Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 16(2) 2023: 182-193
http://ejournal.undip.ac.id/index.php/media_statistika

BREAST CANCER CLASSIFICATION USING SUPPORT VECTOR MACHINE (SVM) AND LIGHT GRADIENT BOOSTING MACHINE (LIGHTGBM) MODELS

Puspita Kartikasari, Iut Tri Utami, Suparti, Syair Dafiq Faizur Rahman
Department of Statistics, Universitas Diponegoro, Semarang, Indonesia

e-mail: *puspitakartikasari@live.undip.ac.id*

DOI: https://doi.org/10.14710/medstat.16.2.182-193

Article Info:
Received: 23 October 2023
Accepted: 25 February 2024
Available Online: 18 May 2024

Keywords:
Classification; Breast Cancer; SVM; LightGBM
.

Abstract: This study examines the existence of breast cancer from the perspective of statistics as one alternative solution. From a statistical point of view, breast cancer management can be done with early detection and appropriate and fast treatment measures through diagnosis classification. In conducting early detection, an accurate diagnosis model is needed and can be developed by developing and testing statistical methods, one of which is the classification method. The classification methods used in this study are Support Vector Machine (SVM) and LightGBM. Both methods have a high level of classification accuracy because the algorithm used is robust and sensitive in determining each object in the classification member. Therefore, these two methods classify breast cancer into malignant and benign categories. The results of this study show that the best method to classify breast cancer is the SVM method, with an accuracy rate of 97.9%.

1. INTRODUCTION

Breast cancer detection has developed very rapidly in the last ten years. The diagnosis of breast cancer was initially only a physical examination. Now, it has developed into genetic and clinical examinations such as histopathology and molecular genetics by cellular microbiology laboratory experts (Pederson & Noss, 2020; Rutgers et al., 2019; Saxena & Gyanchandani, 2020; van der Giessen et al., 2021). These developments have revealed some facts about the causes, ways of diagnosis, and the provision of more effective drugs. However, in reality, breast cancer is still a disease that is increasing in incidence, is still widely prevalent, ranks first of all cancers, and is the leading cause of death for cancer both in Indonesia and in the world, especially for women (Harbeck et al., 2019; KEMENKES RI, 2018; Siegel et al., 2022; WHO, 2021). The problem is closely related to late and inappropriate diagnosis due to the absence of an accurate, standardized diagnosis classification in determining the severity of breast cancer. The absence of certain determinant factors that can provide an accurate picture of the potential risk of breast cancer has implications for the inaccuracy of early detection results and appropriate and rapid treatment measures.

This research examines the existence of breast cancer, which is prevalent, tends to increase, is quite severe, and has a high potential to cause death with a satistical point of view so that an alternative solution to the problem is obtained by developing statistical methods in the form of classification models. Classification methods have been widely used in various fields to provide effective solutions (Bustamam et al, 2019; Khandezamin et al, 2020; Kurniawan et al, 2018; Setiawan et al, 2015; Shi et al, 2009; Wood et al, 2019). Currently, classification modeling has developed a lot, which has implications for increasing model accuracy (Aditya et al, 2015; Hanmastiana et al, 2022; Marianto et al, 2020; Nugraha et al, 2020; Umma et al, 2021; Wardani et al, 2020). The development of methods in this classification model can provide an accurate picture of breast cancer classification based on its severity.

This research focuses on detecting breast cancer by classifying it into malignant and benign. This research compares two methods, namely Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM), in classifying breast cancer into malignant and benign categories. Both methods have a high level of classification accuracy because the algorithms used are powerful and sensitive in determining each object in the classification members. The SVM method uses a hyperplane as a line or plane that separates two data classes with a maximum margin. In contrast, LGBM uses a combined decision tree to predict the target class. The contribution of this research compared to other research is that through the development of classification methods, a model is obtained that can be used as a reference in the selection of treatment so that there is an increase in healing, long life expectancy, a decrease in cancer incidence, complications, and mortality.

Previous research has been conducted by Setiawan (2023) using the nearest neighbor method and feature selection in detecting breast cancer. This study successfully diagnosed cancer based on five existing features with an accuracy of 85%. Another study was also conducted by Widodo et al. (2021) in predicting breast cancer using the KNN, bagging, and Random Forest methods. The best method obtained in this study is KNN, with an accuracy of 74.37%.

