Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BREAST CANCER DIAGNOSIS AND PROGNOSIS USING STACKING ENSEMBLE TECHNIQUE

NOURHAN M. SWELAM[1], AYMAN E. KHEDR[2], HEND AUDA[3]

[1]Faculty of Commerce and Business Administration, Future University in Egypt (FUE), Cairo, Egypt,

[2]Faculty of Computers and Information Technology, Future University in Egypt (FUE), Cairo, Egypt,

[3]Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

E-mail: [1]Nourhan.Swelam@fue.edu.eg,[2]Ayman.Khedr@fue.edu.eg,[3]HendAuda@commerce.helwan.edu.eg

ABSTRACT

Breast Cancer is the most common type of cancers in Egypt, early diagnosis can help to lower the risks. For many physicians, predicting a cancerous tumor remains a challenging task also deciding which treatment plan would help the most. The availability of new medical technologies and the massive amount of patient data had motivated the basis of emergence of new strategies in the prediction and detection of cancer. Data mining analysis and Machine Learning (ML) techniques can help to develop tools that can be used as effective mechanism for early diagnosis and prognosis of breast cancer, which will greatly enhance patients' survival rate. The main objective of this paper is to compare between the performance of supervised learning classification algorithms and the performance of combination of these algorithms using stacking ensemble learning approach in terms of the classification accuracy, precision, recall and ROC. We conducted the experiments on breast cancer dataset collected from University of California, San Francisco. The results demonstrate that the proposed stacking ensemble learning model outperforms individual algorithms.

Keywords: *Breast cancer, Stacking, Classification; J48, KNN, Naïve Bayes, Support vector machine*

1. INTRODUCTION

Globally breast cancer is one of the most crucial diseases, the most widespread of all cancers, and the primary cause of cancer deaths in women [1]. breast cancer has superseded lung cancer as the most common cancer according to the International Agency for Research for Cancer, with an estimation of 2.3 million new cases (11.7% of total 19.3 million new cancer cases) worldwide [2]. There is more than 1.6 percent of all fatalities, with case fatality rates highest in low-income nations [3]. In Egypt, female breast cancer is the most commonly diagnosed type of cancer. Breast cancer affects 38.8% of Egyptian women, according to the National Cancer Registry Program of Egypt; in the last thirty years of the twentieth century, the burden of cancer has more than doubled globally and it is expected to increase again to almost triple by 2030. Cancer incidence is expected to almost double in the next two decades, being 456,000 new cases in 2010 to nearly 861,000 in 2030, which is the highest relative increase among all WHO regions. These estimates are based only on the effect of population growth and ageing, but the additional effect of increasing exposures to cancer risk factors, such as smoking, unhealthy regimen, environmental pollution, and low rate of physical activity, will lead to an even bigger rise in the burden of cancer. [4][5].

The human body is full of Trillions of live cells. Normal body cells multiply, grow, and die in an organized sequence. Normal cells divide more rapidly in the early years of a person's life to help the person to grow. Once a person reaches a certain age, most cells divide only to replace worn-out, damaged, or dead cells. Cancer develops when cells in a specific area of the body begin to multiply uncontrollably. There are many different types of cancer, but they all begin with the uncontrollable growth of cancerous or abnormal cells. Breast cancer is more common in women, but males can have it as well. [3]

Breast cancer risks can be decreased by early diagnosis; according to the American Cancer Society, early diagnosis of breast cancer risks can help to lower the chances of tumor progression and growth. There are many ways for Breast cancer

detection techniques include medical examination by a doctor, self- examination, and mammography. The recovery rate might reach up to 98% if breast cancer is diagnosed in its early stages [6]. In the latest research, Data mining has become a current technique, particularly for healthcare sector applications. It has been usually employed as a research tool for medical researchers, through using large datasets of medical records, they could recognize and develop patterns and interactions among a huge number of variable attributes and use them to predict and analyze the results. The widespread use of data mining enables doctors in detecting diseases early, developing patient results, cutting the cost of medical treatments, improving clinical studies, and allowing results analysis [7-8].

Data mining and analysis has considerable potential in the medical field. By applying data mining and analytics in a systematic manner, the healthcare system was able to specify inefficiencies and best practices [9]. Health systems seek to enhance care while lowering costs; experts believe that current possibilities to decrease costs and improve healthcare may save up to 30% of overall healthcare spending. This might result in a win-win situation for both of us [10]. ML methods have been widely applied in the healthcare industry as a valuable diagnostic tool, assisting clinicians in evaluating existing data and creating medical expert systems [11] As well in intelligent healthcare systems within the last few decades, particularly for breast cancer diagnosis and prognosis [12]. Early diagnosis can greatly enhance the treatment plan results and prognosis and increase the survival rate by promoting immediate clinical treatment to patients.[13] More accurate diagnosis of benign tumors can save patients from having to undergo unneeded procedures, examinations and treatments. As a result, significant research is being conducted to determine the proper diagnosis and prognosis for breast cancer and the classification of individuals into malignant or benign categories. Data mining and machine learning are widely regarded as the preferred approach for detecting significant features in complicated breast cancer datasets [14-15].

1.1 Research objectives

This research aims to propose a framework for disease detection and prognosis using combined analytical data mining techniques, so it could solve many problems by identifying the patterns and disease risk factors which lead to earlier and more accurate disease detection and better prognosis, suggest effective treatments and best practices, help doctors to make better decisions, increase survival rate and improve care provided while cutting cost.

Also, it aims to Illustrate the differences among the different methodologies and techniques of data analytics and data mining used for massive medical volumes of data in terms of accuracy, complexity and scalability then state the appropriate one and explore the expected obstacles that might threaten a proper implementation and prevent benefits of applying the proposed framework.

1.2 Research Question

the main research question is whether the stacking proposed framework for disease detection and prognosis using data analytics and machine learning techniques would address the research objectives and contribute the most to the healthcare sector.

