Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics
Turkiye Klinikleri J Biostat. 2026;18(1):14-26

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/biostatic.2025-116724

Morphological Feature Selection and Classification of Breast Cancer Samples with Machine Learning Methods: Analytical Research

Makine Öğrenme Yöntemleriyle Meme Kanseri Örneklerinin Morfolojik Özellik Seçimi ve Sınıflandırılması: Analitik Araştırma

Berkalp TUNCA[a], Sinan SARAÇLI[b]

[a]Eskişehir Osmangazi University Institute of Science, Department of Statistics, Doctoral Programme, Eskişehir, Türkiye
[b]Balıkesir University Faculty of Medicine, Department of Biostatistics, Balıkesir, Türkiye

ABSTRACT Objective: Breast cancer is one of the most common malignancies among women worldwide, and early detection is crucial for improving clinical outcomes. This study aims to classify benign and malignant breast tumors using machine learning approaches based on cytomorphological features and to identify the most influential variables contributing to classification performance. **Material and Methods:** The dataset, obtained from the Kaggle repository, comprised 699 patient records with 9 cytomorphological features scored from 1 to 10. Data preprocessing and feature selection procedures were applied prior to model development. Logistic regression, Random Forest, and XGBoost models were implemented, and their performances were evaluated using accuracy, sensitivity, F1-score, and receiver operating characteristic (ROC)-area under the curve (AUC) metrics. SHAP (SHapley Additive explanations) analysis was employed to interpret model predictions and assess feature-level contributions. **Results:** Feature importance and SHAP analyses identified Uniformity of Cell Size, Uniformity of Cell Shape, Bare Nuclei, and Bland Chromatin as the most influential predictors. These findings highlight the critical role of cellular and nuclear morphological abnormalities in malignancy discrimination and enhance the interpretability and clinical relevance of the models. **Conclusion:** Among the evaluated models, XGBoost achieved the best overall performance, demonstrating the highest accuracy, sensitivity, and ROC-AUC values. This indicates that ensemble-based learning methods provide superior discriminative capability for breast tumor classification compared to alternative approaches.

Keywords: Machine learning; feature selection; SHAP; breast cancer

ÖZET Amaç: Meme kanseri, dünya genelinde kadınlar arasında en sık görülen malignitelerden biridir ve erken tanı klinik sonuçların iyileştirilmesinde kritik öneme sahiptir. Bu çalışmanın amacı, sitomorfolojik özelliklere dayalı olarak makine öğrenmesi yöntemleri kullanarak benign ve malign meme tümörlerini sınıflandırmak ve sınıflandırma performansına en fazla katkı sağlayan değişkenleri belirlemektir. **Gereç ve Yöntemler:** Çalışmada kullanılan veri seti Kaggle deposundan elde edilmiş olup, 1-10 aralığında puanlanmış 9 sitomorfolojik özelliğe sahip 699 hasta kaydını içermektedir. Model geliştirme öncesinde, veri ön işleme ve özellik seçimi prosedürleri uygulanmıştır. Lojistik regresyon, "Random Forest" ve XGBoost modelleri kullanılmış; model performansları doğruluk, duyarlılık, F1-skoru ve alıcı operatör karakteristiği [receiver operating characteristic (ROC)-eğri altında kalan alan [receiver operating characteristic (AUC)] ölçütleri ile değerlendirilmiştir. Model tahminlerini yorumlamak ve değişken düzeyindeki katkıları incelemek amacıyla SHAP (SHapley Additive explanations) analizi uygulanmıştır. **Bulgular:** Özellik önem analizi ve SHAP sonuçları, Hücre Boyutu Homojenliği (Uniformity of Cell Size), Hücre Şekli Homojenliği (Uniformity of Cell Shape), Çıplak Nükleuslar (Bare Nuclei) ve Bland Kromatin (Bland Chromatin) değişkenlerinin en etkili belirleyiciler olduğunu göstermiştir. Bu bulgular, hücresel ve nükleer morfolojik düzensizliklerin malignite ayrımında kritik rol oynadığını ortaya koymakta ve modellerin yorumlanabilirliğini ile klinik geçerliliğini artırmaktadır. **Sonuç:** Değerlendirilen modeller arasında XGBoost, en yüksek doğruluk, duyarlılık ve ROC-AUC değerlerini elde ederek en iyi genel performansı göstermiştir. Bu durum, topluluk (ensemble) tabanlı öğrenme yöntemlerinin meme tümörü sınıflandırmasında alternatif yaklaşımlara kıyasla daha üstün ayırt edici güce sahip olduğunu göstermektedir.

Anahtar kelimeler: Makine öğrenmesi; özellik seçimi; SHAP; meme kanseri

TO CITE THIS ARTICLE:
Tunca B, Saraçlı S. Morphological feature selection and classification of breast cancer samples with machine learning methods: Analytical research. Turkiye Klinikleri J Biostat. 2026;18(1):14-26.

Correspondence: Berkalp TUNCA
Eskişehir Osmangazi University Institute of Science, Department of Statistics, Doctoral Programme, Eskişehir, Türkiye
E-mail: berkalptunca@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 11 Dec 2025 ***Received in revised form:*** 18 Feb 2026 ***Accepted:*** 18 Feb 2026 ***Available online:*** 28 Apr 2026

2146-8877 / Copyright © 2026 by Türkiye Klinikleri. This in an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Breast cancer is among the most prevalent malignant diseases affecting women worldwide and is associated with high mortality rates when not diagnosed at an early stage.[1] According to World Health Organization data from 2023, approximately 2.3 million women are diagnosed with breast cancer annually, and nearly 685,000 of these cases result in death. Conventional diagnostic methods such as mammography, ultrasonography, and biopsy rely on imaging and pathological assessment; however, their dependence on subjective interpretation and intensive labor may introduce diagnostic inaccuracies.[2] Consequently, statistical and computational analysis of pathological data has gained increasing importance in improving early diagnostic accuracy.

