSE, ConcavitySE, ConcavePointsSE, TumorSize, LymphNodeStatus. These eight features were used to predict the class containing Outcome(i.e Recurrence or Non-Recurrence). To begin predicting the class using the dataset, it needed to be processed for further computation. [9]First, the instances with NA values were dropped, later on after Feature Scaling and Label Encoding the dataset was split using the Leave One Out method. The leave one out method splits the dataset in

$$\frac{n-1}{1} \quad \frac{TrainingSet}{TestingSet} \tag{13}$$

where the number of instances is given by n. The training set was passed to the model which used the RBF kernel the corresponding hyperplane that was found and the results were analysed using 3 metrics Precision, Recall and Accuracy Score. These metrics were represented using the Confusion Matrix and Classification Report.

C. Decision Tree Model

The dataset used for this model was the pruned data used in the regression model. The features used for the model were Outcome, Time, MeanRadius, MeanPerimeter, MeanArea, SmoothnessSE, ConcavitySE, ConcavePointsSE, TumorSize and LymphNodeStatus. For the decision tree, it was essential that the data was converted into a categorical dataset. This is the essential step for building any decision tree. Thus, conversion of the dataset was the first step which was performed. Label encoding and Categorical encoding was done to do this.[10] The conversion of the dataset was done by grouping the column data into different categories like NoRisk, NotoLowRisk, LowRisk, MedRisk and HighRisk. This was done by judging the values of individual columns and the way they affect the target variable. Then the dataset was subject to splitting into training and testing data by using random split. Now the model was trained using two metrics- Gini Index and entropy.[11] These metrics were calculated using the formulas mentioned above (Equation 4 and 5). Using these values, when the model was trained, the accuracy of the model was calculated. This was represented using Confusion Matrix, Report and Accuracy. The report contained further metrics like Precision, Recall, F1 Score and Support. Results were calculated using these scores for both Entropy and Gini Index. These results are mentioned in the following section.

D. Results

When the results were obtained, it was found out that SVM-RBF gave the best result with an accuracy of 97.93%. In contrast, the decision tree algorithm performed poorly with an accuracy of 76%. The multiple linear regression model gave a Mean Squared Error of 0.162. The full results are mentioned in the figure. The results display clearly that SVM with an RBF kernel gives the maximum accuracy, in fact, an extraordinary accuracy score. It is clear from the table that the precision(eq no 8) for SVM-RBF, that is, the ability to correctly label recurrence samples. Also, the recall for recurrence(eq no 9) that is, the ability of the model to find out the recurrence samples was extremely high Any hybrid models should be made by keeping SVM with RBF kernel as the base model. More work needs to be done to use the decision tree model for breast cancer recurrence prediction. The results indicate that the best factors for prediction of breast cancer are 'previous tumour size', 'area of the previous tumour' and' the number of affected lymph nodes'. The results are clearly defined in the following tables.

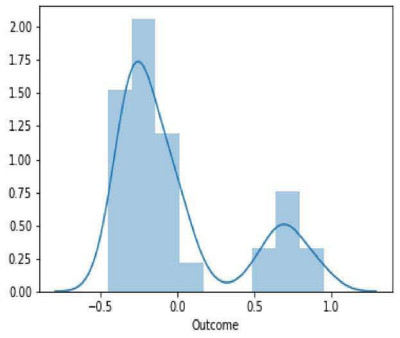

Figure 2. **Distance Plot for Multiple Linear Regression Model**

Table I
COMPARISON OF RESULTS FOR SVM AND DECISION TREE MODELS

Metric	SVM-RBF	Decision Tree
Accuracy	97.93	76.72
Precision	93.36	0.78
Recall	0.91	0.90

V. ABBREVIATIONS AND ACRONYMS

CDF Cumulative Distribution Function
EM Expectation Maximization
KNN K Nearest Neighbour
MRI Magnetic Resonance Imaging
NA Not Available
PAM Prediction Analysis For Microarrays
RBF Radial Basis Function
ROC Receiver Operating Characteristics
SVM Support Vector Machine

VI. CONCLUSIONS

Recurrence in breast cancer can happen anytime in patients. To accurately diagnose the breast cancer recurrence , there should be a way to predict this beforehand. This will help physicians give diagnosis accurately. Having self-learning models will give faster and more accurate results. In future, the machine learning models used in this research to predict the recurrence of breast cancer could be worked upon to increase the accuracy of the predictive model by combining them with ensemble learning techniques. The results could be modelled by constructing a software system. However, for this, more data would be required.

Table II
RESULTS FOR MULTIPLE LINEAR REGRESSION MODEL

Mean Absolute Error(MAE)	0.3217131827
Mean Squared Error(MSE)	0.1619535398
Root Mean Squared Error(RMSE)	0.4024345162

REFERENCES

[1] Rana, Mandeep, Pooja Chandorkar, Alishiba Dsouza and Nikahat Kazi. "BREAST CANCER DIAGNOSIS AND RECURRENCE PREDICTION USING MACHINE LEARNING TECHNIQUES." (2015).