2. LITERATURE REVIEW

2.1. Breast Cancer

Breast cancer diagnosis was initially only a physical examination by a doctor. However, now it has developed into genetic and clinical examinations such as histopathology and molecular genetics (tumor gene expression signature) by cellular microbiology laboratory experts (Pederson & Noss, 2020; Saxena & Gyanchandani, 2020). The development of breast cancer examination has revealed some secrets about the causes, ways of diagnosis, and more effective drug administration. However, the reality shows that breast cancer is still a disease that is increasing in incidence, is still widespread in prevalence, ranks first of all cancers, and is the leading cause of death for cancer both in Indonesia and on the world map, especially for women (KEMENKES RI, 2018; WHO, 2021).

2.2. Classification Analysis

Classification analysis is a method in statistics that classifies variables or objects. The basis of the classification results is the similarity of characteristics possessed by each variable or object with specific indicators. Currently, the modeling process in classification analysis has developed a lot, starting from methods not only based on regression, discriminant or multidimensional but also based on artificial neural networks and algorithms used such as Naïve Bayes, K-Nearest Neighbors, K-Fold Cross Validation, decision trees which have implications for increasing model accuracy (Aditya et al., 2015; Hanmastiana et al., 2022; Marianto et al., 2020; Nugraha et al., 2020; Umma et al., 2021; Wardani et al., 2020)

Classification, in some cases, is used as a decision-making procedure for a case. Therefore, it is necessary to assess the goodness of the classification results to see the accuracy, sensitivity, and specificity values. However, the accuracy value is not an appropriate value that can be used to measure the goodness of the classification results if the observation data occurs in imbalanced class problems. The data imbalance problem will give a misleading accuracy value so that the value cannot be used to assess the goodness of the classification results. Therefore, further tests are carried out with the development of survival analysis in handling this matter.

2.3. Support Vector Machine (SVM)

The basic concept of Support Vector Machine (SVM) combines decades of computational theories, such as hyperplane margins. The classification concept with SVM can be explained simply as finding the best hyperplane. The hyperplane is the best separator between the two data classes, which can be determined by measuring the maximum point of the hyperplane margin. The margin is the distance between the hyperplane and the closest data in each class. The data closest to the best hyperplane is called the support vector (Ispriyanti & Hoyyi, 2016; Prasetyo, 2012).

2.4. Nonlinear Classification

In the real-world problem, the data obtained is rarely linear. Many are non-linear (Nugroho et al., 2003). In SVM, there is a kernel function, which is used to solve non-linear problems. The kernel function allows the implementation of a model in a higher dimensional space (feature space).

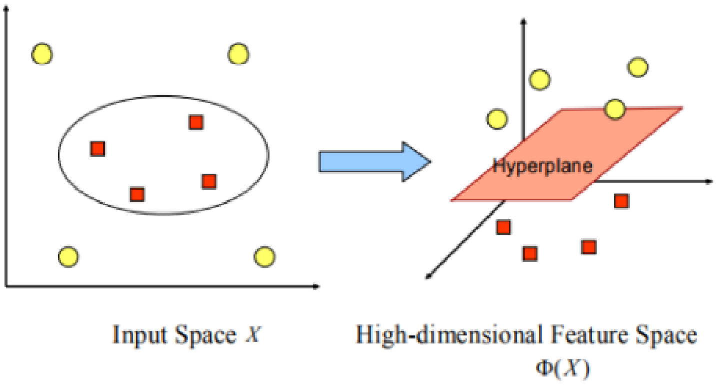

Figure 1. Kernel Functions implementations
Source: (Nugroho et al., 2003)

The kernel function used for mapping is denoted by the symbol (Φ). Suppose for n data samples as Equation (1):

$$\left((\Phi(\boldsymbol{x}_1).\, y_1); (\Phi(\boldsymbol{x}_2).\, y_2); \dots; (\Phi(\boldsymbol{x}_n).\, y_n)\right) \tag{1}$$

