JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 8, No. 3. Juni 2026 DOI: https://doi.org/10.34288/jri.v8i3.534
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

BREAST TUMOR CLASSIFICATION USING RANDOM FOREST WITH FEATURE SELECTION AND GRIDSEARCHCV OPTIMIZATION

Priscilia Amanda Leza[-1], Mutaqin Akbar[-2]

Program Studi Informatika
Fakultas Teknologi Informasi
Universitas Mercu Buana Yogyakarta
221110004@student.mercubuana-yogya.ac.id[-1], mutaqin@mercubuana-yogya.ac.id[-2]

Abstract

Breast tumor classification into benign and malignant categories is an important challenge in the medical field because diagnostic errors can lead to delayed treatment or unnecessary medical procedures. This study aims to analyze the performance of Random Forest and evaluate the effects of feature selection and GridSearchCV hyperparameter optimization on breast tumor classification. The study used the Wisconsin Breast Cancer Diagnostic Dataset, consisting of 569 samples with 30 numerical features extracted from Fine Needle Aspiration (FNA) examinations. Four sequential Random Forest model configurations were compared: baseline Random Forest, Random Forest with feature selection, Random Forest with GridSearchCV optimization, and the integration of feature selection with GridSearchCV. Feature selection was performed using feature importance scores with ROC-AUC-based cross-validation to determine the optimal feature subset. Model evaluation was conducted using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and train-test gap. The results showed that all models achieved the same accuracy of 97.37%, precision of 1.0000, recall of 0.9286, and F1-score of 0.9630. However, the integrated model achieved the highest ROC-AUC of 0.9977 with the smallest train-test gap of 0.0241 while reducing the number of features from 30 to 15. These findings indicate that integrating feature selection and GridSearchCV improves model stability, efficiency, and discriminative capability without reducing classification performance, addressing the limitation of prior studies that applied these techniques separately.

Keywords: Breast Tumor Classification; Random Forest; Feature Importance; Hyperparameter Optimization; GridSearchCV

Abstrak

Klasifikasi tumor payudara ke dalam kategori jinak dan ganas merupakan tantangan penting dalam bidang medis karena kesalahan diagnosis dapat menyebabkan keterlambatan penanganan maupun prosedur medis yang tidak diperlukan. Penelitian ini bertujuan untuk menganalisis kinerja algoritma Random Forest serta mengevaluasi pengaruh seleksi fitur dan optimasi hyperparameter GridSearchCV terhadap klasifikasi tumor payudara. Dataset yang digunakan adalah Wisconsin Breast Cancer Diagnostic Dataset yang terdiri dari 569 sampel dengan 30 fitur numerik hasil pemeriksaan Fine Needle Aspiration (FNA). Penelitian membandingkan empat konfigurasi model Random Forest secara berurutan, yaitu baseline Random Forest, Random Forest dengan feature selection, Random Forest dengan optimasi GridSearchCV, serta integrasi feature selection dan GridSearchCV. Seleksi fitur dilakukan menggunakan feature importance dengan cross-validation berbasis ROC-AUC untuk menentukan subset fitur optimal. Evaluasi model dilakukan menggunakan accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, dan train-test gap. Hasil penelitian menunjukkan bahwa seluruh model memperoleh accuracy sebesar 97,37%, precision 1,0000, recall 0,9286, dan F1-score 0,9630. Namun, model integrasi menghasilkan ROC-AUC tertinggi sebesar 0,9977 dengan train-test gap terkecil sebesar 0,0241 serta berhasil mengurangi jumlah fitur dari 30 menjadi 15 fitur. Hasil tersebut menunjukkan integrasi feature selection dan GridSearchCV mampu meningkatkan stabilitas, efisiensi, dan kemampuan diskriminasi model tanpa menurunkan performa klasifikasi, sekaligus menjawab keterbatasan penelitian sebelumnya yang masih menerapkan kedua teknik tersebut secara terpisah.

Kata kunci: Klasifikasi Tumor Payudara; Random Forest; Seleksi Fitur; Optimasi Hyperparameter; GridSearchCV

INTRODUCTION

Distinguishing between benign and malignant breast tumors is currently a major challenge in the medical field, which demands high diagnostic accuracy to avoid fatal clinical risks (I. I. Lestari & Homaidi, 2024). Inaccuracies in the classification process can have serious implications for patient safety, where a benign tumor misdiagnosed as malignant triggers unnecessary medical procedures and anxiety, while a malignant tumor misdiagnosed as benign (false negative) risks causing delays in the treatment of a deadly disease (Hulaifah Al Abrori & Subhiyakto, 2025; Paepke dkk., 2018).

