https://doi.org/10.51885/3134-8025_JICS_2026_2_3
SRSTI 28.23.15

BREAST CANCER DETECTION USING LOGISTIC REGRESSION: A COMPARATIVE STUDY OF OPTIMIZATION AND FINE-TUNING METHODS

ЛОГИСТИКАЛЫҚ РЕГРЕССИЯНЫ ҚОЛДАНА ОТЫРЫП, СҮТ БЕЗІ ОБЫРЫН АНЫҚТАУ: ОҢТАЙЛАНДЫРУ ЖӘНЕ ДӘЛ БАПТАУ ӘДІСТЕРІН САЛЫСТЫРМАЛЫ ЗЕРТТЕУ

ВЫЯВЛЕНИЕ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ С ПОМОЩЬЮ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ: СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МЕТОДОВ ОПТИМИЗАЦИИ И ТОНКОЙ НАСТРОЙКИ

A.B. Aryn [1*], **O.Zh. Mamyrbaev** [2]
[1]Caspian University, Almaty, Kazakhstan
[2]Institute of Information and Computing Technologies, Almaty, Kazakhstan
*Corresponding author: Aray Aryn, e-mail: arun_arai@mail.ru

Keywords:
breast cancer, logistic regression, hyperparameter tuning, fine-tuning, optimization methods, machine learning.

ABSTRACT
Breast cancer is one of the most prevalent and life-threatening diseases worldwide, making accurate early diagnosis critical for improving patient outcomes. This study investigates the effectiveness of logistic regression models for breast cancer detection using the Wisconsin Breast Cancer (Diagnostic) Dataset. Various optimization techniques, including Ordinary Least Squares, Gradient Descent, Newton's Method, L2 Regularization, Stochastic Gradient Descent (SGD), and GridSearch-based fine-tuning, were applied to determine their impact on model performance. Standard evaluation metrics—Accuracy, Precision, Recall, and F1-score—were used to compare methods. The results demonstrate that GridSearch-based fine-tuning consistently yields the highest overall performance, highlighting the importance of systematic hyperparameter optimization in enhancing model robustness and predictive accuracy. These findings emphasize that even classical models like logistic regression can achieve state-of-the-art results in medical data analysis when combined with appropriate optimization strategies.

Түйінді сөздер:
сүт безінің қатерлі ісігі, логистикалық регрессия, гиперпараметрлерді баптау, дәл баптау, оңтайландыру әдістері, машиналық оқыту.

ТҮЙІНДЕМЕ
Сүт безі қатерлі ісігі бүкіл әлемде ең көп таралған және өмірге қауіп төндіретін аурулардың бірі болып табылады, бұл пациенттердің нәтижелерін жақсарту үшін дәл ерте диагнозды өте маңызды етеді. Бұл зерттеу Висконсиндегі Сүт безі Қатерлі ісігінің Диагностикалық Деректер Жинағын (WBCD) пайдалана отырып, сүт безі обырын анықтау үшін логистикалық регрессия үлгілерінің тиімділігін зерттейді. Оңтайландырудың әртүрлі әдістері, соның ішінде қарапайым ең кіші квадраттар әдісі, градиенттік түсу, Ньютон әдісі, L2- регуляризациясы, стохастикалық градиенттің түсуі (SGD) және

олардың модель өнімділігіне әсерін анықтау үшін GridSearch-ге негізделген дәл баптау қолданылды. Әдістерді салыстыру үшін бағалаудың стандартты көрсеткіштері – дәлдік, толықтық, сезімталдылық және F1-көрсеткіші қолданылды. Нәтижелер GridSearch-ге негізделген дәл баптау модельдердің сенімділігі мен болжау дәлдігін арттыру үшін гиперпараметрлерді жүйелі оңтайландырудың маңыздылығын көрсете отырып, ең жоғары жалпы өнімділікті дәйекті түрде қамтамасыз ететінін көрсетеді. Бұл нәтижелер логистикалық регрессия сияқты классикалық модельдерде тиісті оңтайландыру стратегияларымен үйлескенде медициналық деректерді талдауда заманауи нәтижелерге қол жеткізе алатынын көрсетеді.

Ключевые слова:
рак молочной железы, логистическая регрессия, настройка гиперпараметров, тонкая настройка, методы оптимизации, машинное обучение.

