ARTICLE INFO

AUTHOR'S AFFILIATIONS
Graduate Programme of Medical Technology Laboratory Science, Universitas Muhammadiyah Semarang, Indonesia[1,3]
Medical Laboratory Technology, Faculty of Health Sciences, Universitas Muhammadiyah Kudus, Indonesia[2]
Department of Informatics, Universitas Muhammadiyah Semarang, Indonesia[4]
Regional General Hospital of SIMO Boyolali, Centre Java[5]

CORRESPONDING AUTHOR
Mudyawati Kamaruddin, Graduate Programme of Clinical Laboratory Science, University of Muhammadiyah Semarang
E-mail: mudyawati@unimus.ac.id

Article history

Received	29-09-2025
Revised	21-03-2026
Accepted	31-03-2026
Available online	31-03-2026

Please cite this article in APA 7th edition style as:
Kamaruddin, M., Ekawati. S.N., Iswara, A., Ilham, A., & Wardhani. A.A. (2026). Integrating Clinical, Stage, and IHC Data for Breast Cancer Classification with SVM-RBF. *Jurnal Ilmiah Kedokteran Wijaya Kusuma, 15(1), 117-130*

https://doi.org/10.30742/jikw.v15i1.4908

Integrating Clinical, Stage, and IHC Data for Breast Cancer Classification with SVM-RBF

Mudyawati Kamaruddin[1], Sherly Nur Ekawati[2], Arya Iswara[3], Ahmad Ilham[4], Astri Aditya Wardhani[5]

Abstract

Background: Breast cancer is one of the leading causes of cancer-related deaths among Indonesian women. Early detection and classification of molecular subtypes are crucial for determining appropriate therapy. Accurate determination of biological subtypes of breast cancer is essential for selecting optimal treatment strategies. **This research aims** to build and evaluate a breast cancer subtype classification model using the SVM with an RBF kernel. The subtypes classified include Luminal A, Luminal B, HER2+, and Triple Negative Breast Cancer, utilizing a combination of patient clinical data (age, tumor size, and tumor location), cancer stage, and the expression status of hormonal receptors ER and PR. **The methodological** steps include data preprocessing, feature selection, model training with cross-validation, and performance evaluation using metrics such as accuracy, precision, recall, F1-score, and the ROC-AUC curve. **The results** showed that the majority of patients' ages were in the range of 40–60 years, with dominant tumor sizes between 1 and 3 cm. Luminal A and B subtypes were more frequently observed in patients aged ≥50 years and at early stages, whereas HER2+ and TNBC were mostly observed in patients under 50 years with advanced stages. The established baseline SVM-RBF model achieved high accuracy (91%) but performed poorly at detecting minority subtypes, such as HER2+, with a recall = 0 and an F1-score = 0, indicating model bias toward the majority class. This study demonstrates that the SVM algorithm with the RBF kernel is effective for modeling breast cancer subtype classification using clinical data, cancer stage, and immunohistochemistry results.

Keywords: Breast_cancer_classification, clinical_data, immunohistochemistry, RBF kernel, SVM.

Original Research Article

INTRODUCTION

Breast cancer is the leading cause of morbidity and mortality among women worldwide. According to GLOBOCAN 2020 data, breast cancer accounts for approximately 25% of all cancer cases in women. Further research shows that early diagnosis and classification of cancer subtypes are crucial for selecting the appropriate therapeutic strategy (Sung, H. et al., 2021). Molecular-based breast cancer classification is often based on hormone receptor expression, including Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor 2 (HER-2). Breast cancer

subtypes that are positive for ER and PR generally respond better to hormone therapy and have a better prognosis compared to negative subtypes (Kittaneh, M. et al 2013). Early detection and accurate diagnosis are crucial for improving prognosis and determining effective treatment strategies. In the last decade, the application of machine learning methods, particularly Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, has gained significant attention in breast cancer classification (Abdurrahman, G., 2023). This method can process clinical data, cancer stage, and immunohistochemistry results of ER and PR receptors to provide more accurate predictions (Ilham, A. et al., 2025a).

SVM is a machine learning algorithm designed to solve classification and regression problems. SVMs work by finding the optimal hyperplane to separate the dataset into two classes (Ilham, A. et al., 2025b). In the medical field, SVM has been widely used to classify various types of biological data, including genetic data and medical images (Choudhury, P. et al., 2021). The RBF kernel is a kernel used to handle data that cannot be separated linearly. This kernel maps the data into a higher-dimensional space, allowing for more complex separation between data classes. The RBF kernel function is a parameter that determines the influence range of a data point. Recent studies have reported that applying the RBF kernel in SVM can significantly improve model accuracy, especially when applied to clinical data and breast cancer radiology results (Huang, S. et al., 2018; Kamaruddin, M., 2023).

Clinical data and cancer staging significantly improve the performance of SVM models. Clinical data provide important information about the patient's initial condition, while the cancer stage gives an overview of the extent of disease spread. The combination of these two types of data allows the SVM model to perform more accurate classification. Other studies have also shown that integrating this data improves the predictive ability of SVMs in overall breast cancer diagnosis (Bilal, A. et al., 2024). In addition to clinical data and cancer stage, ER and PR immunohistochemistry results are important for classification. ER and PR status are the main indicators in determining breast cancer subtypes and response to hormone therapy. Recent studies show that SVM models integrating this immunohistochemical information achieve higher accuracy in predicting patient prognosis. This indicates that combining clinical data, cancer staging, and immunohistochemistry results can create a more reliable diagnostic model and support clinical decision-making (Onitilo, A. A., et al., 2009).