In recent years, machine learning (ML) and artificial intelligence (AI)-based approaches have demonstrated substantial success in medical data classification and clinical decision support systems.[3] These methods enable reliable differentiation between malignant and benign cells by capturing complex patterns within high-dimensional datasets. In this context, the Breast Cancer Wisconsin (Benign or Malignant) dataset has become a widely accepted benchmark in breast cancer classification research.[4]

Numerous studies in the ML literature have compared different algorithms for breast cancer diagnosis. Polat and Güneş achieved an accuracy of 98% using Support Vector Machines (SVM), while Sahan et al. reported a 99% accuracy rate with a hybrid fuzzy immune system-based approach.[5,6] These findings collectively indicate that both appropriate feature selection and model parameter optimization play a critical role in improving diagnostic performance, highlighting the importance of variable selection alongside algorithm choice.

Feature selection constitutes a key preprocessing step that enhances model accuracy and generalization performance.[7,8] In this study, Recursive Feature Elimination (RFE) based on logistic regression, Random Forest-based feature importance, and Mutual Information (MI) analysis were applied. These approaches facilitate the identification of variables most strongly associated with the target outcome, enabling systematic determination of the most informative cytomorphological features from both statistical and biological perspectives.

To improve model interpretability, SHAP (SHapley Additive exPlanations) analysis was employed.[9] SHAP addresses the “black box” limitation of ML models by quantifying the contribution of individual features to prediction outcomes, thereby enhancing clinical interpretability, particularly in biomedical research settings.[10]

Morphological assessment of cellular and nuclear characteristics has long played a central role in breast cancer diagnosis and prognosis. Advances in digital pathology have enabled these features to be quantitatively analyzed using computerized morphometric techniques. Beck et al. demonstrated that automatically extracted morphological features, including stromal characteristics, were significantly associated with patient survival, emphasizing the prognostic value of tumor micromorphology.[11]

Nuclear morphometric parameters such as nuclear size, shape, and chromatin organization have been shown to be strongly associated with malignancy. Early studies by Stenkvist et al. established the diagnostic relevance of computerized nuclear morphometry, while subsequent research reported significant associations between increased nuclear area, tumor grade, and poor prognosis.[12-14] These findings suggest that objective morphometric measurements complement traditional pathological grading systems.

Recent research has expanded beyond nuclear features to include tissue architecture and cell cluster organization. Studies have reported significant associations between nuclear orientation, shape variations, hormone receptor status, and patient risk stratification.[15-17] The integration of ML with explainable artificial intelligence (AI) techniques, particularly SHAP, has further enhanced the clinical interpretability of morphological data.[9] Saednia et al. demonstrated that combining quantitative histomorphometric features with ML models enabled accurate prediction of pathological complete response to chemotherapy.[18] Overall, the literature supports the high discriminative and clinical value of morphology-based features in explainable ML frameworks. Table 1 summarizes selected studies relevant to this field.

TABLE 1: Summary of the literature review

Author(s)	Study title	Method(s) used	Findings
Beck et al.[11]	Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival	Digital pathology, automated morphological feature extraction, statistical modeling	Stromal morphological features are significantly associated with survival; the organization within the tumor microenvironment carries prognostic information.
Stenkvist et al.[12]	Computerized Nuclear Morphometry as an Objective Method in Breast Cancer Diagnosis	CNM, linear discriminant analysis	Nuclear size and shape measurements are directly related to malignancy; CNM reduces subjective interpretation.
Abdalla et al.[13]	Correlation of Nuclear Morphometry of Breast Cancer with Prognostic Factors	Nuclear morphometry, correlation and regression analysis	Increase in mean nuclear area is associated with poor prognosis; morphometry serves as a prognostic indicator.
Kashyap et al.[14]	Role of Nuclear Morphometry in Breast Cancer and Its Correlation with Clinicopathologic Parameters	Microscopic image analysis, statistical correlation	A strong positive relationship has been found between nuclear morphometry and tumor grade.
Rawat et al.[15]	Correlating Nuclear Morphometric Patterns with Estrogen Receptor Status in Breast Cancer	ML (SVM, LDA), H&E-based nuclear morphometry	Nuclear shape and orientation are associated with ER negativity.
Whitney et al.[16]	Quantitative Nuclear Histomorphometry Predicts Oncotype DX Risk Categories	ML (Random Forest), nuclear feature extraction	Nuclear morphometric measurements accurately distinguished Oncotype DX risk groups.
Lu et al.[17]	Nuclear Shape and Orientation Features from H&E Images Predict Prognosis of Breast Cancer	Computerized image analysis, morphological feature extraction, survival analysis	Nuclear shape and orientation metrics are significantly related to survival; tissue microarchitecture has prognostic value.
Saednia et al.[18]	Quantitative Digital Histopathology and Machine Learning to Predict Pathological Complete Response	ML (Random Forest, XGBoost), histomorphometric features	Quantitative histomorphometry combined with ML accurately predicted pCR to chemotherapy.
Bansal et al.[19]	Grading Systems in the Cytological Diagnosis of Breast Cancer	Cytological analysis, comparative morphometric grading	Standardization of morphological criteria is necessary; nuclear morphometry provides objective diagnosis.
Ibrahim et al.[20]	Artificial Intelligence in Digital Breast Pathology	Artificial intelligence review, CNN-based analytical evaluation	Interpretability of morphological features through AI-assisted systems has improved; clinical reliability has been strengthened.

CNM: Computerized nuclear morphometry; ML: Machine learning; SVM: Support Vector Machines; LDA: Linear Discriminant Analysis; ER: Estrogen receptor; pCR: Pathological complete response; CNN: Convolutional Neural Network

MATERIAL AND METHODS

DATASET AND VARIABLES

The data set used in this study was obtained from the open-access resource "Breast Cancer Wisconsin".[21] The dataset consists of morphological measurements obtained from cell samples collected at the University of Wisconsin Medical Center and includes a total of 699 patient observations. Each observation includes 9 morphological features and a class label (benign/malignant), graded on a scale of 1 to 10.

