Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 3, No. 2, 2025, 86-91
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Optimizing Breast Cancer Prediction by Applying Machine Learning

Vina Nurmadani[1], Indah Suciati[2], Yoga Aji Sukma[3], Linda Rassiyanti[4]

[1,2,3,4] Institut Teknologi Sumatera, Indonesia
*corresponding author: vina.nurmadani@sd.itera.ac.id

Received July 31, 2025; Received in revised August 10, 2025; Accepted August 12, 2025

Abstract. In 2015, breast cancer ranked among the most prevalent and fatal cancers affecting women globally. Artificial intelligence is urgently needed to help medical professionals make more accurate decisions, reduce overdiagnosis, and streamline the diagnostic process. This study will implement and perform a comparative study of selected machine learning techniques algorithms, with a focus on SVM, XGBoost, and ANN, with various parameter combinations on the breast cancer dataset. Performance metrics such as accuracy, precision, recall, and F1-score were employed to evaluate and compare the algorithms. The results of this study show that the best model for predicting chronic breast cancer disease, which can help medical professionals predict chronic disease so that it can be treated quickly and accurately, is the SVM method using 8 parameters without the mitosis parameter: Clump thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, and Normal Nuclei, with an accuracy value of 0.96 and a sensitivity value of 0.98.
Keywords: ANN; breast cancer; machine learning; SVM; XGBoost

INTRODUCTION

Breast cancer was one of the most common and deadly types of cancer for women worldwide in 2015 (Sun et al., 2017). According to the World Health Organization (WHO), in 2022, there will be 2.3 million women diagnosed with breast cancer and as many as 670,000 deaths worldwide (WHO: BREAST CANCER, 2024). In 2017, an estimated 30% of new cancer cases (252,710) in women were breast cancer (Siegel et al., 2017). Breast cancer is a cancer that forms in breast tissue, where breast tissue cells change and divide uncontrollably, usually resulting in a mass or lump (Houssein et al., 2021). In more severe stages, these tissue cells can spread through the lymph nodes to other organs of the body.

The high number of deaths due to breast cancer requires early detection to reduce the risk of death. Artificial intelligence is beginning to be developed in the healthcare sector, playing a role in disease identification. Medical personnel urgently need AI to make more accurate decisions, reduce misdiagnosis, and streamline the diagnostic process. Machine learning can be applied to detect diseases using classification and clustering methods for decision-making (Mahesh, 2020). Numerous studies have investigated the use of machine learning to detect breast cancer using Decision Tree, Random Forest, Logistic Regression, Naive Bayes, K-nearest Neighbor, and Support Vector Machine methods, achieving the best

Volume 3, No. 2, 2025, 86-91
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

accuracy of 96.82% with the Random Forest method (El_Rahman, 2021). Another breast cancer study using the Naive Bayes method achieved 80% accuracy with 116 datasets (Mubarog et al., 2021). This study will apply and compare several machine learning algorithms, such as ANN, SVM, and XGBoost, with a combination of parameters on the breast cancer dataset. Model evaluation will be conducted through performance assessment based on accuracy, precision, recall, and F1-score to determine the best-performing algorithm. Therefore, the results of this study are expected to facilitate the progress of a fast and accurate medical decision support system for diagnosing breast cancer.

RESEARCH METHODS

This study utilizes the Breast Cancer Dataset sourced from the UCI Machine Learning Repository (Wolberg, 1992). This dataset is general and can be used freely for research purposes. The number of data used is 699 data points, consisting of 458 (65.50%) benign cancers and 241 (34.50%) malignant cancers. The dataset contains 11 parameters, namely Id Number, Clump thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nuclei, Mitoses, and Class. The class parameter is the target in this dataset, where Breast cancer cases labeled as 0 are benign, whereas those labeled as 1 are malignant.