A hybrid approach that combines the RBF kernel and SVM has also been used for mammography image classification (Maulana, H. et al., 2026). Research shows that this method not only improves accuracy but also enhances sensitivity and specificity in detecting breast cancer. By integrating various types of data into the model, this approach provides promising results to support medical decision-making and breast cancer therapy planning (Wang, D. et al., 2023). The application of the SVM method with an RBF kernel, integrating clinical data, cancer stage, and immunohistochemistry results, offers a promising approach to improve classification accuracy in breast cancer. The development of such models can support more effective clinical decisions and improve therapeutic outcomes for breast cancer patients. Previous studies by Wang et al. (2019) have shown that the use of clinical and molecular data in the SVM model increases sensitivity by up to 15% compared to traditional methods (Wang et al., 2019). In addition, this approach can help reduce diagnostic error rates caused by manual evaluation by pathologists.

The use of immunohistochemistry data for breast cancer assisted by machine learning algorithms, such as SVM, adds value to diagnosis, prognosis, and treatment recommendations, making it a potential solution for personalizing medical care (Singh, B.K., 2019). Based on the background, this research aims to build and evaluate a breast cancer subtype classification model using the SVM with an RBF kernel.

MATERIAL AND METHODS

The research sample used is a dataset consisting of 480 breast cancer patient data containing clinical information, including age (years), tumor size (cm), tumor location (left/right), cancer stage (I, II, III, IV), and immunohistochemical results (ER, PR, and HER2) (positive/negative). The number of samples

in the dataset is based on total sampling from January 2023 to December 2024, referring to 1) patients who have been diagnosed with breast cancer who came to RSUP Dr. Soeradji Tirtonegoro Klaten and RSU PKU Muhammadiyah Delanggu, and 2) patients with complete clinical data, including clinical data (age, tumor size, and location), cancer stage, and immunohistochemical results (ER, PR, and HER2). Patients with breast tumors that are not cancer, such as benign tumors (fibroadenoma), are not included in this study.

Data Preprocessing Using Exploratory Data Analysis (EDA)

The purpose of data preprocessing is to optimize model performance by cleaning, normalizing, and encoding data before use in classification. The data preprocessing steps are as follows:

a. Handling Missing Values by:
   1. Identifying variables with missing values, especially in numerical (age, tumor size) and categorical (ER, PR) variables.
   2. Imputing missing data with the mean or median for numerical variables and the mode for categorical variables.
   3. If the number of missing values is greater than 30% in a certain feature, that feature can be removed to maintain data quality.

b. Data Normalization

   Numerical variables such as tumor size may have a very different range of values from other variables, which can cause the model to prioritize variables with a larger scale. The SVM model is sensitive to data scale, so normalization is necessary to ensure that all features are in the same range. Min-Max Scaling: transforms feature values to a 0-1 range to make it easier for the model to process. Standardization (Z-score Normalization): Converts values into a distribution with mean = 0 and standard deviation = 1.

c. Data Distribution Analysis

   Displaying data distribution using visualizations such as histograms, boxplots, and count plots for age, tumor size, tumor location, cancer stage, immunohistochemistry ER and PR examination results, and breast cancer subtypes. Performing correlation analysis between numerical variables and a crosstab between categorical variables to understand the pattern of relationships between features.

d. Categorical Data Encoding

   The SVM model cannot work directly with categorical data, such as tumor location (right/left), cancer stage (I, II, III, IV), and immunohistochemical results for ER and PR (Positive/Negative). Therefore, encoding is required to convert this data to a numerical format for use in the model. There are two common methods used:

   - One-Hot Encoding: Converts categories into binary form (0 and 1):

     ER (+) → 1, ER (-) → 0

     PR (+) → 1, PR (-) → 0
   - Label Encoding: Converts categories into sequential numbers (ordinal). For example:

     Stage I → 1, Stage II → 2, Stage III → 3, Stage IV → 4

   In the subsequent analysis, the ER and PR variables are not displayed in the confusion matrix because they serve as input features in the SVM modeling process; instead, the confusion matrix lists the subtype classification target labels as the output classes of the breast cancer model (Luminal A, Luminal B, TNBC, and HER2+).

e. Feature Selection (FS)

   The goal of FS is to filter the most relevant features to improve model performance and efficiency with the recursive Feature Elimination (RFE) technique, which iteratively removes the

least important features to improve model accuracy using the optimal feature subset, and the Correlation Analysis (CA) to identify features that have a strong linear relationship with each other. Highly correlated features can provide redundant information and may cause overfitting. By removing a highly correlated feature, the model can become simpler and more efficient without losing important information.

Modeling Construction

a. k-fold cross-validation
   The data was split at an 80:20 ratio, with 80% for training and 20% for testing. To enhance the model's reliability, a 5-fold cross-validation approach was used, with each fold serving as validation data. This process was repeated k times, so each part of the dataset has an equal chance of serving as test data. In this study, 5-fold cross-validation was used, meaning the dataset was divided into 5 parts. The model was trained using four parts and tested using the remaining one. The evaluation results from the five iterations are then averaged to obtain a stable performance metric that is independent of any particular subset.
b. SVM Implementation
   This study used the SVM algorithm with an RBF kernel as the primary method for classifying breast cancer subtypes. The RBF kernel was chosen for its ability to handle data patterns that cannot be linearly separated by mapping the data into a higher-dimensional space, enabling an optimal separating hyperplane to be found. The SVM algorithm operates under the maximum-margin principle, which seeks a hyperplane that maximizes the distance to the nearest data points (support vectors).

Model Evaluation

a. Confusion Matrix, used to see the distribution of correct and incorrect predictions between classes.
   1) Accuracy measures the proportion of correct predictions out of all data:

$$\text{Accuracy} = (TP+TN)/((TP+TN+FP+FN)$$

   where TP is True Positive (the model predicts the patient has a low risk of breast cancer); TN is True Negative (the model predicts the patient has a moderate or high risk of breast cancer); FP is False Positive (the patient actually has a moderate or high risk, but the model predicts the patient as low risk); and FN is False Negative (the patient actually has a low risk, but the model predicts the patient as moderate or high risk).