2. RELATED WORK

The research in (Chaurasia, V., & Pal, S., 2021) designed stacking learning ensemble approach with feature selection techniques for measuring the algorithm's overall performance and to make a comparative analysis of breast cancer dataset with reduced and the entire attributes. Using machine learning algorithms as sub models such as SVM, k nearest neighbor, Nave Bayes, and perceptron to be trained then combined to create stacking framework using logistic regression algorithm and predict the accuracy for the whole model. The accuracy of stacked model with 10 features is 98.968 %. On the other hand, the stacked model accuracy with 5 features is 99.968%. Therefore, the research reveals that the dataset with fewer attributes has higher accuracy [16].

The research in (Abdar, et al, 2020) have used machine learning algorithms and data mining approaches to explore automated breast cancer prediction. A nested ensemble methodology is established to utilize stacking and voting as combined classifier techniques in the ensemble methods for discriminating benign from malignant breast cancer. Wisconsin Diagnostic Breast Cancer dataset is used. the proposed two-layer nested ensemble classifiers have been compared to each single algorithm (i.e., Bayes Net and Naive Bayes) as well as the model result to the previous studies. The research results proved that the proposed two-layer nested ensemble model is better than any single classifiers and the majority of previous studies with 98.07% accuracy [17].

The research in (Kumar, et al, 2017), compared the performance of supervised learning classification algorithms and their combinations

using the voting classifier approach. In this study, the performance of SVM, Naive Bayes, and J48 have been examined to improve predictive models for breast cancer prognosis. The results of all three algorithms have been aggregated, to attain a high accuracy rate. Voting employs a combination rule of majority voting, which is applied to these algorithms to boost the accuracy percentage. Concluding that all three algorithms combination through a vote technique is the most effective method for breast cancer prediction [1].

The research in (Rafaqat, et al, 2017) have used the rapid miner tool to apply different classification algorithms with feature selection and generation algorithms. The algorithms have been applied on Wisconsin Breast Cancer dataset with 569 instances and 32 distinct attributes. Additionally, a 10-fold cross validation has been performed which results revealed that the Logistic Regression, Linear Regression, and SVM algorithms performed better in terms of classification accuracy. The previously mentioned algorithms achieved 98.24 %, 98.24 %, and 98.07 %, respectively, than the aforementioned classification techniques [18].

3. METHODOLOGY

3.1 Proposed Approach

The proposed framework is a stacking ensemble technique. Stacking is one of the ensembles learning approaches. The predictions of various learning algorithms are combined by a trained learning algorithm. [16]. the advantage of stacking is that it may use the capabilities of a variety of high-performing models on a classification problem to produce predictions that outperform any individual model in the ensemble.

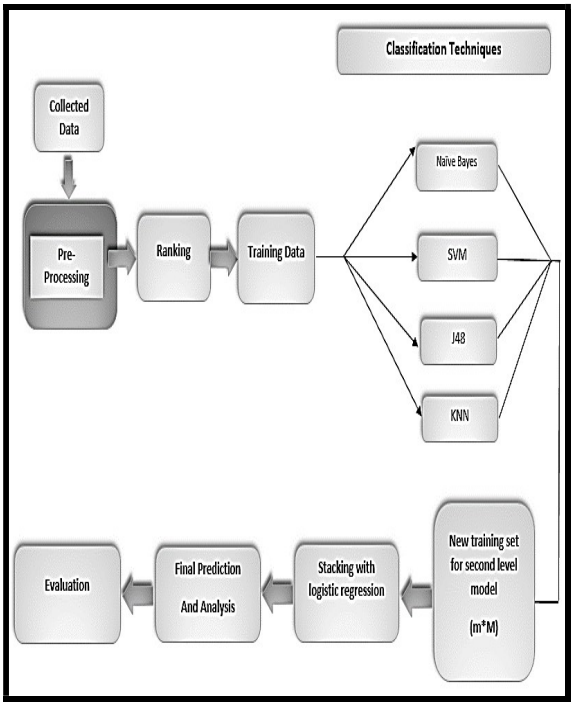

Figure 1: Proposed Stacking Framework

Briefly, in the pre-processing stage, we find and replace all the missing values and eliminate redundant attributes and make some filtration and discretization. Continuing with a ranking that indicates which attributes with low ranks and participating less in tumor prediction. Therefore, the pre-processed dataset is trained and tested with 10-fold cross-validation. In the first level of stacking, all the individual algorithms are trained on the dataset producing a new training set for the second level model, and then a combiner algorithm is trained to make a final prediction seeking better results and a higher accuracy rate. Finally, we use evaluation criteria to analyze the results and the whole experiment and compare between all individual algorithms' results and stacking results. Each step will be introduced in details in the subsequent sections.

3.2 Methods and Techniques

3.2.1 Stacking

Stacking is an ensemble machine-learning algorithm. It is a general approach in which a learning algorithm is trained to combine the predictions of other individual learning algorithms. The individual learners are referred to as the first-level learners, whereas the combiner is referred to as the second-level learner, or a meta-learner. To produce a successful ensemble model, it is often

assumed that the base learners should be as accurate and varied as feasible. [19]

As shown in figure 2, the architecture of the stacking ensemble learning consists of:

- All base learning classifiers are trained using the available data
- IT creates a number of models and forms a new feature matrix from all of the predictions.
- This matrix is then used for the second level model, which is meta-classifier layer to make the final prediction. [20]

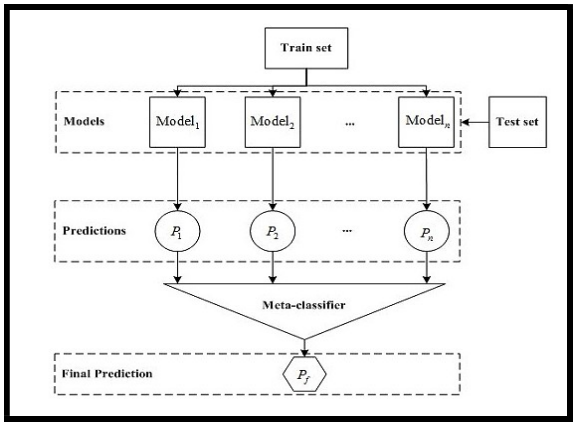

Figure 2: The architecture of the stacking ensemble learning [20]

