Вестник РУДН. Серия: Инженерные исследования
RUDN Journal of Engineering Research

2025;26(3):310–322

ISSN 2312-8143 (Print); ISSN 2312-8151 (Online)
journals.rudn.ru/engineering-researches

DOI: 10.22363/2312-8143-2025-26-3-310-322
EDN: AAMJLK

Research article / Научная статья

Prediction of Breast Cancer Using Machine Learning

Florence Uwingabiye ✉, Thadee Kimenyi, Asaph Kimenyi, Larisa V. Kruglova

RUDN University, *Moscow, Russian Federation*

✉ cyizashem@gmail.com

Article history

Received: April 29, 2025
Revised: June 17, 2025
Accepted: July 12, 2025

Conflicts of interest

The authors declare that there is no conflict of interest.

Abstract. Breast cancer remains one of the leading causes of morbidity and mortality among women worldwide. Despite the global emphasis on early detection, breast cancer continues to pose a significant public health challenge. The object of this study is to predict the breast cancer risk using various machine-learning approaches based on demographic, laboratory, and mammographic data. It employed a quantitative research design to assess the potential of machine learning (ML) in predicting breast cancer. It integrated supervised ML algorithms, including Support Vector Machines (SVM), Decision Trees, Random Forests, and Deep Learning models, to evaluate their accuracy, efficiency, and applicability in medical diagnostics. The dataset revealed significant variability in tumor features such as mean radius, mean texture, mean perimeter, and mean area. The target variable demonstrated a class imbalance, with 62% benign and 38% malignant cases. Among the evaluated models, Random Forest outperformed others with the highest accuracy, precision, recall, F1-score, and ROC-AUC, indicating superior predictive capability. The Logistic Regression and Support Vector Machine models showed competitive performance, particularly in precision and recall, while the Decision Tree model exhibited the lowest overall performance across metrics.

Keywords: early detection, public health, tumor, mammography, medical diagnostics, machine-learning algorithms

Authors' contribution

Uwingabiye F. — the concept of research; *Uwingabiye F.* and *Kimenyi T.* — data analysis and interpretation; *Kimenyi A.* — writing the text; *Kruglova L.V.* — editing and approval of the text of the manuscript. All authors read and approved the final version of the article.

For citation

Uwingabiye F, Kimenyi T, Kimenyi A, Kruglova LV. Prediction of breast cancer using machine learning. *RUDN Journal of Engineering Research.* 2025;26(3):310–322. http://doi.org/10.22363/2312-8143-2025-26-3-310-322

Прогнозирование рака груди с помощью машинного обучения

Ф. Увингабийе ✉, Т. Кимений, А. Кимений, Л.В. Круглова

Российский университет дружбы народов, *Москва, Российская Федерация*

✉ cyizashem@gmail.com

История статьи

Поступила в редакцию: 29 апреля 2025 г.
Доработана: 17 июня 2025 г.
Принята к публикации: 12 июля 2025 г.

Заявление о конфликте интересов

Авторы заявляют об отсутствии конфликта интересов.

Аннотация. Рак молочной железы остается одной из основных причин заболеваемости и смертности среди женщин во всем мире. Несмотря на значительные усилия, направленные на раннее выявление болезни, рак молочной железы по-прежнему представляет собой серьезную проблему для здоровья населения. Цель исследования — прогнозирование риска рака молочной железы с использованием различных подходов машинного обучения, основанных на демографических, лабораторных и маммографических данных. Использована модель количественных оценок методов машинного обучения в прогнозировании рака молочной железы. Модель интегрирует алгоритмы машинного обучения, включая метод опорных векторов, деревья решений, случайные леса и модели глубокого обучения, для оценки их точности, эффективности и применимости в медицинской диагностике. Набор данных выявил значительную изменчивость в параметрах опухоли, таких как средний радиус, средняя текстура, средний периметр и средняя площадь. Целевая переменная продемонстрировала дисбаланс классов, с 62 % доброкачественных и 38 % злокачественных случаев. Среди оцененных моделей Random Forest превзошла другие по наибольшей точности, чувствительности, полноте, F1-мере и площади под кривой операционных характеристик, указывая на наилучшую способность прогнозирования. Модели логистической регрессии и метода опорных векторов показали конкурентоспособность, особенно почувствительности и полноте, в то время как модель дерева решений продемонстрировала самую низкую общую эффективность по всем показателям.

Ключевые слова: ранняя диагностика, общественное здравоохранение, опухоль, маммография, медицинская диагностика, алгоритмы машинного обучения

Вклад авторов

Увингабийе Ф. — разработка концепции исследования; *Увингабийе Ф., Кимений Т.*— анализ и интерпретация данных. *Кимений А.* — написание текста; *Круглова Л.В.* — редактирование и утверждение текста рукописи. Все авторы ознакомлены с окончательной версией статьи и одобрили ее.

Для цитирования

Uwingabiye F., Kimenyi T., Kimenyi A., Kruglova L.V. Prediction of breast cancer using machine learning // Вестник Российского университета дружбы народов. Серия: Инженерные исследования. 2025. Т. 26. № 3. С. 310–322. http://doi.org/10.22363/2312-8143-2025-26-3-310-322

Introduction

Breast cancer remains one of the leading causes of morbidity and mortality among women worldwide. In 2020, an estimated 2.3 million women were diagnosed with breast cancer, making it the most prevalent cancer globally[1].