   2) Precision measures the model's accuracy in predicting positive classes:

$$\text{Precision}=TP/(TP+FP)$$

   3) Recall (Sensitivity), measures how many actual positive data points are successfully identified correctly by the model.

$$\text{Recall}=TP/(TP+FN)$$

   4) F1-Score is the harmonic mean of precision and recall:

$$F1=2 \times (\text{Precision} \times \text{Recall})/(\text{Precision}+\text{Recall})$$

RESULTS

1. Data Preprocessing Results

The dataset comprises 480 records of breast cancer patients, including clinical information such as age, tumor size (in cm), tumor location (left/right), cancer stage (I, II, III, or IV), and ER and

PR immunohistochemistry results. In the EDA stage, initial analysis was performed to understand the data characteristics:

a. Age Distribution

Patient age distribution is an important variable in breast cancer analysis, as age plays a role as one of the main determinants in the incidence of breast cancer. Evaluation of age distribution enables the identification of age groups with the highest prevalence, which can serve as the basis for policy-making related to screening and early detection (**Fig. 1**).

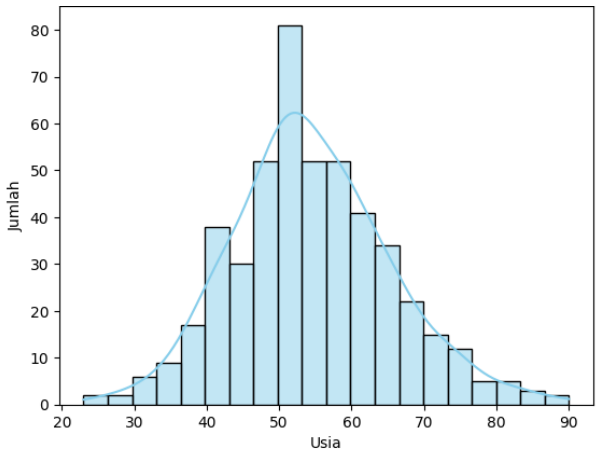

Figure 1. Age Distribution

As shown in **Figure 1**, the age distribution of breast cancer patients is presented as a histogram with a Kernel Density Estimation (KDE) curve. The histogram illustrates that most patients fall within the age range of 40 to 60 years, with the highest concentration occurring around 50 years of age. This indicates that the middle-aged group is the most dominant population in the analyzed data.

b. Tumor Size Distribution

The tumor size variable is one of the important factors influencing the diagnosis and prognosis of breast cancer patients. Tumor size is a crucial indicator in determining cancer stage and selecting appropriate treatment strategies. The distribution of tumor size will be visualized using a histogram, which will show how tumor size is distributed among breast cancer patients. The illustration of the tumor size distribution is shown as follows:

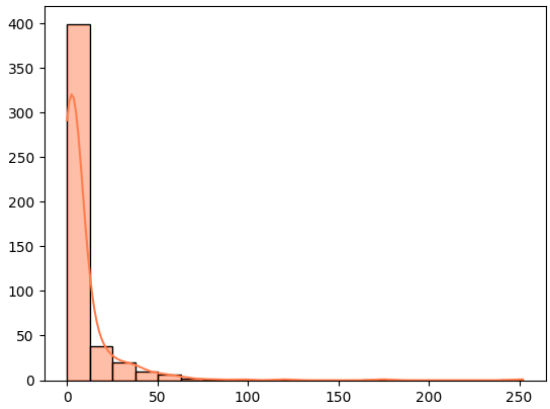

Figure 2. Tumor size distribution

As shown in **Figure 2,** the tumor size distribution of breast cancer patients shows an asymmetrical pattern with a right-skewed tendency. The majority of patients have small tumors, with the peak of the distribution below 10 mm, while a small number have very large tumors, exceeding 100 mm.

c. Tumor Location Distribution

Tumor location is one of the important variables in breast cancer studies, as several studies have shown differences in the frequency of occurrence between the left (sinistre) and right (dextra) sides of the breast. The analysis of tumor location distribution in this dataset aims to determine tumor laterality patterns and potential asymmetries that can serve as a basis for further epidemiological and clinical studies. The description of the tumor location distribution is shown as follows:

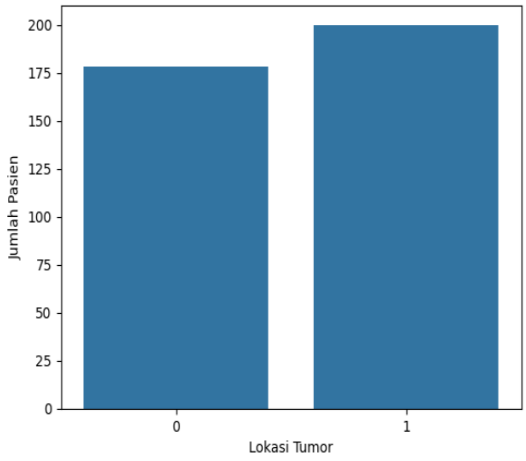

Figure 3. Tumor location distribution. 0; Dextra (right), 1: Sinistra (left)

As shown in **Figure 3,** the distribution of tumor locations in this study indicates that the majority of tumors were found in the left breast (SINISTRE) with 200 cases (52.9%), while the remaining were in the right breast (DEXTRA) with 178 cases (47.1%).

d. Distribution of Cancer Stages

Breast cancer stage is an important clinical indicator that reflects the disease's progression and significantly influences treatment selection and patient prognosis. The staging classification system is typically divided into four levels, namely stages I-IV, which reflect the progression of cancer from early stages to metastatic conditions (**Figure 4**).