АННОТАЦИЯ

Рак молочной железы является одним из наиболее распространенных и опасных для жизни заболеваний во всем мире, что делает точную раннюю диагностику критически важной для улучшения результатов лечения пациентов. В этом исследовании изучается эффективность моделей логистической регрессии для выявления рака молочной железы с использованием набора данных по диагностике рака молочной железы штата Висконсин (WBCD). Для определения их влияния на производительность модели были применены различные методы оптимизации, включая обычные методы наименьших квадратов, градиентный спуск, метод Ньютона, регуляризацию L2, стохастический градиентный спуск (SGD) и тонкую настройку на основе GridSearch. Для сравнения методов использовались стандартные показатели оценки – точность, полнота, чувствительность и показатель F1. Результаты демонстрируют, что точная настройка на основе GridSearch неизменно обеспечивает высочайшую общую производительность, подчеркивая важность систематической оптимизации гиперпараметров для повышения надежности модели и точности прогнозирования. Эти результаты подчеркивают, что даже классические модели, такие как логистическая регрессия, могут обеспечить самые современные результаты анализа медицинских данных в сочетании с соответствующими стратегиями оптимизации.

INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, necessitating accurate and timely diagnostic tools (Ghasemi et al., 2024; Дёмин & Гермашев, 2024). Traditional machine learning approaches, particularly logistic regression, have been widely employed for binary classification tasks due to their simplicity, interpretability, and efficiency (Aryn et al., 2025; Watrianthos et al., 2025). However, the predictive performance of logistic regression models can be limited if parameters are not carefully optimized or thresholds are not properly calibrated (Benghazouani et al., 2025; Won et al., 2025). Watrianthos et al. (2025) proposed a threshold-optimized and calibrated logistic regression approach, demonstrating that systematic adjustment of model parameters and decision thresholds can significantly improve classification accuracy in breast cancer detection. This approach underscores the potential benefits of combining classical statistical models with fine-tuning techniques to enhance predictive reliability in clinical applications (Watrianthos et al., 2025).

Recent advances in hyperparameter optimization (HPO) have highlighted the critical role of efficient tuning in improving machine learning model performance (Won et al., 2025; Ahmad et al., 2026). Won et al. (2025) provide a comprehensive review of multi-fidelity HPO, emphasizing approaches that leverage low-fidelity evaluations to reduce computational costs while maintaining effective model optimization. Their study categorizes existing HPO methods according to underlying strategies, discusses the benefits of considering multiple fidelity levels, and identifies practical challenges that remain unaddressed. These insights underscore the value of systematic hyperparameter tuning in machine learning models, particularly when optimizing performance metrics such as accuracy, precision, recall, and F1-score (Won et al., 2025; Benghazouani et al., 2025; Ahmad et al., 2026).

Benghazouani et al. (2025) investigated the enhancement of logistic regression models through hybrid nature-inspired optimization techniques, combining genetic algorithms and swarm intelligence to optimize model weights and improve feature selection. Their proposed method, EGASI-RL, demonstrated high predictive performance on breast and lung cancer datasets, achieving substantial improvements in accuracy, precision, recall, and F1-score compared to standard logistic regression. The study highlights the potential of integrating global search and adaptive optimization strategies to overcome limitations of weight estimation in high-dimensional or noisy datasets. These findings reinforce the importance of systematic optimization of logistic regression parameters and feature selection, providing a conceptual foundation for applying fine-tuning techniques such as GridSearch and regularization in predictive modeling tasks.

Ghasemi et al. (2024) conducted a systematic scoping review on the application of explainable artificial intelligence (XAI) in breast cancer detection and risk prediction. Their study emphasizes the importance of interpretability and transparency in AI models used in medical settings, highlighting techniques such as SHapley Additive exPlanations (SHAP) for explaining predictions of complex models. While their focus is primarily on tree-based ensemble models, the findings underscore a broader principle relevant to any predictive modeling, including logistic regression: optimizing model performance through careful tuning of parameters and weights should be complemented by strategies that maintain model interpretability and trustworthiness. This perspective supports the rationale for systematically evaluating different optimization methods, regularization, and fine-tuning approaches to improve both predictive accuracy and reliability in breast cancer classification tasks.

Demin and Germashev (2024) review the application of machine learning techniques in breast cancer diagnosis, highlighting their utility in supporting diagnostic tasks, patient state assessment, and risk prediction. The study emphasizes that classical machine learning methods, such as support vector machines, random forests, and convolutional neural networks, have been effectively employed for tasks ranging from image segmentation in mammography to predictive modeling based on clinical and genomic data. The review demonstrates that these approaches can improve diagnostic accuracy, efficiency, and reliability, and underscores the importance of selecting appropriate features and optimizing model parameters. Such findings provide a foundation for systematically applying logistic regression and exploring optimization techniques, including weight tuning, regularization, and fine-tuning, to enhance predictive performance in breast cancer detection tasks.

