## Research and Development Plan

## Project Title:

# Prediction of axillary lymph node status in women with breast cancer using machine learning models

## Organization Name:

**Omid hospital of Mashhad**

**Ferdowsi University of Mashhad**

## Project Manager:

**Sanaz Haghighi**

**Date: 2019-2020**

- ## Introduction and statement of the problem

According to the latest official announcement of the World Health Organization, breast cancer is the second most common cancer in the world in recent years, and it also ranks fifth among all types of cancer in terms of the number of deaths. In Iran this type of cancer, ranks first in terms of incidence and sixth in terms of death rate. Unfortunately, in breast cancer patients, the disease can spread to other parts of the body through the axillary lymph nodes. So, testing the status of lymph node involvement in this disease is a necessary task. In addition, investigating the involvement of the axillary lymph nodes is also a significant factor in determining the most suitable treatment [1].

In the past, the status of lymph node involvement was determined, only with the surgery [1]. After that, imaging methods such as mammography and ultrasound, with minor side effects and non-invasive properties, became more popular in predicting the status of lymph nodes. Magnetic Resonance Imaging (MRI) is a medical imaging technique employed in radiology to generate images of the anatomy and physiological processes within the body. Generally considered safe, MRI may pose risks if safety procedures fail or due to human error, leading to potential injuries. [2]. Another method is biopsy or sentinel lymph node dissection (SLND). In this method, as samples are taken only from small portions of the tissue, if the biopsy result is negative, it does not entirely rule out the possibility of cancer cells in the axillary tissue. On the other hand, when the final result indicates involvement of the lymph nodes, it can be concluded that cancer has spread to the axillary region. [1].

Based on past studies, the accuracy of each of the mentioned methods is given in Table (1):

**Table (1): Analyzing the results of examining surgical, imaging, and biopsy data of patients [1] [2] [3]**

| Method | accuracy while reporting involvement | accuracy while reporting non-involvement |
|---|---|---|
| **Surgery** | Between 76.9% to 97.6% | Between 76.9% to 97.6% |
| **Imaging** | Between 55% to 78% | Between 52% to 70% |
| **Biopsy** | 100 % | Between 62% to 85% |

According to Table (1), it appears that surgery is generally more reliable than other methods. However, there is a growing trend in utilizing less invasive techniques like sentinel lymph node (LN) biopsy or non-invasive methods for predicting axillary lymph node (ALN) status. This shift is motivated by the complications and morbidity associated with conventional axillary surgery, including issues such as lymphedema, range-of-motion restriction, and arm paresthesia and pain [9].

The growing popularity of machine learning in forecasting stems from the limitations other methods face, as the three mentioned methods struggle to provide reliable and accurate results. Approaches such as artificial neural networks

(ANN) and support vector machines (SVM) have shown satisfactory performance in similar studies. Machine learning, trained on datasets that include historical and real-time data, excels at interpreting new data and increases its predictive accuracy. In summary, the widespread adoption of machine learning is driven by factors such as the availability of significant datasets, advanced computational capabilities, continuous algorithmic developments, the prevalence of open source frameworks, and its tangible impact on optimizing business processes and decision-making worldwide. Various industries.

To sum up, in current study, the prediction of lymph node involvement status in breast cancer patients is explored. In addition to traditional methods like imaging and biopsy, five machine learning models are employed for their increased accuracy and efficiency. The primary objective of this research is to identify the most effective approach for predicting lymph node involvement status in women with breast cancer.

The innovations of this research are the use of different machine learning models, considering the results of the patient's biopsy and surgery as input variables to the models, and also examining different modes of combining models to improve prediction accuracy with the stacking approach.

## Research objectives

1. Providing an optimal combination of models to predict the status of axillary lymph node involvement
2. Identifying key variables associated with the status of axillary lymph node involvement
3. Designing and evaluating basic models
4. Fusing models with stacking approach to improve accuracy
5. Investigating the effect of considering the results of patient biopsy and surgery as input variables to the models

## Research questions

1. Which clinical, radiological, or genetic features have the greatest impact on predicting lymph node status?
2. Can stacking and combining models increase prediction accuracy?
3. Can considering patient biopsy and surgical results as input variables for models increase prediction accuracy?

## Research assumptions

1. Using multidimensional data can improve lymph node status prediction.
2. Combining models with stacking will perform better than individual models.

## Literature review and research background

As previously mentioned, the primary focus of this study is the diagnosis of involvement or non-involvement of axillary lymph nodes in cancer cells in women with breast cancer. In the same direction, the PubMed database contains 67 articles specifically addressing the prediction of lymph node status in patients with breast cancer. Out of these articles, only three studies have used machine learning approaches that all of them have used machine learning for interpreting the imaging of axillary. So, there is no study that use demographic and pathological data as inputs of machine learning methods. The oldest paper on this data base, is entitled "Prediction of prognosis in patients with axillary lymph node-positive breast cancer: a statistical study" 1984. This work has identified variables of importance to short- and long-term prognosis in 97 node-positive breast cancer patients followed for a minimum of 98 months. The diameter of the primary tumor, categorized as less than or equal to 3 cm or greater than 3 cm, emerges as a crucial prognostic variable. When combined with the presence/absence of tumor cells in the efferent nodal vessels and the mean nuclear area of the tumor cells, it yielded accurate predictions of disease outcomes 60 and 98 months' post-operation, achieving success rates of 83% and 80% for the respective time frames. While the number of tumor-bearing nodes remains a significant variable, the tumor diameter provided additional valuable information in predicting outcomes [4].

Over the following years, studies in this field have improved and become more sophisticated to achieve better results with higher accuracy. In the same direction, some surveys aim to introduce a model to predict the status of axillary lymph nodes.

In 2005 Tanja Fehm et.al. showed that it is possible to predict axillary lymph node status, with a model based on tumor biological parameters obtained in the primary tumor. Incorporating additional parameters offers the potential for further enhancing the model, aiming to prevent unnecessary axillary lymph node surgery in low-risk women. The predictive accuracy of the model currently stands at 70% [5].

In 2012, Stephanie A. Valente et al. conducted a retrospective review involving 244 consecutive patients diagnosed with invasive breast carcinoma. These patients underwent a comprehensive assessment, including physical examination of the axilla, digital mammography, axillary ultrasonography, and contrast-enhanced breast MRI. Subsequently, histopathologic evaluation was performed on one or more axillary lymph nodes. Ultimately, their conclusion highlighted that the combination of physical examination and multimodal imaging proves valuable for preoperative axillary staging and treatment planning. However, it was emphasized that these methods still fall short as definitive predictors of axillary lymph node involvement [6].