The variables employed in this study are quantitative measures that reflect the microscopic structural characteristics of cells, and their descriptions are provided below:

- **Clump Thickness:** Represents the degree of cell group density; higher values indicate more tightly clustered cells.
- **Uniformity of Cell Size:** Indicates the degree of similarity among cell sizes; decreased uniformity increases the likelihood of malignancy.
- **Uniformity of Cell Shape:** Measures the variation in cell shape; higher variation indicates disrupted cellular differentiation.
- **Marginal Adhesion:** Refers to the degree of adhesion between cells; lower adhesion is indicative of invasive tumor behavior.
- **Single Epithelial Cell Size:** Reflects the average size of individual epithelial cells.

- **Bare Nuclei:** Denotes the number of nuclei separated from the cytoplasm; higher values suggest a loss of cellular integrity.
- **Bland Chromatin:** Represents the degree of chromatin smoothness; irregular chromatin indicates genetic instability.
- **Normal Nucleoli:** Expresses the normality of nucleoli in terms of visibility and number.
- **Mitoses:** Indicates the frequency of cell division; increased mitotic activity reflects accelerated cellular proliferation.

The target variable consists of 2 categories: benign (0) and malignant (1). Missing values were imputed using the median for numerical variables and the mode for categorical variables. Numerical variables were scaled using the StandardScaler method, while categorical variables were encoded through the One-Hot Encoding approach.[22]

MACHINE LEARNING METHODS

Four different classification algorithms were applied to the dataset: Logistic regression, Random Forest, Support Vector Machines with Radial Basis Function kernel (SVM-RBF), and XGBoost.

i. ***Logistic Regression*** is a classical probability model based on a linear decision boundary, modeling the probability of prediction through the logit transformation of the variables.[23]

ii. ***Random Forest*** is an ensemble model that constructs multiple decision trees using random samples and variable subsets, thereby reducing error variance.[24]

iii. ***SVM-RBF*** is designed for non-linearly separable data and employs a RBF kernel to find the optimal separating hyperplane in a high-dimensional feature space.[25]

iv. ***XGBoost*** is an algorithm based on the principle of gradient boosting, training weak classifiers sequentially to minimize classification errors.[26]

In all models, the dataset was split into 80% training and 20% testing subsets. During training, 5-fold Stratified K-Fold cross-validation was implemented. Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and receiver operating characteristic (ROC)-area under the curve (AUC) metrics. ROC curves visually demonstrated the balance between the true positive rate and the false positive rate for each model.

FEATURE SELECTION APPROACHES

To enhance overall model performance, prevent overfitting, and strengthen interpretability, three different feature selection methods were applied:

i. **Recursive Feature Elimination:** Based on logistic regression, variables contributing the least to the model were sequentially eliminated according to model coefficients, and the optimal subset of variables yielding the highest ROC-AUC value was determined.[7]

ii. **Model-Based Feature Selection (Random Forest Importance-RF):** Using the feature importance coefficients provided by the Random Forest model, variables with the highest information gain (Gini impurity reduction) were prioritized.[24]

iii. **Mutual Information:** The amount of information shared between each variable and the target variable was calculated, allowing for the consideration of nonlinear relationships.[8]

The combined evaluation of these 3 approaches identified the morphological variables with the highest levels of importance. Based on the outcomes of these methods, the most powerful biomarkers in determining malignancy were established.

MODEL INTERPRETABILITY: SHAP ANALYSIS

The SHAP analysis was conducted to enhance the interpretability of the model's decision-making mechanism. SHAP, a method grounded in game theory, quantitatively expresses the contribution of each feature to the prediction for each observation, either positively (toward malignancy) or negatively (toward benignity).[9]

In this study, SHAP analysis was applied to the XGBoost model to determine which features contributed most significantly to the model's output. The purpose of the analysis was to identify, both statistically and biologically, the most distinctive factors associated with malignancy. Furthermore, SHAP visualizations were expected to demonstrate that the model not only achieved high accuracy but also provided clinically interpretable explanations of its decision-making process.

ANALYSIS ENVIRONMENT AND IMPLEMENTATION

All analyses were conducted in a Python 3.10 environment. For data processing, the pandas and numpy libraries were utilized; for modeling, scikit-learn and XGBoost libraries were used; and for visualization, matplotlib and seaborn libraries were employed. The SHAP analysis was implemented using the SHAP library (version 0.43.1).

RESULTS

In the initial stage of the study, the correlation results of the morphological features for the examined variables were presented as a heatmap, as shown in Figure 1.

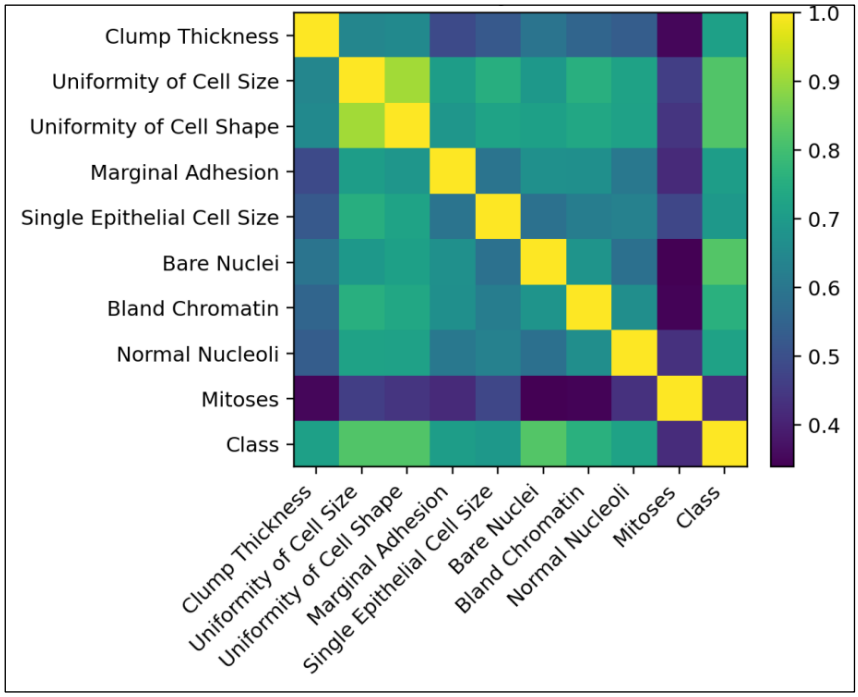


FIGURE 1: Correlation matrix

Examination of the correlation matrix in Figure 1 indicates that most variables exhibit moderate to high positive correlations. Notably, the strong correlations among Uniformity of Cell Size, Uniformity of Cell Shape, and Clump Thickness suggest that morphological homogeneity across these parameters follows a consistent pattern. This finding implies that cell morphology-related features tend to increase concurrently and reflect closely interrelated structural characteristics of tumor cells. In contrast, the Mitoses variable

