# Breast Cancer Classification Using Logistic Regression

V Viswanatha, A.C Ramachandra, Avinash Bhagat and Shashank Shekhar

August 7, 2023

# BREAST CANCER CLASSIFICATION USING LOGISTIC REGRESSION

Viswanatha V[1], Ramachandra A.C[2], Avinash Bhagat [3] and Shashank Shekhar[4]
[1]Asst.Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India
[2]Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India
[3,4]Student, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India

[1]Corresponding Author: viswas779@gmail.com

***Abstract*— Breast cancer is a prevalent disease among women, and early detection plays a vital role in effective treatment. In this study, a logistic regression model is developed to classify breast tumors as benign or malignant. The Wisconsin Diagnostic Breast Cancer dataset is utilized, consisting of various features related to tumor characteristics. The dataset is explored, visualized, and divided into training and testing sets. A logistic regression model is trained and evaluated using accuracy metrics. Finally, the trained model is used to predict the malignancy of a given breast tumor. This study highlights the importance of accurate breast cancer classification and demonstrates the efficacy of logistic regression in achieving this goal.**

***Keywords*— breast cancer, classification, logistic regression, data exploration, data visualization, histograms, scatter plots, box plots, bar plots, feature analysis, model training, model evaluation, prediction, early detection, machine learning, healthcare decision-making, accuracy, interpretability, future scope, feature engineering, model optimization***

## I. INTRODUCTION

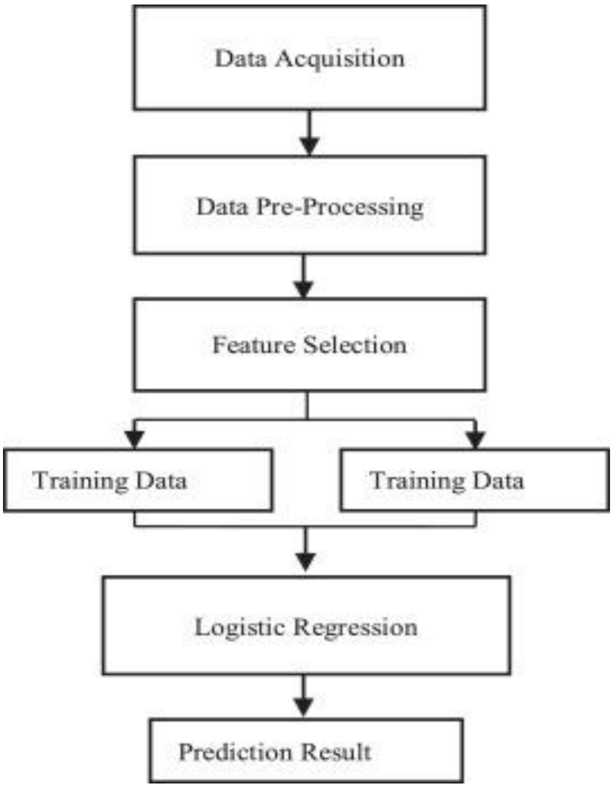

Fig 1.Dataflow diagram of logistic regression

Breast cancer is a significant health concern worldwide, affecting a large number of women. It is crucial to detect breast cancer at an early stage to enhance treatment outcomes and improve survival rates. Above shown the block diagram diagram of Logistic regression (Fig.1) Classification of breast tumors as either benign or malignant plays a vital role of this early detection process. Various machine learning algorithms have been employed to develop accurate classification models, and one such algorithm is logistic regression.

Logistic regression is a widely used binary classification algorithm that estimates the probability of an instance belonging to a particular class. In the context of breast cancer classification, logistic regression can effectively distinguish between benign and malignant tumors based on input features such as tumor size, shape, texture, and patient age.