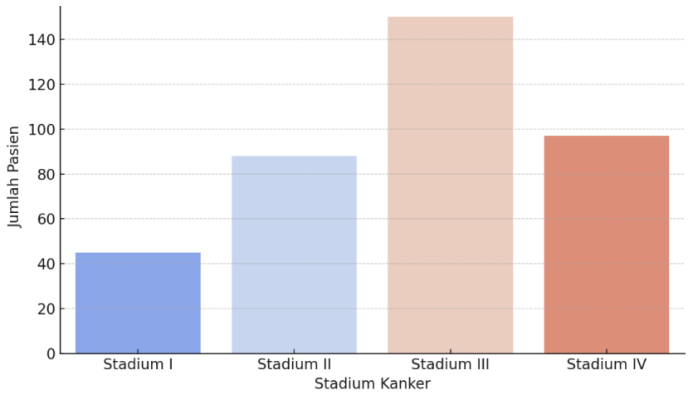

Figure 4. Cancer Stages distribution

As shown in **Figure 4**, the distribution of breast cancer patients by disease stage (stage I to IV). Based on the visualization, it is evident that the majority of patients are diagnosed at an

advanced stage, with Stage III being the most prevalent category, followed by Stage IV. Meanwhile, Stage I and II, which reflect the early stages of cancer development, were recorded in lower numbers.

e. ER and PR Distribution

The status of ER and PR hormonal receptors is a crucial indicator in determining breast cancer subtypes and treatment direction, particularly for hormonal therapy. Patients with positive expression for these receptors tend to have a better response to treatment. Therefore, analyzing ER and PR distribution provides an initial overview of the biological characteristics of patients in the dataset (**Figure 5**).

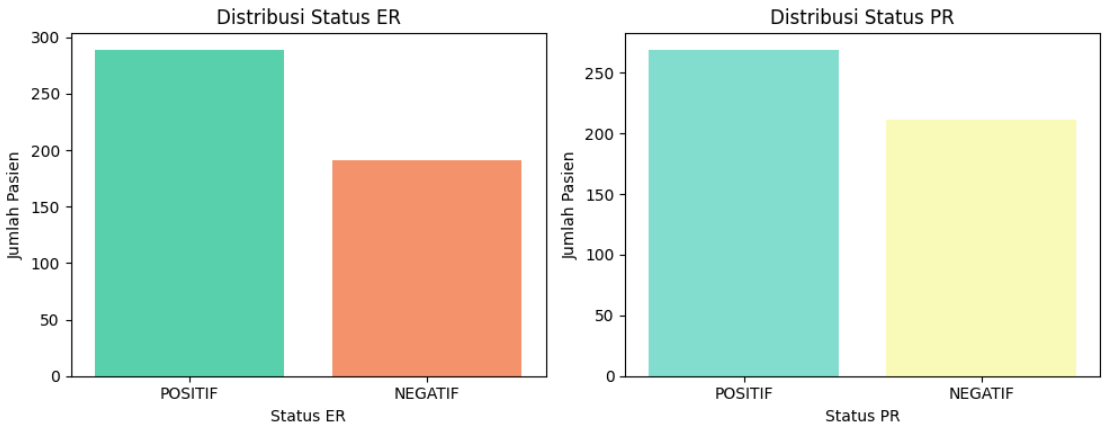

Figure 5. ER and PR distribution

As shown in **Figure 5**, among 290 patients, 60.4% (n=290) were ER-positive, and 39.6% (n=190) were ER-negative. For PR status, 265 patients (55.2%) showed positive expression, and 215 patients (44.8%) showed negative expression.

The confusion matrix for the baseline model uses testing data (96 data points) obtained from the dataset division using the train-test split method, with the following calculation:

$$480 \times 20\% = 96$$

Meaning:

384 data → training and 96 data → testing

2. SVM Classification Results with RBF Kernel Function

The classification model was built using an SVM with an RBF kernel to classify breast cancer subtypes based on patient clinical data, cancer stage, and immunohistochemistry results. The classification process was carried out without data balancing or parameter tuning, and ollowing image presents the confusion matrix of the baseline SVM model's prediction results, which shows the model's success rate in classifying each breast cancer subtype based on the true label and the predicted label:

As shown in **Figure 6** and **Table 1**, the SVM model with an RBF kernel achieved an overall accuracy of 91%. However, the per-class evaluation showed performance imbalance. However, according to the confusion matrix, the model correctly classified Luminal A (52/52) and TNBC (31/31) subtypes with perfect accuracy. The Luminal B subtype was identified in only 4 of 6 cases, and HER2+ was not fully classified; all 7 cases were incorrectly predicted as TNBC. These results are reflected in the evaluation metrics: Luminal A and TNBC have a recall of 1.00, while HER2+ has a precision, recall, and F1-score of 0.00. The macro-average values for precision and recall were 0.69 and 0.67, respectively, indicating a bias towards the majority class (Abdel-Hafiz, H., 2017).

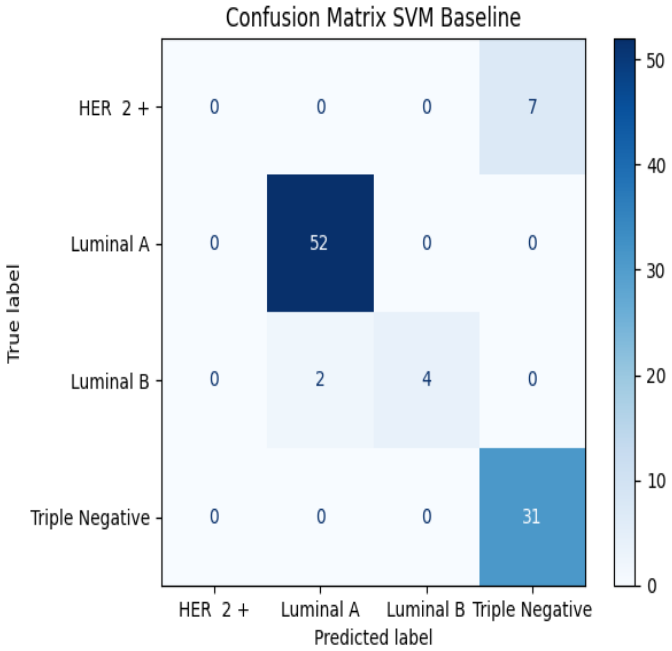

Figure 6. Confusion Matrix SVM Baseline

Table 1. True Label Confusion Matrix

True Label	Pred: Luminal A	Pred : Luminal B	Pred: HER2+	Pred: TNBC
True: LA	52	0	0	0
True: LB	2	4	0	0
True: HER2+	3	1	0	3
True TNBC	0	0	0	31