This issue is supported by the clinical fact that benign breast tumors (Breast Benign Diseases) occur 10 times more frequently than malignant tumors; however, some types of benign tumors still have the potential to develop into malignancies in the future, so every mass of cells requires very careful identification of parameters (N. S. Lestari dkk., 2023). The importance of this research is further reinforced by national data showing that breast cancer ranks first in terms of the highest number of cancer cases in Indonesia, with 68,858 new cases (16.6%), of which approximately 70% to 80% are already detected at an advanced stage due to the low effectiveness of early detection (Fauzi dkk., 2020; Kementerian Kesehatan RI, 2022). Breast tumors are clinically classified into two main categories: benign, which grow slowly, are well-defined, and do not invade surrounding tissue, and malignant, which are invasive and have the potential to metastasize to other organs (Jamil dkk., 2024; Suparna & Kartika, 2022). This situation indicates that the ability to identify and classify tumors at an early stage accurately is a key determinant of improving patient survival and reducing mortality, which exceeds 22,000 deaths (Fauzi dkk., 2020).

Conventional diagnostic methods such as physical examinations, mammography, and fine-needle aspiration (FNA) biopsies often face efficiency challenges due to time-consuming laboratory processes and the need for high-precision interpretation by specialists (Chazar & Widhiaputra, 2020; Kementerian Kesehatan RI, 2018; Zulkarnain & Delyuzar, 2017). In response to these limitations, machine learning-based methods have emerged as promising tools that assist medical professionals in automatically, objectively, and comprehensively detecting unique patterns in medical data (Fauzi dkk., 2020; Magda dkk., 2025).

Advances in machine learning offer solutions for disease classification based on medical data by leveraging patterns in the available data and improving accuracy as the learning process progresses (Ali dkk., 2023). This is relevant because medical data contains many complex, interrelated variables that require effective analytical methods (Nasution & Juledi, 2025). Various algorithms have been used in breast tumor classification, such as Support Vector Machine (SVM), Naïve Bayes, Decision Tree, and Random Forest (Armoogum dkk., 2024; Tjengharwidjaja dkk., 2024), where Random Forest is known as a highly reliable ensemble learning method for tumor classification due to its ability to handle high-dimensional data and features with non-linear relationships without easily suffering from overfitting (Misdiantoro & Susanto, 2025; Yaqoob dkk., 2025). Random Forest works by building a large number of decision trees independently during the training phase and then aggregating the final classification results via majority voting (Ali dkk., 2023; Bimo, 2024; Nguyen dkk., 2013).

Several studies have explored the use of Random Forests for breast tumor classification, using various approaches. The study conducted by (Minnoor & Baths, 2023) utilized the Wisconsin Breast Cancer dataset and applied feature subset selection and model optimization, achieving an accuracy of over 99%. Another study by (Jinana dkk., 2025) used Random Forests with cross-validation and achieved approximately 95% accuracy in distinguishing between benign and malignant tumors. Furthermore, a study by (Kabir dkk., 2023) used feature importance to identify the most influential features in breast tumor classification, thereby providing insight into each feature's contribution to the model. Meanwhile, (Ali dkk., 2023) demonstrated that hyperparameter optimization using GridSearchCV can substantially enhance the predictive capability of classification algorithms. A study by (Nur Fauzi dkk., 2025) reported that in credit risk classification using Random Forest, feature engineering increased accuracy from 92.56% to 97.81%. In contrast, subsequent GridSearchCV tuning produced only a minimal gain to 97.94%. This indicates that the effectiveness of hyperparameter tuning can vary depending on data characteristics, especially since breast cancer diagnostic data has unique distributional characteristics compared to financial data, making specific testing on medical data still relevant.

(Setiawan dkk., 2026) demonstrated that hyperparameter tuning of Random Forest without

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.534
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

feature selection yields relatively unchanged performance, whereas (Premalatha dkk., 2025) applied feature selection without hyperparameter optimization. Both approaches are still applied separately, so it remains unknown whether integrating them, specifically feature selection based on feature importance followed by GridSearchCV hyperparameter optimization, can yield a more effective and stable Random Forest model on medical data. Therefore, this study proposes a breast tumor classification model using the Wisconsin Breast Cancer Diagnostic dataset, integrating feature-importance-based feature selection and Random Forest hyperparameter optimization via GridSearchCV. This integration aims to produce a more stable, efficient, and reliable model for distinguishing benign from malignant tumors. The main scientific contribution of this study is the integration of feature-importance-based feature selection and GridSearchCV optimization within a Random Forest framework for breast tumor classification, demonstrating that the combined approach can improve model stability and discriminative capability while reducing feature dimensionality without sacrificing classification accuracy