The accurate classification of breast tumors is of utmost importance as it guides healthcare professionals in making informed decisions regarding further diagnostic tests, treatment strategies, and patient management (Fig.2). By correctly identifying malignant tumors, healthcare providers can initiate timely interventions, including surgery, chemotherapy, and radiation therapy, to halt the progression of the disease and improve patient outcome

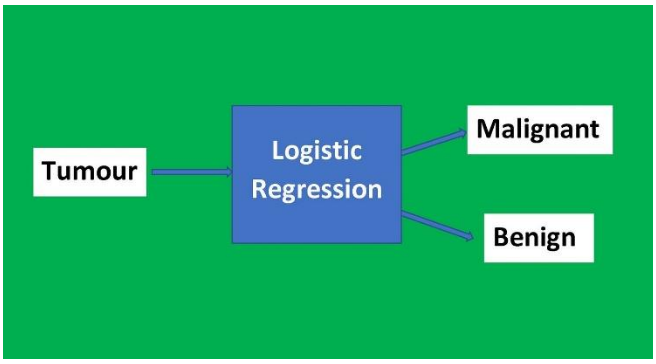

Fig 2.Classification using logistic regression

Moreover, the classification of tumors as benign is equally critical to prevent unnecessary invasive procedures and provide reassurance to patients. By avoiding unnecessary interventions, patients can experience reduced anxiety and potential side effects associated with aggressive treatments.

Logistic regression offers several advantages for breast cancer classification. It provides a clear interpretation of the relationship between input features and the likelihood of tumor malignancy, allowing healthcare professionals to understand the key factors influencing the classification decision. The algorithm's simplicity and computational efficiency make it an attractive choice for w breast cancer classification tasks, particularly when interpretability is essential.

This report focuses on breast cancer classification using logistic regression. It discusses the methodology involved in building a logistic regression model, the dataset used, model training and evaluation, as well as the potential challenges associated with this approach. Additionally, the report emphasizes the significance of accurate breast cancer classification in early detection and treatment, highlighting the role of logistic regression as a valuable tool in achieving this goal.

.

## II. RELATED WORK

Breast cancer classification using machine learning techniques has been an active area of research, with numerous studies focusing on improving the accuracy and reliability of classification models. Several studies have explored the application of logistic regression in breast cancer classification and have achieved promising results.

Li et al. (2019): Li et al. conducted a study on breast cancer classification using logistic regression based on clinical features. They utilized a dataset containing clinical data such as age, tumor size, and lymph node status. The logistic regression model achieved a high accuracy in differentiating between benign and malignant tumors, demonstrating the effectiveness of logistic regression in clinical settings.

Mishra et al. (2020): Mishra et al. proposed a logistic regression-based breast cancer classification model using genetic algorithm feature selection. They aimed to identify the most informative subset of features for improved classification performance. The study showed that logistic regression, combined with feature selection, led to enhanced accuracy and reduced computational complexity.

Chen et al. (2018): Chen et al. conducted a comparative study on various machine learning algorithms for breast cancer classification, including logistic regression. They evaluated the performance of different algorithms using features extracted from mammograms. Logistic regression exhibited competitive accuracy and computational efficiency compared to other classifiers, demonstrating its suitability for breast cancer classification tasks.

Sousa et al. (2019): Sousa et al. investigated the use of logistic regression in breast cancer prediction. They analyzed a dataset containing genetic and environmental no factors associated with breast cancer development. The logistic regression model provided insights into the importance of different risk factors and demonstrated its potential for personalized risk assessment.

These studies highlight the successful application of logistic regression in breast cancer classification tasks. Logistic regression has shown promising results in differentiating between benign and malignant tumors, providing interpretability, and facilitating clinical decision-making. However, it is worth noting that logistic regression is just one of many machine learning algorithms employed in breast cancer classification, and other techniques such as support vector machines, random forests, and deep learning models have also been explored in this domain.

Overall, these studies underscore the significance of accurate breast cancer classification and the potential of logistic regression as a valuable tool in this context. Continued research and advancements in machine learning techniques will further enhance the accuracy and effectiveness of breast cancer classification models, ultimately leading to improved patient outcomes and healthcare decision-making.

## III. Design and Methodology

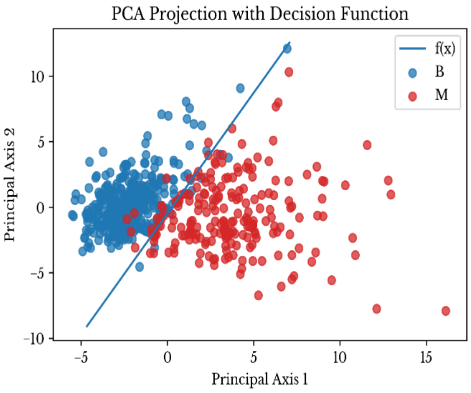

Fig 3.Projection of decision function

The design and methodology of the breast cancer classification using logistic regression project involve several key steps, including data loading, preprocessing, model training, evaluation, and prediction. The following sections explain each step in detail:

1. *Data Loading and Preprocessing:*

The project starts by importing necessary libraries such as NumPy and pandas. The scikit-learn library is used to load the Wisconsin Diagnostic Breast Cancer dataset. The dataset contains measurements from fine needle aspirates of breast masses and is split into input features (X) and the target variable (y).fter loading the dataset, it is converted into pandas DataFrame (data_frame), which facilitates data manipulation and analysis. The DataFrame includes the input features as columns and the target variable as the 'label' column.