3.2.2 K-Nearest Neighbors (K-NN)

K-NN one of the most important ML algorithms and non-parametric lazy learning technique in classification. The objects can be classified by their "k" nearest neighbors. K-NN algorithm analyses the object's neighbors rather than the underlying data distribution. There is also no training phase using the training data. K-NN is a straightforward algorithm that retains all existing instances and classifies new instances based on the majority of its K neighbors. The new instance assigned to the class is the most frequent among its K nearest neighbors as calculated by a distance function. These distance calculation functions can be Euclidean, Manhattan, Minkowski, or Hamming. Variables should be normalized since KNN is computationally costly. Otherwise, greater range of variables may provide misleading outcomes. [15, 21]

3.2.3 Support Vector Machine (SVM)

SVM is a simple method for request and backslide problems. It is capable of dealing with both direct and indirect issues, as well as various logical issues. The scope of SVM is obvious; a line that separates the data to classes is created by count. Regarding to the SVM assessment, the elements from both classes that are closest to the line. These points are referred to as support vectors. We eventually find out how to split the line and the support vectors. This splitting is referred to as the edge. We will almost certainly expand the edge. The optimal hyper-plane is the one with the most remarkable edge. Therefore, SVM attempts to select a decision boundary with the objective of reaching the greatest possible separation between the two classes. [16]

3.2.4 J48

A decision tree is a type of machine learning algorithm that is often used in data analysis and pattern identification. The purpose is to create a predictive model for target variable based on previous input variables. For prediction process, each internal node representing one of the input variables. For each of the input variable's possible values, there are edges to children. A leaf represents the target variable's value; hence, the route from the root to the leaf represents the values of the input variables.

C4.5, an extension of the ID3 tree learning method, is used as a decision tree algorithm. In C4.5, decision trees are generated from a set of training data using the principles of information entropy. Let S=s1, s2... sn be the training set, which consists of some previously classified samples. Each sample si=x1, x2... xn is a vector in which x1, x2... xn represent sample characteristics. Another vector is provided for the training data. C=c1, c2... cn, where c1, c2... cn is considered the class of each training sample. In order to construct a sub-tree at a certain node, C4.5 selects a data feature that would be most effective for dividing its samples' set into subsets boosted in one class or other. We use J48, which is a decision tree algorithm C4.5 implementation. [22]

3.2.5 Naïve Bayes

One of the simplest and oldest classifiers is the Naïve Bayes classifier. It has been commonly used for problem solving in a wide range of disciplines, including pattern recognition, natural language processing, and information retrieval. There is a presumption about naïve Bayes algorithm such as feature conditional independence, which is why it does not result in very well for many practical cases. The naïve Bayes classification is based on the Bayes theorem. A Bayesian classifier is both a

statistical classifier and a supervised learning technique. It is very useful for huge data sets. Given the class variable, the naïve Bayes approach assumed that the value of one feature is independent of the value of any other feature. When compared to numerical variables, it performs better with categorical input variables. [1]

3.3 Dataset and Platform

3.3.1 Dataset

The dataset shown in table 1 is collected by the Breast Imaging Research Program at The University of California, San Francisco; it was a clinical trial to examine if MRI might predict treatment response and risk of recurrence in patients with stage two or three breast cancer who were receiving new treatment of chemotherapy [23]. This study of imaging and tissue-based biomarkers for predicting two outcomes which are pathologic complete response (PCR class) which diagnose if the patient became cancer free or still a cancer patient after receiving neoadjuvant chemotherapy while Residual Cancer Burden class (RCB class) quantifies residual disease in breast or lymph nodes after receiving neoadjuvant chemotherapy. Pathologic complete response (PCR) refers to the exclusion of any residual cancer cells or lymph nodes, Achieving PCR after neoadjuvant chemotherapy is a preferable outcome [24]. The RCB index is a valid and accurate tool to estimate patient prognosis and predict the chances of breast cancer recurrence. Further, this prognostic risk assessment could be employed to define disease progression accurately and advise all subtypes of breast cancer patients with treatment plans and choices [25].

RCB has four categories of residual disease related to prognosis [26]:

- Class RCB-0 for those with no residual disease who achieved pathologic complete response.
- Class RCB-I for those with minimal residual disease.
- Class RCB-II for those with moderate residual disease.
- Class RCB-III those with extensive residual disease.

Table 1. Breast cancer dataset

No.	Variable Name	Variable Description
1	SUBJECTID	Patient ID
2	DataExtractDt	Date clinical data was downloaded from the CALGB database
3	Age	Patient Age
4	race_id	Patient Race 1=Caucasian 3=African American 4=Asian 5=Native Hawaiian/Pacific Islander 6=American Indian/Alaskan Native 50=Multiple race
5	ERpos	Estrogen Receptor Status 0=Negative 1=Positive 2=Indeterminate
6	PgRpos	Progesterone Receptor Status 0=Negative 1=Positive 2=Indeterminate
7	HR Pos	Hormone Receptor Status, pre-treatment 0=Negative for both ER and PR 1=Positive if either ER or PR was Positive 2=Indeterminate if both ER and PR were Indeterminate
8	Her2MostPos	Her2 Status 0=Negative 1=Positive Blank= indeterminate or not done
9	HR_HER2_CATEGORY	3-level HR/Her2 category pre-treatment 1=HR Positive, Her2 Negative 2=Her2 Positive 3=Triple Negative
10	HR_HER2_STATUS	3-level HR/Her2 status pre-treatment HRposHER2neg = HR Positive, Her2 Negative HER2pos = Her2 Positive TripleNeg =Triple Negative
11	BilateralCa	Does the patient have bilateral breast cancer prior to neoadjuvant therapy? 0=No 1=Yes
12	Laterality	Index Tumor Laterality 1=Left 2=Right
13	Baseline	Timepoint 1= Pre-Treatment baseline
14	1-3d AC	Timepoint 2= 1-3days after start of AC (Early Treatment Day1, cycle 2)
15	InterReg	Timepoint 3= Inter-regimen
16	PreSurg	Timepoint 4= Pre-Surgery