This research utilizes machine learning classifiers, namely Support Vector Machine, XGBoost, and ANN. The following is an explanation of each method used:

1. SVM
   SVM is a type of supervised learning algorithm that can be applied to both classification and regression problems (Aldhyani et al., 2020). SVM works by determining the hyperplane that maximizes the margin between different classes in the dataset (Aldhyani et al., 2020).
2. XGBoost
   Developed by Tianqi Chen in 2014, eXtreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm that enhances the performance of predictive models through gradient boosting techniques (Dangeti, 2017). XGBoost aims to prevent overfitting and optimize computational capabilities (Intan Permata & Esther Sorta Mauli Nababan, 2023).
3. ANN
   ANN models the relationship between a series of incoming and outgoing signals using a model inspired by biological nervous systems, specifically human brain cells, that process information (Dangeti, 2017).

The accuracy of breast cancer prediction is measured using metrics derived from a confusion matrix**,** which serves as a tool to assess classification performance in supervised learning models (Purbolaksono et al., 2021). In machine learning, a confusion matrix is frequently employed to assess prediction accuracy by classifying outcomes into four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

In this paper, accuracy, sensitivity, recall, and F1-score will be used to measure disease prediction performance.

1. Accuracy
   Accuracy reflects the ratio of correctly predicted instances to the total number of predictions made by the model (Villavicencio et al., 2021). The formulation can be seen below.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 3, No. 2, 2025, 86-91
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

Accuracy measurement is a criterion often used to measure classification performance, but when working on unbalanced classes, this criterion is less appropriate because the minority class will have a small contribution.

2. Sensitivity
   Sensitivity refers to the model's ability to correctly identify instances of the positive class (True Positives). It indicates how effectively the model captures actual positive cases. The formula is presented below (Houssein et al., 2021).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

   Sensitivity describes a model's success in retrieving information. The sensitivity indicator becomes particularly important when the risk of false-negative (FN) diagnoses/predictions is high.

3. Precision
   This measures the positive class prediction (TP), which is a subset of the positive class, that is, how many predictions for the positive class are valid. The formulation can be seen below (Obaido et al., 2024)

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

   Precision can be used as a measure of model performance when the risk of false positive (FP) diagnosis/prediction is very dangerous, while the risk of false negative (FN) is considered less dangerous.

4. F1-Score
   The F1-score measurement is the harmonic mean between precision and sensitivity. The formulation can be seen below (Al-Harahsheh et al., 2021).

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision}*\text{Sensitivity}}{\text{Precision+Sensitivity}} \quad (4)$$

This measurement takes into account the importance of the risks of false positive (FP) and false negative (FN) values. The F1-score is better used when the classification distribution is unbalanced.

RESULT AND DISCUSSIONS

The dataset includes 11 features that describe the characteristics of breast cancer cells. During the preprocessing phase, it was identified that 16 entries contained missing values in the *Bare Nuclei* attribute. To handle these missing values, imputation was performed using the mean value of the parameter.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 3, No. 2, 2025, 86-91
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