As shown in **Figure 7**, the ROC curve illustrates the SVM model's performance with an RBF kernel for classifying each breast cancer subtype using the One-vs-Rest (OvR) approach. Based on the curve results, the highest Area Under the Curve (AUC) was observed for the Luminal A subtype, with a score of 0.98, indicating that the model can distinguish Luminal A from other subtypes with near-perfect accuracy. The Triple Negative subtype also demonstrated excellent performance, with an AUC of 0.92, indicating high sensitivity and specificity in its recognition. Furthermore, the model's performance for the Luminal B subtype was quite good, with an AUC value of 0.86. However, there is a possibility of overlapping characteristics with the Luminal A subtype, given that both subtypes exhibit estrogen receptor expression (Arnold, et al., 2022).

Meanwhile, the lowest AUC value was obtained for the HER2+ subtype, at 0.78. This suggests that the model's ability to detect HER2+ remains quite limited. This suboptimal performance is most likely due to the smaller amount of HER2+ data compared to other subtypes or limitations in the features used, specifically the exclusion of HER2 expression from the classification process. Overall, AUC values above 0.75 across all subtypes indicate that the SVM-RBF model has strong classification capabilities and is suitable for clinical use, particularly for distinguishing subtypes with distinct molecular characteristics (Bhoo-Pathy, N. et al., 2015).

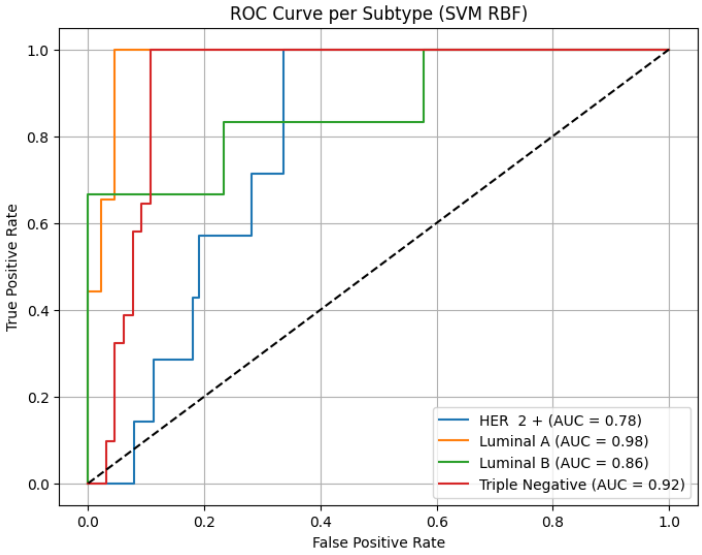

Figure 7. ROC Curve Visualization

Nevertheless, the results of this study need to be understood in the context of several methodological limitations that may affect the validity and interpretation of the classification model. First, the clinical data do not include complete information on the N element (lymph nodes) in the TNM system, so staging was performed directly based on the stage recorded in the medical records. This may potentially introduce classification bias, especially if there is inconsistency between the stage and the actual nodal status. Second, the data do not include information on M (distant metastasis), so although some patients are recorded as stage IV, it is not known exactly how many have experienced metastasis, a key determinant of prognosis and therapy selection.

DISCUSSION

Development of an SVM-based breast cancer subtype classification model with an RBF kernel, designed using clinical data including age, tumor size, tumor location, cancer stage, and ER and PR hormonal expression status. The results of the EDA data analysis on the age distribution of breast cancer patients show that most patients were in the 40–60 years age range, with a peak at 50 years, indicating that the middle-aged group is the most affected. This finding supports screening policies that emphasize the importance of early detection in this age group, given that earlier breast cancer diagnosis at this age can increase the chances of successful therapy and survival (Cai, S. et al.,2020). The 40 to 60-year age range is a significant biological transition phase for women, whether they are approaching or entering menopause. This biological transition phase involves drastic hormonal changes, particularly a decrease in endogenous estrogen levels, which can affect breast epithelial cell proliferation. Several studies show that increases and decreases in estrogen levels from pre menopause to post menopause can trigger an imbalance in cell growth regulation (Desantis, C.E. et al., 2017)). In addition to biological factors, demographic and lifestyle factors also contribute to an increased risk of breast cancer. The 40 to 60-year age group is the late productive age group, where most women have accumulated risk exposure throughout their lives, such as high-fat diets, sedentary lifestyles, exposure to environmental carcinogens, and a history of long-term hormone therapy (Torre, L.A. et al., 2016).

This data also shows differences in the distribution of subtypes across age groups. The majority of patients with Luminal A and B subtypes are found in middle to older age groups, namely 45–65 years. Conversely, a number of observed HER2+ and TNBC-positive patients appear more frequently. These cases often occur in patients in a relatively younger age range. Most large breast cancer patients with Luminal A and B subtypes are in the age group of ≥ 50 years. In contrast, HER2 and TNBC subtypes

are more commonly found in patients < 50 years old (Tittmann, J. et al., 2024). The tumor size distribution shows a rightward skew, indicating that although most patients have small tumors (≤3 cm), a few have very large tumors (>5 cm), especially in HER2+ and TNBC subtypes. Several studies have observed typical tumor sizes after diagnosis from 819,647 patients diagnosed with breast cancer between 1990 and 2014, 93% of women had tumors with a diameter under 50 mm (Irawan, H.W. I., 2025). In another study of 490 breast cancer patients diagnosed from 2016-2017, the average tumor size after cancer diagnosis was 1.4 cm for women who underwent annual mammograms and 1.8 cm for women who underwent examinations only once every two years (Sopik, V and Narod, S. A., 2018).