After mapping with the kernel, the following training process is the same as the linear SVM classification. This mapping process requires the dot product of two pieces of data in the new feature space. The dot product of two vector ($\boldsymbol{x}_i$) and ($\boldsymbol{z}$) is denoted as $\Phi(\boldsymbol{x}_i).\,\Phi(\boldsymbol{z})$. Without knowing the transformation Φ, the dot product value can be calculated using the components of the two vectors in the original dimension, such as Equation (2):

$$K(\boldsymbol{x}_i, \boldsymbol{z}) = \Phi(\boldsymbol{x}_i).\,\Phi(\boldsymbol{z}) \tag{2}$$

where the value of $K(\boldsymbol{x}_i, \boldsymbol{z})$ is a kernel function that shows a non-linear mapping on the feature space and $\Phi(\boldsymbol{x}_i).\,\Phi(\boldsymbol{z})$ is the dot product of two vectors ($\boldsymbol{x}_i$) and ($\boldsymbol{z}$). The prediction of a data set with new feature dimensions is formulated by (Prasetyo, 2012). The hyperplane function for non-linear classification using the kernel function is as in Equation (3):

$$
\begin{aligned}
f(\Phi(\boldsymbol{z}) &= \text{sign } (\boldsymbol{w}.\Phi(\boldsymbol{z}) + b) \\
&= \text{sign}\left(\textstyle\sum_{i=1}^{p} \alpha_i\, y_i \Phi(\boldsymbol{x}_i).\Phi(\boldsymbol{z})\right) + b \\
f(\boldsymbol{z}) &= \text{sign } \left(\textstyle\sum_{i=1}^{p} \alpha_i\, y_i K(\boldsymbol{x}_i, \boldsymbol{z}) + b\right) \qquad (3)
\end{aligned}
$$

where

$$\boldsymbol{w} = \textstyle\sum_{i=1}^{p} \alpha_i\, y_i \Phi(\boldsymbol{x}_i) \qquad b = -\frac{1}{2}\,(\boldsymbol{w}\boldsymbol{x}_{-1} + \boldsymbol{w}\boldsymbol{x}_{+1})$$

with p is number of support vector data; $\boldsymbol{x_i}$ is support vector; $\boldsymbol{z}$ is data to predict**.**
Table 1 provide popular and frequently used kernel functions.

Table 1. Kernel Functions Commonly Used

Kernel Function	Function Formula
Linear	$K\ (\boldsymbol{x}_i, \boldsymbol{z}) = \left(\boldsymbol{x}_i^T.\boldsymbol{z}\right)$
Polynomial	$K\ (\boldsymbol{x}_i, \boldsymbol{z}) = \left(\boldsymbol{x}_i^T.\boldsymbol{z} + 1\right)^{\mathrm{d}}$
Radial Basis Function (RBF)	$K\ (\boldsymbol{x}_i, \boldsymbol{z}) = \exp\left(-\gamma\left\|\boldsymbol{x}_i - \boldsymbol{z}\right\|^2\right)$

2.5. Light Gradien Boosting Machine (LightGBM)

LightGBM algorithm designed by Microsoft Research Asia uses the Gradient Boosting Decision Tree (GBDT) framework. (Ke et al., 2017). The goal is to improve computational efficiency so that problems with big data can be solved efficiently (Liang et al., 2020). LightGBM has several advantages compared to other GBDT methods: faster training speed, higher efficiency, lower memory usage, better accuracy rate, ability to handle data on a large scale, and support for parallel learning and GPU (Rufo et al., 2021)

The raw data set with $N = \{1, 2, \dots, n\}$ examples and the LightGBM model having $T = \{1, 2, \dots, t\}$ trees are assumed to be generated. After t iterations, the final prediction equals the sum of the first $(1 - t)$-th and t-th. The iteration process is described in Equation (4).

$$\hat{y}_i{}^{(t)} = \hat{y}_i{}^{(t-1)} + f_i(\boldsymbol{x}_i) \qquad (4)$$

$\hat{y}_i{}^{(t)}$ is the prediction value of the i-th example at the t-th iteration. $\hat{y}_i{}^{(t-1)}$ denotes the previously generated tree model and $f_i(\boldsymbol{x}_i)$ denotes the newly built model.