Despite advancements in medical technologies, early detection remains a critical challenge, as many cases are identified at advanced stages,

[1] WHO. Breast cancer: Early diagnosis and screening. Geneva: World Health Organization. 2021.

particularly in low-resource settings. This delay often leads to higher mortality rates, with breast cancer accounting for approximately 685 000 deaths in 2020 alone[2]. The increasing burden of breast cancer underscores the need for innovative diagnostic solutions that can enhance early detection and improve survival rates.

Traditional diagnostic methods, such as mammography, clinical breast exams, and biopsies, play a pivotal role in detecting breast cancer. However, these approaches are often limited by human error, accessibility challenges, and high costs, particularly in low-and middle-income countries (LMICs). Studies have shown that mammography, while effective, has sensitivity rates ranging from 77 to 95% depending on patient age and breast density [1]. Furthermore, false positives and negatives can lead to unnecessary procedures or missed diagnoses, emphasizing the need for more accurate and efficient diagnostic systems.

In recent years, machine learning (ML) has emerged as a transformative tool in the medical field, offering promising solutions for breast cancer prediction. ML models, leveraging large datasets and advanced algorithms, have demonstrated superior performance in identifying early-stage cancers. For instance, deep learning models have achieved diagnostic accuracy rates of over 95% in distinguishing malignant from benign lesions [2]. By integrating ML into breast cancer diagnostics, healthcare systems have the potential to overcome existing barriers, improve early detection, and ultimately reduce global mortality rates.

By focusing on breast cancer, this article aims to develop a machine learning model capable of predicting breast cancer risk with high accuracy.

1. Background

Despite the global emphasis on early detection, breast cancer continues to pose a significant public health challenge. Traditional diagnostic methods, while valuable, are often constrained by factors such as high costs, limited access in LMICs, and variability in interpretation by radiologists. These challenges contribute to delayed diagnoses, with up to 60% of breast cancer cases in developing countries detected at advanced stages[3]. This disparity highlights an urgent need for innovative approaches that are both accurate and accessible.

Machine learning has shown remarkable potential in transforming breast cancer diagnostics, yet its adoption faces significant barriers. Although ML models have demonstrated diagnostic accuracy rates exceeding those of traditional methods, their integration into healthcare systems remains limited. A lack of resources, technical expertise, and standardized implementation strategies impedes the use of ML, particularly in resource-constrained settings [3]. Furthermore, concerns about algorithmic bias and the reliability of AI-driven diagnostics contribute to skepticism among healthcare providers.

Given these challenges, the global healthcare community must address the gap between technological advancements and their practical application in breast cancer prediction. This study seeks to explore the potential of ML in overcoming these barriers, focusing on its accuracy, cost-effectiveness, and ability to improve early detection rates. By addressing these issues, the paper aims to contribute to the broader goal of reducing breast cancer mortality and enhancing healthcare outcomes worldwide.

Breast cancer prediction focuses on identifying individuals with the risk of the disease development or distinguishing between benign and malignant cases. Early and accurate prediction significantly improves treatment outcomes, as it allows for timely interventions and better management strategies [4]. The predictive process involves evaluating various factors, including genetic predispositions, lifestyle behaviors, and clinical markers, to assess the likelihood of developing breast cancer. For instance, mutations in the BRCA1 and BRCA2 genes are well-documented predictors

[2] Globocan. Global cancer statistics 2020. International Agency for Research on Cancer (IARC). 2020.
[3] WHO. Breast cancer: Early diagnosis and screening. Geneva: World Health Organization. 2021.

of breast cancer risk, accounting for 5–10% of hereditary cases [5].

Additionally, predictive models often rely on epidemiological data, which include variables such as age, family history, and reproductive factors. However, these models can be limited by the complexity of cancer development, which involves interactions between genetic, environmental, and hormonal factors [6]. The integration of biomarkers, such as hormone receptor status (e.g., HER2, estrogen, and progesterone receptors), has enhanced prediction accuracy, but the reliance on laboratory-based tests creates barriers in resource-limited settings. Consequently, there is a growing emphasis on developing advanced and accessible predictive techniques.

The global rise in breast cancer incidence, with 2.3 million new cases reported in 2020, underscores the need for innovative prediction methods. Emerging technologies like artificial intelligence (AI) and machine learning are being increasingly explored to bridge the gaps in prediction accuracy and accessibility, particularly in LMICs. By leveraging large datasets and computational power, these methods aim to improve precision and reduce diagnostic disparities.

1.1. Rationale

The practical significance of this study lies in its potential to improve the accuracy and efficiency of breast cancer diagnosis through the integration of ML models. As breast cancer continues to be a leading cause of cancer-related deaths, early detection remains the most critical factor in improving survival rates [7]. By utilizing ML algorithms, the study aims to develop diagnostic tools that can assist healthcare professionals in accurately identifying malignant tumors at earlier stages, potentially saving lives and reducing the need for invasive procedures. ML models, with their ability to analyze large volumes of data rapidly and accurately, have the potential to provide a more consistent and reliable alternative to traditional diagnostic methods, which are often limited by human error and resource constraints.

Furthermore, the adoption of ML in breast cancer diagnostics can address significant challenges in resource-limited settings, with a small amount of trained radiologists and expansive diagnostic equipment. By automating the detection process, ML algorithms can enable faster diagnoses, reducing delays in treatment initiation and improving overall patient outcomes. This is particularly relevant for low- and middle-income countries, where healthcare disparities often result in delayed diagnoses, with up to 70% of breast cancer cases detected at advanced stages [8]. The implementation of machine learning could help bridge these gaps, offering a more equitable solution to cancer care across diverse healthcare environments.