In 2015, Su Hyun Yoo et al. endeavored to develop a pathologic nomogram capable of predicting axillary lymph node metastasis (LNM) for each intrinsic subtype of breast cancer based on histologic characteristics observed in breast core needle biopsy (CNB) for routine clinical application. The study included 534 CNBs with invasive ductal carcinoma categorized into 5 intrinsic subtypes. Eighteen clinic pathological characteristics and 8 molecular markers employed in CNB were assessed to construct the most effective predictive model for LNM [7].

In 2015, P. M. Ravdin et al. utilized a training set comprising 5963 patients to build predictive models. These models employed stepwise logistic regression, both with and without first-order interactions. Since all models demonstrated similar performance, the study opted for the simplest model, namely the LR models without interaction terms. The performance of these models in the patient test set was evaluated based on the predicted number of nodes [8].

In 2018, Woo Kyung Moon et al. developed a computer-aided prediction (CAP) model for predicting axillary lymph node (ALN) metastasis in breast cancers using breast ultrasound (US) images. The study involved 249 malignant tumors obtained from 247 female patients, with ages ranging from 20 to 84 years and a mean age of 55 ± 11 years. The tumors were categorized into non-metastatic (130) and metastatic (119) groups based on various features. Following semi-automatic tumor segmentation, 69 quantitative features were extracted, encompassing the morphology and texture of tumors within a region of interest (ROI) in breast US images. Through backward feature selection and linear LR, a prediction model was constructed to estimate the likelihood of axillary lymph node (ALN) metastasis for each collected sample. The study concluded that the proposed Computer-Aided Prediction (CAP) model, incorporating both textural and morphological features of the primary tumor, proves to be a valuable method for determining ALN status in patients with breast cancer [9].

In 2020, Li-Qiang Zhou et al. employed deep learning for predicting Lymph Node Metastasis from primary breast cancer US images. The study concluded that deep learning models, when applied to US images of patients with primary breast cancer, can effectively predict clinically negative axillary lymph node metastasis. This suggests that AI has the potential to offer an early diagnostic strategy for identifying lymph node metastasis in patients with breast cancer who initially present with clinically negative lymph nodes [10].

In one of the latest studies conducted in 2022, Shirin Magdar et al. conducted a univariate analysis to identify significant variables related to lymph node status. According to the findings, tumor grade, ER status, lymph vascular invasion (LVI) status, mean KI-67 index, PT-SUV max, PT-SUV mean, PT-TLG, and TL ratio emerged as significant factors. The study established a relationship between these variables and the SLNB (Sentinel Lymph Node Biopsy) positive and negative groups [11].

In a general overview, studies on predicting lymph node involvement in women with breast cancer can be broadly categorized into two main groups. These studies primarily focus on either creating models to predict the status of lymph node involvement by considering various prognostic factors or accurately identifying tumors in imaging methods such as mammography, breast ultrasound, and magnetic resonance imaging (MRI) or biopsy.

To facilitate a comprehensive comparison, a summarized overview of the previously mentioned papers, along with new additions and the current study, is presented in Table (2). Based on Table (2), some points are noticeable. First, the number of variables in our study is the highest among all studies. It consists of both demographic (like age) and pathological (like HER-2) indicators. The highest number in studies was 8 before. Second, in our study, we have used 5 different machine learning models to diagnose the status of lymph nodes and also used the stacking method to examine the result of combining all these models. The last and most important point is the percent of accuracy which is the most in our study. In result section we discuss about this, comprehensively and completely. In addition to the items compared in the table, the following points can be mentioned as other strengths of this research:

- Conducting a distinct analysis of each model using both pathological and demographic variables as inputs.
- Incorporating imaging and biopsy results into the existing input variables and refitting the models.
- using the most robust machine learning models
- Evaluating the combination of results from all models through a stacking approach.
- Employing the Weighted by Relief method for feature selection.

**Table (2): Comparison between information and results of similar studies**

| | **Number of patients** | **Method of prediction** | **Reported accuracy** | **Number of variables** | **Modeling with clinic pathological variables** | **Modeling with imaging or biopsy** |
|---|---|---|---|---|---|---|
| [5] | 200 | multivariate logistic regression | 70% | 8 | * | |
| [6] | 244 | PE/ MMG/ US/ MRI (imaging methods) | US= 82.8 | - | | * |
| [7] | 534 | core needle biopsy | 79% | 8 | * | |
| [8] | 11964 | Logistic regression models | - | 7 | * | |
| [9] | 247 | a computer-aided prediction model for interpreting ultrasound | 75.1% | - | | * |
| [10] | 1055 | ANN for interpreting US images | 82% | - | | * |
| [11] | 66 | Probabilistic neural network | 71% | 5 | * | |
| [12] | 198 | deep learning-based model | 88% | 3 | | * |
| [13] | 266 | Random Forest/ XGBoost/ANN | ANN=72% | 4 | | * |
| [14] | 3701 | multi-modal and multi-instance deep learning model | 75% | - | | * |
| [15] | 988 | Convolutional neural network for interpreting MRI | 89.2% | - | * | |
| **this study** | **235** | **ANN/SVM/KNN/Logistic/Random Forest (and their combination by stacking approach)** | **93.01%** | **15** | ***** | ***** |

Based on our search, no paper in PubMed has integrated or compared the results of imaging or biopsy with prediction models incorporating prognostic factors. In this study, to explore novel approaches using patient data, biopsy and imaging results were collected along with other available patient information. Subsequently, five different machine learning models were applied to predict the status of lymph node involvement. Finally, employing a stacking approach, various combinations of the five models were examined in two distinct situations: first, without the results of biopsy and imaging, and second, by including them.

## • Methods and models

Research type: Descriptive-analytical using clinical, radiological and genetic data.

### Patients and data sets

In the initial phase of this study, recorded data was collected from 235 women with breast cancer, who were hospitalized in Omid Hospital in Mashhad between 2012 and 2014. Finally, out of 235 patients, 157 patients with breast cancer were included in the study. because The other patients had not performed lymph node removal surgery, and the status of their lymph node involvement was not known. Sixty-four patients had involvement of lymph nodes due to metastasis; in the other 93 patients, no signs of involvement were observed after lymph node removal surgery. Demographic and clinical information, as well as some information about treatment methods and other variables, were extracted from the patient's files. The range of age was, 25_89 years (average: 50 years) and the range of the size of the tumor was 0.2_14 cm (average: 3.84 cm). Other demographic data is shown in Table (3) and a flowchart describing the research process is shown in Figure (1).