The largest tumors were detected in the HER2+ and TNBC subtypes, which are known to be more aggressive. Patients suffering from Luminal A tend to have smaller tumor sizes at diagnosis. This is consistent with previous findings that tumor size correlates positively with prognosis, with HER2+ and TNBC tumors consistently larger at initial diagnosis. This data is sufficient to support the idea that tumor size can be a very important predictor in subtype modeling, especially for early detection. In Indonesia, almost 48.3% of breast cancer patients have T4 tumor sizes, and about 31.7% are in stage IIIB, with a dominance of HER2+ and TNBC in advanced stages. This indicates the importance of early detection. This is important in high-risk populations (Irawan, H.W.I., 2025). The distribution of tumor locations between the left (SINISTRA) and right (DEXTRA) sides did not show a significant difference in subtype distribution; the data were almost evenly balanced and showed no strong correlation. Previous research indicates that tumor location (right or left) is not significantly related to the distribution of molecular subtypes. Luminal A, Luminal B, HER2+, and TNBC subtypes were found almost evenly in both the right and left breasts. These results are also consistent with several previous studies that reported no significant relationship between tumor location and the biological or molecular distribution of breast cancer subtypes (Mallapasi, M.N. et al., 2021).

However, it is important to note that although tumor location does not directly affect the subtype, anatomical location can indirectly influence the technical aspects of clinical management and prognosis. Some studies show that tumors located in the left breast (sinistra) have a higher risk of cardiac exposure during radiotherapy, especially with conventional techniques that do not include cardiac protection. In the long term, this can increase the risk of cardiovascular complications, particularly in patients receiving high-dose radiation in the mediastinal area (Mourman, S.E. M. et al., 2021). Thus, although epidemiologically tumor location is not directly related to severity or subtype, it remains important in clinical management, especially for planning radiotherapy and surgery and for protecting vital organs such as the heart and lungs.

The distribution of cancer stages showed that most HER2+ and TNBC cases were diagnosed at stages III and IV, which is consistent with their clinical characteristics as fast-growing and invasive subtypes. The distribution of cancer stages indicates that HER2+ and TNBC tend to be found at advanced stages, reinforcing the importance of earlier screening approaches for these subtypes. Conversely, Luminal A patients were more frequently found at stage I or II, reflecting slower tumor growth and earlier detection due to a good response to estrogen. The tendency towards advanced stages in this study reflects the difficulty in early detection of breast cancer in various developing countries, including Indonesia. Several contributing factors include delays in diagnosis, limited access to healthcare facilities, lack of early screening programs, and a lack of public understanding of the early symptoms of breast cancer (Tittamnn, J. et al., 2024). Nevertheless, in this study, although the stage distribution showed a dominance of advanced stages (III and IV), the molecular subtyping results showed that the majority of patients were classified into Luminal A and B subtypes, which are generally associated with early stages. This discrepancy indicates an inconsistency between clinical data and subtype classification results, possibly because some patients are at an advanced stage but classified as Luminal, reflecting cases detected late, even though their tumor type tends to grow slowly. This is inseparable from the challenges in developing countries like Indonesia, where diagnostic delays are still common due to limited access to healthcare services, lack of routine screening, and low public awareness of early breast cancer symptoms. Therefore, the advanced-stage distribution in this study

is not solely due to the aggressiveness of the subtype, but is also influenced by systemic and social factors that delay detection across all subtypes, including Luminal A and B.

These results are consistent with research published in The Lancet Global Health journal, which noted that over 50% of breast cancer patients in low- and middle-income countries are diagnosed at stage III or IV, while in developed countries this figure is less than 30%. This is related to differences in referral systems, health literacy, and screening coverage (Huang, 2018). The distribution of ER and PR expression levels is a key determinant in subtype classification. Luminal A and B subtypes have positive ER and PR status, while HER2+ and TNBC are largely negative for both receptors. These results are consistent with ER and PR expression, along with HER2, being the main basis for molecular classification of breast cancer. ER and PR play an important role in response to hormonal therapy, and are strong differentiating markers between Luminal and non-Luminal groups. The baseline SVM model with an RBF kernel achieved 91% accuracy. However, the confusion matrix results showed that the model's performance was very uneven across classes. None of the HER2+ cases were classified, with precision, recall, and F1-score values of 0.00. Conversely, the classification performance for Luminal A and TNBC was excellent (recall = 1.00). This indicates a model bias towards the majority class, a common phenomenon in classification with imbalanced data (Huang).

In addition, the classification model in this study struggled to recognize the HER2+ subtype, most likely because HER2 expression data were not included among the features. The lack of HER2 data reduces the informativeness of distinguishing between the Luminal B and HER2-enriched subtypes, thereby affecting the model's accuracy in detecting this aggressive subtype. Similarly, Ki-67 expression data, an important indicator of tumor proliferation and a key for distinguishing between Luminal A and B, were unavailable, leading to incomplete molecular information.

SVM has long been used in breast cancer classification, especially in identifying subtypes based on clinical and genomic data. SVMs with an RBF kernel can achieve high accuracy in breast cancer classification, but often struggle to detect minority classes. Imbalanced data is a major challenge in real-time anomaly detection. A dataset is considered imbalanced if one of its classes has a very large dominance over the others (Huang, D.S/, et al. 2018). The most common way to handle imbalanced data is to use resampling methods, such as random undersampling or oversampling, that rebalance the majority and minority classes. Predicting machine failure is a challenge because datasets are often imbalanced.