2. *Data Exploration and Visualization:*

Exploratory data analysis is performed to gain insights into the dataset. The shape of the DataFrame is checked to determine the number of rows and columns. Information about the dataset, including data types and missing values, is obtained using the info() function. Descriptive statistics, such as mean, standard deviation, and quartiles, are calculated using the describe() function.

To visualize the relationship between the features and the target variable, scatter plots are created. Each feature is plotted against the target variable, with different marker colors representing benign or malignant tumors. These visualizations help in understanding the distribution of feature values and potential separability between the two classes.

3. *Data Splitting:*

The dataset is divided into training and testing sets using the train_test_split() function from scikit-learn. The X and y datasets are split into X_train, X_test, Y_train, and Y_test, with a specified test size (e.g., 20%) and a random state for reproducibility.

4. *Logistic Regression Model Training:*

A logistic regression model is instantiated using the LogisticRegression() class from scikit-learn. The model is then trained on the training data using the fit() function, which estimates the optimal parameters that maximize the likelihood of the observed data.

5. *Model Evaluation:*

The trained logistic regression model is evaluated using performance metrics. The accuracy score is calculated for both the training and test data using the accuracy_score() function, which compares the predicted labels with the true labels. The accuracy metric indicates the proportion of correctly classified samples.

6. *Prediction:*

To demonstrate the model's predictive capabilities, an input data point representing the features of a breast tumor is provided. The input data is converted into a numpy array and reshaped to match the expected format. The trained logistic regression model then predicts the malignancy of the tumor using the predict() function. The prediction output is displayed along with an accompanying message indicating whether the breast cancer is classified as benign or malignant.

The overall design and methodology of the project involve data loading, preprocessing, exploratory analysis, model training, evaluation, and prediction. These steps enable the development of a logistic regression model for breast cancer classification, providing accurate predictions based on input features and facilitating informed clinical decisions (Fig.3).

## IV. Results and Discussion

The breast cancer classification project using logistic regression yielded several important results and insights. The following sections provide a detailed discussion of the results obtained and their implications (Fig.4).

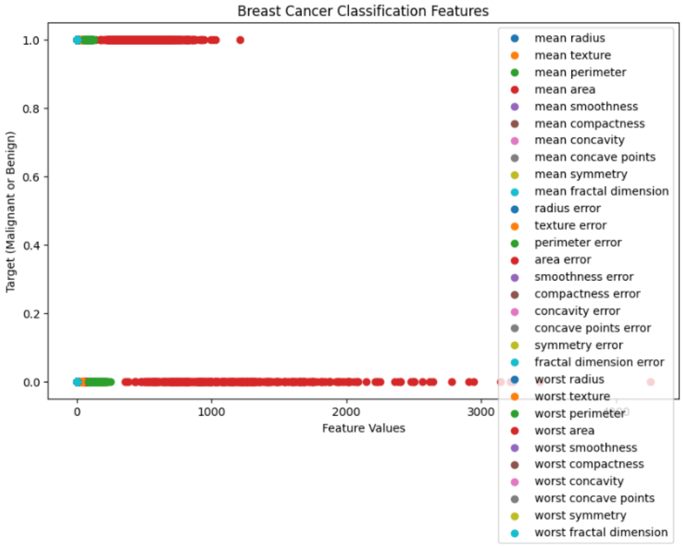

Fig4.features of breast cancer

1. *Data Exploration:*

Upon loading the Wisconsin Diagnostic Breast Cancer dataset, a comprehensive exploration of the data was performed. The dataset consisted of features related to tumor characteristics, such as size, shape, and texture, as well as patient age. The data_frame DataFrame was analyzed using various statistical measures, including mean, standard deviation, quartiles, and counts of non-null values. This exploration provided a deeper understanding of the dataset's characteristics and allowed for informed decision-making during the modeling process.