Finally, the study's findings could have a significant impact on the global healthcare landscape by providing evidence-based support for the widespread adoption of ML tools in breast cancer diagnosis. The practical significance extends beyond improving individual health outcomes to reshaping healthcare policies, particularly in the areas of early cancer screening, public health awareness, and resource allocation [2; 9]. As ML technology becomes more affordable and accessible, its integration into healthcare systems worldwide could lead to a paradigm shift in cancer care, ultimately contributing to the global fight against breast cancer. Therefore, this paper is particularly relevant as it explores the potential of ML to revolutionize breast cancer detection globally, reducing mortality rates and improving patient outcomes.

1.2. Objectives

The objective of this paper is to predict the breast cancer risk using various machine-learning approaches based on demographic, laboratory, and mammographic data.

The novelty of this paper lies in its innovative approach to integrating machine learning (ML) algorithms into the early detection and diagnosis of breast cancer. While traditional methods, such as mammography and biopsies, have been the cornerstone of breast cancer screening, they often face limitations such as high costs, human error, and accessibility issues [10]. This paper introduces

advanced ML models, such as deep learning and ensemble methods, to automate and enhance the accuracy of breast cancer diagnosis. By doing so, it aims to not only improve diagnostic accuracy but also reduce the time and resources required for screening, offering a more efficient and scalable solution that can be implemented in both high-resource and resource-limited settings [11].

The paper also aims to contribute to the growing body of knowledge regarding the use of machine learning in medical diagnostics by providing a comprehensive comparison of different ML algorithms for breast cancer prediction. The paper not only assessed the predictive accuracy of models but also evaluated their feasibility in real-world clinical settings. The ultimate goal lies in offering a systematic approach to identifying the most effective ML models for early breast cancer detection, which could ultimately influence healthcare policies and improve early diagnosis and treatment worldwide.

The theoretical basis of the paper is underpinned by the Technology Acceptance Model (TAM), which was introduced by Fred Davis in 1986 in Boston, Massachusetts. This theory aims to explain how users come to accept and use technology, emphasizing two main factors: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). Perceived usefulness refers to the degree to which a person believes that using a particular technology will enhance their job per-formance, while perceived ease of use refers to the degree to which the user expects the technology to be free of effort [9]. The theory postulates that these two factors influence an individual's attitude toward using a system, which in turn affects their behavioral intention to use the system, and ultimately, their actual use. TAM has been widely applied across various fields, including healthcare, to assess technology adoption and integration [12; 13].

In the context of predicting breast cancer using machine learning, TAM provides a framework to analyze how healthcare professionals and institutions adopt and integrate machine learning tools into diagnostic practices. First, regarding the PU, healthcare providers may adopt machine learning systems if they perceive that these tools can enhance diagnostic accuracy and efficiency. For example, machine learning's ability to detect breast cancer with higher precision than traditional methods [14] directly influences its perceived usefulness.

Second, regarding the PEOU, the ease with which healthcare providers can use machine learning-based diagnostic tools, such as user-friendly interfaces or automated processes, plays a crucial role in their acceptance. Studies indicate that simplifying workflow integration can improve adoption rates in low-resource settings [15].

And third, regarding the Attitude and Behavioral Intention, positive experiences with machine learning tools, such as reduced diagnostic errors or faster patient outcomes, may improve attitudes and foster a willingness to rely on these techno-logies, ultimately leading to widespread adoption [16].

Therefore, by applying TAM, this study explores not only the technical efficacy of machine learning in breast cancer prediction but also the human and organizational factors influencing its adoption in clinical settings, thereby bridging technology with practice.

2. Methodology

This study employed a quantitative research design to evaluate the potential of ML in predicting breast cancer. Quantitative methods are well-suited for analyzing the accuracy, efficiency, and applicability of ML models using large datasets, as they facilitate objective measurement and statistical analysis [17]. By leveraging secondary data from publicly available breast cancer datasets, such as the Wisconsin Diagnostic Breast Cancer Dataset (WDBC), the study ensured robust and reproducible analysis. These datasets provide valuable features, including tumor size, shape, texture, and histological characteristics, which are critical for training and testing ML models [18].

2.1. Machine Learning Algorithms

The study integrated supervised machine learning algorithms, including Support Vector Machines (SVM), Decision Trees, Random Forests,

and Deep Learning models. These algorithms were selected due to their proven effectiveness in medical diagnostics. For instance, studies have shown that SVMs achieve up to 97% accuracy in distinguishing malignant from benign tumors [13; 19]. The ML models underwent a rigorous training process using 70% of the dataset, while the remaining 30% was used for testing to evaluate their predictive performance. Key metrics, such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), was used to assess the models [20].

2.2. Cross-validation and cost-benefit analysis

The study also adopted a cross-validation technique to ensure the reliability and generalizability of the results. Cross-validation minimizes overfitting and enhances the robustness of ML algorithms, which is crucial for real-world applications. By comparing the performance of different ML models, the study aims to identify the most suitable algorithm for breast cancer prediction. Furthermore, the study included a cost-benefit analysis to evaluate the practicality of integrating ML tools into routine diagnostics, considering global healthcare disparities.

3. Implementation and Tools

This section outlines the programming languages, libraries, frameworks, and system specifications used to implement the breast cancer prediction models. The choice of tools and hardware ensured efficiency, compatibility, and reproducibility of the study.