RESEARCH METHODS

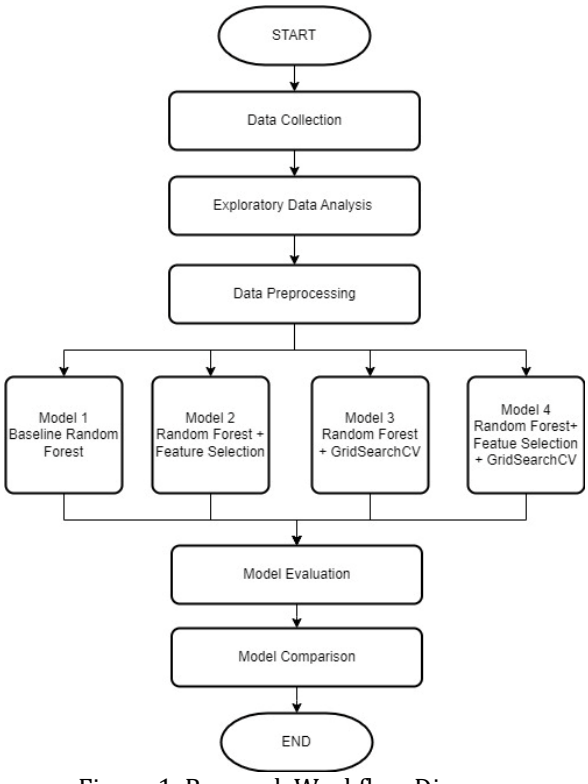

Figure 1. Research Workflow Diagram

This study is a comparative experimental study that systematically compares four Random Forest model configurations in sequence. This design allows the impact of each technique to be measured separately before the two approaches are combined: Model 1 serves as the baseline, implementing a standard Random Forest classifier with default hyperparameters, without any feature selection or optimization; Model 2 applies feature selection without hyperparameter optimization; Model 3 applies hyperparameter optimization without feature selection; and Model 4 integrates both approaches as the final model.

Table 1. Experimental Design of Four Model Configurations

Model	Features	FS	GSCV	Score
M1: Baseline RF	30	No	No	-
M2: RF + FS	15	Yes	No	-
M3: RF + GSCV	30	No	Yes	Recall
M4: RF + FS + GSCV	15	Yes	Yes	Recall

Based on the experimental design shown in Table 1, each model configuration was evaluated sequentially to analyze the individual and combined effects of feature selection and GridSearchCV optimization on Random Forest performance.

Dataset

The dataset used in this study is the Wisconsin Breast Cancer Diagnostic Dataset, sourced from the UCI Machine Learning Repository and available on the Kaggle platform at uciml/breast-cancer-wisconsin-data. This dataset contains 569 samples from Fine Needle Aspiration (FNA) examinations. It is widely used in machine learning research in the medical field due to its high data quality and ability to represent the characteristics of breast tumor cell nuclei in detail (Minnoor & Baths, 2023; Setiawan dkk., 2026). The dataset comprises 30 numerical features, consisting of 357 benign cases (62.7%) and 212 malignant cases (37.3%). The target variable is the diagnosis,

with the classes malignant (M) and benign (B). The dataset contains no missing values.

Data Preprocessing

The preprocessing steps include removing irrelevant columns, namely the ID column and Unnamed: 32, checking for missing values, and encoding the target variable, where M is converted to 1 and B to 0. The dataset was split into 80% training and 20% test data using train_test_split with stratify=y to preserve class proportions and random_state=42 to ensure reproducibility. Normalization was not applied because Random Forest is not sensitive to feature scale.

Feature Selection Based on Feature Importance

Feature selection was performed using the feature importance scores generated by Model 1. Features were sorted by score from highest to lowest, and the optimal number of features was determined using cross-validation (StratifiedKFold, k=5) with the ROC-AUC metric over the range n=5 to n=20. The value of n that yielded the highest average CV AUC was set as the optimal number of features and used as input for Models 2 and 4.

GridSearchCV Hyperparameter Optimization

Model 3 and Model 4 use identical GridSearchCV configurations, deliberately set to ensure that any performance difference between the two models is attributable solely to the presence or absence of feature selection. The parameter n_estimators was set to 100–300, as a greater number of trees produces more stable predictions; (Jinana dkk., 2025) demonstrated that n_estimators=100 and max_depth=10 yield good performance in breast cancer classification. The parameter max_depth was set to [10, 20, None] to allow trees sufficient depth to capture important patterns, while min_samples_split and min_samples_leaf were limited to [2, 5] and [1, 2] respectively to prevent overfitting and maintain generalization, supported by (Setiawan dkk., 2026) who showed that Random Forest on breast cancer data achieves high performance without overly strict tree constraints. The parameter max_features was included, as (Premalatha dkk., 2025) emphasize that appropriate feature selection improves model performance. The criterion parameter tests both Gini and Entropy impurity functions, as (Zhu dkk., 2025) explain that the Gini Index is commonly used for optimal node splitting in Random Forest. Finally, class_weight='balanced' was included because breast cancer datasets tend to be imbalanced, with benign cases outnumbering malignant ones, making the model more sensitive to malignant tumors (Hulaifah Al Abrori & Subhiyakto, 2025).

Recall-based scoring was applied in both models, as false negatives are more dangerous than false positives in a medical context, and high recall is critical for detecting all malignant tumor cases (Jinana dkk., 2025). This identical scoring also ensures a fairer comparison between Model 3 and Model 4, enabling the determination of whether feature selection contributes meaningfully to model performance. The complete parameter configurations are shown in Table 2.

Table 2. GridSearchCV Parameter Configuration on Model 3 and Model 4

Parameter	Value
n_estimators	[100, 200, 300]
max_depth	[10, 20, None]
min_samples_split	[2, 5]
min_samples_leaf	[1, 2]
max_features	['sqrt', 'log2']
criterion	['gini', 'entropy']
class_weight	[None, 'balanced']
Scoring	Recall
Cross-Validation	StratifiedKFold (k=5)

Model Evaluation

Each model was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and the train-test gap. The best model was determined based on a combination of performance, stability, and efficiency, not solely on the highest accuracy. In a medical context, models with high recall and high ROC-AUC are prioritized to minimize the risk of misdiagnosis in malignant tumors.