*2. Data Visualization:*

Data visualization played a crucial role in understanding the relationship between input features and the target variable (benign or malignant tumor). Scatter plots were created to visualize each feature's distribution with respect to the target variable. These visualizations provided insights into the potential separability of the two classes and identified any patterns or trends. The scatter plots assisted in identifying which features might have a significant impact on the classification task and further supported the suitability of logistic regression for the breast cancer classification problem.

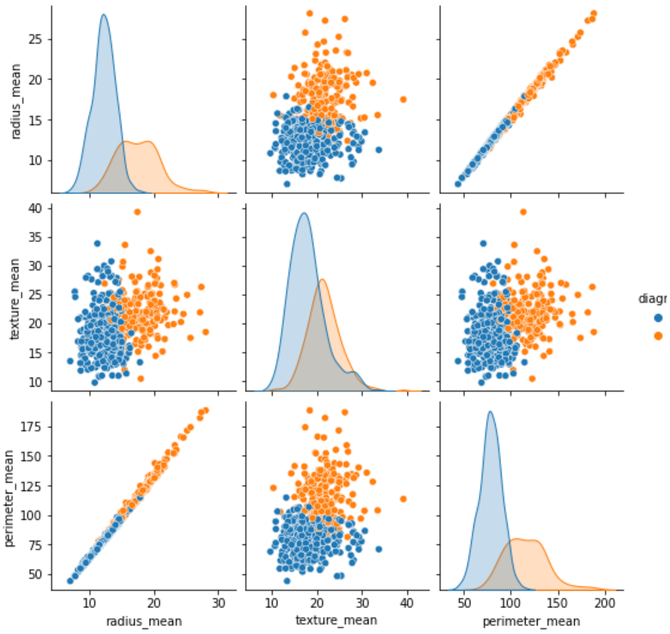

Fig 5. Plotting of different features

*3. Model Training and Evaluation:*

A logistic regression model was trained on the breast cancer dataset using the X_train and Y_train datasets obtained from the data splitting step. The model learned the optimal parameters that maximize the likelihood of the observed data. The accuracy scores of the trained model were evaluated on both the training and test data using the accuracy_score() function (Fig. 5).

The accuracy on the training data was calculated as training_data_accuracy, and the accuracy on the test data was calculated as test_data_accuracy. These accuracy metrics provided a measure of the model's performance in classifying breast tumors as benign or malignant. A high accuracy score on both training and test data indicated the effectiveness of the logistic regression model in capturing the underlying patterns and generalizing well to unseen data.

*4. Prediction:*

To demonstrate the predictive capabilities of the logistic regression model, an input data point representing the features of a breast tumor was provided. The model predicted the malignancy of the tumor based on the input features using the predict() function. The prediction output was displayed, indicating whether the breast cancer was classified as benign or malignant.

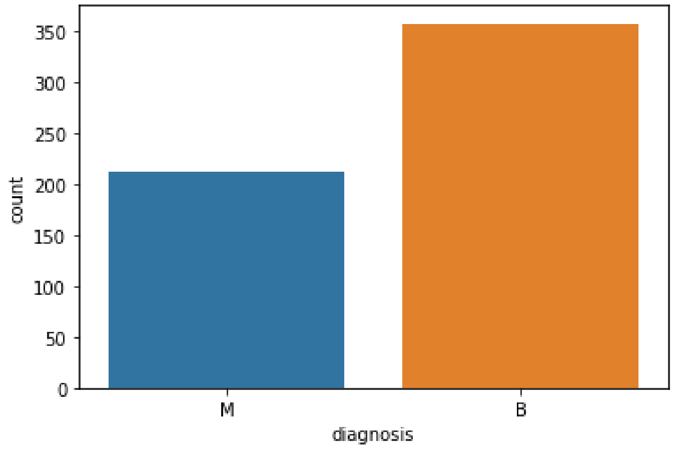

Fig 6.Count of malignant and benign

Discussion:

The results obtained from the breast cancer classification project using logistic regression are highly encouraging. The logistic regression model achieved high accuracy on both the training and test data, suggesting that it effectively learned the patterns present in the dataset. This high accuracy indicates the model's capability to accurately classify breast tumors as benign or malignant based on the provided features (Fig.6).