A common approach to handle classification with imbalanced data is to balance the data using sampling methods such as random undersampling, random oversampling, or SMOTE (Huang, 2018). These results support the conclusion that SMOTE improves recall for minority classes such as HER2+. The application of the SMOTE method increased recall and F1-score for minority classes, though it was accompanied by a slight decrease in overall accuracy to 83%. The combination of SMOTE with GridSearchCV yielded a model with the best class-balanced F1-score (0.66), demonstrating the effectiveness of combining data balancing and hyperparameter optimization to improve model generalization on real-world imbalanced data (ilham, A. et al., 2025). SMOTE has proven effective in balancing imbalanced datasets for various applications, including cancer detection. SMOTE demonstrates that this technique can improve recall for the minority class without sacrificing overall accuracy. These research results support your finding that by using SMOTE, HER2+ recall significantly increased, although total accuracy slightly decreased (Huang., 2018). SMOTE helps increase the number of minority samples, thereby reducing data imbalance. However, SMOTE has limitations, namely that SMOTE performs linear interpolation between minority data, which risks generating synthetic samples that do not adequately represent natural variability. SMOTE does not account for complex feature distributions, so it can introduce noise and overfitting. To address data imbalance among breast cancer subtypes, especially the relatively small numbers of HER2+ and TNBC subtypes, this study uses the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an interpolation-based oversampling technique that generates new synthetic data from minority classes by taking two minority data points and creating new samples between them. Unlike conventional oversampling

techniques that merely duplicate data, SMOTE maintains data diversity and reduces the risk of overfitting. This technique is widely used in medical classification because it can improve model sensitivity to minority classes (Huang).

SMOTE works well for datasets with simple distributions but is less effective for complex clinical datasets. SVMs find the optimal hyperplane that maximizes the margin between the data classes, making them well-suited for handling non-linear and high-dimensional data. SVM performance is highly dependent on the selection and tuning of hyperparameters, such as C and gamma in the RBF kernel. Improper hyperparameter settings can result in a suboptimal model, which can, in turn, affect classification accuracy.

Other limitations include small sample sizes in minority groups and the absence of patient outcome data, such as therapy response or survival rates, so the model has not yet been evaluated in the context of long-term clinical prediction. Therefore, although the model demonstrates good statistical performance, these results need to be interpreted with caution, and further studies with more comprehensive data are recommended, including additional molecular data and the integration of clinical outcomes to improve the model's practical utility in real medical practice. It is recommended to integrate additional features such as HER-2 and KI-67 expression. In addition, ensemble model approaches such as Random Forest or deep learning-based methods can be used to improve sensitivity to complex patterns in breast cancer clinical data.

CONCLUSION

1. The distribution of clinical characteristics shows that most patients are in the 40 to 60 years age group, with dominant tumor sizes between 1 and 3 cm, and tumor location tends to be balanced between the left and right sides. Luminal A and B subtypes are more frequently found in older age and at early stages, while HER2+ and TNBC subtypes are more commonly found in younger age and at advanced stages. The majority of patients showed positive expression for ER and PR receptors.
2. A SVM classification model with an RBF kernel was successfully built and was able to effectively map breast cancer subtypes. The baseline model yielded a high accuracy of 91%, but showed weaknesses in classifying minority classes (HER2+ and TNBC) due to data imbalance.

CONFLICT OF INTEREST

The author(s) declare that there is no conflict of interest regarding the publication of this article.

ACKNOWLEDGEMENTS

The authors would like to thank the Directorate of Research and Community Service (DPPM) for financial support in Fiscal Year 2025, the Research and Community Service Institute (LPPM), and the Graduate School of Universitas Muhammadiyah Semarang, as well as the microbiota Team for their valuable assistance in technical support, data analysis, and manuscript preparation.

REFERENCES

Abdurrahman, G. (2023). Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid. *JUSTIFY : Jurnal Sistem Informasi Ibrahimy*, 20;2(1):74–80. https://doi.org/10.35316/justify.v2i1.3370

Abdel-Hafiz H. (2017). Epigenetic Mechanisms of Tamoxifen Resistance in Luminal Breast Cancer. *Diseases,* 6;5(3):16. https://doi.org/10.3390/diseases5030016

Arifiyanti AA, Wahyuni ED. (2020). SMOTE: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining. Scan - *Jurnal Teknologi Informasi dan Komunikasi*, 28;15(1). https://doi.org/10.33005/scan.v15i1.1850

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., et al. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast*, 1;66:15–23. https://doi.org/10.1016/j.breast.2022.08.010

Bhoo-Pathy, N., Verkooijen, H., Tan, EY. *et al.* (2015). Trends in presentation, management and survival of patients with *de novo* metastatic breast cancer in a Southeast Asian setting. *Sci Rep* **5**, 16252. https://doi.org/10.1038/srep16252

Cai, S., Zuo, W., Lu, X., Gou, Z., Zhou, Y., Liu, P., et al. (2020). The Prognostic Impact of Age at Diagnosis Upon Breast Cancer of Different Immunohistochemical Subtypes: A Surveillance, Epidemiology, and End Results (SEER) Population-Based Analysis. Front Oncol. 23;10.

Choudhury, P., Foroughi, C., Larson, B. (2021). Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Management Journal,* 1;42(4):655–83. https://doi.org/10.1002/smj.3218.

DeSantis, C.E., Ma, J., Goding, Sauer, A., Newman, L.A., Jemal, A. (2017). Breast cancer statistics racial disparity in mortality by state. CA Cancer J Clin. 67(6):439–48.

Doren, A., Vecchiola, A., Aguirre, B., Villaseca, P. (2018). Gynecological–endocrinological aspects in women carriers of BRCA1/2 gene mutations. *Climacteric. Taylor and Francis Ltd*; 2018, 21: 529–35. https://doi.org/10.1080/13697137.2018.1514006

Eliyatkin, N., Yalcin, E., Zengel, B., Aktaş, S., Vardar, E. (2015). Molecular Classification of Breast Carcinoma: From Traditional, Old-Fashioned Way to A New Age, and A New Way. Journal of Breast Health, 7 (11): 59–66. DOI: 10.5152/tjbh.2015.1669

Huang, S., Nianguang, C.A.I., Penzuti, P.P., Narandes, S., Wang, Y., Wayne, X.U. 2018. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics and Proteomics. *International Institute of Anticancer Research*, 15: 41–51. doi: 10.21873/cgp.20063

Ilham, A., Nagara, T. A. P., Kamaruddin, M., Khikmah, L., & Mantoro, T. (2025a). Fetal Health Risk Classification using Important Feature Selection and CART Model on Cardiotocography Data. *Informatica*, *49*(1). https://doi.org/10.31449/inf.v49i1.5658

Ilham, A., Kamaruddin, M., and Khikmah, L., (2025b). Dari Data Ke Diagnosis: Teori Dan Implementasi *Support Vector Machine* untuk Diagnosis Medis Yang Lebih Cepat Dan Akurat. Zahir Publishing, Sleman Yogyakarta.