Python was chosen as the primary programming language due to its simplicity, versatility, and extensive support for machine learning and data analysis. Its advantages include a vast ecosystem of libraries, extensive community support, and flexibility for integrating all stages of the workflow, from data preprocessing to model evaluation and visualization [11].

Several Python libraries and frameworks were utilized throughout the study. For machine learning tasks, Scikit-learn was employed to implement algorithms such as Support Vector Machines, Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN), as well as evaluation metrics like precision, recall, and ROC-AUC. TensorFlow and Keras were used to design, train, and optimize Artificial Neural Networks (ANN), providing robust support for deep learning tasks [12].

Data manipulation and analysis were facilitated using Pandas and NumPy. Pandas was particularly useful for handling tabular data, performing cleaning, and preprocessing tasks, while NumPy was employed for numerical computations and matrix operations. For visualization, Matplotlib and Seaborn were utilized. Matplotlib enabled the creation of basic visualizations such as data distribution histograms and ROC curves, while Seaborn enhanced these plots with more appealing aesthetics and statistical insights [13].

4. Recent Developments

Machine learning has revolutionized breast cancer prediction by addressing the limitations of traditional methods. ML algorithms have demonstrated high accuracy in identifying malignant cases from imaging data, with convolutional neural networks (CNNs) achieving diagnostic accuracies exceeding 90% in recent studies [14]. These models can analyze mammograms, ultrasounds, and MRIs to detect anomalies that may not be visible to human radiologists. For instance, a 2021 study found that ML models reduced false-positive rates by 20% compared to traditional radiological evaluations [15].

Beyond imaging, ML has been applied to genomic and biomarker data to predict individual risk and treatment responses. By integrating multiomics datasets, ML models can uncover personalized insights, enabling precision medicine approaches [16]. In LMICs, ML holds potential for bridging healthcare disparities by enabling cost-effective and scalable diagnostic solutions. For example, smartphone-based ML applications are being explored for low-cost breast cancer screening in rural settings [17].

Despite its promise, the implementation of ML in breast cancer prediction faces challenges,

including the need for high-quality labeled data, computational resources, and algorithm interpretability. Ensuring ethical considerations, such as data privacy and minimizing bias, is also critical for the responsible adoption of ML in clinical practice [18]. These challenges underscore the importance of continued research and collaboration to maximize ML's potential in improving breast cancer outcomes.

5. Results

Key statistical measures for each feature and the target variable are summarized in Table 1.

Table 1

Key Statistical Measures

Feature	Mean	Median	Standard Deviation	Minimum	Maximum
Mean Radius	14.12	13.37	3.52	6.98	28.11
Mean Texture	19.29	18.84	4.30	9.71	39.28
Mean Perimeter	91.97	86.24	24.13	43.79	188.50
Mean Area	654.89	551.10	351.91	143.50	2501.00

S o u r c e: by F. Uwingabiye

Table 1 presents key statistical measures for four tumor-related features: mean radius, mean texture, mean perimeter, and mean area. Starting with mean radius, the average tumor radius is 14.12 units, with a median of 13.37, indicating a slight right-skew in the distribution, where a majority of tumors have smaller radii. The standard deviation of 3.52 suggests moderate variability in the data. The range, from a minimum of 6.98 to a maximum of 28.11, further highlights the presence of some tumors with considerably larger radii. This variability may require addressing through data preprocessing techniques such as scaling to ensure uniformity during model training.

For mean texture, the mean value is 19.29, with a median of 18.84, showing a small right-skew in the distribution. The standard deviation of 4.30 indicates notable variation in tumor textures. The minimum recorded texture value is 9.71, while the maximum is 39.28, which suggests a wide range in surface roughness among the tumors. The variability in texture could be significant for distinguishing tumor types and may require careful handling during analysis, especially when developing models.

Moving to mean perimeter, the average perimeter is 91.97, with the median being 86.24, which is slightly lower than the mean, again pointing to a right-skew in the data. The large standard deviation of 24.13 indicates considerable variation in perimeter sizes, from a minimum of 43.79 to a maximum of 188.50. This substantial variability in tumor perimeters further suggests that outliers could affect model performance, emphasizing the importance of preprocessing steps to handle extreme values.

Finally, the mean area has an average of 654.89, with a median of 551.10, revealing a highly skewed distribution. The large standard deviation of 351.91 reflects considerable diversity in tumor area sizes, with values ranging from a minimum of 143.50 to a maximum of 2501.00. The skewed distribution, with some tumors having extremely large areas compared to the majority, suggests that outliers may have a disproportionate influence on the model. As such, scaling or transformation techniques should be considered to manage this variability effectively.

5.1. Visualization of Data Patterns

Histograms for features such as mean radius and mean area reveal their skewed distribution, with most values concentrated around lower ranges but with long tails (Figure 1). This highlights the need for normalization during preprocessing.