Research Tools and Environment

The research was implemented in Python via Google Colaboratory, using scikit-learn as the core modeling and evaluation library, with pandas and numpy handling data processing, and matplotlib, along with seaborn, for graphical visualization.

RESULTS AND DISCUSSION

Feature Selection Results

Figure 2 and Figure 3 present the results of feature subset evaluation and feature importance ranking.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.534
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

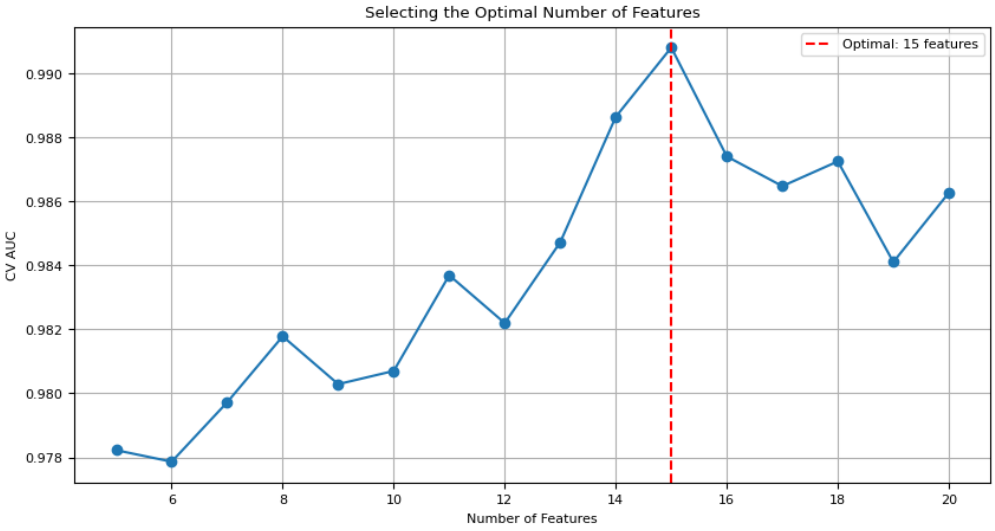

Figure 2. Cross-validation AUC scores for each feature subset size, with peak performance at n=15

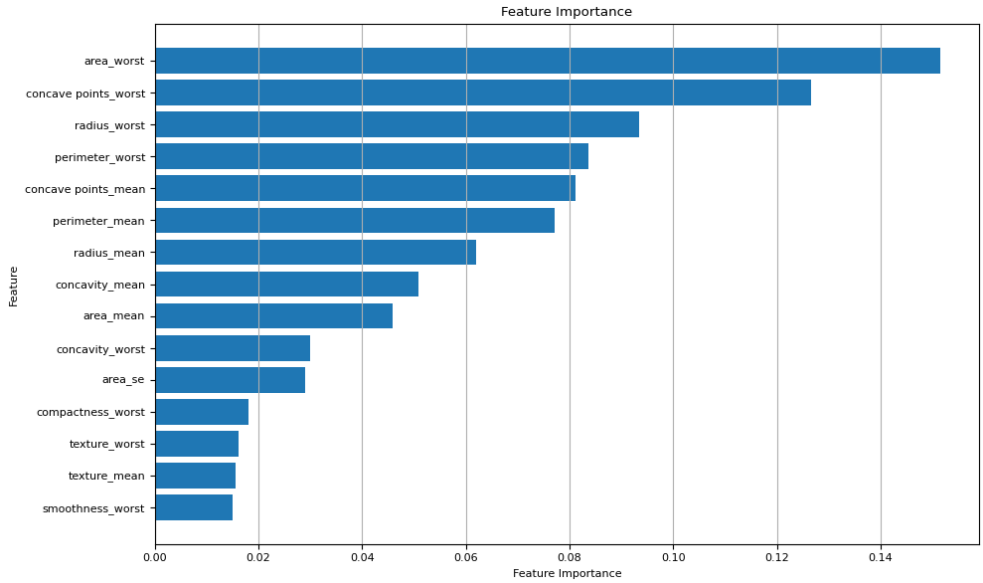

Figure 3. The 15 selected features' feature importance scores are ranked from highest to lowest.

The most influential feature was area_worst (0.1514), followed by concave points_worst and radius_worst, as shown in Figure 2. Clinically, area_worst represents the largest tumor cell nucleus area recorded in Fine Needle Aspiration (FNA) examination, where malignant tumors generally exhibit larger nuclear areas than benign tumors due to abnormal cell proliferation (Girdhar dkk., 2023). Similarly, concave points_worst reflects the severity of concave portions on the nucleus contour, while radius_worst measures the maximum nucleus radius. Both features are important morphological indicators associated with malignancy, as malignant tumors tend to exhibit larger nuclear size and greater boundary irregularity than benign tumors (Attique & Khan, 2025). Therefore, these features provide clinically relevant information for distinguishing benign and malignant breast tumors. These selected features were also used as inputs for Model 4 to evaluate the effect of feature selection consistently across different model configurations.