**Table (3): some Demographic data for 157 patients**

| characteristic | Data set |
|---|---|
| **No. Of patients** | 157 |
| **involvement** | 64 |
| **Non-involvement** | 93 |
| **Age** | |
| **<40 y** | 47 |
| **40_49 y** | 55 |
| **50_59 y** | 27 |
| **60_69 y** | 22 |
| **≥70** | 6 |
| **Involved breast** | |
| **Right** | 64 |
| **Left** | 93 |
| **Tumor size** | |
| **≤0.2 Cm** | 62 |
| **2.1_4.0 Cm** | 74 |
| **>4 Cm** | 21 |

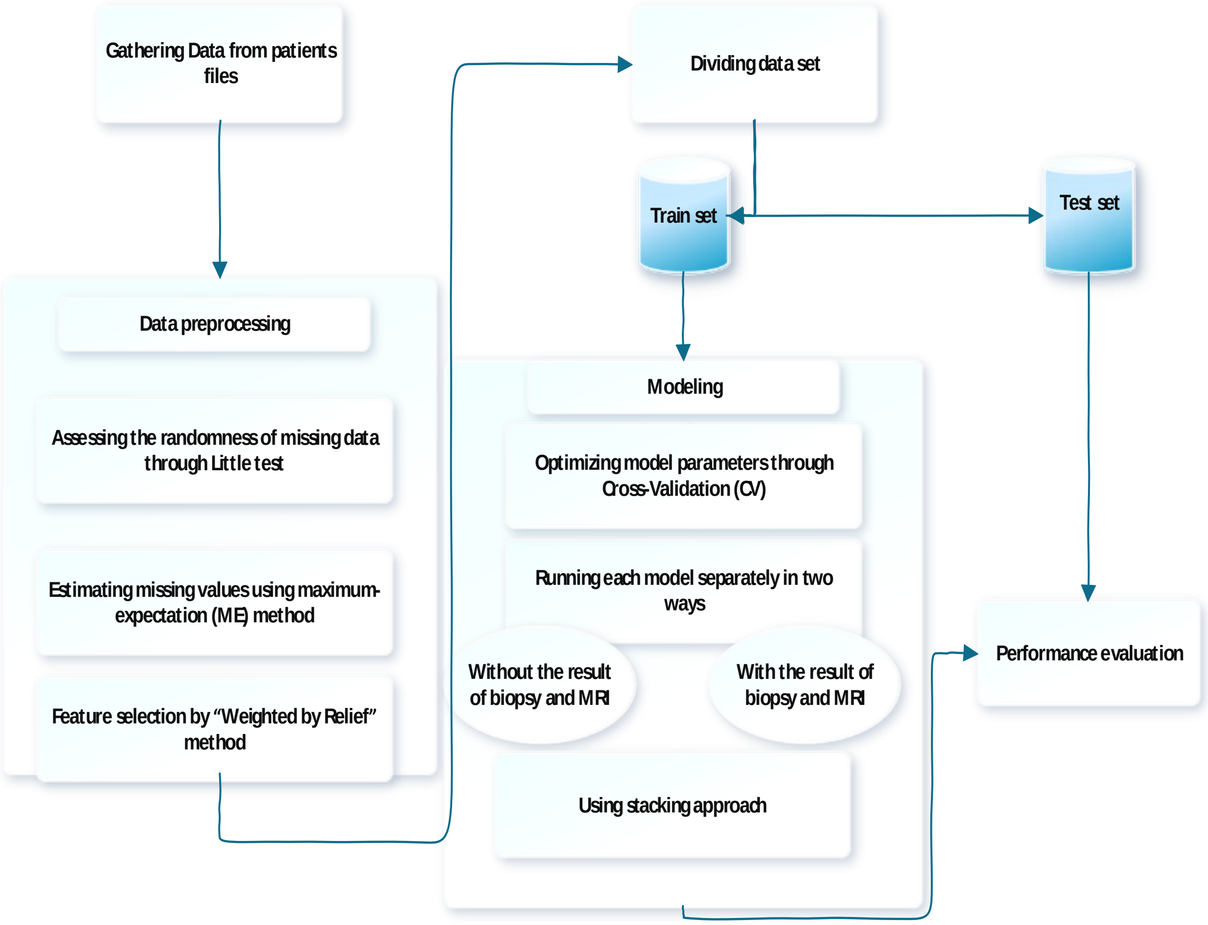
Gathering Data from patients files
Dividing data set
Train set
Test set
Data preprocessing
Assessing the randomness of missing data through Little test
Estimating missing values using maximum-expectation (ME) method
Feature selection by "Weighted by Relief" method
Modeling
Optimizing model parameters through Cross-Validation (CV)
Running each model separately in two ways
Without the result of biopsy and MRI
With the result of biopsy and MRI
Using stacking approach
Performance evaluation

**Figure 1: Flowchart of procedures in the data processing and development and evaluation of machine learning models**

## Data preprocessing

In data management, the most effective approach is to ensure comprehensive data collection, minimizing or avoiding missing data. However, the presence of missing data in the data set is almost inevitable. In this research, the data related to some variables had missing values [16]. Before selecting the suitable method for handling missing data, it was essential to examine the data for complete randomness. In the context of multivariate quantitative data analysis, Little's test was employed to assess this criterion [17]. The results of the test showed that the missing data were entirely random. So, for estimating the missing values, the Expectation-Maximization method was used [18]. In the next step, it was time to choose the suitable variables to enter the model [19]. For this purpose, the Weighted by relief method was used on the training set. Based on this method, among the 15 collected variables, only the following 8 variables applied to the final phase of the research: 1. The involved breast (left, right), 2. the initial condition in the first visit (Primary patient or Recurrence), 3. the degree of tumor malignancy (1,2,3), 4. the status of the estrogen and, 5. progesterone hormone receptors, 6. human growth factor 2 (positive or negative), 7. the period between observing the symptoms by the patient to start the treatment process, 8. the Triple negative of breast cancer (positive or negative). So, it can be concluded that all the demographic variables have been excluded and only pathological variables have been detected effective on the status of lymph nodes.