[2] Ojha, Uma and Savita Goel. "A study on prediction of breast cancer recurrence using data mining techniques." 2017 7th International Conference on Cloud Computing, Data Science Engineering - Confluence (2017): 527-530.

[3] Mathappan, Nivaashini Rs, Soundariya. "Deep Boltzmann Machine based Breast Cancer Risk Detection for Healthcare Systems". International Journal of Pure and Applied Mathematics (2018). 119.

[4] Roshani, Faezeh, I. Burhan Türksen, Mohammad Hossein Fazel Zarandi and Maede Maftooni. "Fuzzy expert system for prognosis of breast cancer recurrence." 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC) (2015): 1-5.

[5] Almuhaidib, Daad Abdullah, Hadil Ahmed Shaiba, Najla Ghazi Alharbi, Sara Muhammad Alotaibi, Fatima Moteb Albusayyis, Mashael Abdulalim Alzaid and Reem Mohammed Almadhi. "Ensemble Learning Method for the Prediction of Breast Cancer Recurrence." 2018 1st International Conference on Computer Applications Information Security (ICCAIS) (2018): 1-6.

[6] Xu, Tanjin and Stephen A. Ramsey. "Exploration of regression models for cancer noncoding mutation recurrence." BCB (2016).

[7] Richter, Aaron N. and Taghi M. Khoshgoftaar. "Predicting Cancer Relapse with Clinical Data: A Survey of Current Techniques." 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI) (2016): 369-376.

[8] Dr Prof. Neeraj, Sakshi Sharma, Renuka Purohit, Pramod Singh Rathore, Prediction of Recurrence Cancer using J48 Algorithm, Proceedings of the 2nd International Conference on Communication and Electronics Systems (ICCES 2017).

[9] Sayed, Shabina, Shoeb Ahmed and Rakesh Kumar Poonia. "Holo entropy enabled decision tree classifier for breast cancer diagnosis using wisconsin (prognostic) data set." 2017 7th International Conference on Communication Systems and Network Technologies (CSNT) (2017): 172-176.

[10] Abreu, Pedro, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade and Daniel Castro Silva. "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review." ACM Comput. Surv. 49 (2016): 52:1-52:40..

[11] Jacob, Shomona Gracia and R. Geetha Ramani. "Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques." (2012).

[12] Jhajharia, Smita, Seema Verma and Rajesh Kumar. "Predictive Analytics for Breast Cancer Survivability: A Comparison of Five Predictive Models." ICTCS '16 (2016).

[13] Umesh, D. R. and Bharathkumar Ramachandra. "Association rule mining based predicting breast cancer recurrence on SEER breast cancer data." 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (2015): 376-380.

[14] Mahrooghy, Majid, Ahmed Bilal Ashraf, Dania Daye, Elizabeth S. McDonald, Mark Rosen, Carolyn Mies, Michael D. Feldman and Despina Kontos. "Pharmacokinetic Tumor Heterogeneity as a Prognostic Biomarker for Classifying Breast Cancer Recurrence Risk." IEEE Transactions on Biomedical Engineering 62 (2015): 1585-1594.

[15] Pritom, Ahmed Munshi, Md. Ahadur Sabab, Shahed Shihab, Shihabuzzaman. Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique. (2016).

[16] Eltalhi, Saria Kutrani, Huda. "Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review." (2019).

[17] Ferroni, Patrizia Zanzotto, Fabio Massimo Riondino, Silvia Scarpato, Noemi Guadagni, Fiorella Roselli, Mario. Breast Cancer Prognosis Using a Machine Learning Approach. (2019).

[18] Foo, Jasmine and Kevin Leder. "Rare events in cancer recurrence timing." Proceedings Title: Proceedings of the 2012 Winter Simulation Conference (WSC) (2012): 1-10.

[19] Sehhati, Mohammadreza, Alireza Mehri Dehnavi, Hossein Rabbani and Meraj Pourhossein. "Stable Gene Signature Selection for Predic-

tion of Breast Cancer Recurrence Using Joint Mutual Information." IEEE/ACM Transactions on Computational Biology and Bioinformatics 12 (2015): 1440-1448.
[20] Khan, Rafaqat Alam, Nasir Ahmad and Nasru Minallah. "Classification and Regression Analysis of the Prognostic Breast Cancer using Generation Optimizing Algorithms." (2013).
[21] You, Haowen and George Rumbe. "Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data." IJIMAI 1 (2010): 5-12.
[22] Bennett, Kristin P., Ayhan Demiriz and Richard Maclin. "Exploiting unlabeled data in ensemble methods." KDD (2002).
[23] McGuire, Andrew, James A L Brown, Carmel Malone, Ray McLaughlin, Michael J. Kerin and Jonas Cicenas. "Effects of Age on the Detection and Management of Breast Cancer." Cancers (2015).
[24] Data Set: Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian, Breast Cancer Wisconsin (Prognostic) Data Set UCI Repository, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)
[25] Medical News Today, 2019, https://www.medicalnewstoday.com/articles/37136.php
[26] Sci-kit Learn documentation, 2019, https://scikit-learn.org/stable/
[27] Stanford Open Classroom, 2012, openclassroom.stanford.edu/