Cristea et al. (2025) investigated optimized logistic regression models for analyzing complex gastroenterological images, comparing their performance against state-of-the-art approaches. The study emphasizes that careful tuning of model parameters can significantly enhance predictive accuracy, even when using relatively simple models like logistic regression, highlighting the continued relevance of classical methods when optimized properly. Similarly, Khan et al. (2026) explored hybrid machine learning techniques for breast cancer prediction, combining multiple algorithms and optimization strategies to maximize model performance across various metrics. Both studies converge on the insight that systematic parameter optimization and hybridization strategies can substantially improve the robustness and accuracy of predictive models in medical applications, demonstrating the importance of fine-tuning, feature selection, and integration of complementary methods in high-stakes diagnostic tasks.

Recent studies have demonstrated that systematic hyperparameter tuning can substantially enhance the performance of logistic regression models across diverse datasets (Ahmad et al., 2026). By carefully adjusting parameters such as regularization strength, solver type, and iteration limits, predictive accuracy and robustness can be significantly improved. Inspired by these findings, this study explores the application of logistic regression for breast cancer detection, incorporating multiple optimization techniques and fine-tuning strategies to maximize model performance and reliability (Ahmad et al., 2026). This study provides a comprehensive evaluation of logistic regression models for breast cancer diagnosis, highlighting the benefits of systematic optimization and fine-tuning (Watrianthos et al., 2025; Benghazouani et al., 2025; Ahmad et al., 2026). The findings aim to inform future applications of logistic regression in medical data analysis, demonstrating how classical models can achieve high predictive performance when coupled with appropriate optimization strategies (Cristea et al., 2025; Khan et al., 2026).

MATERIALS AND METHODS

Dataset Description

The Breast Cancer Wisconsin (Diagnostic) Dataset serves as a standard benchmark in medical machine learning research. It comprises 569 samples derived from fine needle aspiration (FNA) of breast masses. Due to its structured feature space and moderate dimensionality, the dataset is highly suitable for the evaluation of linear and probabilistic classifiers, including logistic regression. Furthermore, the clinical interpretability of its features facilitates a rigorous comparative analysis of optimization strategies and regularization techniques.

The structural attributes of the dataset, encompassing 30 clinical features alongside identification and diagnostic labels, are detailed in Table 1. This summary includes the non-null count and the data representation format for each attribute, ensuring data integrity for subsequent preprocessing.

Table 1. Detailed dataset attributes

№	Column Name	Data Type
0	id	int64
1	diagnosis	object
2	radius_mean	float64
3	texture_mean	float64
4	perimeter_mean	float64
5	area_mean	float64
6	smoothness_mean	float64
7	compactness_mean	float64
8	concavity_mean	float64
9	concave points_mean	float64
10	symmetry_mean	float64
11	fractal_dimension_mean	float64
12	radius_se	float64
13	texture_se	float64
14	perimeter_se	float64
15	area_se	float64
16	smoothness_se	float64
17	compactness_se	float64
18	concavity_se	float64

End of Table 1

№	Column Name	Data Type
19	concave points_se	float64
20	symmetry_se	float64
21	fractal_dimension_se	float64
22	radius_worst	float64
23	texture_worst	float64
24	perimeter_worst	float64
25	area_worst	float64
26	smoothness_worst	float64
27	compactness_worst	float64
28	concavity_worst	float64
29	concave points_worst	float64
30	symmetry_worst	float64
31	fractal_dimension_worst	float64
32	Unnamed: 32	float64

The dataset defines a binary classification task where samples are categorized as either malignant (M) or benign (B). The feature space comprises 30 real-valued numerical attributes representing the morphological characteristics of cell nuclei derived from digitized images. These attributes—including radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension—are quantified using three statistical measures: the mean, standard error (SE), and the "worst" (mean of the three largest values) measurement.

During the preprocessing phase, the unique identifier and an auxiliary null column were excluded from the feature set, as they provide no predictive utility. To ensure compatibility with the learning algorithms, the categorical target labels were numerically encoded.

The class distribution exhibits a moderate imbalance, as illustrated in the sample breakdown: 357 observations (62.7%) are classified as benign, while 212 observations (37.3%) are malignant. Although the dataset is not perfectly balanced, the representation of the minority class is statistically sufficient to support robust model training and evaluation without the requirement for synthetic resampling or data augmentation techniques.