According to equation (4), each new prediction is generated by the residual and the previous prediction. The complete training process is described in equation (5), and the regulation term can be calculated by equation (6), which is used to reduce the complexity of the model and can be used to improve the usability of other datasets.

$$\begin{cases} \hat{y}_i{}^{(0)} = 0 \\ \hat{y}_i{}^{(1)} = f_1(\boldsymbol{x}_i) = \hat{y}_i{}^{(0)} + f_1(\boldsymbol{x}_i) \\ \hat{y}_i{}^{(2)} = f_1(\boldsymbol{x}_i) + f_2(\boldsymbol{x}_i) = \hat{y}_i{}^{(1)} + f_2(\boldsymbol{x}_i) \end{cases} \qquad (5)$$

$$\textstyle\sum_{t=1}^{T} L(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(\boldsymbol{x}_i) + \Omega(f_i)\right)$$

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \qquad (6)$$

y_i is the actual value, $y_i^{(t)}$ is the predcited value. $\sum l$ represent the losses between each group y_i and $y_i^{(t)}$. $\Omega(f_i)$ is the regulation term. T represents the number of leaves, ω is the leaf weight, λ and γ are coefficients, with default values set for $\gamma = 0$ and $\lambda = 1$.

LightGBM has a characteristic that distinguishes it from other tree-boosting algorithms is that it splits the tree lengthwise (leaf-wise tree growth) with the best fit, while other tree-boosting algorithms split the tree depth-wise or level-wise (level-wise tree growth). Therefore, when growing on the same leaves in LightGBM, the leaf-wise algorithm can reduce more losses than the level-wise algorithm and produce much better accuracy that is not met by other boosting algorithms (Rufo et al, 2021). However, the leaf-wise algorithm is more prone to overfitting.

2.6. Measures of Model Goodness

Classification accuracy is a measure of classification accuracy that shows the overall performance of the classification technique (Nugroho et al., 2003). The higher the classification accuracy, the better the performance of the classification technique. It is common to measure classification performance using a confusion matrix. The confusion matrix contains information that compares the classification results performed by the system with the classification results that should be.

Table 2. Confusion Matrix

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	TP (True Positive)	FN (False Negative)
Actual Negative Class	FP (False Positive)	TN (True Negative)

True Positive (TP) is the total positive data records that are classified into positive values; False Positive (FP) is the total negative data records that are classified into positive values; False Negative (FN) is the total positive data records that are classified into negative values; True Negative (TN) is the total negative data records classified into negative values.

There are various measures in evaluating model performance based on the confusion matrix, including accuracy, precision, and F1 score. Accuracy in the confusion matrix is the percentage of correctness of the prediction results on the testing data. Precision is a measure of the proportion of TP predictions. Accuracy, specificity, precision, and F1 score can be obtained as Equations (7).

$$\begin{gathered}\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \\ \text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}} \\ \text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}\end{gathered} \tag{7}$$

Datasets with unbalanced classes and the same error cost can use a performance measure of error rate: error rate = 1- accuracy. However, the error rate could be a better performance measure in classifying unbalanced class data. In such cases, the performance measure of classification results is usually measured by recall as in equation (7) (Sokolova & Lapalme, 2009)

3. MATERIAL AND METHOD

3.1. Data Source

The data used in this research is secondary data taken from the UCI Machine Learning website that obtained in 1995 (Wolberg et al., 1995). This data contains information on cell nuclei characteristics in breast mass images. This data has 569 total observations with 30 columns as features (X) and 1 column as targets (Y) consisting of 2 classes, namely benign (negative cancer) and malignant (positive cancer).

3.2. Research Variables

The research variables used in this study consist of independent (X) and dependent (Y) variables. The Y variable of this study is the cancer category, which consists of 2 types: negative cancer (benign) and positive cancer (malignant). There are 10 independent variables obtained from features calculated based on cell images of breast masses, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These ten features were calculated based on the mean, maximum, and standard error values, resulting in 30 features. These features are commonly used in medical imaging analysis and have been found to be effective in distinguishing between benign and malignant tumors (Martellini et al., 2003; Hu et al., 2015).

3.3. Analysis Steps

The analysis steps of modeling using SVM and LightGBM are as follows: (1) Features in the dataset that have high collinearity with other features will be removed. Collinearity testing in this study uses the Pearson Correlation method with a high collinearity limit of 0.9; (2) The dataset is then divided into two parts: training data and testing data. The proportion used is 75% for training data and 25% for testing data; (3) Perform standardization to change the data to have an average value of 0 and a standard deviation of 1; (4) Search for the best parameters for each model using the Random Search method, which is evaluated using cross-validation; (5) Training each model using the appropriate parameters on a subset of the training data; (6) Evaluating model performance using a subset of testing data; (7) Interpret the performance results of each model and determine the best model.