## Models

In this section, all the models used in this study are introduced. Each model has its own characteristics and advantages. The following are the reasons why these models are suitable for medical data, especially in the field of cancer diagnosis:

### 1. Artificial neural network (ANN) model

Artificial neural networks (ANN) are a fundamental component of machine learning, inspired by the structure and function of the human brain. These networks consist of interconnected nodes or "neurons" that process information in layers, each layer helping to extract complex patterns and features from the input data. Artificial neural networks consist of an input layer, hidden layers, and an output layer, and the connections between neurons are assigned weights that are adjusted during the learning process. Learning occurs through a training phase where the network modifies its weights based on the data provided, allowing it to make accurate predictions or classifications.

The strength of artificial neural networks lies in their ability to model complex relationships in data, which makes them particularly effective in tasks such as image recognition, natural language processing, and predictive analytics. The hierarchical structure and adaptability of neural networks enable them to recognize subtle patterns and increase their performance in handling diverse and complex data sets. As a versatile tool in the machine learning landscape, ANNs continue to advance in various fields and provide solutions to complex problems through their capacity to learn and generalize from diverse datasets [20].

**Why it is suitable for this study?**

Flexibility and learning power: Artificial neural networks are able to simulate complex and nonlinear relationships in data, which is useful in many medical problems such as cancer diagnosis and prediction of its metastases.

Ability to process large and complex data: ANN has the ability to process large amounts of data (such as biopsy and imaging results), which is suitable for more accurate diagnosis and medical decision-making.

Fusion of different data: This model can make good use of different data (clinical, imaging, and biopsy data) to provide more accurate predictions.

### 2. Support vector machine (SVM) model

SVMs are powerful machine learning algorithms designed for classification and regression tasks. The essence of SVM lies in its ability to find an optimal hyperplane that separates the data into distinct classes in the feature space. "Support vectors" are data points that are closest to the decision boundary and help determine the optimal hyperplane. SVMs are particularly effective in high-dimensional spaces, making them suitable for tasks such as image classification and text classification. In addition, SVMs can handle linear and non-linear relationships between features through the use of different kernel functions, providing flexibility in capturing complex patterns in data. One of the notable strengths of SVMs is their ability to generalize well, even in scenarios with limited training data. The goal of SVMs is to increase robustness against changes and noise in the dataset by maximizing the margin between classes. This makes SVMs valuable in real-world applications where data sets are sparse or exhibit overlapping patterns. The versatility, efficiency, and generalization capabilities of support vector machines contribute to their widespread adoption in various fields, including finance, healthcare, and bioinformatics [21].

**Why it is suitable for this research?**

High accuracy in data separation: SVM is very effective in separating complex and nonlinear data, especially when the classification boundaries are not clearly separated. This feature is very important for cancer prediction and lymph node involvement, which require high accuracy.

Use of additional data: SVM is well able to use additional data such as biopsy and imaging results, and significantly increase the prediction accuracy.

### 3. Nearest neighbor (KNN) model

KNN is a simple yet effective machine learning algorithm used for classification and regression tasks. The main idea behind KNN is to predict the class or value of a data point based on the majority class or average of its nearest neighbors in the feature space. The "K" in KNN represents the number of neighbors considered in the prediction, and the algorithm calculates distances, often using the Euclidean distance, to identify the closest data points. KNN operates on the assumption that similar instances in the feature space tend to share similar results, making it particularly suitable for scenarios where local patterns are essential. A notable feature of KNN is its lazy learning approach - the model does not explicitly learn from the training data during the training phase. Instead, it remembers the data set and performs the calculations at forecast time. This adapts the KNN to changes in the dataset and is particularly useful in situations where underlying patterns may evolve over time. While KNN is intuitive and easy to implement, its performance can be affected by data dimensionality and dataset size. Nevertheless, KNN remains a valuable tool, especially in applications such as recommender systems, image recognition, and anomaly detection [22].

**Why is it suitable for this research?**

Simplicity and efficiency in small problems: KNN is a simple and efficient model that can perform well when the number of data is limited. This model is especially suitable for medical data, which sometimes cannot be collected due to restrictions on access to complete data.

Ability to process new data: KNN can easily update its performance as new data is entered, without the need to retrain the model, which can be useful in clinical environments with up-to-date and continuous data.

### 4. Random forest (RF) model

RF is a powerful ensemble learning algorithm that is widely used for classification and regression tasks. The system works by building a large number of decision trees during the training phase and combining their outputs for robust predictions. Each tree in the random forest is trained on a random subset of features and a bootstrap sample of the training data that introduces variation among individual trees. This diversity contributes to the resilience of the model against overfitting, as the collective decision of multiple trees leads to more accurate and stable predictions than a single tree. One of the strengths of Random Forest lies in its ability to handle high-dimensional data, large datasets, and a combination of categorical and numerical features. This algorithm is less sensitive to outliers and noise and provides feature importance ranking and helps to identify the most influential variables. Random Forest's versatility and robustness make it a popular choice in a variety of fields, including finance, healthcare, and remote sensing, where accurate predictions and interpretability are essential [23].

**Why it is suitable for this research?**

Robust and interpretable model: Random Forest uses a combination of multiple decision trees for prediction, which makes this model very robust and error-resistant.

Ability to deal with heterogeneous features: In medical data, there are a variety of features (e.g. clinical data, biopsy, imaging) that may be interdependent. Random Forest can process and analyze these features well.

Improved accuracy with more data: The Random Forest model usually improves its accuracy when more data is input, which can be seen in this study by adding imaging and biopsy data.

### 5. Logistic regression (LR) model

LR is a widely used statistical method for binary classification problems where the outcome variable has two possible classes. Despite its name, logistic regression is used for classification rather than regression. The algorithm models the probability of a sample belonging to a particular class using a logistic function, which ensures that the predicted probabilities lie between 0 and 1. The model calculates the coefficients for each input feature, and the weighted sum of these features is converted into probabilities with intercepts through the logistic function. A decision boundary is then applied to classify the samples into one of two classes based on their predicted probabilities. One of the notable features of logistic regression is its simplicity and interpretability. The coefficients obtained from the model provide insights into the effect of each feature on the probability of belonging to a particular class. Logistic regression is particularly useful when the relationship between the characteristics and the binary outcome is linear. While logistic regression is basic, it is versatile, and modifications such as multinomial logistic regression extend its applicability to scenarios with more than two classes. This algorithm finds application in fields such as healthcare to predict disease outcomes, marketing to predict customer churn, and various other fields that require binary classification tasks [24].