17	RFS	Recurrence-free survival time
18	RFS_ind	Recurrence-free survival indicator 1=event (local or distant progression or death) 0=censor at last follow-up
Outcomes (Classes)		
19	PCR	"Pathologic Complete Response, post-neoadjuvant (no residual invasive disease in breast or lymph nodes; presence of only in situ disease are considered disease free):" 0= No (did not achieve PCR) (cancer patient) 1= Yes (cancer free)
20	RCB Class	Residual Cancer Burden class: 0= 0, RCB index 0 1= I, RCB index less than or equal to 1.36 2= II, RCB index greater than 1.36 or equal to 3.28 3= III, RCB index greater than 3.28

3.3.2 Platform

WEKA, an open-source data-mining platform, is used for this research experimental work. WEKA was created at New Zealand's University of Waikato. It is a repository of machine learning techniques for data mining problems. It provides tools for data pre-processing, classification, regression, clustering, association rules mining, and visualizations, allowing creating and applying machine-learning algorithms to real-world data mining tasks. [27]

4. EXPERIMENTAL SETUP

In proposed framework, the purpose is to obtain the highest accuracy and outstanding results by following a stacking ensemble learning approach, it requires training a logistic regression algorithm to aggregate the predictions of four different learning algorithms such as (J48, SVM, KNN and Naive Bayes) to achieve a strong ensemble model. Each ML algorithm has its own advantage, which will be beneficial to the other algorithms. Logistic regression is used in stacking to learn how to optimally combine the predictions from the various contributing algorithms. In this research, two types of experiments will be conducted over the dataset:

4.1 Diagnosis Experiment

In the diagnosis experiment, RCB attribute will be excluded and the PCR attribute will be considered as a class. All attributes are trained over the dataset to predict whether the patient after receiving the new chemotherapy became cancer free or still a cancer patient and needs subsequent treatment plans.

4.2 Prognosis Experiment

In the prognosis experiment, PCR attribute will be included in this experiment and the RCB will be considered as a class. All attributes are trained over the dataset to predict which RCB category the patient would have after receiving the new chemotherapy.

4.3 Experiment stages

The proposed model has the typical four stages; each stage consists of many steps and criteria to be followed. The four stages are the following:

- Preprocessing
- Data Mining
- Result Validation
- Result Evaluation

4.3.1 Phase one: Pre-processing

Preprocessing datasets is a critical stage in the data mining process. Data collection methods are most likely not completely under control. As a result, the data set may contain missing values, values that are out of range, or data combinations. Analyzing such a dataset without first preparing it may result in inaccurate findings. Data preprocessing methods includes many categories such as Data cleaning, Data integration, Data Transformation and Data Reduction.

So, in order to improve classification performance while also lowering storage and computing costs, a feature ranking approach must be used. It is the process of determining a subset of the available features in a dataset by assessing each feature's efficiency. As a result, feature ranking helps to minimize the problems associated with huge datasets, resulting in more general and easier to interpret models. [28]

4.3.2 Phase two: Data mining

The main concept in the data mining phase is to apply each data mining algorithm on the training set to detect the pattern, but in the stacking learning approach we can go an extra mile, it has a first-level model and a second-level model.in the first level model, the training dataset goes through different algorithms (J48, SVM, KNN and Naive Bayes). Then we take the predictions of these

algorithms and combine them to form a new matrix of size M*m; Capital M is a number of algorithms and small m is the training data set. So, the predictions from the first level are used as features for the second level model.

4.3.3 Phase three: Result Validation

Result validation is to use the testing set to verify the detected pattern. Prediction may be obtained using the model on the test instances, and the accuracy rate is calculated by comparing the previous label value of the test data to the predicted value by the algorithm. So, Accuracy rate is calculated for each algorithm in the first level model to form the new training dataset and to be compared later after applying the stacking ensemble-learning algorithm. Then in the second level model, logistic regression is trained to make the final predictions on the test set and find the accuracy rate that will lead us to the final step, which is result evaluation and analysis.

4.3.4 Phase four: Result Evaluation

Evaluating the performance of a data mining technique is an essential aspect in machine learning. Evaluation method is the benchmark to test the efficiency and performance of any model. The following methods for performance evaluation of the proposed model and classifiers:

- **Confusion matrix**

Confusion matrix is also called as error matrix. An NxN table summarizes how successful a classification model's predictions were. It mainly reports true positive, true negative, false positive and false negative.

- **Accuracy**

Accuracy is a metric which used for assessing classification models. Informally, it is the percentage of correct predictions made by the model. Formally, accuracy is the number of correct predictions divided by the total number of predictions [1]. This is achieved by applying the following formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Recall**

Recall is the proportion of positive cases that are labeled as positive (true positive rate). Recall seeks to address the following question: What percentage of true positives were successfully identified? The higher the recall, the lower the false negative rate. It is referred to as sensitivity as well. [28]. this is achieved by applying the following formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

- **Precision**

Precision is defined as the proportion of relevant instances found among the retrieved instances. Precision seeks to address the following question: What percentage of positive identifications were correct? The higher the precision, the lower the false positive rate [28]. This is achieved by applying the following formula:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

- **Kappa Statistic**

The Kappa statistic is a measure of how well the algorithm would perform comparing to how well it would have performed randomly. [1] The higher the model scores, the larger the difference between accuracy and null rate which is definitely preferred. This is achieved by applying the following formula:

$$K = \frac{(Po-Pe)}{(1-Pe)} \quad (4)$$

Where Po represents observed accuracy while Pe represents expected accuracy

- **F1-measure**

The F1-measure is a method of combining a classifier's precision and recall measurements by taking an equally harmonic mean of each. The range is from 0 to 1. The F1 score can reach its best with a maximum of 1 and its worst with a minimum of 0. [16] This is achieved by applying the following formula:

$$F1 = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

- **ROC Curve**

The area under the ROC curve is a two-dimensional measurement of the entire area under the curve. It evaluates how well the parameters differentiate between two diagnostic groups. It is commonly used to assess the validity of classification models and it is obtained by tracing the true positive rate (TPR) and the false positive rate (FPR) [16]. The model performs better if the ROC curve is closer to the upper left corner of the graph.