demonstrates weak correlations with other features, indicating that mitotic activity may represent a relatively independent biological indicator. Overall, the results suggest that morphological consistency is strongly associated with malignancy risk, and that jointly considering these variables may enhance classification performance.

The confusion matrices corresponding to the models employed in the study are presented in Figure 2.

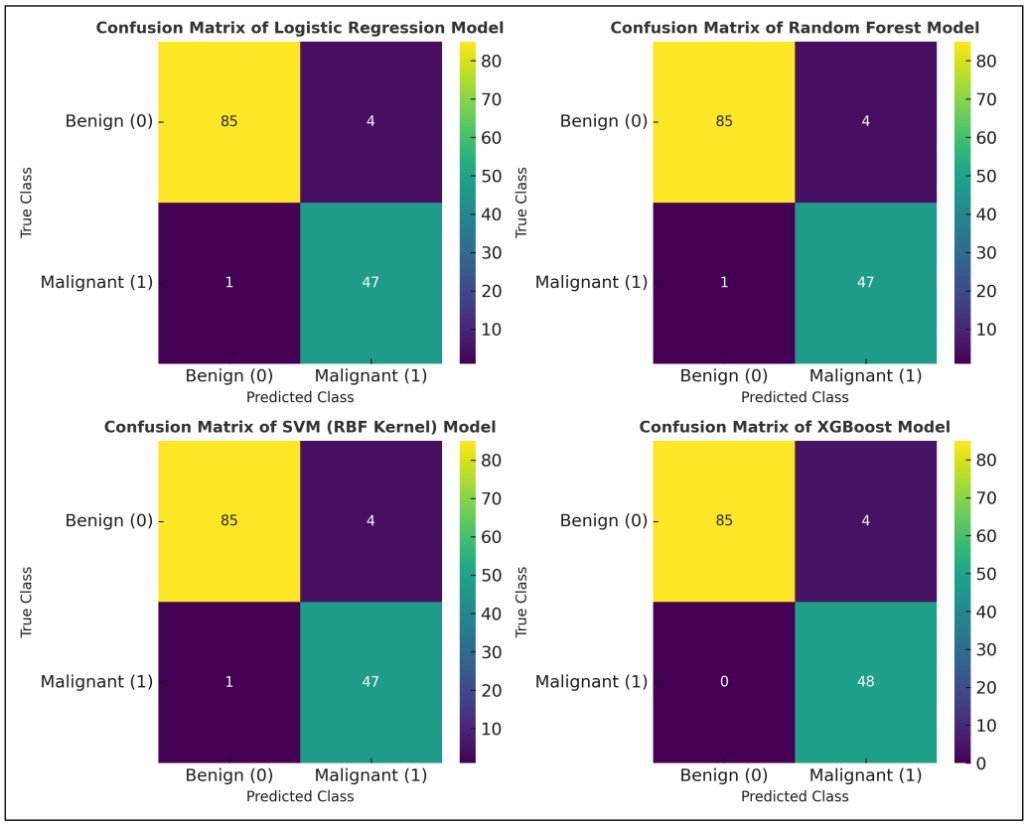


FIGURE 2: Confusion matrices of classification models

SVM: Support vector machines; RBF: Recursive feature elimination

The confusion matrices presented in Figure 2 illustrate the classification performance of four models - logistic regression, Random Forest, SVM (RBF kernel), and XGBoost- in distinguishing benign (0) and malignant (1) cases. Statistically, all models achieved very high overall accuracy with minimal misclassification. Logistic regression, Random Forest, and SVM-RBF yielded identical results, correctly classifying 85 benign and 47 malignant samples, while misclassifying 4 benign and 1 malignant case, reflecting a high sensitivity for the malignant class. The XGBoost model demonstrated superior performance by correctly classifying all 48 malignant cases and 85 benign cases, with only 4 false positives. This result indicates the most favorable balance between recall and precision among the evaluated models. Overall, while all algorithms exhibited excellent classification capability, XGBoost achieved slightly higher predictive precision and sensitivity, minimizing Type II errors (false negatives) and providing the most robust diagnostic reliability for this dataset.

Figure 3 shows the ROC curve graphs for the classification models.