Model Evaluation

Table 3 presents the performance comparison of all four Random Forest model configurations using accuracy, precision, recall, F1-score, ROC-AUC, and train-test gap metrics.

Table 3. Performance Comparison of All Four Models on the Test Set

Model	Features	Test Acc	Train Acc	Gap	Precision	Recall	F1	ROC-AUC
M1: Baseline RF	30	0.9737	1.0000	0.0263	1.0000	0.9286	0.9630	0.9929
M2: RF + FS	15	0.9737	1.0000	0.0263	1.0000	0.9286	0.9630	0.9952
M3: RF + GSCV	30	0.9737	1.0000	0.0263	1.0000	0.9286	0.9630	0.9964
M4: RF + FS + GSCV	15	0.9737	0.9978	0.0241	1.0000	0.9286	0.9630	0.9977

Based on Table 3, all four models produced identical test accuracy, precision, recall, and F1-score values of 0.9737, 1.0000, 0.9286, and 0.9630, respectively, indicating that the Random Forest algorithm already achieves strong classification performance on this dataset even without further optimization. While these threshold-based metrics remain consistent across all configurations, meaningful differences emerge in ROC-AUC values, feature efficiency, and train-test stability.

ROC-AUC improved progressively from 0.9929 in Model 1 to 0.9952 in Model 2, 0.9964 in Model 3, and reached its highest value of 0.9977 in Model 4, suggesting that the combination of feature selection and GridSearchCV consistently enhances the model's probabilistic discriminative capability. Regarding generalization stability, Models 1, 2, and 3 each achieved a perfect training accuracy of 1.0000 with a train-test gap of 0.0263, while Model 4 produced a training accuracy of 0.9978 and the smallest gap of 0.0241. Although the difference is marginal, this trend is consistent with reduced mild overfitting. Additionally, feature selection reduced the number of input features from 30 to 15 without any loss in classification performance, yielding a more efficient model while maintaining full predictive capability.

ROC Curve and Best Model Analysis

The ROC curves for all four models are displayed in two panels in Figure 4: a full-view panel and a zoomed-in panel of the critical area. All models produced ROC curves well above the random classifier line, reflecting excellent classification capability in separating benign from malignant tumor categories. In the zoomed view, with a False Positive Rate (FPR) range from 0 to 0.12, the curve for Model 4 appears closest to the upper-left corner compared to the other models. This indicates that Model 4 has the best discriminative capability at low thresholds, which is clinically relevant because it helps reduce the risk of false negatives in malignant tumor cases. Model 4 also achieved the highest ROC-AUC of 0.9977, surpassing Models 1, 2, and 3, indicating that the combination of feature selection and GridSearchCV optimizes substantially enhances the models discriminative capacity between malignant and benign classes.

Figure 5 presents the confusion matrix for Model 4, the best-performing model. Most benign and malignant samples were classified correctly, with only a small number of misclassifications. This result supports the high recall and ROC-AUC values achieved by Model 4 in distinguishing the two classes.

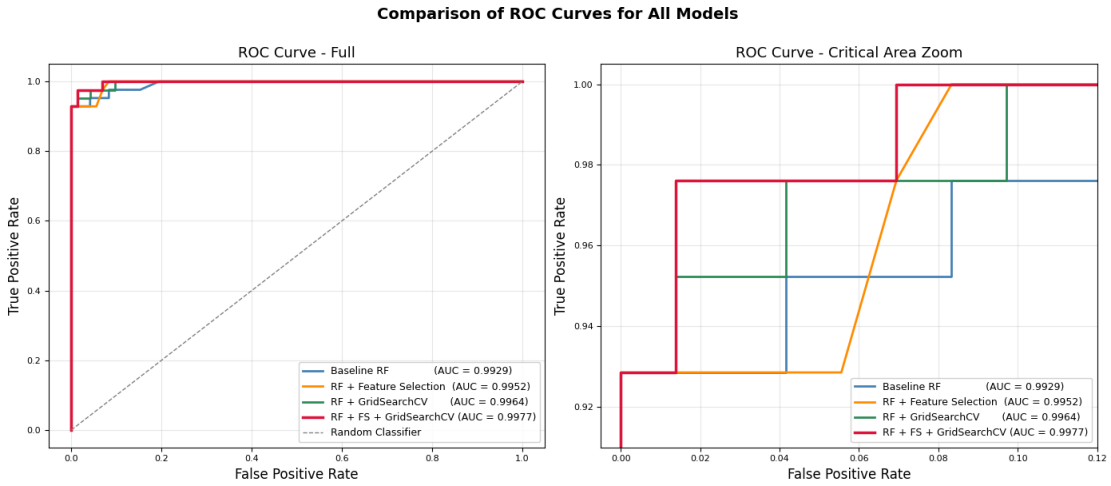

Figure 4. ROC curves of all four models: full view (left) and zoomed view of the critical area (right)