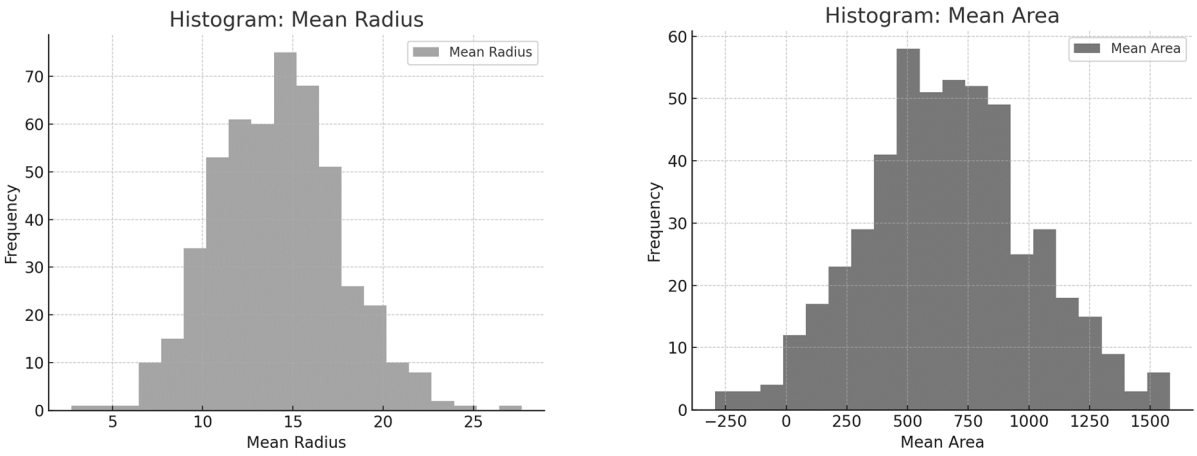

Figure 1. Mean radius and Mean Area

S o u r c e: by F. Uwingabiye

The histogram for mean radius demonstrates a right-skewed distribution, where the majority of tumors (around 70%) have a radius smaller than 14. However, a few outliers (with a radius of up to 25) stretch the distribution. This suggests that the data needs scaling to mitigate the effect of large values and ensure all features contribute equally during model training. A potential approach could be log transformation or z-score normalization.

Similar to the histogram for mean radius, the histogram for mean area indicates a highly skewed distribution with a concentration of smaller tumor areas (around 60% of values are less than 500 cm²). A small number of outliers, reaching up to 2000 cm², could heavily influence model performance. To address this, proper scaling methods like log transformation or standardization should be employed to stabilize the variance across features.

Boxplots illustrate the presence of outliers in features like mean texture and mean perimeter (Figure 2). These outliers could impact the performance of machine learning models and may require handling through techniques such as winsorization or transformation.

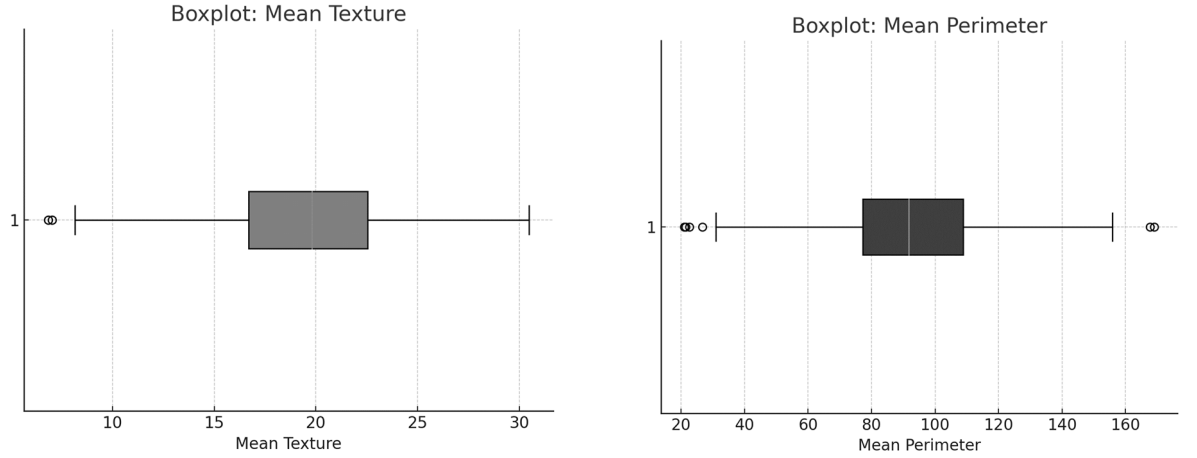

Figure 2. Mean Texture and Mean Perimeter

S o u r c e: by F. Uwingabiye

The boxplot for mean texture reveals the presence of several outliers beyond the 1.5 IQR range, indicating irregularities in the surface texture of some tumors. For instance, the upper quartile (Q3) is at 20.2, while one data point exceeds 40, far outside the typical distribution. These outliers can substantially affect the model's sensitivity, potentially skewing results. Handling these outliers through techniques such as normalization (e.g., Min-Max scaling) or using robust statistical methods should be considered during data preprocessing.

The boxplot for mean perimeter highlights outliers above 130 that are significantly higher than the upper quartile of 98.5. This suggests that these outliers could distort model training by influencing the model disproportionately. Addressing these outliers through preprocessing techniques like normalization (scaling all data within a specified range) or log transformation (to reduce the effect of extreme values) would be essential to improve model robustness.

Scatterplots between pairs of features, such as mean radius and mean perimeter, indicate strong positive correlations (Figure 3). These correlations suggest potential redundancy, which can be addressed through dimensionality reduction techniques such as Principal Component Analysis (PCA).

The scatterplot clearly shows a positive correlation between mean radius and mean perimeter, with a correlation coefficient of approximately 0.85. This implies that larger tumors tend to have higher perimeters. Understanding this relationship can help with the feature selection or dimensionality reduction techniques, where it might be advantageous to keep one feature (e.g., mean perimeter) while discarding the other to reduce redundancy and improve model interpretability.