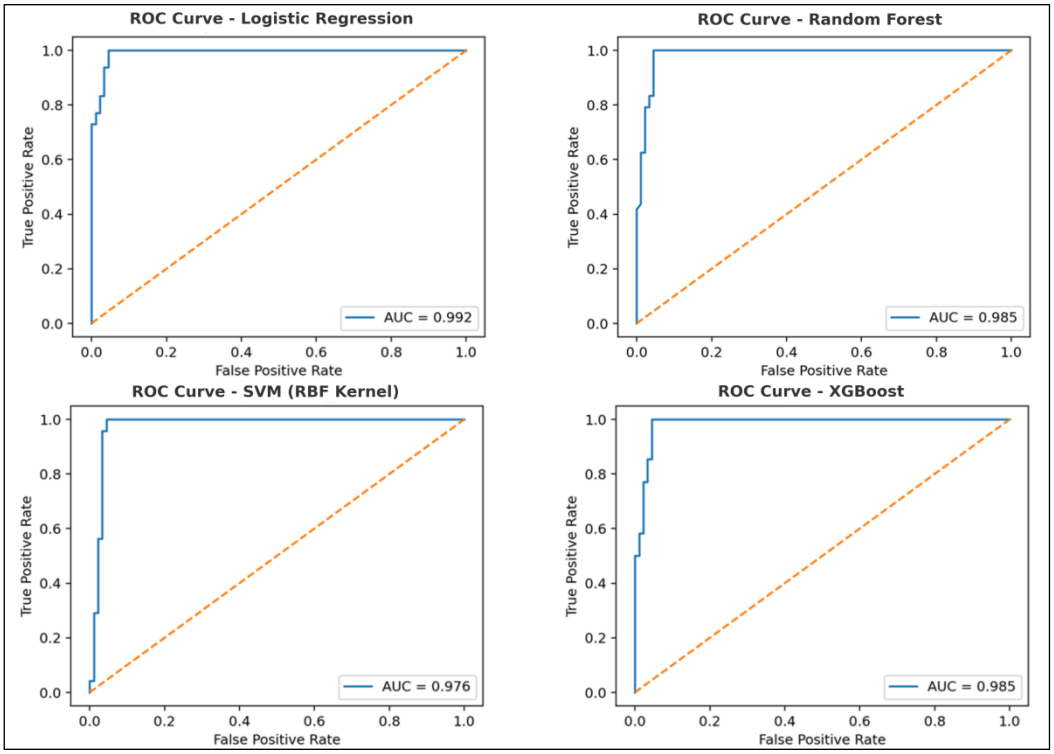


FIGURE 3: ROC curve of classification models

ROC: Receiver operating characteristic; SVM: Support vector machines; RBF: Recursive feature elimination; AUC: Area under the curve

The ROC curves presented in Figure 3 illustrate the discriminatory performance of the 4 models in distinguishing benign and malignant cases. All models achieved notably high AUC values, indicating excellent separability. Logistic regression demonstrated the highest AUC (0.992), reflecting near-perfect classification with a strong balance between sensitivity and specificity. Random forest and XGBoost followed closely, each achieving an AUC of 0.985, indicating robust discriminative ability and generalization performance. The SVM-RBF model yielded a slightly lower but still high AUC of 0.976, confirming effective classification capability. Statistically, AUC values exceeding 0.97 across all models indicate reliable discrimination between malignant and benign cases with minimal misclassification risk. Overall, these results highlight that the applied algorithms, particularly logistic regression and XGBoost, provide high diagnostic accuracy for the morphological dataset.

The performance metrics of the models used in the study are summarized in Table 2 in terms of accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC-AUC criteria.

TABLE 2: Summary of model performances

Model	Accuracy	Precision	Recall (sensitivity)	Specificity	F1	ROC_AUC
Logistic regression	0.964	0.922	0.979	0.955	0.949	0.992
Random forest	0.964	0.922	0.979	0.955	0.949	0.985
SVM_RBF	0.964	0.922	0.979	0.955	0.949	0.976
XGBoost	0.971	0.923	1.000	0.955	0.96	0.985

ROC: Receiver operating characteristic; AUC: Area under the curve; SVM: Support vector machines; RBF: Recursive feature elimination

When Table 2 is examined, it is observed that all classification models demonstrate high and consistent performance. Accuracy values range from 0.964 to 0.971, with logistic regression, Random Forest, and SVM-RBF producing identical results across the main performance metrics (accuracy=0.964, precision=0.922, recall/sensitivity=0.979, specificity=0.955, and F1-score=0.949). This indicates that these models provide balanced classification performance and stable generalization capability. The XGBoost model shows a modest yet noteworthy advantage over the other methods. It achieves the highest accuracy (0.971) and perfect sensitivity (recall/sensitivity=1.000), indicating that all malignant cases were correctly identified. Importantly, the specificity value (0.955) remains comparable to that of the other models, suggesting that the increase in sensitivity does not come at the cost of a higher false-positive rate. From a clinical perspective, where minimizing false-negative classifications is of critical importance, this characteristic represents a meaningful advantage. Overall, while all algorithms exhibit strong discriminative performance, XGBoost provides a slight statistical improvement in sensitivity and overall classification accuracy.

TABLE 3: Feature selection summary table

Feature	Selected by RFE?	RFE rank	Selected by RF?	RF importance	Mutual information
Uniformity of cell size	No	5	Yes	0.289	0.492
Uniformity of cell shape	Yes	1	Yes	0.226	0.473
Bare nuclei	Yes	1	Yes	0.152	0.421
Bland chromatin	Yes	1	Yes	0.108	0.379
Single epithelial cell size	No	4	Yes	0.082	0.352
Normal nucleoli	No	3	No	0.061	0.335
Clump thickness	Yes	1	No	0.047	0.330
Marginal adhesion	Yes	1	No	0.029	0.337
Mitoses	No	2	No	0.007	0.142

RFE: Recursive feature elimination; RF: Random forest

Examination of Table 3 indicates that Uniformity of Cell Shape, Bare Nuclei, Bland Chromatin, Clump Thickness, and Marginal Adhesion were consistently identified as the most influential morphological predictors across multiple feature selection methods. In particular, Uniformity of Cell Shape, Bare Nuclei, and Bland Chromatin ranked among the top variables in both RFE and Random Forest analyses, supported by relatively high MI values (0.473, 0.421, and 0.379, respectively). These results underscore the critical role of morphological uniformity and nuclear integrity in distinguishing malignant from benign cells. Additionally, Uniformity of Cell Size and Single Epithelial Cell Size showed moderate importance, with notable contributions in the Random Forest model (0.289 and 0.082, respectively). In contrast, Mitoses exhibited the lowest importance across all methods (RF importance=0.007; MI=0.142), indicating a limited contribution to predictive performance in this dataset. Overall, the findings demonstrate that features related to cell size and shape uniformity, nuclear smoothness, and adhesion characteristics provide the highest discriminative power for malignancy classification.

