Journal of Physical Sciences: Theories, Methods and Applications
Volume 1, Issue 1 (2026), pp. 79–87
ISSN: xxxx-xxxx (print), xxxx-xxxxx (online)
https://jpstma.org
https://doi.org/10.xxxx/example.2026.001

Research article

Evaluating the effect of feature scaling on L2-regularized logistic regression for breast cancer classification

Oluwaseun Biodun Onuoha[1,*]

[1] *Department of Mathematical Sciences, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria*

Email addresses
[1] oluwaseun.akinduko@aaua.edu.ng

ABSTRACT

Feature scaling is a common preprocessing technique in machine learning but its effect depends on the scaling strategy and the type of data used. In this paper, a performance test of L2-regularized Logistic Regression was carried out on breast cancer classification, based on three feature-scaling conditions, namely, no scaling, Min–Max scaling, and Z-score standardization. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, consisting of 569 observations and 30 predictors was used. By using five-fold cross-validation, all the three conditions were examined based on the same model specification, the same folds and the same evaluation criteria. In order to prevent leakage of data, only the training portion of each fold was used to evaluate the scaling parameters, and later applied to its validation part. The performance of the model was evaluated based on accuracy, sensitivity, specificity, F1-score, receiver operating characteristic area under the curve (ROC-AUC), convergence, and computational measures. The findings of this research showed that Z-score standardization recorded the highest mean accuracy of 0.9737, sensitivity of 0.9436, specificity of 0.9916, F1-score of 0.9638, and ROC-AUC of 0.9937. Min–Max scaling recorded the highest mean specificity of 0.9972, the lowest mean sensitivity of 0.8635 and F1-score of 0.9242, and the least computational cost. Across the five folds, all the three scaling conditions achieved a 100% convergence rate. The findings of this study also revealed that different feature-scaling conditions can perform differently when applied to the same L2-regularized Logistic Regression model, though this is applicable only to the WDBC dataset and the experimental configuration under consideration.

Keywords: Feature scaling; L2 regularization; Logistic Regression; breast cancer classification; Min–Max scaling; Z-score standardization

Mathematics Subject Classification: 62J12; 62J07; 62H30

ARTICLE HISTORY

Submitted
31 October 2026

Accepted
9 November 2026

Published online
14 November 2026

CITE THIS ARTICLE

O. B. Onuoha. Evaluating the effect of feature scaling on L2-regularized logistic regression for breast cancer classification. *J. Phys. Sci., Theor. Methods Appl.*, 1(1), 2026, 79–87.

[*] **Corresponding Author:** oluwaseun.akinduko@aaua.edu.ng

© 2026 The Author(s). Published by the Journal of Physical Sciences: Theories, Methods and Applications (JPSTMA). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

Breast cancer is a significant health problem worldwide and correct classification of the tumour is crucial for diagnosis. The diagnosis of breast lesions as malignant or benign is an integral part of diagnosis. Thus, machine learning methods have been widely investigated for breast cancer classification. Among these approaches, logistic regression (LR) remains an appealing model for binary clinical prediction because of its relatively simple formulation, computational efficiency, and interpretability. Evidence from clinical prediction studies suggests that more complex machine learning techniques do not necessarily outperform LR when valid comparisons are made [1].

The numerical scales of predictor variables can be important when fitting classification models. Common feature-scaling methods include Min–Max scaling and Z-score standardization. Min–Max scaling maps each predictor to a specified range, whereas Z-score standardization centers a predictor at its mean and scales it by its standard deviation. Scaling can be important when regularization is used since the regularization penalty acts on the model coefficients. For LR, there is a well-established form of regularization, namely, ridge estimation [2]. Thus, changing the scaling condition while keeping the classifier and regularization parameter fixed provides a meaningful way to investigate whether preprocessing decisions are associated with differences in model performance.