**Why it is suitable for this study?**

Simplicity and interpretability: One of the main advantages of logistic regression models is the simplicity of interpreting the results. In medicine, interpreting the results is very important for doctors to make correct treatment decisions.

Suitable for two-class problems: In medical diagnostic problems, there is usually a two-class diagnosis (e.g., lymph node involvement or not), for which logistic regression is a suitable method for this type of problem.

**Why are these models suitable for medical data?**

1. Ability to analyze complex data: Medical data is usually complex, nonlinear, and multidimensional. ANN and SVM models are able to simulate more complex patterns.

2. Ability to generalize to new data: These models, especially when properly trained, can generalize well to new data, which is very crucial in medical problems.

3. Consideration of diverse data: Medical data includes various information such as disease history, test results, imaging, and biopsy. These models can use different data simultaneously and achieve more accurate predictions.

Therefore, the choice of these models is quite logical and appropriate due to their power in analyzing medical data, high accuracy, and ability to generalize to new data.

### Stacking approach

Stacking, or Stacked Generalization, represents an ensemble machine learning algorithm that uses a meta-learning approach to determine the optimal way of combining predictions from two or more base machine-learning algorithms. Stacking performs much better than single models because it combines several different forecasting models. In fact, stacking can combine the advantages of several different models and exploit the weaknesses of each model in other models. The reasons for the superiority of stacking are as follows:

1. Combining the advantages of multiple models

In stacking, models are trained in parallel, each providing predictions for the data in turn. Then, a final model (usually a simple model such as logistic regression) combines the predictions of the base models to provide the final forecast.

This combination makes the final model's predictive power greater than the individual abilities of the base models. For example, one model may be better at recognizing certain types of patterns, while another model may be better at other areas. Stacking intelligently combines these differences.

2. Reducing prediction error

Stacking has the advantage of reducing the errors of the base models. When different models have different results, the final model can achieve higher accuracy by combining these results. Especially in medical data that has a lot of complexity and variety, stacking can reduce prediction error.

3. Better performance in complex problems

Medical models often have complex data that may not be able to be analyzed linearly or simply. In these situations, stacking can use the combination of complex and simple models to achieve higher accuracy.

4. Flexibility and better generalization

Stacking can use different models such as SVM, ANN, KNN, and Random Forest, each of which compensates for the weaknesses of the other. This feature allows the final model to be able to best process different data.

The architecture of a stacking model typically consists of two or more base models, commonly referred to as level-0 models. These base models generate predictions that are then aggregated by a meta-model, often termed the level-1 model. The meta-model combines the predictions from the base models, providing a comprehensive and refined outcome that enhances predictive accuracy and generalization across various scenarios.

- Level-0 Models (*Base-Models*): These are models that are trained on the training data, and their individual predictions are collected. These base models serve as the foundation for the ensemble approach.
- Level-1 Model (*Meta-Model*): The meta-model, or level-1 model, is responsible for learning the optimal way to combine the predictions generated by the base models. It takes the outputs of the base models as input and produces a refined and consolidated prediction, enhancing the overall performance of the ensemble.

The meta-model is trained using the predictions generated by the base models. Specifically, data that was not used during the training of the base models is provided to these base models for making predictions. The predictions, along with the corresponding expected outputs, form the input and output pairs of the training dataset used to fit the meta-model. This process allows the meta-model to learn how to effectively combine the predictions from the base models, optimizing the ensemble's overall predictive performance. The typical method for preparing the training dataset for the meta-model involves k-fold cross-validation of the base models. In this approach, the out-of-fold predictions serve as the foundation for constructing the training dataset for the meta-model. Following the preparation of the meta-model's training dataset, the meta-model is trained independently on this data, while the base models are trained on the entire original training dataset. This ensures that the meta-model learns from the diverse predictions generated by the base models, optimizing its ability to effectively combine their outputs and enhance the ensemble's overall predictive performance. The meta-model is frequently designed to be simple, offering a straightforward interpretation of the predictions generated by the base models. Linear models are commonly chosen for this role, with linear regression employed for regression tasks (predicting a numeric value) and LR utilized for classification tasks (predicting a class label). However, while this approach is common, it is not mandatory, and other types of models can be used as the meta-model depending on the specific requirements and characteristics of the problem at hand.

- Regression Meta-Model: Linear Regression.
- Classification Meta-Model: LR (used in this study)

The extent of performance improvement achieved through stacking depends on the complexity of the problem at hand and the adequacy of its representation in the training data. It is particularly effective when the problem is intricate enough that there is substantial knowledge to be gained by combining predictions from multiple models. Additionally, the choice of base models is crucial, and their effectiveness relies on both their individual skill and the degree of correlation in their predictions (or errors).

If a base model demonstrates comparable or superior performance to the stacking ensemble, it is advisable to favor the base model. This preference is attributed to the lower complexity of the base model, making it simpler to describe, train, and maintain. The decision to use stacking or a base model hinges on a nuanced evaluation of problem complexity, data representation, and the performance and correlation of the individual models involved.

### Evaluation criteria

The three markers given in Table (4) have been used as the evaluation criteria of the models. It is necessary to explain that the factors of sensitivity, specificity, and accuracy were calculated from the following methods. The TP index denotes the number of accurately predicted positive cases (indicating involvement), while FP represents the count of incorrectly predicted positive cases (erroneous lymph node involvement predictions). Similarly, TN and FN signify the accurate and inaccurate predictions of negative cases (non-involvement), respectively. In medical terminology, sensitivity refers to the percentage of individuals who test positive for a disease among those who actually have the disease. A highly sensitive test is effective in ruling out individuals who do not have the disease. On the other hand, specificity is the percentage of individuals without the disease who test negative for it. A highly specific test aids in correctly identifying individuals who truly have the disease [5] [6].

**Table (4): the diagnostic performance indices**

| | |
|---|---|
| **sensitivity** | $\boldsymbol{Sensitivity = \frac{TP}{TP + FN}}$ |
| **specificity** | $\boldsymbol{Specificity = \frac{TN}{TN + FP}}$ |
| **Accuracy** | $\boldsymbol{Accuracy = \frac{correct\ predictions}{all\ predictions}}$ |

## • Results

Each machine learning model has several hyper parameters that can be adjusted to optimize its performance. Its behavior is controlled by hyper parameters, which have a big effect on how well the model performs. Therefore, prior to presenting the primary results of this research, the adjusted hyper parameters for each model are presented.