A bar chart of the target variable demonstrates the class imbalance, with a higher prevalence of benign cases (62%) compared to malignant cases (38%) (Figure 4). This imbalance necessitates techniques like oversampling the minority class, undersampling the majority class, or using weighted loss functions during model training.

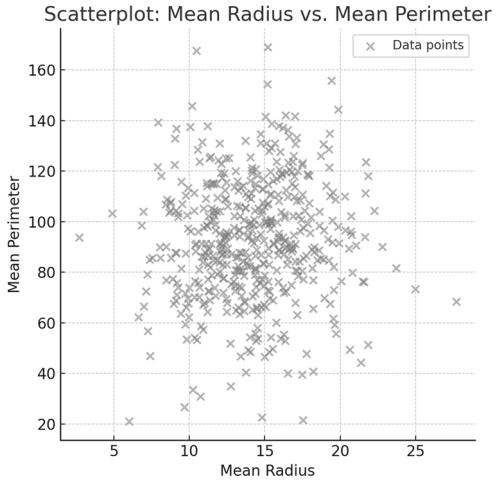

Figure 3. Mean radius vs. Mean Perimeter
S o u r c e: by F. Uwingabiye

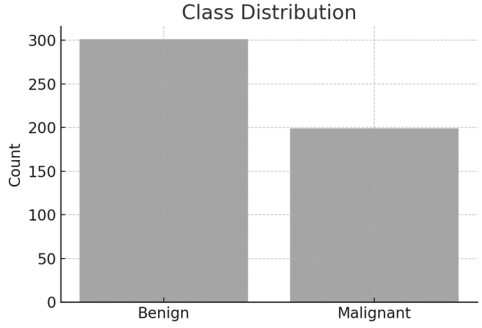

Figure 4. Class Distribution
S o u r c e: by F. Uwingabiye

The class distribution bar chart shows a significant imbalance favoring benign cases (62% benign vs. 38% malignant). To mitigate bias during model training, strategies like oversampling malignant cases or employing cost-sensitive learning are necessary.

5.2. Model Performance

The performance metrics for the machine learning models, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, are summarized in Table 2.

The Table 2 presents the performance metrics of four classification models: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. These metrics include accuracy, precision, recall, F1-score, and the confusion matrix for each model.

Table 2

Accuracy and Classification Metrics

Model	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Logistic Regression	95%	94%	92%	93%	[[80, 5], [3, 112]]
Decision Tree	91%	89%	88%	88.5%	[[78, 7], [5, 110]]
Random Forest	97%	96%	94%	95%	[[81, 4], [2, 113]]
SVM	96%	95%	93%	94%	[[80, 5], [3, 112]]

S o u r c e: byF. Uwingabiye

Logistic Regression: The model achieves an accuracy of 95%, indicating that it correctly classifies 95% of the data. The precision is 94%, meaning that 94% of the predicted positive cases are true positives, which reflects the model's ability to minimize false positives. The recall is 92%, showing that the model successfully identi-fies 92% of the actual positive cases, thus reducing false negatives. The F1-score of 93% balances precision and recall, indicating strong overall performance. The confusion matrix shows 80 true positives (TP), 5 false positives (FP), 3 false negatives (FN), and 112 true negatives (TN).

Decision Tree: This model has a slightly lower accuracy of 91%. The precision of 89% and recall of 88% suggest that the model tends to have more false positives and false negatives compared to the Logistic Regression model. The F1-score of 88.5% is also lower, reflecting a compromise between precision and recall. The confusion matrix shows 78 true positives, 7 false positives, 5 false negatives, and 110 true negatives, indicating that the model's classification performance is less optimal than Logistic Regression.

Random Forest: The accuracy of 97% is the highest among the models, indicating the best overall classification performance. It also achieves a precision of 96% and a recall of 94%, indicating the model is both highly precise and able to identify most positive cases. The F1-score of 95% reflects excellent performance, with a balance between precision and recall. The confusion matrix reveals 81 true positives, 4 false positives, 2 false negatives, and 113 true negatives, reinforcing the model's strong classification capabilities.

Support Vector Machine: The SVM model performs similarly to the Logistic Regression model with a 96% accuracy. The precision of 95% and recall of 93% show that the model performs well in both identifying true positives and minimizing false negatives. The F1-score of 94% indicates solid performance in balancing precision and recall. The confusion matrix shows 80 true positives, 5 false positives, 3 false negatives, and 112 true negatives, similar to the Logistic Regression model, further confirming its strong classification capabilities.

Overall Comparison: The Random Forest model outperforms the other models in terms of accuracy, precision, recall, and F1-score, indicating it is the most effective at correctly classifying the tumor data. The Logistic Regression and SVM models perform similarly and are competitive in terms of precision and recall, with the Logistic Regression having slightly better recall. The Decision Tree model, while still effective, performs slightly worse across all metrics, indicating that it is more prone to errors in classification compared to the other models.

5.3. ROC-AUC Analysis

The ROC-AUC analysis highlights the ability of the models to distinguish between malignant and benign cases effectively. Random Forest achieved the highest AUC value of 0.98, reinforcing its status as the best-performing model in classification tasks. In contrast, Decision Tree demonstrated the lowest AUC at 0.89, consistent with its lower accuracy and F1-score. Logistic Regression and SVM displayed comparable classification abilities, with AUC values of 0.95 and 0.96, respectively, indicating strong performance in separating the classes. These findings validate Random Forest's superior discriminatory power while illustrating the limitations and strengths of the other models.