- **AUC**

AUC is an abbreviation for "Area under the ROC Curve." In other words, AUC measures the full two-dimensional area beneath the whole ROC curve from (0, 0) to (1, 1). AUC is a metric that aggregates performance throughout all classification thresholds. AUC may be interpreted

as the likelihood that the model ranks a random positive sample higher than a random negative example. The model performs better when the area under the curve is closer to 1[29].

5. RESULT AND DISCUSSION

5.1 Diagnosis Experiment Results

Starting with pre-processing to filter and select dataset attributes, duplicated attributes such as (HR-HER2 STATUS) was removed. Then all missing values in the dataset were replaced with the modes and the means of the training dataset by using WEKA. All the continuous attributes shown in table 1, such as 3, 13, 14, 15, 16, and 17 were discretized by 4 intervals. Table 2 shows dataset attributes for predicting PCR class.

Each one of these attributes contributes into predicting PCR Class. Using an attribute evaluator such as the CorrelationAttributeEval technique with a Ranker Search Method to rank all of the attributes in the dataset, so it will help us to prioritize the needed features for better classification and accuracy. As shown in the table 3, the dataset attributes and their attribute evaluator weightage in ranked order.

Table 2. Dataset features for PCR test

No.	Variable Name
1	SUBJECTID
2	DataExtractDt
3	Age
4	race_id
5	ERpos
6	PgRpos
7	HR Pos
8	"Her2MostPos
9	HR_HER2_CATEGORY
10	BilateralCa
11	Laterality
12	Baseline
13	1-3d AC
14	InterReg
15	PreSurg
16	RFS
17	RFS_ind
18	PCR (Class)

Removing low ranks features could improve the predictive accuracy of the classification algorithm but the algorithm could over fit if more features were reduced even after reaching global minimum. Therefore, to avoid model over fitting, we only removed ID and DATE attributes, as they have no benefit and contribution. After that the dataset is divided into a training dataset and testing dataset. The training set is the final product of the pre-processing stage.

Table3. Ranked attributes for PCR class test

Attribute	weightage	Rank
HR Pos	0.33416	1
PgRpos	0.32566	2
Her2MostPos	0.29311	3
ERpos	0.29226	4
HR_HER2_STATUS	0.28502	5
rfs_ind	0.19442	6
MRI LD PreSurg	0.17884	7
MRI LD InterReg	0.1611	8
MRI LD Baseline	0.10535	9
MRI LD 1-3dAC	0.06777	10
RFS	0.0512	11
age	0.04228	12
race_id	0.03199	13
Laterality	0.02226	14
BilateralCa	0.00384	15
SUBJECTID	0.00383	16
DataExtractDt	0	17

The dataset was trained and tested for each classification algorithm before being combined using the stacking ensemble-learning algorithm to achieve better classification results. We used 10-fold cross-validation to reduce problems like over fitting and selection bias. Comparing and evaluating the results and the performance of the four algorithms without the stacking ensemble technique, results revealed that the KNN algorithm has a better performance compared to the other three algorithms with 98.79% Accuracy, then applying the stacking ensemble technique using logistic regression algorithm and comparing the performance again with all other four algorithms and the results showed that the stacked model outscored all other individual algorithms with the highest accuracy 99.08%. All the evaluation results obtained for the dataset are illustrated from table 4 to table 7.

Table4. Individual classifiers accuracy before ranking

	J48	SVM	KNN	Naïve Bayes
Accuracy	74.44 %	98.70%	84.66%	98.23 %
Kappa statistic	0.431	0.9663	0.5968	0.9543
Mean absolute error	0.2613	0.0144	0.1535	0.0179
Root mean	0.4467	0.09	0.3917	0.133

squared error				
Relative absolute error	67.4858 %	3.713 %	39.6331 %	4.6224 %
Root relative squared error	101.5375 %	20.4492 %	89.0372 %	30.2379 %

According to the proposed framework, first, the dataset is trained and tested with the individual algorithms, later the attributes were ranked and the lowest ranked attributes were removed then all four algorithms' outputs were combined in second level of stacking using logistic regression algorithm. In table 4, all individual algorithms accuracies were computed and simulated. Simulation outcomes revealed that J48 has the highest accuracy with 98.7% compared to all other algorithms as J48 has the best performance of all for predicting whether the new chemotherapy was useful, and the patient became cancer free or still cancer patient.

Table5. Individual classifiers accuracy after ranking

	J48	SVM	KNN	Naïve Bayes	Stacked with logistic
Accuracy	75.43 %	98.70 %	84.66 %	98.79 %	99.08%
Kappa statistic	0.43	0.966	0.597	0.969	0.9761
Mean absolute error	0.252	0.015	0.154	0.009	0.0091
Root mean squared error	0.425	0.09	0.392	0.068	0.0688
Relative absolute error	65.01 %	3.71 %	39.63 %	2.38 %	2.3512 %
Root relative squared	96.57 35 %	20.44 92 %	89.03 72 %	15.54 55 %	15.643 2 %

Table6. PCR class performance analysis of the dataset using the mentioned classifiers after ranking

	J48	SVM	KNN	Naïve Bayes	Stacked with logistic
Precision	0.787	0.987	0.844	0.988	0.991
Recall	0.754	0.987	0.847	0.988	0.991
F-Measure	0.764	0.987	0.845	0.988	0.991
ROC Area	0.827	0.999	0.793	1.000	1.000
PRC Area	0.860	0.999	0.795	1.000	1.000

After ranking, the results showed a bit of improvement in terms of accuracy. J48 & SVM have no change but Naïve Bayes & KNN have an improved accuracy rate. However, after ranking KNN has the highest accuracy rate compared to other algorithms with 98.79, outscoring the J48 with only 0.09%. The results obtained for the dataset are illustrated in table 5. When the stacking ensemble technique using logistic regression algorithm was applied to attain highest prediction value than the other algorithms and by comparing the performance again with all other four algorithms, the results showed that the stacked model outscored all other individual algorithms with the highest accuracy 99.08 % and the accuracy has been enhanced by 0.29% from the previous highest accuracy rate. The results obtained for the dataset are illustrated in table 5, 6 and Figure 3.