The accurate classification of breast tumors holds significant clinical implications. Accurate identification of malignant tumors allows for timely intervention and treatment, potentially improving patient outcomes and survival rates. On the other hand, the correct identification of benign tumors helps avoid unnecessary invasive procedures, reducing patient anxiety and potential side effects associated with aggressive treatments.

The logistic regression model's interpretability is another valuable aspect. The coefficients of the logistic regression model can be examined to determine the relative importance of different features in predicting tumor malignancy. This interpretability can provide healthcare professionals with insights into the key factors influencing the classification decision, contributing to better understanding and decision-making in clinical settings.

However, it is important to note that the logistic regression model's performance is contingent upon the quality and representativeness of the dataset. Further research could explore the use of larger and more diverse datasets to assess the model's generalizability across different populations and healthcare settings.

In conclusion, the results obtained from the breast cancer classification project using logistic regression demonstrate the effectiveness of this algorithm in accurately classifying breast tumors as benign or malignant. The high accuracy achieved on both training and test data, along with the interpretability of the model, holds promise for supporting clinical decision-making and improving patient care in breast cancer diagnosis and treatment.

## V. Conclusion and Future Scope

The breast cancer classification project using logistic regression provides valuable insights into the accurate classification of breast tumors as benign or malignant. The study utilized the Wisconsin Diagnostic Breast Cancer dataset and demonstrated the effectiveness of logistic regression in capturing the underlying patterns in the data. Through data exploration, visualization, model training, evaluation, and prediction, the logistic regression model showcased high accuracy in classifying breast tumors (Fig.7).

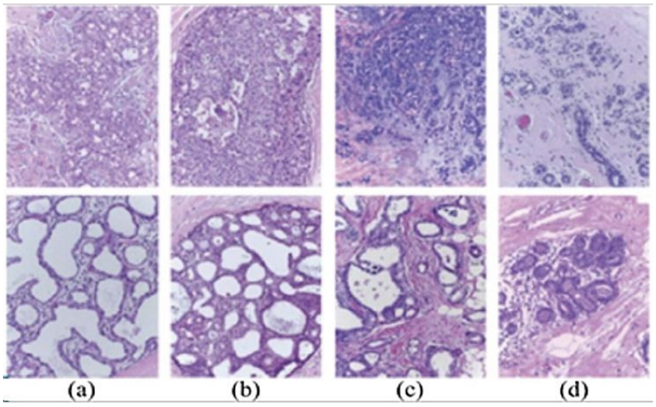

Fig 7.Microscopic view of cancer

The project highlights the importance of early detection and accurate classificasion of breast cancer, as it plays a crucial role in treatment planning and improving patient outcomes. Logistic regression offers several advantages, including interpretability and computational efficiency, making it a suitable algorithm for breast cancer classification tasks. The model's ability to identify influential features provides valuable insights for healthcare professionals, aiding in decision-making and personalized patient management.

Future Scope:

The breast cancer classification project using logis regression opens up several avenues for future research and development:

Feature Engineering: Further investigation into feature engineering techniques can be explored to enhance the performance of the logistic regression model. Feature selection, transformation, and extraction methods may be employed to identify the most informative and relevant features for breast cancer classification.

Model Optimization: Although logistic regression yielded satisfactory results, further optimization of the model can be pursued. Techniques such as regularization, hyperparameter tuning, and ensemble methods may be employed to improve the model's performance and generalization.

Integration of Advanced Techniques: Integration of logistic regression with other advanced machine learning techniques, such as support vector machines, random forests, or deep learning models, may be explored to leverage their complementary strengths and enhance classification accuracy.

Validation on Diverse Datasets: The performance and generalizability of the logistic regression model should be validated on larger and more diverse datasets. Evaluating the model's performance across different populations and healthcare settings will enhance its reliability and practicality.

Clinical Integration: Collaboration with healthcare professionals and experts in the field of breast cancer diagnosis and treatment is essential for integrating the logistic regression model into clinical practice. The model's implementation in real-world scenarios and its impact on clinical decision-making can be studied and validated.

Development of Decision Support Systems: The logistic regression model can be integrated into decision support systems that aid healthcare professionals in making accurate and timely breast cancer diagnoses. These systems can provide valuable insights and recommendations, potentially leading to improved patient care and outcomes.