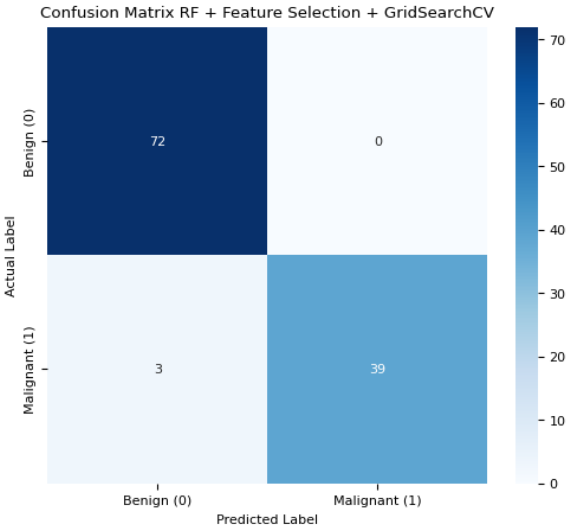

Figure 5. Confusion matrix of Model 4 (RF + Feature Selection + GridSearchCV).

Discussion

This study demonstrates that the Random Forest algorithm achieves very good classification performance on breast tumor data. All models demonstrated strong ability to distinguish benign and malignant tumors, with consistently high accuracy, precision, recall, and F1 scores. A precision of 1.0000 indicates that all predicted malignant cases were correctly classified without generating false positives, while a recall of 92.86% indicates that the model successfully detected most malignant cases. In the medical context, recall is highly important because it directly relates to the model's ability to minimize false-negative errors that may delay breast cancer diagnosis (Jinana et al., 2025). These findings also support the theory that Random Forest can handle high-dimensional medical data and effectively recognize complex patterns through ensemble learning and majority voting (Ali dkk., 2023; Dalfi dkk., 2023; Yaqoob dkk., 2025).

Feature selection based on feature importance reduced the feature set from 30 to 15 without sacrificing classification performance. The ROC-AUC value increased from 0.9929 in Model 1 to 0.9952 in Model 2, and then reached 0.9977 in Model 4. This finding is supported by strong correlations among features such as radius, perimeter, and area, indicating redundancy in the dataset. These results are consistent with the theory that feature selection aims to select relevant features, reduce data complexity, and improve model efficiency (Dalfi dkk., 2023; Sarkaleh dkk., 2026; Yaqoob dkk., 2025). This finding is also in line with (Kabir dkk., 2023), who showed that feature importance can identify influential features in breast tumor classification, and with (Minnoor & Baths, 2023), who demonstrated that reduced feature subsets can still achieve high classification performance on the Wisconsin Breast Cancer dataset.

The implementation of GridSearchCV in Models 3 and 4 improved ROC-AUC values, although no significant increase in accuracy was observed. Accuracy evaluates classification results at a single threshold, whereas ROC-AUC measures the model's ability to distinguish benign and malignant classes across multiple classification thresholds. Therefore, the increase in ROC-AUC indicates that hyperparameter optimization improves discriminative capability and model stability even when accuracy remains unchanged

(Minnoor & Baths, 2023; Setiawan dkk., 2026; Sholeh dkk., 2025). These results align with the observations of (Setiawan dkk., 2026), indicating that hyperparameter tuning of Random Forests did not significantly improve accuracy on breast cancer datasets. Similar findings were reported by (Ali dkk., 2023), who stated that GridSearchCV can improve classification performance by optimizing parameter combinations.

Based on all evaluation results, Model 4 emerged as the optimal model, achieving an ROC-AUC of 0.9977 and the smallest train-test gap of 0.0241 with only 15 features. These results indicate that integrating feature selection with GridSearchCV improves model stability while reducing signs of overfitting. These results directly address the gap outlined earlier, in which prior studies often treated feature selection and hyperparameter tuning as independent processes. (Premalatha dkk., 2025) implemented only feature selection, whereas (Setiawan dkk., 2026) applied only hyperparameter tuning. Therefore, this study demonstrates that integrating both approaches sequentially can produce a more stable, efficient, and optimal Random Forest model for breast tumor classification.

Although the accuracy, precision, recall, and F1-score values of all four models were identical, Random Forest has reached an optimal point on threshold-based metrics. Therefore, improvements are better evaluated through ROC-AUC, train-test gap, and feature efficiency. Nevertheless, this study still has several limitations. The dataset only consisted of 569 samples from the Wisconsin Breast Cancer Diagnostic Dataset, and feature selection based on feature importance may still be influenced by high correlations among features such as radius, perimeter, and area. In addition, this study focused solely on the Random Forest algorithm, without comparison to other machine learning methods. However, the results consistently show that combining feature selection with GridSearchCV improves model stability and discriminative power in distinguishing benign from malignant tumors.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study demonstrates that Random Forest can classify breast tumors into benign and malignant categories with very good performance. The baseline model (Model 1) achieved an accuracy of 97.37%, precision of 1.0000, recall of 0.9286, F1-score of 0.9630, and ROC-AUC of 0.9929 using all 30 features, indicating that Random Forest has strong baseline classification capability even without additional optimization.