Data Preprocessing and Feature Standardization

Prior to model training, the dataset was divided into feature variables and a target variable. All numerical attributes were assigned to the feature matrix X, while the diagnostic outcome (diagnosis) was selected as the target vector y.

The target variable was encoded into numerical format to represent a binary classification problem, where one class corresponds to benign tumors and the other to malignant tumors [9].

The dataset was then split into training and testing subsets in order to evaluate model generalization performance on unseen data. The training set was used for model fitting, while the test set was reserved exclusively for performance evaluation.

Feature standardization was applied to ensure that all input variables contribute equally to the learning process. Since the dataset contains features with different physical scales and magnitudes, standardization is particularly important for optimization-based models such as logistic regression, which rely on gradient-based and second-order optimization methods.

Standardization was performed using z-score normalization, where each feature was transformed by subtracting its mean and dividing by its standard deviation, computed from the training data:

$$X_{std} = \frac{X - \mu}{\sigma}$$

where μ and σ denote the mean and standard deviation of each feature in the training set, respectively. The same transformation parameters were applied to the test set to prevent data leakage.

Models and optimization methods

In this study, a baseline logistic regression model was employed as the primary classification algorithm for breast cancer detection. Logistic regression is a widely used probabilistic model for binary classification tasks and serves as a strong baseline due to its simplicity, interpretability, and solid theoretical foundation.

The model estimates the probability of class membership by applying a sigmoid function to a linear combination of input features. During training, the model learns a set of weight parameters that maximize the likelihood of observing the given class labels.

In the baseline configuration, logistic regression was applied without explicit regularization or hyperparameter tuning, allowing the model to learn weights solely based on the training data. The standardized feature set was used as input to ensure numerical stability and efficient convergence during optimization.

The logistic regression model was trained using the standardized training dataset. After model fitting, predictions were generated for the test dataset to assess classification performance on previously unseen data.

The performance of the logistic regression model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics were computed on the test dataset to assess the predictive performance of the model.

To investigate the effect of different optimization strategies on logistic regression performance, several methods for estimating model weights were applied and evaluated. Each method was trained on the standardized training dataset, and predictions were generated for the test dataset. Model performance was assessed using the same evaluation metrics for consistency.

Ordinary Least Squares (OLS)

As an initial optimization approach, the Ordinary Least Squares (OLS) method was applied to estimate model parameters. Although OLS is traditionally used for linear regression, it was employed in this study as a baseline optimization technique to examine its behavior in a binary classification setting.

After training, predictions were obtained on the test dataset, and classification performance metrics were computed.

Gradient Descent

The Gradient Descent optimization method was used to iteratively update the model weights by minimizing the loss function with respect to the training data. This approach relies on the first-order derivative of the loss function and updates parameters in the direction of the negative gradient.

The optimized model was then evaluated on the test dataset using the selected performance metrics.

Newton Method

The Newton Method was applied as a second-order optimization technique that utilizes both first and second derivatives of the loss function. By incorporating curvature information through the Hessian matrix, this method aims to achieve faster convergence compared to first-order methods.

Model predictions were generated for the test set, and performance metrics were subsequently calculated.

Regularization (L2)

To control model complexity and reduce the risk of overfitting, L2 regularization was applied to the logistic regression model. This technique introduces a penalty term proportional to the squared magnitude of the model weights into the optimization objective.

The resulting predictions were evaluated using the same set of classification metrics.

Stochastic Gradient Descent (SGD)

In addition to batch optimization methods, Stochastic Gradient Descent (SGD) was employed as an alternative optimization strategy. Unlike standard gradient descent, SGD updates model parameters using individual samples or small batches, which can improve computational efficiency and scalability for larger datasets.

Fine-tuning

In the final stage of the study, fine-tuning was applied to the logistic regression model in order to improve its predictive performance. Fine-tuning refers to the systematic optimization of model hyperparameters that are not directly learned during training but significantly influence model behavior and generalization ability.

In this work, fine-tuning was performed using GridSearchCV, which conducts an exhaustive search over a predefined grid of hyperparameter values combined with cross-validation.

The following hyperparameters of the logistic regression model were tuned:

- Regularization strength (C): values ranging from 0.01 to 100 were evaluated. This parameter controls the trade-off between fitting the training data and model complexity. Smaller values of C correspond to stronger regularization, while larger values allow the model to fit the data more closely.

- Optimization solver: three solvers (lbfgs, newton-cg, and saga) were considered. These solvers implement different numerical optimization strategies for estimating model parameters and may affect convergence behavior and performance.