The bar chart illustrates the feature importances according to the Random Forest method which are given in Figure 4.

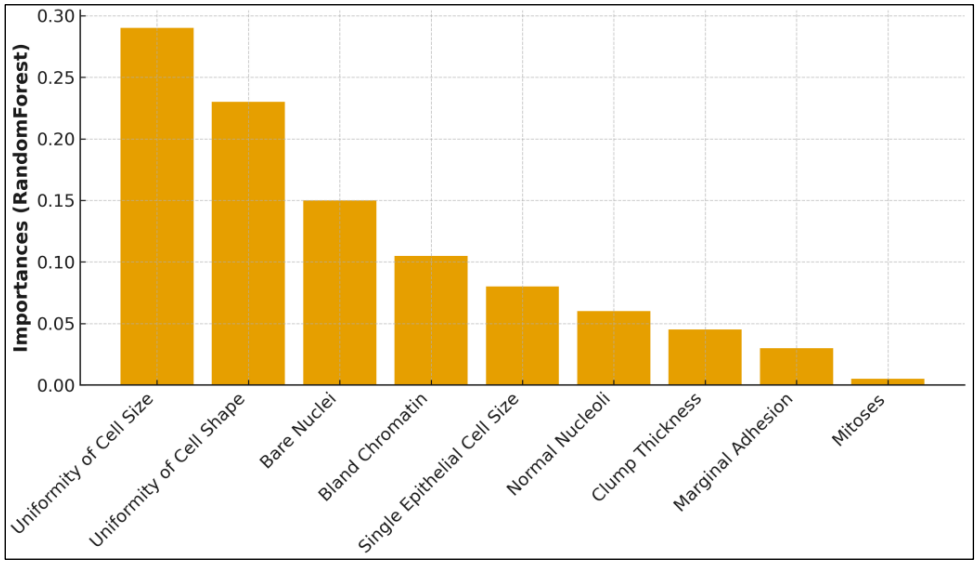


FIGURE 4: Feature importances according to the Random Forest method

Figure 4 illustrates the variable importance scores derived from the Random Forest model. The results indicate that Uniformity of Cell Size and Uniformity of Cell Shape are the most influential predictors, suggesting that loss of morphological uniformity is a strong indicator of malignancy. Bare Nuclei and Bland Chromatin also exhibit high importance, highlighting the critical role of nuclear structure and chromatin organization in tumor characterization. Variables such as Single Epithelial Cell Size, Normal Nucleoli, and Clump Thickness contribute at a moderate level, whereas Mitoses shows the lowest contribution to the model. Overall, the findings emphasize that morphological homogeneity and nuclear features are the primary determinants within the Random Forest decision structure.

SHAP analysis was employed to enhance model interpretability by quantifying the direction and magnitude of each feature's contribution to predictions. Accordingly, Figure 5 presents the SHAP summary plot illustrating the relative influence and effect direction of the morphological variables included in the study.

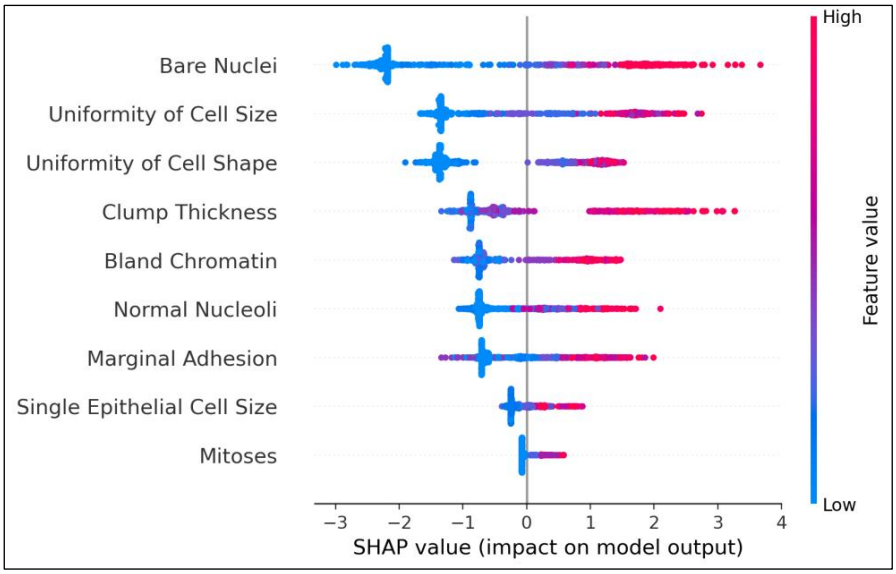


FIGURE 5: The SHAP summary plot for the XGBoost model

Figure 5 presents the SHAP summary plot for the XGBoost model. The results indicate that Bare Nuclei, Uniformity of Cell Size, and Uniformity of Cell Shape exert the strongest influence on model predictions. Higher values of these variables (red points) substantially increase the probability of classification into the malignant class, indicating that nuclear irregularities and loss of cellular morphological uniformity are key drivers of malignancy prediction. In contrast, Mitoses, Single Epithelial Cell Size, and Marginal Adhesion display relatively low SHAP values, suggesting a limited contribution to the model's decision process. Overall, the XGBoost model predominantly relies on features associated with nuclear structure and cellular homogeneity to discriminate malignant from benign cases. The ranking of variables based on their mean SHAP values is presented in Figure 6.