Previous studies have shown that scaling does not necessarily have a consistent effect across different datasets and classifiers. Cao et al. [3] compared scaling methods in biomedical classification tasks, including the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, with LR among the classifiers considered. They reported that classification accuracy could vary across scaling conditions. Henderi et al. [4] compared Min–Max and Z-score normalization based on k-nearest neighbors on WDBC and found that Min–Max normalization recorded higher accuracy.

In breast cancer classification, previous studies have shown the importance of feature scaling because different classifiers like the k-nearest neighbors and LR can have different performances, depending on the scales of the predictors. But different datasets, models, preprocessing methods and experimental setups have been applied to these studies. For example, the Red Wine Quality dataset was used to investigate normalization techniques with LR in Adeyemo et al. [5] and LR was also found to be effective on the WDBC dataset by other studies [6, 7]. In broader machine learning pipelines, other preprocessing techniques, like Min–Max scaling and Z-score standardization have also been used for breast cancer research [8, 9]. Thus, it is hard to determine the impact of different scaling techniques on logistic regression with the WDBC breast cancer dataset. In this study, the LR model and WDBC dataset are not changed, and the direct comparison of no scaling, Min–Max scaling and Z-score standardization is performed to address this gap.

The significance of feature scaling in machine learning applications has been restated in recent studies. Pinheiro et al. [10] studied the impact of 12 feature-scaling methods on 14 datasets and 16 machine learning algorithms, and found that the impact of feature scaling differed depending on the machine learning algorithm, with Logistic Regression being one of the algorithms that was sensitive to the feature scaling method. In the same way, Watrianthos et al. [11] used Z-score standardization in a Logistic Regression model for breast cancer classification on the WDBC dataset and added threshold optimization and probability calibration to their method. While these studies offer valuable insights into the impact of feature scaling on machine learning and breast cancer classification, they do not directly compare no scaling, Min–Max scaling and Z-score standardization under the same experimental setup and in the same L2-regularized Logistic Regression framework. This provides further motivation for the present study.

In terms of optimisation, the difference between Z-score standardization and Min–Max scaling could affect the numerical properties of L2-regularized LR. If the predictor variables are on very different scales, the contributions of the variables to the gradient and Hessian of the objective function can be very different, which can impact the conditioning of the optimisation problem. Z-score standardization is a method for scaling predictors to have a mean of zero and a variance of one, which puts them on a more comparable scale and may make them more numerically conditioned. Min–Max scaling, however, is based on the minimum and maximum values observed, and is sensitive to extreme observations that can cause the range of most observations to be compressed. Outliers also affect the mean and standard

deviation of the data, but they have a different effect on the Z-score standardization. Moreover, the L2 penalty directly affects the model coefficients, and changes in the scales of the predictors can affect the effective regularization applied to the model coefficients when the same penalty parameter is used. The above considerations give a theoretical justification for studying if Z-score standardization and Min–Max scaling give different predictive and computational performance in L2-regularized LR.

The WDBC dataset contains digitized images of fine-needle aspirates of breast masses. The dataset has 569 observations and 30 real-valued predictors. It can be used for the comparison of scaling techniques in a controlled manner. The data is available from the UCI Machine Learning Repository [12] and its nuclear feature information are provided by Street et al. [13]. Earlier research has examined scaling and normalization either separately or as part of a larger LR analyses. However, no study has compared the three scaling conditions (no scaling, Min–Max scaling, and Z-score standardization) under the same experimental condition and for the same L2-regularized LR model. Consequently, the aim of this paper is not to discuss normalization itself, but to answer the methodological question.

The WDBC dataset from the breast cancer classification problem is used throughout this paper to analyse the effect of three preprocessing conditions (no scaling, Min–Max scaling and Z-score standardization). The three conditions were evaluated with an identical classifier, regularization parameter, data partitions, and evaluation criteria. The research used stratified five-fold cross-validation for the main analysis to avoid leakage of data. Thus, each training fold was used to evaluate the scaling parameters and then applied to the corresponding validation fold. Some commonly used performance measures are accuracy, sensitivity, specificity, F1-score and receiver operating characteristic area under the curve (ROC-AUC). The aim of this study, therefore, is to provide an empirical comparison of the most commonly used scaling options, rather than proposing a new classifier or a more complex preprocessing method.