- Penalty type: L2 regularization was selected to ensure numerical stability and compatibility with all chosen solvers.

A 5-fold cross-validation strategy was employed during the grid search to ensure robust estimation of model performance across different training subsets. The F1-score was used as the optimization criterion, as it provides a balanced measure of precision and recall and is well suited for binary classification problems with moderate class imbalance.

RESULTS AND DISCUSSION

This section presents the experimental results obtained using logistic regression models and different optimization methods.

To evaluate and compare the performance of the models, the following classification metrics were used: Accuracy, Precision, Recall, and F1-score.

The results obtained for logistic regression models trained using various weight optimization techniques are summarized in Table 2. These techniques include the method of least squares (OLS), gradient descent (GD), Newton's method, L2 regularization, stochastic gradient descent (SGD), as well as the baseline logistic regression model.

The OLS and gradient descent–based models demonstrated identical performance, achieving an accuracy of 0.9649 and an F1-score of 0.95. Newton's method resulted in lower classification performance, with an accuracy of 0.9035 and an F1-score of 0.8642.

The logistic regression model with L2 regularization achieved an accuracy of 0.9649, a precision of 0.975, and an F1-score of 0.9512, matching the performance of the baseline logistic regression model. The stochastic gradient descent–based model showed comparable results, with an accuracy of 0.9561 and an F1-score of 0.9398.

The quantitative results for all optimization methods are summarized in Table 2.

For better visualization, Figure 1 presents a graphical comparison of the optimization methods based on the selected evaluation metrics.

Table 2. Performance comparison of logistic regression models using different weight optimization methods

index	Accuracy	Precision	Recall	F1-score
OLS	0.9649122807017544	1.0	0.9047619047619048	0.95
GD	0.9649122807017544	1.0	0.9047619047619048	0.95
Newton	0.9035087719298246	0.8974358974358975	0.8333333333333334	0.8641975308641975
L2 Regularization	0.9649122807017544	0.975	0.9285714285714286	0.9512195121951219
SGD	0.956140350877193	0.9512195121951219	0.9285714285714286	0.9397590361445783
Logistic Regression (baseline)	0.9649122807017544	0.975	0.9285714285714286	0.9512195121951219

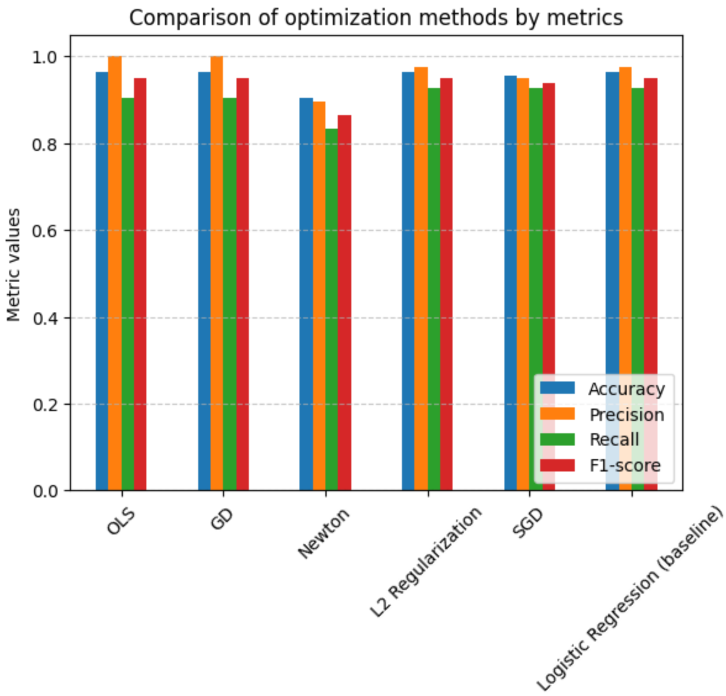

Figure 1. A visual comparison of the evaluated optimization methods across the selected performance metrics

Note – compiled by the authors

To improve the performance of the logistic regression model, a fine-tuning procedure was applied using grid search with cross-validation. The optimization was performed by varying key hyperparameters of the logistic regression model, including the regularization strength, optimization solver, and regularization type.

The base model was defined as follows:

maximum number of iterations: max_iter = 1000

random seed: random_state = 42

The hyperparameter search space included:

C (inverse regularization strength): {0.01, 0.1, 1, 10, 100}

solver: {lbfgs, newton-cg, saga}

penalty: L2 regularization

A 5-fold cross-validation strategy was employed, and the F1-score was selected as the optimization criterion. The grid search was conducted on the standardized training dataset, and the best hyperparameter configuration was selected based on the cross-validated performance.