Feature selection based on feature importance in Model 2 successfully reduced the number of features from 30 to 15 without decreasing classification performance. ROC-AUC improved from 0.9929 to 0.9952, with the most influential features being area_worst, concave points_worst, and radius_worst, indicating that some eliminated features likely contained redundant information. GridSearchCV optimization in Model 3 further improved ROC-AUC to 0.9964, demonstrating that hyperparameter optimization enhances discriminative capability rather than directly improving accuracy, while mild signs of overfitting persisted without feature selection.

The integration of feature selection and GridSearchCV in Model 4 achieved the best overall performance with the highest ROC-AUC of 0.9977 and the smallest train-test gap of 0.0241. Model 4 was also the only model in which the training accuracy did not reach 1.0000 (0.9978), indicating reduced overfitting. Using only 15 features, Model 4 proved more efficient, stable, and optimal for breast tumor classification than the other three models.

Suggestion

The limitations identified in this study suggest several potential improvements. Expanding the dataset to include medical records from multiple hospitals could enhance model generalizability across broader patient populations. Given the susceptibility of conventional feature importance measures to multicollinearity, methods such as RFE and SHAP-based evaluation offer promising avenues for more robust variable selection. Additionally, comparing Random Forest with other machine learning algorithms, such as Support Vector Machines, XGBoost, or LightGBM, under similar feature selection and hyperparameter tuning configurations would provide a more comprehensive evaluation. Finally, class imbalance handling techniques such as SMOTE or class_weight='balanced' can be explored further to improve recall on the malignant class, given the significant clinical consequences of false negatives in breast tumor detection.

REFERENCES

Ali, N. M., Besar, R., & Aziz, N. A. A. (2023). A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation.

Bulletin of Electrical Engineering and Informatics, *12*(2), 1047–1054. https://doi.org/10.11591/eei.v12i2.4838

Armoogum, S., Dewi, D. A., Kezhilen, M., & Trinawarman, D. (2024). Breast cancer prediction using metrics-based classification. *Journal of Applied Data Sciences*, *5*(3), 1508–1519. https://doi.org/10.47738/jads.v5i3.351

Bimo, A. A. (2024). *Pemanfaatan Decision Tree pada Algoritma Random Forest untuk Klasifikasi Kanker Payudara*.

Chazar, C., & Widhiaputra, B. E. (2020). Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine. *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, *12*, 67–79. https://doi.org/10.37424/informasi.v12i1.48

Dalfi, M. A. H., Chaabouni, S., & Fakhfakh, A. (2023). Breast Cancer Detection Using Random Forest Supported by Feature Selection. *International Journal of Intelligent Systems and Applications in Engineering*, *2024*(2s), 223–238.

Fauzi, A., Supriyadi, R., & Maulidah, N. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest. *Jurnal Infortech*, *2*(1), 96–101. https://doi.org/10.31294/infortech.v2i1.8079

Hulaifah Al Abrori, Z. Z., & Subhiyakto, E. R. (2025). Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression. *Jurnal Algoritma*, *22*(1), 300–311. https://doi.org/10.33364/algoritma/v.22-1.2164

Jamil, A. R., Hadi, J., & Munandar, I. (2024). Tumor Mammae. *SCIENA*, *3*(6), 398. https://doi.org/10.56260/sciena.v3i6

Jinana, Y., Kusrini, & Kusnawi. (2025). Klasifikasi Random Forest Terhadap Diagnosa Penyakit Kanker Payudara Berdasarkan Status Keganasan. *TEKNIMEDIA: Teknologi Informasi Dan Multimedia*, *6*(1), 129–134. https://doi.org/10.46764/teknimedia.v6i1.259

Kabir, S. S., Ahmed, M. S., Siddique, M. M., Ema, R. R., Rahman, M., & Md. Galib Syed. (2023). Breast Tumor Prediction And Feature Importance Score Finding Using Machine Learning Algorithms. *Radioelectronic and Computer Systems*, (4), 32–42. https://doi.org/10.32620/REKS.2023.4.03

Kementerian Kesehatan RI. (2018). *Keputusan Menteri Kesehatan Republik Indonesia*.

Kementerian Kesehatan RI. (2022, Februari 2). *Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan*. Kementerian Kesehatan RI. https://kemkes.go.id/id/kanker-payudaya-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan

Lestari, I. I., & Homaidi, A. (2024). Komparasi Algoritma Naive Bayes Dan Random Forest Pada Klasifikasi Kanker Payudara. *Gudang Jurnal Multidisiplin Ilmu*, *2*(12), 778–785. https://doi.org/10.59435/gjmi.v2i12.1206

Lestari, N. S., Kuncoro, A. N., & Ngadikun. (2023). Deteksi Tumor Payudara (Breast Benign Diseases) Berdasar Interaksi Eritrosit Akibat Perubahan Ion-Ion Dalam Darah-Edta Menggunakan Spektrofotometer Uv-Vis. *Journal for Physics Education and Applied Physics*, *5*(1), 47–61.