### 1. Adjust the hyper parameters of the models

In this study, the trial and error process in conjunction with K-fold CV has been used to adjust the hyper parameters of the models in Rapid Miner software, and their values are given in Table (5). In this process, multiple trials with different parameter values are performed, and for each set of parameters, the model's performance are evaluated using CV. This allows iteratively refine the parameter values based on the model's performance on validation data, ultimately leading to the selection of optimal parameters that generalize well. Moreover, using CV, helps prevent overfitting by providing a more realistic estimate of the model's performance on unseen data. It adds a systematic evaluation component to the trial and error process, ensuring that the selected parameters lead to a model that performs well on a broader range of data.

**Table (5): Optimal parameters of the models**

**ANN**

| Parameter | optimal amount |
|---|---|
| Training cycle | 40 |
| Learning rate | 0.2 |
| Momentum | 0.9 |
| Decay | No |
| Shuffle | Yes |
| normalized | No |

**SVM**

| Parameter | optimal amount |
|---|---|
| Kernel Type | multi quadratic |
| Kernel Cache | 5 |
| C | 14.3 |
| Scale | No |

**Random forest**

| Parameter | optimal amount |
|---|---|
| Number of Trees | 21 |
| Apply Pruning | No |
| Apply pre pruning | No |
| Max depth | 35 |
| Criteria | Gain Ratio |

**Logistic regression**

| Parameter | optimal amount |
|---|---|
| Kernel Type | Dot |
| Kernel Cache | 5 |
| C | 50 |
| Scale | No |

**KNN**

| | |
|---|---|
| **K** | **1** |

### 2. Models Performance

After determining the optimal values for the model's hyper parameters, initially, the eight selected variables were individually applied to the models. Subsequently, the output of each model was assessed based on the criteria outlined in Table (4). The results of these evaluations are presented in Table (6)-A. Following this, patients' MRI and biopsy results were incorporated as additional variables alongside the previous inputs, and all models were re-executed. The outcomes are provided in Table (6)-B. It is evident that this step has led to improvements in all indicators across all models.

At this stage, an evaluation of the accuracy of the imaging and biopsy methods has been conducted based on the information recorded in patients' files regarding the results of mammography of the axillary lymph nodes at the beginning of the treatment and the outcomes of surgical lymph node removal. The results are presented in Table (7).

**Table (6): models Performance**

| | | | | | |
|---|---|---|---|---|---|
| **without mammography and biopsy results** | | | | | |
| **ANN** | **SVM** | **KNN** | **RF** | **LR** | **Indicator**<br>**Approach** |
| 70.31 | 56.32 | 68.75 | 64.71 | **71.11** | **Sensitivity** |
| 70.31 | 63.41 | 68.75 | 66.67 | 61.45 | **Specificity** |
| **70.38** | 58.4 | 68.85 | 65.51 | 64.81 | **Accuracy** |
| **considering mammography and biopsy results** | | | | | |
| 83.56 | **87.88** | 85.71 | 83.56 | 79.66 | **Sensitivity** |
| **94.55** | 90.32 | 93.10 | **94.55** | 75.36 | **Specificity** |
| 88.40 | **89.17** | **89.17** | 88.14 | 77.5 | **Accuracy** |

**Table (7): biopsy and imaging results**

| **Method** | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **Imaging** | 65% | 63% | 64% |
| **Biopsy** | 100% | 68% | 84% |

Based on the results shown in Table (6), In the first case, the highest accuracy belongs to the ANN model. But when the model reports the involvement, the LR model is more accurate. the SVM model is the least accurate one, which is only slightly more than 50%. Then after adding biopsy and MRI results to the models, evaluation indicators improved significantly in all models, while the difference between the two indicators Sensitivity and Specificity decreases. The SVM accuracy has improved the most and it reached from 58.4% to 89.17%, which is also the accuracy rate for KNN model. However, if the outcome of the model is involvement, the SVM model with a probability of 87.88% is more reliable. But if the model report is non -involvement, the result of the ANN and RF models are 94.55% which are the best performing. Compared to the information presented in Table (7), the SVM model is the only one exhibiting lower accuracy when contrasted with biopsy and imaging methods. Therefore, separately running all models with pathology variables has better results than running only biopsy or only imaging.

In the case of lymph nodes in breast cancer, as previously stated, if it is not involvement and is diagnosed with error, a vain surgery with complications such as lymphoma and limitation of hand movement is imposed on the patient. On the other hand, if the cancer has reached the lymph nodes, but this is not correctly diagnosed, the cancer may affect the patient's body with wider metastases and threaten her life. Therefore, the correct diagnosis of involvement is more important. Accordingly, the SVM model is recommended.

### 3. **Ensemble Method – Stacking**

In Table (8), the results of implementing the stacking approach and the different combinations of 5 machine learning models are presented, incorporating the results of biopsy and mammography as additional variables in the models. For example, in composition number 17, In the context of combining Support Vector Machines (SVM) and Artificial Neural Networks (ANN) using stacking, the process typically unfolds as follows:

1. Base models (SVM and ANN): Initially, separate SVM and ANN models are trained on the dataset to make predictions independently. Each model captures different aspects of the underlying patterns in the data.
2. Meta-model formation: The predictions of SVM and ANN models serve as input features for a meta-model. These predictions become new input features for the meta-model, which is often a simpler model such as LR (in this study) or another algorithm capable of combining diverse predictions.
3. Meta-model training: The meta-model is trained on the dataset using SVM and ANN predictions as input features. During this training, the meta-model learns how to weight and combine the predictions of the base models to optimize the overall performance.
4. Final Prediction: After the meta-model is trained, it can be used to make final predictions on new, unseen data.

**Table (8): Stacking implementation results for different modes of combining models (Test set)**

| NO | ANN | SVM | KNN | R.F | Logistic | Accuracy |
|---|---|---|---|---|---|---|
| 1 | * | * | * | * | * | **90.70** |
| 2 | * | * | * | * | | **88.40** |
| 3 | * | * | * | | * | **88.40** |
| 4 | * | * | | * | * | **89.39** |
| 5 | * | | * | * | * | **88.40** |
| 6 | | * | * | * | * | **88.33** |
| 7 | * | * | * | | | **87.63** |
| 8 | * | * | | * | | **89.74** |
| 9 | * | * | | | * | **90.71** |
| 10 | * | | * | * | | **88.27** |
| 11 | * | | * | | * | **88.40** |
| 12 | * | | | * | * | **91.35** |
| 13 | | * | * | * | | **88.27** |
| 14 | | * | * | | * | **89.87** |
| 15 | | * | | * | * | **88.46** |
| 16 | | | * | * | * | **88.40** |
| **17** | ***** | ***** | | | | **93.01** |

| | | | | | | |
|---|---|---|---|---|---|---|
| 18 | * | | * | | | **87.63** |
| 19 | * | | | * | | **90.62** |
| 20 | * | | | | * | **90.71** |
| 21 | | * | * | | | **89.06** |
| 22 | | * | | * | | **89.10** |
| 23 | | * | | | * | **88.46** |
| 24 | | | * | * | | **88.33** |
| 25 | | | * | | * | **89.87** |
| 26 | | | | * | * | **88.40** |

Based on Table (8), the highest accuracy (93.01 %) is related to the combination of the two models ANN and SVM. Therefore, the results of implementing this combination are in Table (9).