After identifying the optimal hyperparameters, the tuned model was used to generate predictions on the test dataset. Figure 2 shows the performance metrics of the baseline and GridSearch-optimized logistic regression models.

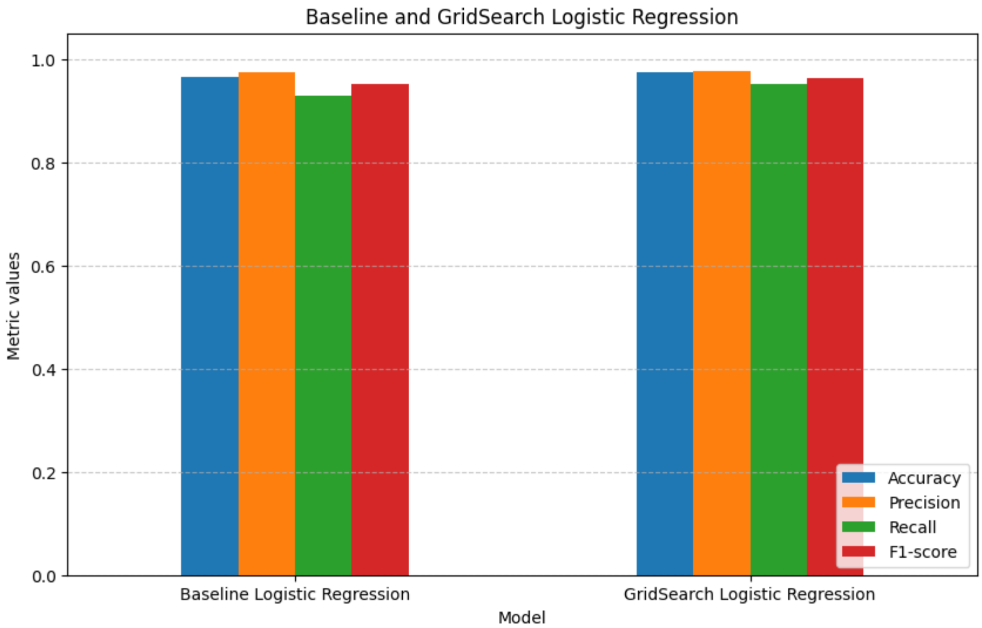

Figure 2. Comparison of baseline and fine-tuned logistic regression models
Note – compiled by the authors

To provide a rigorous comparative assessment of the implemented optimization paradigms, the performance metrics for each methodology were aggregated and analyzed. The experimental framework evaluated the following optimization approaches: Ordinary Least Squares (OLS), Gradient Descent (GD), Newton's Method, Baseline Logistic Regression (LR), LR with L2 Regularization, Stochastic Gradient Descent (SGD), and Fine-tuned LR via GridSearchCV.The comparative performance results are tabulated in Table 3, which details the Accuracy, Precision, Recall, and F1-score achieved by each algorithm on the sequestered test dataset. These metrics provide a multidimensional view of the classifiers' ability to generalize to unseen clinical data.

Table 3. Performance comparison of optimization methods for breast cancer detection

Methods	Accuracy	Precision	Recall	F1-score
OLS	0.9649122807017544	1.0	0.9047619047619048	0.95
Gradient Descent	0.9649122807017544	1.0	0.9047619047619048	0.95
Newton Method	0.9035087719298246	0.8974358974358975	0.8333333333333334	0.8641975308641975
Logistic Regression (Baseline)	0.9649122807017544	0.975	0.9285714285714286	0.9512195121951219
Logistic Regression + L2	0.9649122807017544	0.975	0.9285714285714286	0.9512195121951219
Logistic Regression + SGD	0.956140350877193	0.9512195121951219	0.9285714285714286	0.9397590361445783

End of Table 1

Methods	Accuracy	Precision	Recall	F1-score
Logistic Regression + GridSearch	0.9736842105263158	0.975609756097561	0.9523809523809523	0.963855421686747
Note – compiled by the authors				

Figure 3 illustrates a comparative visualization of the evaluation metrics obtained for all optimization methods considered in this study.

The chart enables a clear comparison of model performance across different optimization strategies using Accuracy, Precision, Recall, and F1-score.