It is observed that stacking algorithm has the highest accuracy rate with least error rate compared to the performance of all individual algorithms; also, it is observed that stacking algorithm curve in Fig.3 performs well in terms of sensitivity and specificity as the curve is closer to y-axis. The stacking classifier with logistic algorithm that combined the results of four other algorithms has an AUC of 0.9998, which indicates how the stacking model is performing well and accurate.

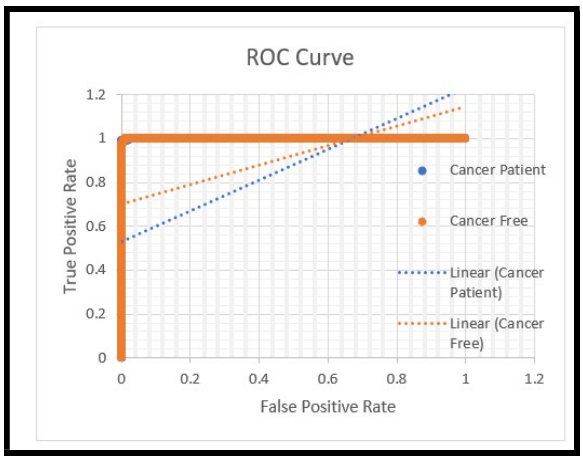

Figure 3: ROC curve for cancer patient and cancer free classes using stacking classifier

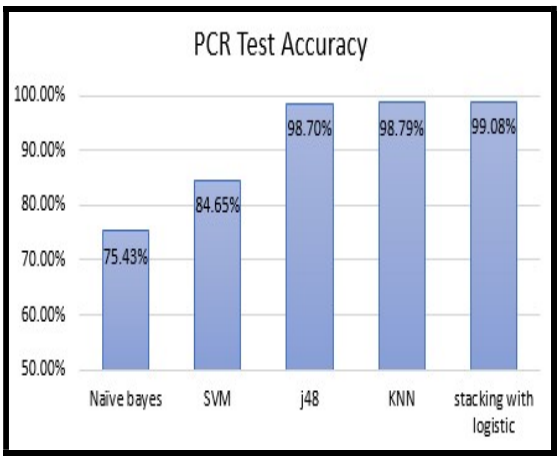

Figure 4: Accuracy percentage of all classifiers for diagnosis Test

Using stacking model confusion matrix obtained in table 7 and the results calculated in tables 5 and 6, we may conclude that 3723 instances are actually cancer patients that are correctly classified and 14 instances are not correctly classified, also 1313 instances are actually not cancer patients that are correctly classified and 21 instances are not correctly classified.

Table7. Stacking classifier confusion matrix for PCR Test

Total no. of instances		Predicted class	
		0 (Cancer patient)	1(Cancer free)
Actual class	0 (Cancer patient)	3723	26
	1(Cancer free)	21	1313

Since confusion matrices provide enough information to compute a variety of performance metrics, including Accuracy, precision and recall, the efficiency and the performance of the stacked model were tested. From Figure 4, the stacked model has the highest accuracy that it could make accurate predictions about whether the patient after neoadjuvant chemotherapy became cancer free or still a cancer patient with 99.08% accuracy and has the least error rate and the highest rate of precision, recall, F-measure and kappa statistics. The WEKA tool is used to simulate all of the values and values are illustrated in table 5 and 6. Hence, the study reaches the conclusion that using stacking ensemble approach provide the best performance among all individual algorithms.

5.2 Prognosis Experiment Results

The same preprocessing steps will be applied, and it results in table 8 Shows the dataset attributes and their attribute evaluator weightage in ranked order for prediction RCB Class and ID and DATE attributes will be removed as well.

The dataset was trained and tested for each classification algorithm to predict which RCB category the patient would be, before being combined using the stacking ensemble-learning algorithm. Comparing and evaluating the results and the performance of the four algorithms without the stacking ensemble technique, the results revealed that the KNN algorithm has a better performance compared to the other three algorithms with 97.52% Accuracy, then applying the stacking ensemble technique using logistic regression algorithm and comparing the performance again with all other four algorithms and the results showed that the stacked model surpassed all other individual algorithms with the highest accuracy 98.05% Shown in table 10. All the evaluation results obtained for the dataset are illustrated from table 9 to table 12.

Table8. Ranked attributes for RCB class test

Attribute	weightage	Rank
PCR	0.57297	1
Her2MostPos	0.19723	2
HR Pos	0.19416	3
HR_HER2_STATUS	0.17635	4
ERpos	0.17633	5
PgRpos	0.17386	6
MRI LD PreSurg	0.12904	7
MRI LD InterReg	0.11474	8
rfs_ind	0.11329	9
MRI LD 1-3dAC	0.08069	10
Laterality	0.08057	11
RFS	0.07582	12

MRI LD Baseline	0.0541	13
race_id	0.03694	14
age	0.02932	15
BilateralCa	0.02861	16
SUBJECTID	0.00223	17
DataExtractDt	0	18

Table9. Individual classifiers accuracy before ranking for RCB

	J48	SVM	KNN	Naïve Bayes
Accuracy	75.55 %	97.46%	82.12 %	96.4 %
Kappa statistic	0.6411	0.9631	0.7345	0.9477
Mean absolute error	0.1602	0.0147	0.2681	0.0183
Root mean squared error	0.2971	0.0902	0.339	0.1341
Relative absolute error	46.5139 %	4.2603 %	77.8247 %	5.3068 %
Root relative squared error	71.5863 %	21.7345 %	81.673 %	32.315 %

According to the proposed model, first, the dataset is trained and tested with the individual algorithms, later the attributes were ranked and the lowest ranked attributes were removed and then all four algorithms' outputs were combined in second level of stacking using logistic regression algorithm. In table 9, all individual algorithms accuracies were computed and simulated. Simulation outcomes revealed that J48 has the highest accuracy with 97.46% compared to all other algorithms as J48 has the best performance of all for predicting which RCB category the patient would have after receiving the neoadjuvant chemotherapy.