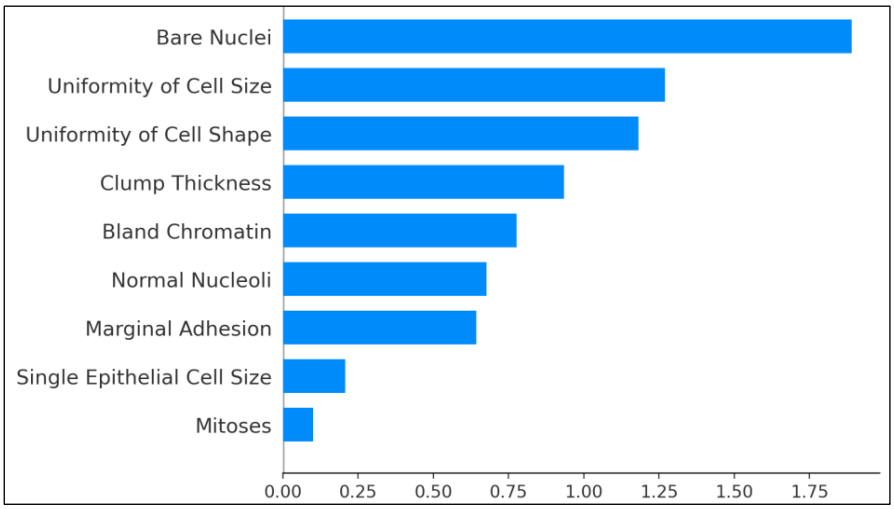


FIGURE 6: Variable importance ranking based on mean SHAP values

Figure 6 presents the variable importance ranking for the XGBoost model based on mean SHAP values, reflecting the average contribution of each feature to the model output. Bare Nuclei exhibits the highest mean SHAP value, indicating its dominant role in the decision-making process. This is followed by Uniformity of Cell Size and Uniformity of Cell Shape, highlighting that structural irregularities in cells are the primary factors associated with increased malignancy risk. Clump Thickness and Bland Chromatin demonstrate moderate influence, whereas Single Epithelial Cell Size and Mitoses contribute minimally. Overall, the results indicate that features related to nuclear structure and morphological uniformity are the most clinically relevant determinants in distinguishing malignant from benign cases.

DISCUSSION

In this study, logistic regression, Random Forest, SVM-RBF, and XGBoost were applied to classify benign and malignant breast tumors using cytomorphological features. Despite relying on different statistical principles, all models demonstrated high accuracy, sensitivity, and generalization performance, with ROC-AUC values exceeding 0.97, indicating strong clinical discriminative capability. Among the evaluated algorithms, XGBoost achieved the highest performance (97.1% accuracy and 100% sensitivity), highlighting the advantage of ensemble-based gradient boosting methods in capturing complex feature interactions.

Logistic regression provided a strong and stable baseline model, while Random Forest benefited from ensemble aggregation to reduce variance. The SVM-RBF model effectively handled nonlinear decision boundaries; however, XGBoost exhibited the most notable improvement due to its sequential error-minimization framework. Feature selection analyses using RFE, Random Forest importance, and MI pro-

duced complementary results, consistently identifying Uniformity of Cell Size, Uniformity of Cell Shape, Bare Nuclei, and Bland Chromatin as the most influential predictors. In contrast, Mitoses showed limited standalone importance, suggesting that mitotic activity is informative primarily when evaluated in conjunction with other morphological indicators.

The Random Forest importance scores and SHAP analysis of the XGBoost model further reinforced these findings. SHAP results demonstrated that Bare Nuclei, Uniformity of Cell Size, and Uniformity of Cell Shape exert the strongest influence on malignancy prediction, confirming that nuclear structure and cellular homogeneity are central determinants in classification. These results are consistent with earlier studies on the Wisconsin Breast Cancer dataset, which also identified Bare Nuclei and Uniformity of Cell Shape as key discriminative features.[4-6]

Beyond model performance, this study specifically aimed to clarify how different feature selection strategies converge or diverge in identifying clinically relevant morphological indicators. Unlike many previous studies that rely on a single feature importance measure, the simultaneous use of RFE, Random Forest-based importance, and MI allowed for a more robust assessment of feature stability across methodological perspectives. Such multi-method feature selection strategies have been emphasized in the literature as a way to reduce method-specific bias and improve the reliability of selected predictors.[8,27] The consistency observed among these methods in the present study strengthens the robustness of the identified morphological features.

From a biological standpoint, the prominence of features related to cell size uniformity, cell shape irregularity, and nuclear structure aligns well with established histopathological evidence. Increased variability in cell and nuclear morphology reflects loss of differentiation and genetic instability, which are recognized hallmarks of malignant transformation.[12,14] The persistence of these features across different models and explainability analyses suggest that the ML algorithms capture biologically meaningful patterns rather than spurious statistical associations.

Beyond predictive accuracy, the main contribution of this study lies in the joint evaluation of multiple feature selection techniques and the integration of SHAP-based explainability, enabling transparent interpretation of model decisions and addressing the “black box” limitation of ML in medical applications.

Nevertheless, several limitations of the present study should be acknowledged. The analysis was conducted on a single, well-established dataset containing a limited number of cytomorphological variables, without incorporating additional clinical, molecular, or imaging-derived features. Moreover, the study did not include large-scale simulation experiments or external validation on independent datasets. Previous studies have highlighted the importance of multi-dataset validation and richer feature spaces to improve generalizability and clinical impact of ML models in biomedical applications.[28] These limitations should therefore be considered when interpreting the findings.

Recent studies that aim to conduct comprehensive comparisons of ML methods typically employ a large number of classifiers, extensive simulation designs, and multi-dataset validation frameworks. For example, performed a broad comparative analysis of supervised classification models for heart disease prediction by incorporating multiple algorithms and external dataset validation strategies.[29] In contrast, the present study does not aim to provide an exhaustive benchmarking of ML methods. Instead, it focuses on the consistency of feature selection results, interpretability of model decisions, and the biological relevance of cytomorphological predictors within a limited but representative set of classification models. Future research may extend the proposed framework by incorporating broader algorithmic comparisons, simulation-based evaluations, and multi-dataset validation approaches to further assess the generalizability of explainable ML models in breast cancer diagnosis.

Future studies should consider larger and more diverse datasets and explore deep learning or hybrid models, as well as complementary explainability methods such as Local Interpretable Model-agnostic Explanations or Integrated Gradients, to further enhance robustness and interpretability.