Irawan, H. W. I. ketut. (2020). Clinical and Subtypes of Breast Cancer in Indonesia. *281Asian Pacific Journal of Cancer Care• Vol 5• Issue 4apjcc.Waocp.Com*, *5*. DOI:10.31557/APJCC.2020.5.4.281

Kamaruddin, M. (2023). Can Genomics of Gut Microbiota in Stool Samples be Analysed by MERLIN?. *Journal of Intelligent Computing & Health Informatics*, *3*(2), 30-34. https://doi.org/10.26714/jichi.v3i2.11289

Kamaruddin, M., Marzuki, I., Burhan, A., & Ahmad, R. (2021, February). Screening acetylcholinesterase inhibitors from marine-derived actinomycetes by simple chromatography. In IOP Conference Series: Earth and Environmental Science (Vol. 679, No. 1, p. 012011). IOP Publishing.

Kittaneh, M., Montero, A. J., Glück, S. (2013). Molecular Profiling for Breast Cancer: A Comprehensive Review. *Biomark Cancer*, Jan 5. DOI: 10.4137/BIC.S9455

Marandi M, Hossein Abadi S. Aqueous synthesis of colloidal CdSexTe1-x – CdS core–shell nanocrystals and effect of shell formation parameters on the efficiency of corresponding quantum dot sensitized solar cells. Solar Energy. 2020 Oct 1;209:387–99. https://doi.org/10.1016/j.solener.2020.08.059

Mallapasi, M.N., Kusumanegara, J., Kabo, P., Usman, U., Mulyono, M.T., Faruk, M. (2021). Cardiac metastasis of triple-negative breast cancer mimicking myxoma: A case report. Int J Surg Case Rep. 1;88.

Maulana, H., Kamaruddin, M., Suyanto, A., & Rabban, A. (2026). Identification of Neisseria gonorrhoeae Bacteria Using a Convolutional Neural Network (CNN) Based on Image

Classification. *PHARMACOLOGY, MEDICAL REPORTS, ORTHOPEDIC, AND ILLNESS DETAILS*, *5*(1), 21–35. https://doi.org/10.55047/comorbid.v5i1.1963

Moorman, S.E.H., Pujara, A.C., Sakala, M.D., Neal, C.H., Maturen, K.E., Swartz, L., et al. (2021). Annual screening mammography associated with lower stage breast cancer compared with biennial screening. American Journal of Roentgenology. 1;217(1):40–7.

Onitilo, A.A., Engel, J.M., Greenlee, R.T., Mukesh, B.N. (2009). Breast Cancer Subtypes Based on ER/PR and HER2 Expression: Comparison of Clinicopathologic Features and Survival. *Clin Med Res*, 1;7:4–13. DOI: 10.3121/cmr.2009.825

Pramesh CS, Badwe RA, Bhoo-Pathy N, Booth CM, Chinnaswamy G, Dare AJ, et al. (2022). Priorities for cancer research in low- and middle-income countries: a global perspective. Nat Med., 28:649–57. https://doi.org/10.1038/s41591-022-01738-x

Rahman MdM, Rahman, A., Akter, S., Pinky, S.A. (2023). Hyperparameter Tuning Based Machine Learning Classifier for Breast Cancer Prediction. *Journal of Computer and Communications*, 11(04):149–65. DOI: 10.4236/jcc.2023.114007

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. In *Computers in Biology and Medicine* (Vol. 149). Elsevier Ltd. https://doi.org/10.1016/j.compbiomed.2022.106043

Sopik, V., Narod, S.A. (2018) The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer. Breast Cancer Res Treat. 2018 Aug 1;170(3):647–56.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*., 71(3):209–49. DOI: 10.3322/caac.21660

Tittmann, J., Ágh, T., Erdősi, D., Csanády, B., Kövér, E., Zemplényi, A., Kovács, S., & Vokó, Z. (2024). Breast cancer stage and molecular subtype distribution: real-world insights from a regional oncological center in Hungary. *Discover Oncology*, *15*(1). https://doi.org/10.1007/s12672-024-01096-9.

Torre, L.A., Siegel, R.L., Ward, E.M., Jemal, A. (2016). Global cancer incidence and mortality rates and trends - An update. Vol. 25, Cancer Epidemiology Biomarkers and Prevention. American Association for Cancer Research Inc., p. 16–27.

Trihardianingsih, L., Santos, L. G., kunci-GridSearrch, C.V.K., Udara K. (2024). Optimasi Hyperparameter GridSearchCV pada Klasifikasi Kualitas Udara menggunakan Support Vector Machine [Internet]. Vol. 1, Jurnal Informasi dan Teknologi). Available from: https://data.jakarta.go.id/

Wang, D., He, H., Wei, C. (2023). Cellular and potential molecular mechanisms underlying transovarial transmission of the obligate symbiont Sulcia in cicadas. *Environ Microbiol*, 2023, 1;25(4):836–52. DOI: 10.1111/1462-2920.16310

Wang, R., Zhu, Y., Liu, X., Liao, X., He, J., Niu, L. (2019). The Clinicopathological features and survival outcomes of patients with different metastatic sites in stage IV breast cancer. *BMC Cancer*, 12;19(1). DOI: 10.1186/s12885-019-6311-z