4. RESULTS AND DISCUSSION

4.1. Parameter Search

The search for the best parameters used in this study uses the Random Search method with 100 iterations in both methods. The best parameters are determined based on the highest accuracy evaluation results on the cross-validation subset taken with the Repeated Stratified Cross Validation method with the number of folds of 3 and 5 iterations.

4.1.1. SVM Parameter Search

The SVM parameters searched in this study are the C parameter and the kernel used. Tuning the C parameter is essential as it controls SVM's regularization strength. A smaller C value allows for a more flexible decision boundary, potentially capturing intricate patterns in the data. On the other hand, a larger C value imposes a stricter margin, focusing on well-defined patterns.

The search space for the C parameter has a minimum value of 0.001 and a maximum value of 100 The range of 0.001 to 100 is choosen for the model to adapt with varying degrees of complexity in the dataset. at the same time, the kernel parameter consists of RBF (Radial basis Function), Polynomial, and Linear.

Table 3. SVM parameter search result

Step	Accuracy	C	Kernel
53	0.9417	33.0835	Polynomial
82	0.9300	0.1570	Linear
80	**0.9639**	**3.1065**	**RBF**

Table 3 shows that the SVM model with RBF (Radial Based Function) kernel and C value of 3.1065 has the best performance with an accuracy of 0.9639. The SVM model can achieve this accuracy after the 80th search using the Random Search method.

4.1.2. LightGBM Parameter Search

The LightGBM parameters searched in this research are the number of estimators, maximum binning, learning rate, and number of leaves. The search space for the number of estimators has a value of 500 to 2000, maximum binning has a value of 256 to 1024, the learning rate has a value of 0.00001 to 0.02, and the number of leaves has a value of 2 to 256.

Table 4. LightGBM Parameter Search Result

Step	Accuracy	learning rate	max bin	n estimator	n leaves
0	0.9324	0.0016	987	1000	154
8	0.9498	0.0079	985	600	208
12	0.9507	0.0081	398	1300	199
38	0.9521	0.0091	632	700	63
91	**0.9531**	**0.0076**	**552**	**2000**	**232**

Based on Table 4, LightGBM utilized the 'n_estimator' parameter, which refers to the number of assembled trees, with a total of 2000 trees, and each tree had a maximum of 232 leaves. The model yielded the best results with a maximum number of 552 bins and a learning rate of 0.0076, achieving an accuracy of 0.9531. The LightGBM model attain this accuracy after the 91st search using the Random Search method.

4.2. Perfomance Result

SVM and LightGBM models with tuned parameters are then evaluated using a subset of test data. This test measures how the two models perform on datasets that the model has never seen. Performance measurements can be seen through the confusion matrix results and other metric evaluations such as accuracy, F1 score, precision, and recall.

Based on Figure 3, for malignant as the positive class, it can be determined that SVM has fewer false negatives than LightGBM. The SVM model has three false negatives, and LightGBM has six. Whereas in false positives, both models have no predictions that show false positives.

Based on the metric evaluation results in Table 5, the SVM model has better performance than the LightGBM model in the benign class as a positive class with metric values of Accuracy, F1-Score, and Recall of 0.9790, 0.9836, and 0.9677. As for the Malignant class as a positive class, SVM excels in Accuracy, F1-score, and Recall metrics with values of 0.9790, 0.9709, and 0.9434.