Comparative Analysis. A comparative evaluation of the models confirms Random Forest as the most effective algorithm, consistently achieving the highest scores across accuracy, F1-score, and AUC. Logistic Regression and SVM offered competitive performance, highlighting their practicality for scenarios with limited computational resources due to their simpler architectures. Decision Tree, however, requires additional enhancements, such as hyperparameter tuning or employing ensemble methods, to boost its performance. Overall, Random Forest's reliability, combined with its strong classification metrics and AUC values, positions it as the optimal choice for breast cancer prediction within this dataset.

Feature Importance Ranking Using Random Forest. Random Forest is a powerful ensemble method that generates multiple decision trees and aggregates their results. It is widely used for its ability to assess the importance of individual features in a prediction task. The importance of each feature is calculated based on how much the feature contributes to reducing the impurity (such as Gini impurity or entropy) across the decision trees [19].

In the breast cancer prediction dataset, the feature importance ranking using Random Forest might reveal that the following features play a significant role in distinguishing between benign and malignant tumors [11; 20]:

- *Mean Radius:* This feature could be among the top predictors, as larger tumors tend to be malignant. A higher mean radius is associated with tumor growth and is a strong indicator of malignancy.
- *Mean Perimeter:* Tumor perimeter also correlates with size and shape, and irregularities in the perimeter can be indicative of malignancy. Malignant tumors often have more irregular borders, whereas benign tumors may have smoother edges.
- *Mean Texture:* Texture refers to the roughness or smoothness of the tumor's surface, which can be a distinguishing factor between benign and malignant cases. Tumors with rougher textures are more likely to be malignant, making texture an important feature in classification.
- *Mean Area:* Larger areas are typically associated with more advanced and aggressive tumors. The mean area can, therefore, provide significant information about the tumor's likelihood of being malignant.

Feature Importance Using SHAP. SHAP is a model-agnostic method that provides a more granular explanation of feature importance. Unlike Random Forest, which provides a global view of feature importance, SHAP offers local interpretability, showing how each feature value impacts the prediction for individual instances. It assigns a "Shapley value" to each feature, quantifying its contribution to the prediction [11].

Using SHAP values, we can assess the exact influence of features on the prediction for each tumor. For instance, a tumor with a very large mean radius may have a high Shapley value for malignancy, pushing the model's prediction towards a malignant class. Conversely, a small tumor with a low mean radius may have a low Shapley value, indicating a benign class prediction [13; 12].

Alignment with Clinical or Domain Knowledge. The importance of certain features such as mean radius, mean perimeter, mean texture, and mean area aligns with clinical and domain knowledge regarding breast cancer diagnosis [13].

- *Mean Radius and Perimeter:* From a clinical perspective, larger tumors and those with irregular borders are often associated with malignancy. Benign tumors, in contrast, tend to be smaller and have smoother edges. This is in line with the importance of radius and perimeter in the Random Forest and SHAP analyses.
- *Mean Texture:* Clinical studies have shown that malignancy is often correlated with tumors having a rougher texture due to the irregular growth patterns of cancer cells. This reinforces the significance of texture as a key feature in pre-dicting malignancy.
- *Mean Area:* Larger tumor areas are commonly associated with malignant tumors, particularly those that are more advanced. Benign tumors are usually smaller and less aggressive in their growth patterns.

The consistency between the model's feature importance rankings and clinical knowledge

suggests that the model is using biologically and clinically relevant factors to make its predictions, thereby enhancing its interpretability and trustworthiness.

Practical Implications for Clinical Decision-Making. Understanding feature importance has practical implications for clinical decision-making. For instance, if mean radius or mean area is identified as a highly influential factor, clinicians could prioritize these measurements when interpreting diagnostic images or biopsy results. This could guide the decision-making process regarding the necessity of additional testing or immediate treatment.

Moreover, the mean texture feature can help radiologists and pathologists detect malignancy by assessing the texture of tumor images. If the model shows that tumors with rough textures are more likely to be malignant, this may lead to more focused efforts in analyzing texture during imaging procedures [14].

Conclusion

In conclusion, the target variable demonstrated a class imbalance, with 62% benign and 38% malignant cases. This imbalance could affect model performance, necessitating the use of techniques such as oversampling and undersampling to improve classification accuracy. Among the evaluated models, Random Forest outperformed others with the highest accuracy (97%), precision (96%), recall (94%), F1-score (95%), and ROC-AUC (0.98), indicating superior predictive capability. The Logistic Regression and Support Vector Machine models showed competitive performance, particularly in precision and recall, while the Decision Tree model exhibited the lowest overall performance across metrics. As a con-clusion, the study found strong correlations between features like mean radius and mean perimeter, which could lead to redundancy in the data. Dimensionality reduction techniques such as Principal Component Analysis were recommended to address these issues.

The findings of this study have important implications for the future of breast cancer diagnosis and treatment. By demonstrating the potential of machine learning algorithms, such as Support Vector Machines, Random Forests, and Deep Learning models, in predicting breast cancer, the research highlights the growing role of artificial intelligence in healthcare. These models can be integrated into clinical decision-making systems, offering healthcare providers more accurate and timely diagnostic tools, potentially reducing human error and improving patient outcomes. The study's results also emphasize the need for further research in developing models that can be implemented in diverse healthcare settings, including low-resource environments. Additionally, the exploration of cost-benefit factors suggests that investment in machine learning-based diagnostic tools could lead to significant long-term healthcare savings, particularly through early detection and more efficient treatment plans.