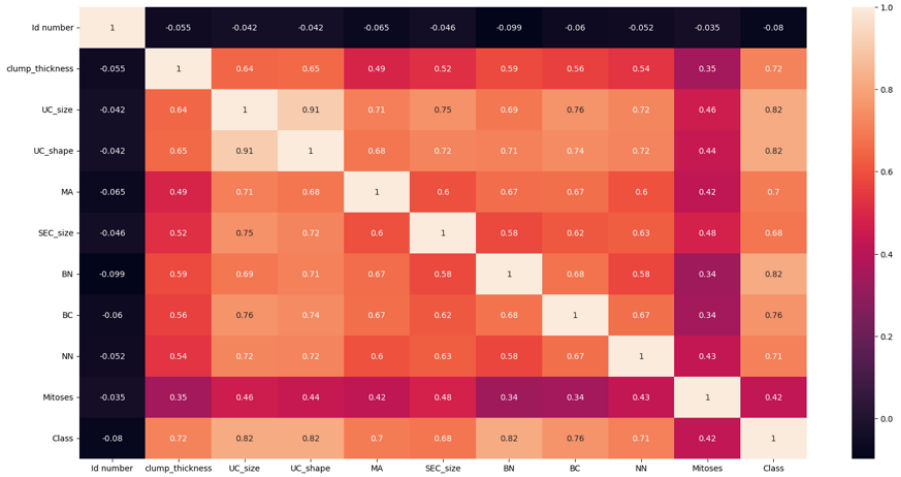

Figure 1. Correlation between parameters in the breast cancer dataset

Figure 1 illustrates the correlation among parameters in the breast cancer dataset, which helps identify the degree of association between two variables. A higher correlation value indicates a stronger relationship between parameters. Based on the correlation matrix, the **ID Number** parameter exhibits the weakest correlation and was therefore excluded from the modeling process. The highest correlation is observed between **UC_size** and **UC_shape,** with a coefficient of **0.82**, suggesting that these features play a significant role in machine learning model performance. In contrast, the **Mitoses** parameter has a relatively low correlation value of **0.42**, second only to ID Number. Based on the correlation analysis, several combinations were conducted to identify the most effective model for breast cancer prediction, namely a combination with all parameters, without the mitoses parameter, and the mitoses parameter and sec_size (Single Epithelial Cell Size). Subsequently, the dataset was split into 8:2 for training and testing to facilitate model development and evaluation. The classification outcomes for breast cancer under various parameter configurations are shown in the table below:

Table 1. The classification outcomes for breast cancer under various parameters

Scenario	Method	Accuracy	Precision	Sensitivity	F1-Score
Classification results of breast cancer disease with all parameters	ANN	0.94	0.88	0.93	0.91
	SVM	0.96	0.91	0.97	0.94
	XGBoost	0.96	0.93	0.93	0.93
The findings of the classification of breast cancer with all correlation parameters using 8 parameters, without the method	ANN	0.95	0.91	0.93	0.91
	SVM	0.96	0.91	0.98	0.94
	XGBoost	0.96	0.93	0.93	0.93
The results of breast cancer disease classification with all correlation parameters using 7 parameters, without method, and sec_size	ANN	0.95	0.91	0.93	0.92
	SVM	0.96	0.91	0.95	0.93
	XGBoost	0.96	0.93	0.93	0.93

Volume 3, No. 2, 2025, 86-91
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

According to the scenario, Classification results of breast cancer disease with all parameters, both the SVM and XGBoost models demonstrated the highest levels of accuracy, each with a score of **0.96**. The **XGBoost** model also recorded the **highest precision**, with a value of **0.93.** Meanwhile, the highest performance values for sensitivity and F1-score are obtained from the SVM method at 0.97 and 0.94.

The next parameter combination is *the findings* of the classification of breast cancer with all correlation parameters using 8 parameters, without the method. The SVM and XGBoost algorithm models show the highest accuracy value of 0.96. The highest Precision performance measurement is in the XGBoost method at 0.93. The highest performance values for sensitivity and F1-score are in the SVM method at 0.98 and 0.94.

Scenario The results of breast cancer disease classification with all correlation parameters using 7 parameters, without method, and sec_size show that the SVM and XGBoost algorithm models have the highest accuracy value of 0.96. The highest precision performance measurement is in the XGBoost method at 0.93. The highest performance value in sensitivity is the SVM method at 0.95.

Classification performance is measured not only by the highest accuracy value but also by other metrics such as precision, sensitivity, and F1-score. The highest accuracy values were obtained by the SVM and XGBoost methods with a combination of all parameters. Therefore, additional evaluation using other performance metrics is required. Classification performance can be measured by the best accuracy value. The highest accuracy values are found in the SVM and XGBoost methods with all parameter combinations, so it must be observed using other performance measurements such as Precision, Sensitivity, and F1-Score. Sensitivity measurement was chosen because it is used as an indicator of False Negative (FN) predictions, which are riskier than False Positive (FP), namely when the results of the breast cancer prediction are benign, but the actual data indicates a more dangerous cancer. Therefore, the SVM method utilizing eight parameters—excluding the mitosis parameter was selected as the most effective model for predicting breast cancer, achieving an accuracy of 0.96 and a sensitivity of 0.98.