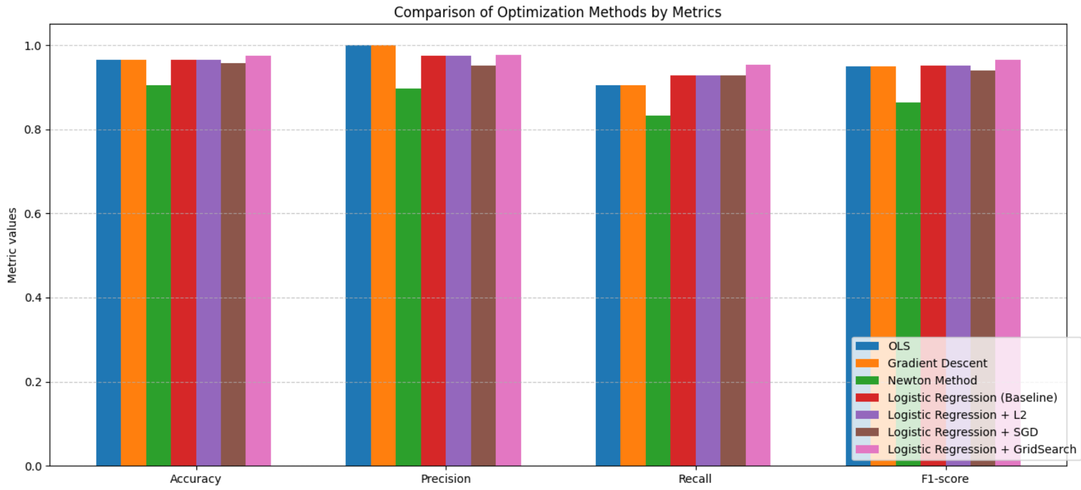

Figure 3. Comparison of Optimization Methods for Breast Cancer Detection

Note – compiled by the authors

Fine-tuning via GridSearch yielded the best overall performance, achieving the highest values across all evaluation metrics (Accuracy = 0.9737, Recall = 0.9524, F1-score = 0.9639), which highlights the importance of systematic hyperparameter optimization.

Ordinary Least Squares (OLS) and Gradient Descent demonstrated strong and stable performance, both achieving Accuracy = 0.9649 and F1-score = 0.95, with perfect precision (Precision = 1.0) and slightly lower recall (Recall = 0.9048), making them comparable to the baseline logistic regression model.

Newton's Method was the least effective approach in this study, resulting in noticeably lower performance (Accuracy = 0.9035, Recall = 0.8333, F1-score = 0.8642) compared to other optimization techniques.

L2 regularization and stochastic gradient descent (SGD) improved model robustness and maintained competitive performance; however, their results (F1-score = 0.9512 for L2 and 0.9398 for SGD) did not surpass those obtained with fine-tuned logistic regression.

Overall, logistic regression proved to be a robust and effective model for breast cancer detection, particularly when combined with appropriate optimization strategies and fine-tuning, consistently achieving high accuracy and balanced precision–recall trade-offs.

CONCLUSION

This study investigated the effectiveness of logistic regression for breast cancer detection using the Breast Cancer Wisconsin (Diagnostic) dataset, with a particular focus on optimization techniques and fine-tuning strategies for model weights.

A baseline logistic regression model was first implemented and evaluated, demonstrating strong classification performance in terms of accuracy, precision, recall, and F1-score. To further analyze the impact of different optimization strategies, several methods for weight estimation were applied, including ordinary least squares, gradient descent, Newton’s method, L2 regularization, and stochastic gradient descent. The experimental results showed that optimization methods such as ordinary least squares and gradient descent achieved comparable performance to the baseline logistic regression model, while Newton’s method exhibited relatively lower performance, indicating potential sensitivity to data characteristics or numerical stability in this setting.

Regularization techniques and stochastic optimization provided stable results and helped control model complexity; however, they did not significantly outperform the baseline model in terms of overall classification metrics. The most notable improvement was achieved through fine-tuning using GridSearch with cross-validation, which optimized key hyperparameters of the logistic regression model. This approach resulted in the highest overall performance, particularly in terms of accuracy and F1-score, demonstrating the importance of systematic hyperparameter optimization.

Overall, the findings confirm that logistic regression remains a strong and reliable baseline for breast cancer detection, especially when combined with appropriate data preprocessing and fine-tuning techniques. The comparative analysis highlights that while different optimization methods may lead to similar performance levels, hyperparameter tuning plays a crucial role in achieving optimal classification results. These results emphasize the value of combining classical statistical models with modern optimization and validation techniques in medical data analysis.hypotheses need to be tested, or which applied aspects may be implemented in the future.