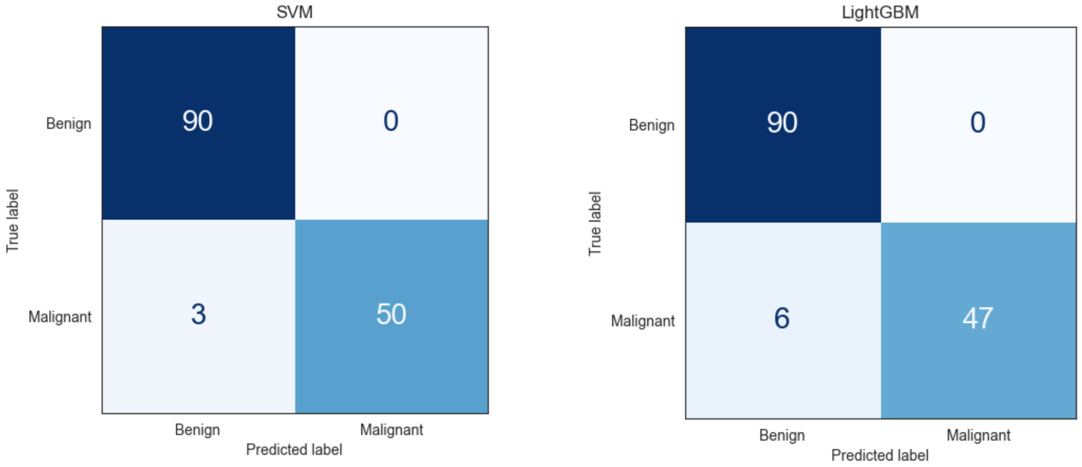

Figure 3. Confusion Matrix Model SVM and LightGBM

Table 5. Classification Report

Class	Metrics	SVM	LightGBM
Benign	Accuracy	**0.9790**	0.9580
	F1 Score	**0.9836**	0.9677
	Precision	**0.9677**	0.9375
	Recall	1.0000	1.0000
Malignant	Accuracy	**0.9790**	0.9580
	F1 Score	**0.9709**	0.9400
	Precision	1.0000	1.0000
	Recall	**0.9434**	0.8868

Based on the evaluation results above, the SVM model on malignant as a positive class has a smaller false negative error than the LightGBM model. This is indicated by the SVM recall value of 0.9434, higher than LightGBM, with a value of 0.8868. However, the SVM model does not differ much from the LightGBM model in the benign class as a positive class. This can be seen through the metric value in the benign class, which has a pretty close difference value. Based on the analysis, the SVM model performs better than the LightGBM model. This is indicated by the higher value of the Accuracy, F1-Score, Precision, and Recall metrics tested on each model.

According to the evaluation metrics in Table 5, the SVM model outperforms the LightGBM model in the benign class as a positive class, with accuracy, F1-Score, and recall values of 0.9790, 0.9836, and 0.9677, respectively. As for the malignant class, SVM demonstrates better accuracy, F1-score, and recall metrics with values of 0.9790, 0.9709, and 0.9434. The above results indicate that the SVM model has a lower false negative error than the LightGBM model for the malignant class as a positive class. This is evident from the SVM recall value of 0.9434, higher than LightGBM's value of 0.8868. However, the difference between the SVM and LightGBM models is not significant in the benign class as a positive class, as the metric values are pretty similar. Based on the analysis, the SVM model performs better than the LightGBM model. This is indicated by the higher accuracy, F1-Score, precision, and recall metrics obtained from testing each model.

The LightGBM and SVM models have better performance when compared to the previous research model, namely KNN, with an accuracy score of 85% and 74.37%. Modelling using Lightgbm and SVM has a good performance, with a lightGBM score of

95.80% and an SVM of 97.90%. This score can be achieved due to the addition of the Random Search method to find the best parameters in each model. LightGBM and SVM have their respective advantages; in this study, LightGBM has a faster training time than SVM. On the other hand, LightGBM has quite a lot of parameters to tune when compared to SVM. If the dataset later becomes larger, researchers recommend using LightGBM because it has a shorter training time.

5. CONCLUSION

Analysis and discussion of the comparison of SVM and LightGBM models on breast cancer classification using datasets from UCI Machine Learning shows that the SVM model has an accuracy value of 97.9%, F1-Score of 0.9836 for Benign as a negative class and 0.9709 for Malignant as a positive class. In comparison, the LightGBM model has an accuracy value of 97.9%, F1-Score of 0.9677 for Benign as a positive class and 0.9400 for Malignant as a positive class. Based on these results, the SVM model is more appropriate to be used as a model for breast cancer classification because it has better performance than the LightGBM model.

The potential of SVM to improve medical practice by providing reliable and precise decision support systems to medical practitioners is demonstrated by its application in the breast cancer classification. By the utilization of SVM, medical professionals can improve the patient care and management by enhancing their ability to classify and describe breast cancer. Moreover, applying SVM to the classification of breast cancer fits in with the overarching objective of utilizing technology and data-driven methods to tackle healthcare issues, espescially in the areas of cancer diagnosis and detection.