2 Materials and Methods

2.1 Dataset

This study used the WDBC dataset from the UCI Machine Learning Repository [12], comprising 569 observations and 30 real-valued predictors derived from digitized fine-needle aspirates (FNA) of breast masses [13]. There were 212 malignant and 357 benign cases. Benign cases were assigned 0, and malignant cases were assigned 1. No feature construction, dimensionality reduction, feature selection, or resampling was applied, and the same 30 predictors were retained for all experimental conditions.

2.2 Logistic Regression and Scaling Conditions

The classifier used was binary logistic regression. For observation $\mathbf{a}_i$, where $\mathbf{a}_i$ denotes the vector of predictor values for the ith observation, the probability of the positive class is defined as

$$p_i = \frac{1}{1 + \exp[-(\gamma_0 + \mathbf{a}_i^T \boldsymbol{\gamma})]}, \tag{1}$$

In the above equation, $\boldsymbol{\gamma}$ represents the vector of feature coefficients and γ_0 represents the intercept. To estimate the model parameters, the L2-regularized logistic loss given by

$$J(\gamma_0, \boldsymbol{\gamma}) = -\frac{1}{n} \sum_{i=1}^{n} [z_i \log(p_i) + (1 - z_i) \log(1 - p_i)] + \frac{\alpha}{2} \|\boldsymbol{\gamma}\|_2^2, \tag{2}$$

was minimized, where n is the number of observations, $z_i \in \{0, 1\}$ is the class label, and $\alpha = 0.01$ is the regularization parameter. The L2 penalty was not applied to the intercept γ_0. The L2 formulation is consistent with the standard ridge approach to LR [2]. Initial values for the model parameters were set to zero, and the parameters were estimated using a deterministic damped Newton method with Armijo backtracking. The maximum number of iterations was 100, and convergence was defined as the infinity norm of the gradient being less than or equal to 10^{-8}. The backtracking factor and Armijo

parameter were set to 0.5 and 10^{-4}, respectively. All scaling conditions were evaluated using the same model specification, regularization parameter, optimization settings, and convergence criteria.

The observations were categorized as positive when the predicted probability was ≥ 0.5 and as negative when the predicted probability was < 0.5. The three preprocessing conditions were compared. The first used the original predictors (no scaling). The second applied Min–Max scaling,

$$a_{ij}^{*} = \frac{a_{ij} - a_{j,\min}}{a_{j,\max} - a_{j,\min}}, \tag{3}$$

and the third used Z-score standardization,

$$a_{ij}^{*} = \frac{a_{ij} - \mu_j}{\sigma_j}. \tag{4}$$

Here, a_{ij} denotes the original value of the jth predictor for the ith observation, and a_{ij}^{*} denotes the corresponding scaled value. The quantities $a_{j,\min}$ and $a_{j,\max}$ are the minimum and maximum values of the jth predictor, respectively, while μ_j and σ_j denote its mean and standard deviation. Only the training data is used to estimate the scaling quantities for each fold and thereafter applied without change to the corresponding validation data.

2.3 Stratified Five-Fold Cross-Validation

To evaluate performance, stratified five-fold cross-validation was used. Across the five folds, the class ratio of malignant to benign was about the same, with stratification. Four folds were used for training and the fifth used for validation for each iteration. All three scaling conditions were applied with the same fold assignments for a fair comparison.

The scaling parameters were estimated in each training fold to avoid leakage of data. No data was added to the calculation of the minimum, maximum, mean, and standard deviation used for preprocessing in the validation fold. A fixed random seed of 20260817 was used to ensure reproducibility. There was no addition of oversampling, feature selection or other classifiers.