CONCLUSION AND SUGGESTIONS

The analysis conducted in this research indicates that the best model for predicting chronic breast cancer, which can assist medical personnel in predicting chronic disease and enabling rapid and accurate treatment, is the SVM method using eight parameters, excluding the mitosis parameter: Clump thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, and Normal Nuclei. The accuracy value is 0.96, and the sensitivity value is 0.98.

Recommendations based on this research result include increasing the data used to predict disease, conducting tests with other algorithms and parameters to improve the quality of the prediction model, and using original data from the relevant region to develop a more appropriate chronic disease prediction system.

REFERENCE

Aldhyani, T. H. H., Alshebami, A. S., & Alzahrani, M. Y. (2020). Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2020*. https://doi.org/10.1155/2020/4984967

Volume 3, No. 2, 2025, 86-91
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Al-Harahsheh, H., Al-Shraideh, M., & Al-Sharaeh, S. (2021). Performance of Malware Detection Classifier Using Genetic Programming in Feature Selection. *Informatica (Slovenia)*, *45*(4), 517–529. https://doi.org/10.31449/INF.V45I4.3819

Dangeti, Pratap. (2017). *Statistics for Machine Learning*. Packt Publishing.

El_Rahman, S. A. (2021). Predicting breast cancer survivability based on machine learning and features selection algorithms: a comparative study. *Journal of Ambient Intelligence and Humanized Computing*, *12*(8), 8585–8623. https://doi.org/10.1007/s12652-020-02590-y

Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. In *Expert Systems with Applications* (Vol. 167). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2020.114161

Intan Permata, & Esther Sorta Mauli Nababan. (2023). Application Of Game Theory In Determining Optimum Marketing Strategy In Marketplace. *JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM*, *2*(2), 65–71. https://doi.org/10.55606/jurrimipa.v2i2.1336

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. https://doi.org/10.21275/art20203995

Mubarog, I., Setyanto, A., & Sismoro, H. (2021). Sistem Klasifikasi Pada Penyakit Breast Cancer Dengan Menggunakan Metode Naïve Bayes. *Creative Information Technology Journal*, *6*(2), 109. https://doi.org/10.24076/citec.2019v6i2.246

Obaido, G., Achilonu, O., Ogbuokiri, B., Amadi, C. S., Habeebullahi, L., Ohalloran, T., Chukwu, C. W., Mienye, E. D., Aliyu, M., Fasawe, O., Modupe, I. A., Omietimi, E. J., & Aruleba, K. (2024). An Improved Framework for Detecting Thyroid Disease Using Filter-Based Feature Selection and Stacking Ensemble. *IEEE Access*, *12*, 89098–89112. https://doi.org/10.1109/ACCESS.2024.3418974

Purbolaksono, M. D., Tantowi, M. I., Hidayat, A. I., & Adiwijaya. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal RESTI*, *5*(2), 393–399. https://doi.org/10.29207/resti.v5i2.3008

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, *67*(1), 7–30. https://doi.org/10.3322/caac.21387

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P., & Zhu, H. P. (2017). Risk factors and preventions of breast cancer. In *International Journal of Biological Sciences* (Vol. 13, Issue 11, pp. 1387–1397). Ivyspring International Publisher. https://doi.org/10.7150/ijbs.21635

Villavicencio, C. N., Macrohon, J. J. E., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using weka. *Algorithms*, *14*(7). https://doi.org/10.3390/a14070201

WHO : BREAST CANCER. (2024, March 13). WHO. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Wolberg, W. (1992). *Breast Cancer Wisconsin (Original)*. UCI Machine Learning Repository. https://archive-beta.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original