**Table (9): Stacking implementation results for the combination of ANN and SVM**

| Stacking (ANN+SVM) | Indicator |
|---|---|
| 91.04 | **Sensitivity** |
| 95.08 | **Specificity** |
| 93.01 | **Accuracy** |

Based on these results, if the disease involves lymph nodes, the model with 91.04% probability predicts this, correctly . On the other hand, if there is no lymph node involvement, the probability of correct prediction increases to 95.31%.

Therefore, it can be concluded that with modeling using machine learning algorithms, the prediction of lymph node involvement in breast cancer patients can be improved to an acceptable level. When, in addition to the demographic and pathological variables, the results of imaging and biopsy of the patient's breast in the initial visit are applied to the models as input variables, the integration of machine learning models with the accumulation approach, increases the accuracy of prediction. So, for example, in the case of non-involvement reporting, the probability of error is only 4.69%. Usually, for more efficient treatment, the physician determines the time intervals for check-ups, according to the conditions of each patient. In Figure (2), In each stage of the study, the prediction accuracy in reporting involvement and non-involvement is given in this study.

based on these results, it is suggested that if the biopsy results show axillary lymph node involvement, trust it completely. Otherwise, if the biopsy results show non-involvement, using a stacking approach for a combination of two models of ANN and SVM to predict it with the highest accuracy (95.31%) is recommended. It means if the model says that there is no involvement, it is right with a 95.08% probability. However, because there is still a 4.92 percent chance of error, regular check-up periods are necessary for the patient.

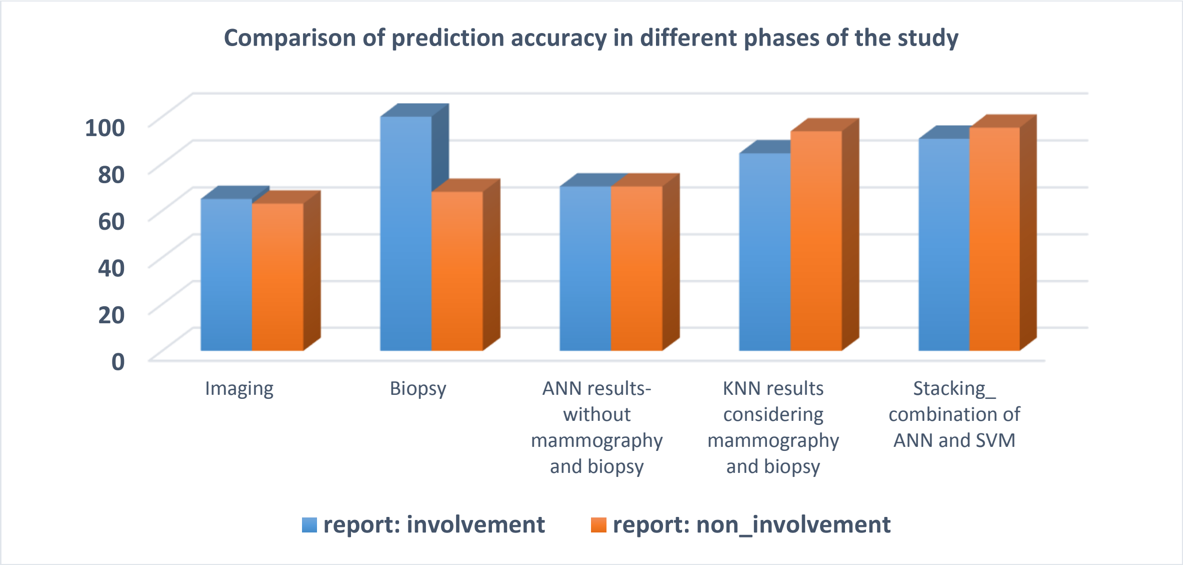

**Figure (2): Comparison of prediction accuracy in different phases of the study**

- ## Project implementation timeline

| Operation | Start time | End time |
|---|---|---|
| Primary research | 2019 November 19 , Tuesday | 2019 December 4 |
| Obtaining approval from the hospital and ethics code | 2019 December 4 | 2019 December 11 |
| Collecting data from hospital patient records | 2019 December 13 | 2020 June 1 |
| Data analysis and preparation | 2020 June 4 | 2020 June 17 |
| Study of similar research | 2020 June 18 | 2020 July 6 |
| Data preprocessing | 2020 July 6 | 2020 July 30 |
| Data processing | 2020 July 30 | 2020 August 7 |
| Modeling in Rapid Miner software:<br>Feature Selection<br>Splitting the data into training and testing sets<br>Training the models<br>Stacking the models<br>Model evaluation | 2020 August 7 | 2020 August 8 |

| | | |
|---|---|---|
| **Final testing** | | |
| **Model execution** | 2020 August 8 | 2020 August 27 |
| **Validation** | 2020 August 29 | 2020 August 27 |
| **Combining models with stacking** | 2020 September 1 | 2020 September 20 |
| **Result review** | 2020 September 22 | 2020 September 30 |
| **Writing the initial version** | 2020 October 1 | 2020 November 11 |
| **Editing the article** | 2020 November 11 | 2020 November 20 |

## Required tools and resources

1. Files of women with breast cancer at Omid Hospital in Mashhad
2. A powerful GPU-based processing system at Ferdowsi University of Mashhad to run models
3. Rapid Miner software
4. SPSS software

## Expected outputs

1. Open source Journal article available for every one
2. Provide results in a single file to the hospital
3. Find the best combination of models to predict lymph node involvement

## Importance and Applications

1. Medical Importance: Helps in more accurate diagnosis and reduces the need for invasive procedures
2. Research Importance: Provides a new approach in using the stacking approach
3. Practical Application: Possibility of using the model in the clinic and hospital

## Proposed Budget

| | |
|---|---|
| Human Resources | 60000000 Rials |
| Software Specialist | 20000000 Rials |
| Hardware Rental | 5000000 Rials |

- **Limitations and Advantages**

Our study has several limitations that need to be considered. First, this was a retrospective study, and the results were dependent on the composition of these limited-size data. Further improvement with larger and prospective studies must be achieved before actual clinical use. Secondly, lymph node metastasis and no metastasis were inherently unstable diagnoses in that their accuracy is dependent on the time of breast surgery. For example, some of the patients with negative lymph nodes, if followed up for a long enough time, may have eventually progressed to have positive lymph nodes. Finally, because our study was a single-center study, accessing data from several scattered centers or even cities is recommended in further studies to extrapolate the results more confidently to all individuals.