2.4 Evaluation Metrics and Experimental Protocol

The main evaluation metrics are accuracy, sensitivity, specificity, F1 score and ROC-AUC. The percentage of observations that are correctly classified is the accuracy, and the percentage of malignant and benign cases that are correctly identified is the sensitivity and specificity, respectively. F1-score gives an alternative measure of positive-class classification performance. The assessment of discrimination across classification thresholds was performed using ROC-AUC [14].

Predictions and performance measures were recorded for each fold and scaling condition. The main results are presented as the mean and standard deviation of the five fold-level values. To obtain aggregate confusion-matrix-based performance measures, pooled out-of-fold predictions were also retained. All experiments were implemented in MATLAB using the same LR implementation and experimental settings for the three conditions. Thus, the main experimental variable was feature scaling.

3 Results and Discussion

3.1 Five-Fold Cross-Validation Results

All the main results were obtained using stratified five-fold cross-validation, and the same folds and LR specification were used for all three scaling conditions. Table 1 shows the mean and standard deviation of the fold-level performance measures.

Table 1: Five-fold cross-validation results for the three scaling conditions; ROC-AUC denotes receiver operating characteristic area under the curve.

Condition	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
No scaling	0.9510 ± 0.0240	0.9204 ± 0.0670	0.9692 ± 0.0117	0.9324 ± 0.0348	0.9897 ± 0.0085
Min–Max	0.9474 ± 0.0160	0.8635 ± 0.0336	0.9972 ± 0.0062	0.9242 ± 0.0239	0.9895 ± 0.0077
Z-score	$\mathbf{0.9737 \pm 0.0163}$	$\mathbf{0.9436 \pm 0.0354}$	0.9916 ± 0.0076	$\mathbf{0.9638 \pm 0.0225}$	$\mathbf{0.9937 \pm 0.0056}$

Z-score standardization produced the highest mean accuracy (0.9737 ± 0.0163), sensitivity (0.9436 ± 0.0354), F1-score (0.9638 ± 0.0225), and ROC-AUC (0.9937 ± 0.0056). The corresponding values for the no-scaling condition were 0.9510, 0.9204, 0.9324, and 0.9897, respectively, which were slightly lower. Min–Max scaling achieved the lowest mean accuracy (0.9474), sensitivity (0.8635) and the highest mean specificity (0.9972), but the lowest mean F1-score (0.9242). Thus, the Min–Max condition showed high specificity for benign observations but lower sensitivity in identifying malignant observations.

The mean performance measures for the three conditions are compactly compared in Figure 1.