ACKNOWLEDGMENT

This research was supported by LPPM Universitas Diponegoro scheme Riset Pengembangan dan Penerapan (RPP), number of contract 609-16/UN7.D2/PP/VIII/2023.

REFERENCES

Aditya, A. R., Suparti, S., & Sudarno, S. (2015). Ketepatan Klasifikasi Pemilihan Metode Kontrasepsi Di Kota Semarang Menggunakan Booststrap Aggregatting Regresi Logistik Multinomial. *Jurnal Gaussian*, *4*(1), 11–20. https://doi.org/10.14710/J.GAUSS.4.1.11-20

Bustamam, A., Bachtiar, A., & Sarwinda, D. (2019). Selecting Features Subsets Based on Support Vector Machine-Recursive Features Elimination and One Dimensional-Naïve Bayes Classifier using Support Vector Machines for Classification of Prostate and Breast Cancer. *Procedia Computer Science*, *157*, 450–458. https://doi.org/10.1016/j.procs.2019.08.238

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., & Cardoso, F. (2019). Breast Cancer. *Nature Reviews Disease Primers*, *5*(1), 66. https://doi.org/10.1038/s41572-019-0111-2

Hanmastiana, I. M., Warsito, B., Rahmawati, R., Yasin, H., & Kartikasari, P. (2022). Classification of Public Opinion on Social Media Twitter concerning the Education in Indonesia Using the K-Nearest Neighbors (K-NN) Algorithm and K-Fold Cross

Validation. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *21*(2), 99–106. https://doi.org/10.29313/statistika.v21i2.297

Hu, Z. D., Zhou, Z. R., & Qian, S. (2015). How to Analyze Tumor Stage Data in Clinical Research. *Journal of Thoracic Disease*, *7*(4), 566–575. https://doi.org/10.3978/j.issn.2072-1439.2015.04.09

Ispriyanti, D. & Hoyyi, A. (2016). Analisis Klasifikasi Masa Studi Mahasiswa Prodi Statistika Undip dengan Metode Support Vector Machine (Svm) Dan Id3 (Iterative Dichotomiser 3). *Media Statistika*, *9*(1), 15–29. https://doi.org/10.14710/medstat.9.1.15-29

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*. https://github.com/Microsoft/LightGBM.

Kemenkes RI. (2018). *Panduan Penatalaksanaan Kanker Payudara*. http://kanker.kemkes.go.id/guidelines/PPKPayudara.pdf

Khandezamin, Z., Naderan, M., & Rashti, M. J. (2020). Detection and Classification of Breast Cancer Using Logistic Regression Feature Selection and GMDH Classifier. *Journal of Biomedical Informatics*, *111*, 103591. https://doi.org/10.1016/j.jbi.2020.103591

Kurniawan, D., Suparti, & Sugito. (2018). Classification Accuracy on the Family Planning Participation Status Using Kernel Discriminant Analysis. *Journal of Physics: Conference Series*, *1025*, 012111. https://doi.org/10.1088/1742-6596/1025/1/012111

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics*, *8*(5), 765. https://doi.org/10.3390/math8050765

Marianto, F. Y., Tarno, T., & Maruddani, D. A. I. (2020). Perbandingan Metode Naïve Bayes Dan Bayesian Regularization Neural Network (Brnn) Untuk Klasifikasi Sinyal Palsu Pada Indikator Stochastic Oscillator (Studi Kasus: Saham PT Bank Rakyat Indonesia (Persero) Tbk Periode Januari 2017 – Agustus 2019). *Jurnal Gaussian*, *9*(1), 16–25. https://doi.org/10.14710/J.GAUSS.9.1.16-25

Nugraha, B. W., Widiharih, T., & Kartikasari, P. (2020). Perancangan Multidimensional Scalling Metrik dengan GUI PYTHON 3.8 untuk Klasifikasi Program Keluarga Berencana. *Jurnal Statistika Universitas Muhammadiyah Semarang*, *8*(2), 114. https://doi.org/10.26714/jsunimus.8.2.2020.114-120

Nugroho, A. S., Witartoo, A. B., & Handoko, D. (2003). Application of Support Vector Machine in Bioinformatics. *Proceeding of Indonesian Scientific Meeting in Central Japan*, December 20, 2023.