On the other hand, there are also some benefits: First, although, most studies have focused on improving imaging methods in predicting the involvement or non-involvement of lymph nodes and in recent studies, modeling has explicitly been done, and then the results have been compared with those obtained from biopsy and imaging, in this study the results of biopsy and MRI were applied directly to the models as two input variables along with other demographic and pathological variables in the models. This is a new process that has led to higher accuracy. The fact that Magnetic resonance imaging (MRI) is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body, and the biopsy uses the information of the living native cells and their physiology to check the condition of the cell, can justify the accuracy improvement. Second, combining all modes of five machine learning models and comparing their results is also a new scenario for predicting the status of lymph nodes.

- **<u>References</u>**

1. *Lazar, A.M., et al., Feasibility of Sentinel Lymph Node Biopsy in Breast Cancer Patients with Axillary Conversion after Neoadjuvant Chemotherapy&mdash;A Single-Tertiary Centre Experience and Review of the Literature. Diagnostics, 2023. **13**(18): p. 3000.*
2. *Tafreshi, N.K., et al., Molecular and functional imaging of breast cancer. Cancer Control, 2010. **17**(3): p. 143-55.*
3. *de Freitas, R., Jr., et al., Accuracy of ultrasound and clinical examination in the diagnosis of axillary lymph node metastases in breast cancer. Eur J Surg Oncol, 1991. **17**(3): p. 240-4.*
4. *Duan, Y., et al., Multimodal radiomics and nomogram-based prediction of axillary lymph node metastasis in breast cancer: An analysis considering optimal peritumoral region. J Clin Ultrasound, 2023. **51**(7): p. 1231-1241.*
5. *Fehm, T., et al., Prediction of axillary lymph node status of breast cancer patients by tumorbiological factors of the primary tumor. Strahlenther Onkol, 2005. **181**(9): p. 580-6.*
6. *Valente, S.A., et al., Accuracy of predicting axillary lymph node positivity by physical examination, mammography, ultrasonography, and magnetic resonance imaging. Ann Surg Oncol, 2012. **19**(6): p. 1825-30.*
7. *Yoo, S.H., et al., A histomorphologic predictive model for axillary lymph node metastasis in preoperative breast cancer core needle biopsy according to intrinsic subtypes. Hum Pathol, 2015. **46**(2): p. 246-54.*
8. *Ravdin, P.M., et al., Prediction of axillary lymph node status in breast cancer patients by use of prognostic indicators. J Natl Cancer Inst, 1994. **86**(23): p. 1771-5.*
9. *Moon, W.K., et al., Computer-aided prediction model for axillary lymph node metastasis in breast cancer using tumor morphological and textural features on ultrasound. Comput Methods Programs Biomed, 2018. **162**: p. 129-137.*

*10. Zhou, L.Q., et al., Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning. Radiology, 2020.* ***294****(1): p. 19-28.*

*11. Mojarad, S., et al., Prediction of nodal metastasis and prognosis of breast cancer by ANN-based assessment of tumour size and p53, Ki-67 and steroid receptor expression. Anticancer Res, 2013.* ***33****(9): p. 3925-33.*

*12. Cattell, R., et al., Preoperative prediction of lymph node metastasis using deep learning-based features. Vis Comput Ind Biomed Art, 2022.* ***5****(1): p. 8.*

*13. Park, S., et al., Application of Machine Learning Algorithm in Predicting Axillary Lymph Node Metastasis from Breast Cancer on Preoperative Chest CT. Diagnostics (Basel), 2023.* ***13****(18).*

*14. Ding, Y., et al., Multi-center study on predicting breast cancer lymph node status from core needle biopsy specimens using multi-modal and multi-instance deep learning. npj Breast Cancer, 2023.* ***9****(1): p. 58.*

*15. Chen, M., et al., Development and validation of convolutional neural network-based model to predict the risk of sentinel or non-sentinel lymph node metastasis in patients with breast cancer: a machine learning study. EClinicalMedicine, 2023.* ***63****: p. 102176.*

*16. Taherdoost, H., Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. International Journal of Academic Research in Management, 2016.* ***5****: p. 18-27.*

*17. Li, C., Little's Test of Missing Completely at Random. The Stata Journal, 2013.* ***13****(4): p. 795-809.*

*18. Molenberghs, G. and G. Verbeke, Multiple Imputation and the Expectation-Maximization Algorithm. 2005.*

*19. Chandrashekar, G. and F. Sahin, A survey on feature selection methods. Computers & Electrical Engineering, 2014.* ***40****(1): p. 16-28.*

*20. Grossi, E. and M. Buscema, Introduction to artificial neural networks. Eur J Gastroenterol Hepatol, 2007.* ***19****(12): p. 1046-54.*

*21. Evgeniou, T. and M. Pontil, Support vector machines: theory and applications, in Machine Learning and Its Applications: advanced lectures. 2001, Springer-Verlag. p. 249–257.*

*22. Guo, G., et al. KNN Model-Based Approach in Classification. in OTM Conferences / Workshops. 2003.*

*23. Ali, J., et al. Random Forests and Decision Trees. 2012.*

*24. Peng, J., K. Lee, and G. Ingersoll, An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES, 2002.* ***96****: p. 3-14.*

*25. Kwon, H., J. Park, and Y. Lee, Stacking Ensemble Technique for Classifying Breast Cancer. Healthc Inform Res, 2019.* ***25****(4): p. 283-288.*

*26. Brar, P., S. Jain, and I. Singh, Complications of Axillary Lymph Node Dissection in Treatment of Early Breast Cancer: A Comparison of MRM and BCS. Indian J Surg Oncol, 2011.* ***2****(2): p. 126-32.*