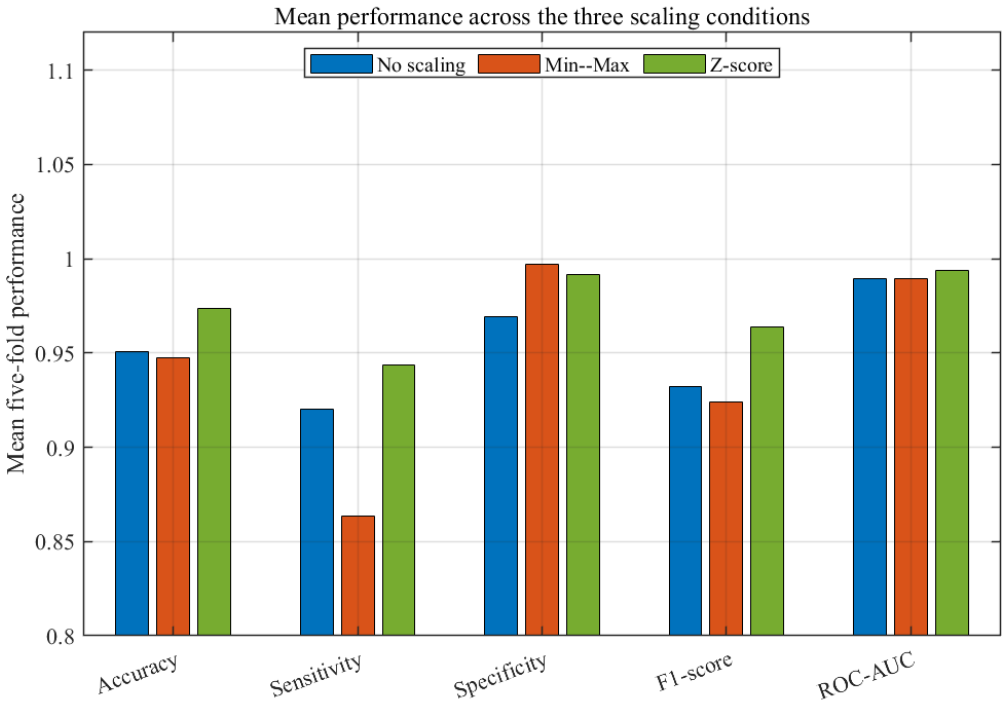


Figure 1: Mean five-fold performance across the three scaling conditions.

The fold-level results also showed variability across the five validation folds. The accuracy with no scaling ranged from 0.9217 to 0.9735. The accuracy with Min–Max scaling ranged from 0.9217 to 0.9646. The accuracy with Z-score standardization ranged from 0.9565 to 1.0000. Correspondingly, sensitivity ranged from 0.8372 to 1.0000 for no scaling, from 0.8140 to 0.9048 for Min–Max scaling, and from 0.9070 to 1.0000 for Z-score. These ranges signify that there are differences in the observed performance across the five validation folds instead of getting report from just one validation fold, respectively.

3.2 Pooled Out-of-Fold Performance

The predictions from the five validation folds were combined to get pooled out-of-fold predictions. There were 346 true negatives, 195 true positives, 11 false positives, and 17 false negatives, for no-scaling condition, leading to an accuracy of 0.9508, a specificity of 0.9692, a sensitivity of 0.9198, an ROC-AUC of 0.9889, and an F1-score of 0.9330.

In the pooled confusion matrix, there were 356 true negatives, 183 true positives, 1 false positive and 29 false negatives for Min–Max scaling. The accuracy was 0.9473, with a specificity of 0.9972, sensitivity of 0.8632, ROC-AUC of 0.9899, and F1-score of 0.9242.

For the Z-score condition, there were 200 true positives, 354 true negatives, 3 false positives, and 12 false negatives. The pooled accuracy was 0.9736, with a sensitivity of 0.9434, specificity of 0.9916, F1-score of 0.9639, and ROC-AUC of 0.9934. Overall, the pooled results showed the best predictive performance for Z-score standardization.

3.3 Convergence and Computational Performance

All five folds converged under all three scaling conditions, giving a convergence rate of 100% for each condition. The mean numbers of iterations were 10.2 for no scaling, 6.0 for Min–Max scaling, and 9.0 for Z-score standardization. The average CPU times for training were 0.0171 s, 0.0037 s, and 0.0054 s, respectively.

In this implementation, the lower number of iterations and shorter execution time for Min–Max scaling indicate lower computational cost. The choice of scaling should not be based on computational speed alone. Min–Max scaling had lower sensitivity and F1-score than both no scaling and Z-score standardization, so it did not provide the best overall predictive performance. Z-score standardization achieved the highest classification accuracy, while all three scaling conditions had a convergence rate of 100%.

3.4 Discussion

The results indicate that feature scaling had an impact on the performance of the L2-regularized LR model on the WDBC dataset, but the impact was not uniform across all evaluation measures. The highest mean accuracy, sensitivity, F1-score and ROC-AUC were achieved by z-score standardization, and the highest specificity and the lowest computational cost were achieved by Min–Max scaling. These results suggest that the scaling strategy can affect the trade-off between various components of classification performance, even with the same classifier and regularization parameter.