After ranking, the results showed a bit of improvement in terms of accuracy. J48 & SVM have no change but Naïve Bayes & KNN have an improved accuracy rate. However, after ranking KNN has the highest accuracy rate compared to other algorithms with 97.52, outscoring the J48 with only 0.06%. The results obtained for the dataset are illustrated in table 10. When the stacking ensemble technique using logistic regression algorithm was applied to attain highest prediction value than the other algorithms and by comparing the performance again with all other four algorithms and the results showed that the stacked model outscored all other individual algorithms with the highest accuracy 98.05 % and the accuracy has been enhanced by 0.53% from the previous highest accuracy rate. The results obtained for the dataset are illustrated in table 10 and 11.

It is observed that stacking algorithm has the highest accuracy rate with least error rate compared to the performance of all individual algorithms; also, it is observed stacking algorithm ROC performs well in terms of sensitivity and specificity as the ROC value is closer to 1. While AUC Value of the stacking classifier with logistic algorithm that combined the results of four other algorithms has a 0.999, which indicates how the stacking model is performing well and accurate.

Table10. Individual classifiers accuracy after ranking for RCB

	J48	SVM	KNN	Naïve Bayes	Stacked with logistic
Accuracy	78.2 %	97.46 %	82.1 %	97.52 %	98.05 %
Kappa statistic	0.67	0.963 1	0.734	0.964	0.9718
Mean absolute error	0.16	0.015	0.268	0.009	0.0094
Root mean squared error	0.29	0.09	0.339	0.069	0.0712
Relative absolute error	45.1 %	4.26 %	77.84 %	2.73 %	2.74 %
Root relative squared error	68.6 %	21.74 %	81.72 %	16.72 %	17.15 %

Table 11. RCB class performance analysis of the dataset using the mentioned classifiers after ranking

	J48	SVM	KNN	Naïve Bayes	Stacked with logistic
Precision	0.779	0.975	0.817	0.975	0.981
Recall	0.782	0.975	0.821	0.975	0.981
F-Measure	0.774	0.975	0.815	0.975	0.981
ROC Area	0.899	0.999	0.894	0.999	1.000
PRC Area	0.788	0.997	0.752	0.999	0.999

Using stacking classifier confusion matrix obtained in table 12 and the results calculated in tables 10 and 11 we may conclude that 1334 instances are actually cancer free that are correctly classified , also 497 instances are actually RCB-I patients that are correctly classified and 9 instances are not correctly classified and 2168 instances are considered RCB-II patients that are correctly classified and 63 instances are not correctly classified and finally 985 instances are considered RCB-III cancer patients that are correctly classified and 27 instances are not correctly classified.

Table 12. Stacking classifier confusion matrix for RCB Test

Total no. of instances		Predicted class			
		0 (RCB -0)	1 (RCB -I)	2 (RCB -II)	3 (RCB-III)
Actual class	0 (RCB -0)	1334	0	0	0
	1 (RCB - I)	0	497	9	0
	2 (RCB - II)	0	16	2168	47
	3 (RCB - III)	0	0	27	985

Since confusion matrices provide enough information to compute a variety of performance metrics, including Accuracy, precision and recall, the efficiency and the performance of the stacked model were tested. As shown in Figure 5, the stacked model has the highest accuracy that could make accurate predictions about which RCB class the patient would have after receiving neoadjuvant chemotherapy with 98.05% accuracy and has the least error rate and the highest rate of precision, recall, and F-measure and kappa statistics. The WEKA tool is used to simulate all of the values and values are illustrated in table 10 and 11. Hence, again the study reaches the conclusion that using stacking ensemble approach provide the best performance among all individual algorithms.

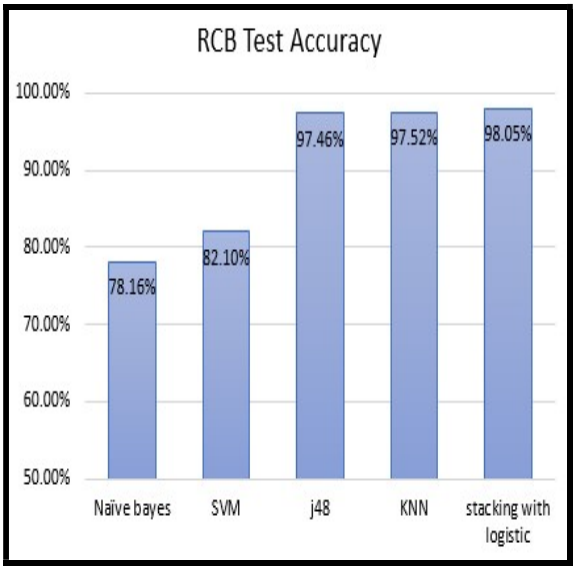

Figure 5: Accuracy percentage of all classifiers for prognosis test

Although there are numerous datamining studies are released every year, health sector and diseases prediction and prognosis, particularly breast cancer, remain a very rich and interesting field. Predicting the diseases accurately and deciding which treatment plan would be most beneficial is still a challenge.

our proposed framework was applied on a new dataset to test treatment response on breast cancer patients. Diagnosis and prognosis and chances of cancer recurrence were predicted. The proposed model of new stacked combination of machine learning algorithms was tested twice for diagnosis process then prognosis process with various evaluation measurements.

6. CONCLUSION AND FUTURE WORK

Accurate and on time diagnosis of any disease is an essential challenge in the healthcare sector and the diagnosis quality is main as well for successful prognosis. Consequently, the better the diagnosis and prognosis, the better treatments outcomes and survival rates for patients. Breast cancer is one of the most commonly diagnosed cancer types for women and becomes more critical if it is not diagnosed early. With the aid of data mining analytics and machine learning, a stacking ensemble learning model is proposed which trains a logistic regression algorithm to combine a several other learning algorithms i.e. (J48, SVM, KNN and Naïve Bayes) using all results and outcomes as an additional input to make final prediction.

We had breast cancer dataset and conducted two experiments on it; one for diagnosis test to identify or classify if the patient become cancer free or still have residual cancerous cells after receiving neoadjuvant chemotherapy and the other experiment is for prognosis test to identify or classify the cancer patients to which RCB class and know which cancer stage they have to decide the subsequent treatment plans. The findings of the stacking model for both experiments had the highest accuracy rate; for diagnosis test, the stacking model achieved 99.08% accuracy which increased by 0.29% from the previous highest accuracy rate of all other compared individual algorithms and for the prognosis test, the stacking model achieved 98.05% accuracy which increased by 0.53% from the previous highest accuracy rate of all other compared individual algorithms. Concluding that achieving highest accuracy rate with least error rates and performing better according all other evaluation parameters indicates that the proposed stacking ensemble model is more reliable and sophisticated for the prediction process.