Pederson, H. J. & Noss, R. (2020). Updates in Hereditary Breast Cancer Genetic Testing and Practical High Risk Breast Management in Gene Carriers. *Seminars in Oncology*, *47*(4), 182–186. https://doi.org/10.1053/j.seminoncol.2020.05.008

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi menggunakan MATLAB*. Yogyakarta: ANDI.

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics*, *11*(9), 1714. https://doi.org/10.3390/diagnostics11091714

Rutgers, E., Balmana, J., Beishon, M., Benn, K., Evans, D. G., Mansel, R., Pharoah, P., Perry Skinner, V., Stoppa-Lyonnet, D., Travado, L., & Wyld, L. (2019). European Breast Cancer Council manifesto 2018: Genetic Risk Prediction Testing in Breast Cancer. *European Journal of Cancer*, *106*, 45–53. https://doi.org/10.1016/j.ejca.2018.09.019

Saxena, S., & Gyanchandani, M. (2020). Machine Learning Methods for Computer-Aided Breast Cancer Diagnosis Using Histopathology: A Narrative Review. *Journal of Medical Imaging and Radiation Sciences*, *51*(1), 182–193. https://doi.org/10.1016/j.jmir.2019.11.001

Setiawan, F. H., Rahmawati, R., & Suparti, S. (2015). Ketepatan Klasifikasi Keikutsertaan Keluarga Berencana Menggunakan Regresi Logistik Biner dan Regresi Probit Biner (Studi Kasus di Kabupaten Semarang Tahun 2014). *Jurnal Gaussian*, *4*(4), 845–854. https://doi.org/10.14710/J.GAUSS.4.4.845-854

Setiawan, Y. (2023). Data Mining berbasis Nearest Neighbor dan Seleksi Fitur untuk Deteksi Kanker Payudara. *Jurnal Informatika: Jurnal Pengembangan IT*, *8*(2), 89–96. https://doi.org/10.30591/jpit.v8i2.4994

Shi, Y., Dai, D., Liu, C., & Yan, H. (2009). Sparse Discriminant Analysis for Breast Cancer Biomarker Identification and Classification. *Progress in Natural Science*, *19*(11), 1635–1641. https://doi.org/10.1016/j.pnsc.2009.04.013

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(1), 7–33. https://doi.org/10.3322/caac.21708

Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Umma, F. N., Warsito, B., & Maruddani, D. A. I. (2021). Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 Di Kabupaten Pemalang. *Jurnal Gaussian*, *10*(2), 221–229. https://doi.org/10.14710/j.gauss.v10i2.29934

van der Giessen, J. A. M., van Dulmen, S., Velthuizen, M. E., van den Muijsenbergh, M. E. T. C., van Engelen, K., Collée, M., van Dalen, T., Aalfs, C. M., Hooning, M. J., Spreeuwenberg, P. M. M., Fransen, M. P., & Ausems, M. G. E. M. (2021). Effect of a Health Literacy Training Program for Surgical Oncologists and Specialized Nurses on disparities in referral to Breast Cancer Genetic Testing. *The Breast*, *58*, 80–87. https://doi.org/10.1016/j.breast.2021.04.008

Wardani, N. S., Prahutama, A., & Kartikasari, P. (2020). Analisis Sentimen Pemindahan Ibu Kota Negara dengan Klasifikasi Naïve Bayes untuk Model Bernoulli dan Multinomial. *Jurnal Gaussian*, *9*(3), 237–246. https://doi.org/10.14710/J.GAUSS.9.3.237-246

WHO. (2021). *Breast Cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Widodo, A. M., Anwar, N., Irawan, B., Meria, L., & Wisnujati, A. (2021). Komparasi Perfomansi Algoritma Pengklasifikasi KNN, Bagging dan Random Forest untuk Prediksi Kanker Payudara. *KONIK (Konferensi Nasional Ilmu Komputer)*, *5*, 367–372.

Woldberg, W., Mangansarian, O., Street, N., & W. Street. (1995). Breast Cancer Winconsin (Diagnostic). *UCI Machine learning Repository*. https://doi.org/https://doi.org/10.24432/C5DW2B

Wood, A., Shpilrain, V., Najarian, K., & Kahrobaei, D. (2019). Private Naive Bayes Classification of Personal Biomedical Data: Application in Cancer Data Analysis. *Computers in Biology and Medicine*, *105*, 144–150. https://doi.org/10.1016/j.compbiomed.2018.11.018