The observed performance pattern with Z-score standardization is in line with the literature, which indicates that feature scaling can influence the performance of classifiers, especially when the numerical ranges of the predictor variables are significantly different. Cao et al. [3] demonstrated that scaling might impact the classification accuracy of Logistic Regression and Support Vector Machine models in biomedical datasets, such as the WDBC dataset. Their study, however, took into account several datasets, scaling methods and classifiers. The present study makes a more specific comparison between the WDBC dataset and an L2-regularized LR model.

The findings of the present study can also be compared with the WDBC study of Henderi et al. [4] which found that the k-nearest neighbours algorithm achieved better accuracy with Min–Max normalization than with Z-score normalization. The difference in the results observed could be attributed to the classifiers used in the two studies. K-nearest neighbours is distance-based, and so depends directly on the distances between observations, while LR is based on a linear decision function with estimated coefficients. The comparison indicates that the impact of a preprocessing method could be different for different classification methods.

Adeyemo et al. [5] also explored normalization techniques for LR, but with a different dataset. They have a similar example of the effect of normalization on LR performance. In breast cancer classification, scaling has also been applied as a part of larger preprocessing steps [8, 9]. In the present study, the scaling condition is varied while keeping the other experimental conditions fixed, which is different from these investigations.

From a diagnostic point of view, sensitivity is of particular interest because a false-negative classification is a case where a malignant case is misclassified as benign. The present experiment showed that Min–Max scaling had higher specificity and lower sensitivity than the other two conditions. This means that the higher specificity was associated with a higher percentage of missed malignant cases. The results highlight the need to take both sensitivity and specificity into account when evaluating the

performance of classification, especially in applications where the failure to classify a malignant case is a major concern.

The computational results revealed differences in the efficiency of the three scaling conditions. The mean number of iterations and the average CPU time were the lowest for Min–Max scaling, and all three conditions converged 100%. The above observations indicate that the Min–Max scaling needed less computational effort under the experimental conditions studied. However, in selecting a scaling strategy, computational efficiency must be taken into account along with predictive performance.

There are some limitations to this study. First, only one benchmark dataset and one classifier were used in the analysis, so the results cannot be extrapolated to other breast cancer datasets, clinical populations or classification algorithms without further investigation. Second, only three scaling conditions were explored, and other scaling methods, nonlinear transformations and feature-selection methods were not considered. In addition, the regularization parameter was set to $\alpha = 0.01$ in all the experiments and the sensitivity of the results to other values of the regularization parameter was not explored. Therefore, the impact of feature scaling on the observed effects might be partly due to the regularization parameter used, and future research should investigate this relationship with a variety of regularization strengths. Lastly, the WDBC dataset is not from a clinical study, and the results are not intended to be used as evidence of clinical efficacy.

The findings support the importance of considering feature scaling when applying L2-regularized LR to the WDBC dataset. This study is an empirical comparison of the relationship between predictive and computational performance of different preprocessing choices, while keeping the classifier, predictors, validation folds and regularization parameter fixed and varying the scaling condition. The study does not propose a new classification algorithm, but rather provides evidence of the effect of feature scaling on the performance of a well-defined LR model.

4 Conclusion

In this study, the Wisconsin Diagnostic Breast Cancer dataset was used to investigate the impact of feature scaling on the predictive and computational performance of L2-regularized logistic regression. The three conditions were: No scaling, Min–Max scaling and Z-score standardization. The results showed that the Z-score standardization had the highest mean accuracy, sensitivity, F1-score and ROC-AUC, while the Min–Max scaling had the highest mean specificity and the lowest computational cost. The convergence rate was 100% for all three conditions.

The results indicate that the scaling strategy may affect the performance of logistic regression in various aspects, even if the classifier, regularization parameter and experimental conditions are fixed. In particular, the results suggest that different scaling approaches could be more effective in terms of predictive performance and computational efficiency, depending on the measure used.