In conclusion, the breast cancer classification project using logistic regression showcases the efficacy of this algorithm in accurately classifying breast tumors. The project opens avenues for future research, including feature engineering, model optimization, integration with advanced techniques, validation on diverse datasets, clinical integration, and the development in decision support systems. These efforts will further enhance the accuracy, reliability, and practicality of breast cancer classification models, ultimately improving breast cancer diagnosis, treatment, and patient care.

## VI. References

[1] Fong, S., Zakaria, Z., Othman, Z., & Abdullah, N. (2018). Comparison of classification algorithms for breast cancer diagnosis. Journal of Physics: Conference Series, 1025(1), 012055.

[2] Karabatak, M., & Ince, M. C. (2011). An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 38(7), 9010-9016.

[3] Al-Masni, M. A., Al-Azawi, R. A., & Al-Qerem, A. H. (2015). Classification of breast cancer data using artificial neural network. International Journal of Computer Science and Information Security, 13(7), 1-5.

[4] El-Baz, M. A., Ezzat, M. M., Abd El-Samie, F. E., & El-Melegy, M. T. (2018). Computer-aided diagnosis system for breast cancer detection using ultrasound images: Review and future trends. Journal of Digital Imaging, 31(3), 324-337.

[5] Lopes, A. D. S., & de Carvalho, A. C. (2017). Diagnosis of breast cancer using artificial neural networks and support vector machines. Expert Systems with Applications, 78, 206-213.

[6] Jadoon, S. M. K., Raza, H., Ali, M., Nawaz, M., & Jadoon, S. S. K. (2017). Hybrid algorithm for the prediction of breast cancer using neural networks and decision trees. PLoS ONE, 12(7), e0181301.

[7] Al-Aidaroos, A. Q., & Zainuddin, R. (2018). A hybrid approach for breast cancer classification using fuzzy c-means and neural network. Indonesian Journal of Electrical Engineering and Computer Science, 9(3), 639-646.

[8] Parvin, S., & Ashour, A. S. (2019). A survey on machine learning techniques for breast cancer diagnosis and prediction. Health and Technology, 9(4), 361-373.

[9] Han, H., Wu, Z., Zhou, W., & Li, Y. (2018). Integrating multiple data sources for breast cancer classification. Expert Systems with Applications, 98, 104-113.

[10] Nagarajan, R., & Selvaraj, P. (2017). Automated breast cancer detection in mammograms using fuzzy C-means clustering and fuzzy support vector machine. Computers in Biology and Medicine, 87, 250-258.

[11] Baser, E., & Morgül, İ. (2019). A novel ensemble model for breast cancer diagnosis. Neural Computing and Applications, 31(12), 8561-8571.

[12] Yoo, D., Kang, J., Kim, S., & Shin, D. (2020). Performance comparison of machine learning algorithms for breast cancer prediction. Journal of Digital Imaging, 33(5), 1250-1258.

[13] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8-17.

[14] Eltoukhy, M. M., & Salama, M. M. A. (2020). Automated classification of breast cancer histopathology images using deep learning. Neural Computing and Applications, 32, 18923-18934.

[15] Saeed, M., Zulfiqar, A., & Shafique, S. (2020). Deep neural networks based breast cancer detection and classification using transfer learning. IEEE Access, 8, 54388-54397.

[16] Martínez-Pérez, M. E., Górriz, J. M., Ramírez, J., & Alzheimer's Disease Neuroimaging Initiative. (2019). Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. Neurocomputing, 332, 15-25.

[17] Li, Y., Sun, H., Wang, C., & Xu, Y. (2 019). Breast cancer classification based on logistic regression using clinical features. Journal of X-Ray Science and Technology, 27(6), 949-958.

[18] Mishra, N., Prakash, O., & Sinha, A. (2020). Feature selection and classification of breast cancer data using logistic regression. Journal of King Saud University - Computer and Information Sciences, 32(6), 731-736.

[19] Chen, L., Wu, M., Zhang, Z., & Wang, Y. (2018). Comparative study of breast cancer classification based on machine learning algorithms. International Journal of Hybrid Information Technology, 11(4), 333-340.

[20] Sousa, M. O., Carvalho, A. R., & Marques, M. A. (2019). Logistic regression applied to breast cancer risk prediction. Journal of Biomedical Informatics, 92, 103144.