The experimental results for the two tests show that the proposed framework accomplishes the research goals by bringing attention to new framework that determine best and accurate results after investigating data mining techniques in previous work. Also, enabling early detection of breast cancer and offering accurate diagnosis and prognosis. As well as determining if the new treatment will be beneficial for the patient or not which reduces cost, save time, and avoid any unnecessary procedures. concluding that the proposed framework has the potential to improve care, save lives, lower costs and make better informed decisions. However, there are some limitations that might threaten a proper implementation and prevent benefits of applying the proposed framework such as data accessibility, data collection and availability, data sample size, complexity of the analysis which increasing the computing time.

In future, various feature selection methods could be applied on forthcoming cancer diagnosis and prognosis applications to handle large and high dimensionally datasets. Additionally, comparative study on ensemble techniques could be conducted in order to model other diseases and evolve diagnostic and prognostic efficiency.

REFERENCES

[1] Kumar, U. K., Nikhil, M. S., & Sumangali, K. (2017, August). Prediction of breast cancer using voting classifier technique. In 2017 IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM) (pp. 108-114). IEEE.

[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May; 71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338.

[3] Chaurasia, V., & Pal, S. (2014). Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), 10-22.

[4] ALORABI, M., ELGHAZAWY, H., Alorabi, M., & Elghazawy, H. Cancer Control in Egypt: Investing in Health.

[5] Ibrahim, A. S., Khaled, H. M., Mikhail, N. N., Baraka, H., & Kamel, H. (2014). Cancer incidence in Egypt: results of the national population-based cancer registry program. Journal of cancer epidemiology, 2014.

[6] American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019.

[7] V Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, *6*(1), 1-25.

[8] Khedr, A. E., Idrees, A. M., Hegazy, A. E. F., & El-Shewy, S. (2018). A proposed configurable approach for recommendation systems via data mining techniques. *Enterprise Information Systems*, *12*(2), 196-217.

[9] Khedr, A. (2013). Business intelligence framework to support chronic liver disease treatment. *International Journal of computers & technology*, *4*(2a2), 307-312.

[10] Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M., Paul, R., Smriti, K., ... & Somani, B. K. (2021). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971-)*, 1-11.

[11] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer

detection and diagnosis. In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) (pp. 1-4). IEEE.

[12] Ali, M. M. R., Helmy, Y., Khedr, A. E., & Abdo, A. (2018, February). Intelligent Decision Framework to Explore and Control Infection of Hepatitis C Virus. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 264-274). Springer, Cham.

[13] Osman, M. A., Darwish, A., Khedr, A. E., Ghalwash, A. Z., & Hassanien, A. E. (2017). Enhanced breast cancer diagnosis system using fuzzy clustering means approach in digital mammography. In Handbook of Research on Machine Learning Innovations and Trends (pp. 925-941). IGI Global.

[14] El Seddawy, A. B., Sultan, T., & Khedr, A. (2013). Enhanced K-mean algorithm to improve decision support system under uncertain situations. International Journal of Computer Science and Network Security, 13(7), 4094-4101.

[15] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[16] Chaurasia, V., & Pal, S. (2021). Stacking-Based Ensemble Framework and Feature Selection Technique for the Detection of Breast Cancer. SN Computer Science, 2(2), 1-13.

[17] Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., & Gururajan, R. (2020). A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters, 132, 123-131

[18] Khan, R. A., Suleman, T., Farooq, M. S., Rafiq, M. H., & Tariq, M. A. (2017). Data mining algorithms for classification of diagnostic cancer using genetic optimization algorithms. Ijcsns, 17(12), 207.

[19] Zhou, Z. H. (2019). Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

[20] Jiang, W., Chen, Z., Xiang, Y., Shao, D., Ma, L., & Zhang, J. (2019). SSEM: A novel self-adaptive stacking ensemble model for classification. IEEE Access, 7, 120337-120349.

[21] Nithya, B., & Ilango, V. (2017). Comparative analysis of classification methods in R environment with two different data sets. Int J Sci Res Comput Sci, Eng Inf Technol, 2(6).

[22] Reddy, G. S., & Chittineni, S. (2021). Entropy based C4. 5-SHO algorithm with information gain optimization in data mining. *PeerJ Computer Science*, *7*, e424.

[23] Welcome To the Cancer Imaging Archive - The Cancer Imaging Archive (TCIA). (n.d.). The Cancer Imaging Archive (TCIA). Retrieved July 27, 2018, from https://www.cancerimagingarchive.net/

[24] Biswas, T., Jindal, C., Fitzgerald, T. L., & Efird, J. T. (2019). Pathologic Complete Response (pCR) and Survival of Women with Inflammatory Breast Cancer (IBC): An Analysis Based on Biologic Subtypes and Demographic Characteristics. International journal of environmental research and public health, 16(1), 124. https://doi.org/10.3390/ijerph16010124

[25] Yau, C., van der Noordaa, M., Wei, J., Osdoit, M., Reyal, F., Hamy, A. S., & Symmans, W. F. (2020, February). Residual cancer burden after neoadjuvant therapy and long-term survival outcomes in breast cancer: A multi-center pooled analysis. In CANCER RESEARCH (Vol. 80, No. 4). 615 CHESTNUT ST, 17TH FLOOR, PHILADELPHIA, PA 19106-4404 USA: AMER ASSOC CANCER RESEARCH.

[26] Hamy, A. S., Darrigues, L., Laas, E., De Croze, D., Topciu, L., Lam, G. T., ... & Reyal, F. (2020). Prognostic value of the Residual Cancer Burden index according to breast cancer subtype: Validation on a cohort of BC patients treated by neoadjuvant chemotherapy. *PloS one*, *15*(6), e0234191.

[27] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[28] Etaiwi, W., & Naymat, G. (2017). The impact of applying different preprocessing steps on review spam detection. *Procedia computer science*, *113*, 273-279.

[29] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.