This study could be expanded to other breast cancer datasets, classification algorithms, and scaling techniques in future studies. It would also be interesting to explore the interaction between feature scaling and various regularisation strengths and optimisation methods. These studies may yield a more general understanding of the influence of preprocessing decisions on regularized classification models.

Declarations

Funding

The author received no specific funding for this study.

Data availability

This study used publicly available data from the UCI Machine Learning Repository at https://doi.org/10.24432/C5DW2B.

Ethics approval

Not applicable.

Competing interests

The author has no conflicts of interest to declare regarding this paper.

Author contributions

O.B.O. was responsible for conceptualization, methodology, software, validation, formal analysis, and writing the original draft. O.B.O. reviewed and approved the manuscript.

References

[1] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models*, J. Clin. Epidemiol., **110** (2019), 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[2] S. le Cessie and J. C. van Houwelingen, *Ridge estimators in logistic regression*, J. R. Stat. Soc. Ser. C (Appl. Stat.), **41**(1) (1992), 191–201. https://doi.org/10.2307/2347628

[3] X. H. Cao, I. Stojkovic, and Z. Obradovic, *A robust data scaling algorithm to improve classification accuracies in biomedical data*, BMC Bioinformatics, **17** (2016), 359. https://doi.org/10.1186/s12859-016-1236-x

[4] Henderi, T. Wahyuningsih, and E. Rahwanto, *Comparison of Min-Max normalization and Z-Score normalization in the K-nearest neighbor (kNN) algorithm to test the accuracy of types of breast cancer*, Int. J. Informatics Inf. Syst., **4**(1) (2021), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

[5] A. Adeyemo, H. Wimmer, and L. M. Powell, *Effects of normalization techniques on logistic regression in data science*, J. Inf. Syst. Appl. Res., **12**(2) (2019), 37–44.

[6] R. Entezari, *Breast cancer diagnosis via classification algorithms*, arXiv:1807.01334 (2018). https://doi.org/10.48550/arXiv.1807.01334

[7] W. Cheng and Z. Yu, *A fast and interpretable logistic regression framework for breast tumor classification using the Wisconsin Diagnostic Dataset*, medRxiv (2025). https://doi.org/10.64898/2025.12.23.25342946

[8] M. Kumar and V. Bhardwaj, *Evaluating label encoding and preprocessing techniques for breast cancer prediction using machine learning algorithms*, Int. J. Comput. Intell. Syst., **18** (2025), 218. https://doi.org/10.1007/s44196-025-00957-7

[9] M. E. Yildirim and Y. B. Salman, *Breast cancer classification using z-score thresholding and machine learning*, J. Electron. Electromed. Eng. Med. Inform., **7**(4) (2025), 1355–1366. https://doi.org/10.35882/jeeemi.v7i4.1165

[10] J. M. H. Pinheiro, S. V. B. de Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. de Souza, R. V. Godoy, L. A. Ambrosio, and M. Becker, *The impact of feature scaling in machine learning: Effects on regression and classification tasks*, IEEE Access, **13** (2025), 199903–199931. https://doi.org/10.1109/ACCESS.2025.3635541

[11] R. Watrianthos, Rayendra, E. Asri, Yuhefizar, and Humaira, *Threshold-optimized and calibrated logistic regression for breast cancer classification*, Salud, Ciencia y Tecnología, **5**(2241)(2025). https://doi.org/10.56294/saludcyt20252241

[12] W. Wolberg, O. Mangasarian, N. Street, and W. Street, *Breast Cancer Wisconsin (Diagnostic) [Dataset]*, UCI Machine Learning Repository (1993). https://doi.org/10.24432/C5DW2B

[13] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, *Nuclear feature extraction for breast tumor diagnosis*, in *Biomedical Image Processing and Biomedical Visualization*, vol. 1905, SPIE, Bellingham, WA, USA (1993), 861–870. https://doi.org/10.1117/12.148698

[14] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognit. Lett., **27**(8) (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010