CONFLICT OF INTEREST: The authors declare that there is no conflict of interest regarding the publication of this article.

FUNDING: Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT: Not applicable.

INFORMED CONSENT STATEMENT: Not applicable. The study did not involve human participants.

DATA AVAILABILITY STATEMENT: The data used in this study are publicly available. This research is based on the Wisconsin Breast Cancer (Diagnostic) Dataset, which is available in the public domain and can be accessed through the Machine Learning Repository of the University of California, Irvine (UCI):

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

ACKNOWLEDGEMENTS: Not applicable.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES: The authors did not use generative artificial intelligence technologies in the preparation of this manuscript. All analyses, interpretations, and writing were performed by the authors.

REFERENCES

Watrianthos, R., Rayendra, R., Asri, E., Yuhefizar, Y., & Humaira, H. (2025). Threshold-optimized and calibrated logistic regression for breast cancer classification. Salud, Ciencia y Tecnología, 5, 8. https://doi.org/10.56294/saludcyt20252241

Won, J., Lee, H. S., & Lee, J. W. (2025). A review on multi-fidelity hyperparameter optimization in machine learning. ICT Express, 11(2), 245–257. https://doi.org/10.1016/j.icte.2025.02.008

Benghazouani, S., Nouh, S., & Zakrani, A. (2025). A robust logistic regression approach enhanced by hyperparameter optimization techniques through swarm intelligence and genetic algorithms: Advancing cancer diagnosis. High-Confidence Computing, 100378. https://doi.org/10.1016/j.hcc.2025.100378

Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., & Shaban-Nejad, A. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. Cancer Innovation, 3(5), e136. https://doi.org/10.1002/cai2.136

Дёмин, К. С., & Гермашев, И. В. (2024). Использование методов машинного обучения при диагностике рака молочной железы // Demin, K. S., & Germashev, I. V. Ispol'zovanie metodov mashinnogo obucheniya pri diagnostike raka molochnoy zhelezy [The use of machine learning methods in breast cancer diagnosis]. Digital Diagnostics, 5(3), 578. https://doi.org/10.17816/DD625866 (In Russ.)

Cristea, D. M., Sima, I., & Iantovics, L. B. (2025). Comparative analysis of optimized logistic regression with state-of-the-art models for complex gastroenterological image analysis. Frontiers in Medicine, 12, 1655612. https://doi.org/10.3389/fmed.2025.1655612

Khan, M. Z., Salahuddin, M., Khan, A. A., Sultan, S. A., Bhatti, S. D., Ullah, S., & Mohamed, M. A. (2026). Innovative hybrid machine learning methods for breast cancer predictive optimization. International Journal of Computational Intelligence Systems. https://doi.org/10.1007/s44196-025-01135-5

Ahmad, M., Javed, N., Muzafar, A., Muzafar, M., Naseer, H., Huang, G., & He, D. (2026). Enhancing logistic regression performance through hyperparameter tuning: A comparative evaluation across datasets. https://doi.org/10.21203/rs.3.rs-8304042/v1

Aryn, A., Mamyrbaev, O., & Pavlov, S. (2025). Классификация рака молочной железы, используя методы машинного обучения // Aryn, A., Mamyrbaev, O., & Pavlov, S. Klassifikatsiya raka molochnoy zhelezy, ispol'zuya metody mashinnogo obucheniya [Breast cancer classification using machine learning methods]. Nauka intellektual'nykh sistem, 1(1), 25–37. https://kazscience.kz/index.php/SIS/article/view/14/3 (In Russ.)

Авторлар туралы мәліметтер
Информация об авторах
Information about authors

Арын Арай Болатқызы – техника ғылымдарының магистрі, Каспий қоғамдық университеті, Алматы қ., Қазақстан
Арын Арай Болатқызы – магистр технических наук, Каспийский университет, г. Алматы, Казахстан
Aryn Aray Bolatkyzy – Master of Technical Sciences, Caspian University, Almaty, Kazakhstan
e-mail: arun_arai@mail.ru,
ORCID: https://orcid.org/0000-0001-7023-0424

Мамырбаев Өркен Жұмажанұлы – PhD, Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан
Мамырбаев Оркен Жумажанович – PhD, Института информационных и вычислительных технологий, г. Алматы, Казахстан
Mamyrbayev Orken Zhumazhanovich – PhD, Institute of Information and Computing Technologies, Almaty, Kazakhstan
e-mail: morkenj@mail.ru,
ORCID: https://orcid.org/0000-0001-8318-3794