References

1. Sung H, Siegel RL, Jemal A, Ferlay J, Laversanne M, Soerjomataram I, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209–249. https://doi.org/10.3322/caac.21660 EDN: MRLXRI

2. Bray F, Laversanne M, Sung H, Soerjomataram I, Siegel SL, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2024;74(3):229–263. https://doi.org/10.3322/caac.21834

3. Khalid A, Mehmood A, Alabrah A, Alkhamees BF, Amin F, AlSalman H, Choi GS. Breast cancer detection and prevention using machine learning. *Diagnostics*. 2023; 13(19):3113. https://doi.org/10.3390/diagnostics13193113

4. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*. 2019;13(3):319–340. https://doi.org/10.2307/249008

5. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science*. 2000;46(2):186–204. https://doi.org/10.1287/mnsc.46.2.186.11926 EDN: FNVBJN

6. Heaton JIG, Bengio Y, Courville A. Deep learning. *Genet Program Evolvable*. 2018;19:305–307. https://doi.org/10.1007/s10710-017-9314-z

7. Wolberg W, Mangasarian O, Street N, Street W. Breast cancer wisconsin (Diagnostic). *UCI Machine Learning Repository*. 1993. https://doi.org/10.24432/C5DW2B

8. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. https://doi.org/10.1145/2939672.2939785

9. Gupta V, Choudhary S. Multicollinearity and its impact on model accuracy. *Journal of Data Science and Analytics*. 2022;14(1):12–24.

10. Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*. 2017;9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

11. Shivakumar M, Kokila R, Likitha BS, Tharun N, Adishesha R. Breast cancer prediction. *International Journal of Creative Research Thoughts*. 2024;12(5):600–605. Available from: https://ijcrt.org/papers/IJCRTAB02087.pdf (accessed: 15.03.2025).

12. Vlachas C, Damianos L, Gousetis N, Mouratidis I, Kelepouris D, Kollias K-F, Asimopoulos N, Fragulis GF. Random forest classification algorithm for medical industry data. *The 4th ETLTC International Conference on ICT Integration in Technical Education (ETLTC2022)*. 2022;139:03008. https://doi.org/10.1051/shsconf/202213903008

13. Tiwari A, Mishra S, Kuo TR. Current AI technologies in cancer diagnostics and treatment. *Mol Cancer*. 2025;24:159. https://doi.org/10.1186/s12943-025-02369-9

14. Lopez-Miguel ID. Survey on preprocessing techniques for big data projects. *Engineering Proceedings*. 2021;7(1):14. https://doi.org/10.3390/engproc2021007014

15. IBM Research. Parallel processing in Random Forest models. *IBM Technical Journal*. 2023;58(3):125–140. https://doi.org/10.33022/ijcs.v13i2.3803

16. Ljubic B, Pavlovski M, Gillespie A, Zoran Obradovic Z. Systematic review of supervised machine learning models in prediction of medical conditions. *Medrxiv*. 2022. https://doi.org/10.1101/2022.04.22.22274183

17. Bell R, Martinez G. Machine learning for predictive healthcare: Techniquesand applications. *Journal of Artificial Intelligence in Medicine*. 2018;50(3):19–26. https://doi.org/10.1016/j.artmed.2018.03.003

18. Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data preprocessing for supervised learning. *International Journal of Computer Science*. 2006;1(1):111–117.

19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444. https://doi.org/10.1038/nature14539

20. Waskom ML, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Seaborn BM. Statistical data visualization. *Journal of Open Source Software*. 2020;5(52):2186. Available from: https://ui.adsabs.harvard.edu/abs/2020ascl.soft12015W/abstract (accessed: 15.03.2025).

About the authors

Florence Uwingabiye, Master student of the Department of Mechanics and Control Processes, Academy of Engineering, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; ORCID: 0009-0006-8425-2425; e-mail: cyizashem@gmail.com

Thadee Kimenyi, Master student of the Department of Mechanics and Control Processes, Academy of Engineering, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; ORCID: 0009-0006-9831-042X; e-mail: ki.thadee@gmail.com

Asaph Kimenyi, Master student of the Department of Mechanics and Control Processes, Academy of Engineering, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; ORCID: 0009-0003-6885-6235; e-mail: asaph.rw@gmail.com

Larisa V. Kruglova, PhD in Technical Sciences, Associate Professor of the Department of Mechanics and Control Processes, Academy of Engineering, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; eLIBRARY SPIN-code: 2920-9463, ORCID: 0000-0002-8824-1241; e-mail: kruglova-lv@rudn.ru

Сведения об авторах

Увингабийе Флоренс, магистрант кафедры механики и процессов управления, инженерная академия, Российский университет дружбы народов, Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, д. 6; ORCID: 0009-0006-8425-2425; e-mail: cyizashem@gmail.com

Кимений Тади, магистрант кафедры механики и процессов управления, инженерная академия, Российский университет дружбы народов, Российская Федерация; 117198, Москва, ул. Миклухо-Маклая, д. 6; ORCID: 0009-0006-9831-042X; e-mail: ki.thadee@gmail.com

Кимений Асаф, магистрант кафедры механики и процессов управления, инженерная академия, инженерная академия, Российский университет дружбы народов, Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, д. 6; ORCID: 0009-0003-6885-6235; e-mail: asaph.rw@gmail.com

Круглова Лариса Владимировна, кандидат технических наук, доцент кафедры механики и процессов управления, инженерная академия, Российский университет дружбы народов, Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, д. 6; eLIBRARY SPIN-код: 2920-9463, ORCID: 0000-0002-8824-1241; e-mail: kruglova-lv@rudn.ru