CONCLUSION

In conclusion, this study demonstrates that combining classical statistical models with modern ML approaches yields high accuracy and strong interpretability in breast cancer classification based on cytomorphological features. The findings confirm that variables related to nuclear structure and morphological uniformity carry the highest diagnostic value. By integrating robust performance evaluation with explainable AI techniques, the study not only corroborates existing literature but also contributes to the development of reliable and clinically interpretable decision-support systems for breast cancer diagnosis.

Importantly, the contribution of this study extends beyond predictive accuracy by emphasizing interpretability and methodological consistency in feature selection. The integrated use of multiple feature selection approaches together with SHAP-based explainability provides a transparent framework for understanding model decisions, addressing the "black-box" limitation commonly associated with ML models in medical research.[9,10] Such transparency has been identified as a key requirement for the clinical adoption of AI-based decision support systems.[28]

Future research should focus on integrating additional clinical, histopathological, and molecular variables, validating the proposed framework on independent and multi-center datasets, and employing simulation-based comparative designs. Furthermore, complementary explainability methods such as LIME or Integrated Gradients may be explored to provide alternative perspectives on model behavior and enhance interpretability.[30] These extensions would further strengthen the translational potential of explainable ML approaches in breast cancer diagnosis.

Ethics Committee Approval

Ethical approval was not required because the study used openly accessible, fully anonymized secondary data and involved no direct human participation.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Berkalp Tunca, Sinan Saraçlı;* ***Design:*** *Berkalp Tunca, Sinan Saraçlı;* ***Control/Supervision:*** *Sinan Saraçlı, Berkalp Tunca;* ***Data Collection and/or Processing:*** *Berkalp Tunca, Sinan Saraçlı;* ***Analysis and/or Interpretation:*** *Berkalp Tunca, Sinan Saraçlı;* ***Literature Review:*** *Berkalp Tunca;* ***Writing the Article:*** *Berkalp Tunca;* ***Critical Review:*** *Sinan Saraçlı.*

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-49. [Crossref] [PubMed]
2. Sugimoto M, Hikichi S, Takada M, Toi M. Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. Annals of Breast Surgery. 2023;7. [Crossref]
3. Chaurasia V, Pal S. A novel approach for breast cancer detection using data mining techniques. International Journal of Innovative Research in Computer and Communication Engineering. 2017;5(1):1036–1041. [Crossref]
4. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci U S A. 1990;87(23):9193-6. [Crossref] [PubMed] [PMC]
5. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007;17(4):694-701. [Crossref]

6. Sahan S, Polat K, Kodaz H, Güneş S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med. 2007;37(3):415-23. [Crossref] [PubMed]
7. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46(1-3):389-422. [Crossref]
8. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Computing and Applications. 2014;24:175-86. [Crossref]
9. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NIPS 2017) 30; 2017. p. 4765-74. [Link]
10. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des. 2020;34(10):1013-26. [PubMed] [PMC]
11. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med. 2011;3(108):108ra113. [Crossref] [PubMed]
12. Stenkvist B, Westman-Naeser S, Holmquist J, Nordin B, Bengtsson E, Vegelius J, et al. Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations. Cancer Res. 1978;38(12):4688-97. [PubMed]
13. Abdalla F, Boder J, Markus R, Hashmi H, Buhmeida A, Collan Y. Correlation of nuclear morphometry of breast cancer in histological sections with clinicopathological features and prognosis. Anticancer Res. 2009;29(5):1771-6. [PubMed]
14. Kashyap A, Jain M, Shukla S, Andley M. Role of Nuclear Morphometry in Breast Cancer and its Correlation with Cytomorphological Grading of Breast Cancer: A Study of 64 Cases. J Cytol. 2018;35(1):41-5. . [Crossref] [PubMed] [PMC]
15. Rawat RR, Ruderman D, Macklin P, Rimm DL, Agus DB. Correlating nuclear morphometric patterns with estrogen receptor status in breast cancer pathologic specimens. NPJ Breast Cancer. 2018;4:32. [Crossref] [PubMed] [PMC]
16. Whitney J, Corredor G, Janowczyk A, Ganesan S, Doyle S, Tomaszewski J, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. BMC Cancer. 2018;18(1):610. [Crossref] [PubMed] [PMC]
17. Lu C, Romo-Bucheli D, Wang X, Janowczyk A, Ganesan S, Gilmore H, Rimm D, Madabhushi A. Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers. Lab Invest. 2018 Nov;98(11):1438-1448. 2018 Jun 29. [Crossref] [PubMed] [PMC]
18. Saednia K, Lagree A, Alera MA, Fleshner L, Shiner A, Law E, et al. Quantitative digital histopathology and machine learning to predict pathological complete response to chemotherapy in breast cancer patients using pre-treatment tumor biopsies. Sci Rep. 2022;12(1):9690. [Crossref] [PubMed] [PMC]
19. Bansal C, Pujani M, Sharma KL, Srivastava AN, Singh US. Grading systems in the cytological diagnosis of breast cancer: a review. J Cancer Res Ther. 2014 Oct-Dec;10(4):839-45. [Crossref] [PubMed]
20. Ibrahim A, Gamble P, Jaroensri R, Abdelsamea MM, Mermel CH, Chen PC, Rakha EA. Artificial intelligence in digital breast pathology: Techniques and applications. Breast. 2020 Feb;49:267-273. 2019 Dec 19. [Crossref] [PubMed] [PMC]
21. University of California, Irvine Machine Learning Repository [Internet]. [Cited: November 12, 2025]. Breast Cancer Wisconsin Data Set. 2024. Available from: [Link]
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-30. [Link]
23. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. New York: John Wiley & Sons; 2013.
24. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32. [Crossref]
25. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97. [Crossref]
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p.785-94. [Crossref]
27. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003;3:1157-82. [Link]
28. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019;17:195. [Crossref]
29. Ogundepo EA, Yahya WB. Performance analysis of supervised classification models on heart disease prediction. Innovations in Systems and Software Engineering. 2023;19(1):129-44. [Crossref]
30. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD Conference; 2016. p.1135-44. [Crossref]