Magda, K., Inonu, O. Y., & Susanto, E. R. (2025). Komparasi Metode Machine Learning Untuk Diagnosis Penyakit Kanker Payudara. *EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi*, *15*(1), 15. https://doi.org/10.36448/expert.v15i1.4313

Minnoor, M., & Baths, V. (2023). Diagnosis of Breast Cancer Using Random Forests. *Procedia Computer Science*, *218*, 429–437. https://doi.org/10.1016/j.procs.2023.01.025

Misdiantoro, D., & Susanto, E. R. (2025). Optimasi Akurasi Prediksi Penyakit Kanker Payudara Menggunakan Metode Random Forest Optimization of Breast Cancer Disease Prediction Accuracy Using Random Forest Method. *Jurnal Pendidikan dan Teknologi Indonesia (JPTI)*, *5*(5), 1407–1416. https://doi.org/10.52436/1.jpti

Nasution, F. A., & Juledi, A. P. (2025). Penerapan Algoritma Random Forest untuk Klasifikasi Tingkat Keparahan Penyakit pada Data Rekam Medis. *Journal of Computer Science and Information Systems*, *6*(3), 371–378.

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, *06*(05), 551–560. https://doi.org/10.4236/jbise.2013.65070

Nur Fauzi, N. P., Khomsah, S., & Putra Wicaksono, A. D. (2025). Penerapan Feature Engineering dan Hyperparameter Tuning untuk Meningkatkan Akurasi Model Random Forest pada Klasifikasi Risiko Kredit. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *12*(2), 251–262. https://doi.org/10.25126/jtiik.2025128472

Paepke, S., Metz, S., Brea Salvago, A., & Ohlinger, R. (2018). Benign Breast Tumours-Diagnosis and Management. *Breast Care*, *13*(6), 403–412. https://doi.org/10.1159/000495919

Premalatha, K., Duraisamy, P. D., & Sivakumar, K. (2025). Machine learning framework for breast cancer detection with feature selection with L2 ridge regularization: Insights from multiple datasets. *Journal of Translational Genetics and Genomics*, *9*(1), 11–34. https://doi.org/10.20517/jtgg.2024.82

Sarkaleh, M. K., Azgomi, H., & Kiani-Sarkaleh, A. (2026). Breast Cancer Classification Using Feature Selection via Improved Simulated Annealing and SVM Classifier. *Diagnostics*, *16*(4). https://doi.org/10.3390/diagnostics16040637

Setiawan, A., Andri Armaginda Siregar, Setiawan, N., Jalaluddin Nasution, Naufal Dhiya Putra Dalimunthe, & Farhan Sardy Abdillah. (2026). Optimasi Performa Model SVM dan Random Forest untuk Klasifikasi Kanker Payudara Menggunakan Penyetelan Hyperparameter. *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)*, *4*(3), 2141–2149. https://doi.org/10.62712/juktisi.v4i3.789

Sholeh, M., Lestari, U., & Andayati, D. (2025). Hyperparameter Optimization Using Grid Search and Random Search to Improve the Performance of Prediction Models with Decision Trees. *Jurnal Riset Multidisiplin dan Inovasi Teknologi*, *3*(03), 453–464. https://doi.org/10.59653/jimat.v3i03.2025

Suparna, K., & Kartika, S. L. M. K. (2022). Kanker Payudara Diagnostik, Faktor Risiko, dan Stadium. *Ganesha Medicina Journal*, *Vol 2*(1), 42–46. https://doi.org/10.23887/gm.v2i1.47032

Tjengharwidjaja, A., Dharma, B., Munenori, Y., & Emmanuel, M. (2024). Klasifikasi Pasien Terkena Breast Cancer Menggunakan Metode Machine Learning. *Computatio: Journal of Computer Science and Information Systems*, *8*(1), 86–95. https://doi.org/10.24912/computatio.v8i1.15174

Yaqoob, A., Verma, N. K., Mir, M. A., Tejani, G. G., Eisa, N. H. B., Mamoun Hussien Osman, H., & Shah, M. A. (2025). SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-95786-1

Zhu, J., Zhao, Z., Yin, B., Wu, C., Yin, C., & Chen, R. (2025). An integrated approach of feature selection and machine learning for early detection of breast cancer. *Scientific Reports*, *15*, 1–12. https://doi.org/10.1038/s41598-025-97685-x

Zulkarnain, C. R., & Delyuzar. (2017). Perbandingan Antara Neoplasma Jinak Dan Ganas Pada Payudara Berdasarkan Pemeriksaan Fisik Diagnostik Dan Biopsi Aspirasi Jarum Halus. *Ibnu Sina Biomedika*, *1*(2).

Attique, Z., & Khan, S. (2025). *Quantitative Analysis of Breast Nuclei Morphology for Cancer Diagnosis Using Supervised Machine Learning*.

Girdhar, A., Raju, K., & P N, S. (2023). Significance of Nuclear Morphometry in Breast Lesions: A Cross-Sectional Study. *Cureus*. https://doi.org/10.7759/cureus.39378