IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.17, No.4, October 2023, pp. 429~440
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.88926

Backward Elimination for Feature Selection on Breast Cancer Classification Using Logistic Regression and Support Vector Machine Algorithms

Salsha Farahdiba*[1], Dwi Kartini[2], Radityo Adi Nugroho[3], Rudy Herteno[4], Triando H. Saragih[5]
[1,2,3,4,5]Department of Computer Science, Faculty of Mathematic and Natural Science, Lambung Mangkurat University, Banjarmasin, Indonesia
e-mail: ***[1]salshafrhdba@gmail.com**, [2]dwikartini@ulm.ac.id, [3]radityo.adi@ulm.ac.id, [4]rudy.herteno@ulm.ac.id, [5]triando.saragih@ulm.ac.id

Abstrak

Kanker payudara merupakan jenis kanker yang paling banyak diderita oleh wanita di seluruh negara di dunia. Salah satu cara untuk mencegah tingginya angka kematian akibat penyakit tersebut adalah dengan sistem deteksi keganasan kanker. Algoritma klasifikasi Regresi Logistik dan Support Vector Machine (SVM) sering digunakan untuk deteksi penyakit kanker payudara, tetapi penggunaan kedua algoritma ini seringkali tidak memberikan hasil yang optimal jika diterapkan pada dataset yang memiliki banyak fitur, sehingga diperlukan algoritma tambahan untuk meningkatkan kinerja klasifikasi dengan menggunakan seleksi fitur Backward Elimination. Perbandingan klasifikasi Regresi Logistik dan SVM dilakukan dengan menerapkan seleksi fitur pada data kanker payudara untuk melihat model terbaik. Dataset kanker payudara memiliki 30 fitur dan dua kelas yaitu Benign dan Malignant. Penerapan Backward Elimination mengurangi fitur dari 30 fitur menjadi 13 fitur, sehingga memengaruhi peningkatan performa kedua model klasifikasi. Klasifikasi terbaik diperoleh menggunakan seleksi fitur Backward Elimination dan SVM kernel linier dengan peningkatan nilai akurasi dari 96,14% menjadi 97,02%, nilai presisi dari 98,06% menjadi 99,49%, nilai recall dari 90,48% menjadi 92,38%, dan nilai AUC dari 0,95 menjadi sebesar 0,96.

Kata kunci*—Klasifikasi Kanker Payudara, Backward Elimination, Regresi Logistik, SVM*

Abstract

Breast cancer is a prevalent form of cancer that afflicts women across all nations globally. One of the ways that can be done as a prevention to reduce elevated fatality due to breast cancer is with a detection system that can determine whether a cancer is benign or malignant. Logistic Regression and Support Vector Machine (SVM) classification algorithms are often used to detect this disease, but using these two algorithms often does not give optimal results when applied to datasets with many features, so an additional algorithm is needed to improve classification performance by using Backward Elimination feature selection. Logistic Regression and SVM algorithms were compared by applying feature selection to breast cancer data to see the best model. The breast cancer dataset has 30 features and two classes, Benign and Malignant. Backward Elimination has reduced features from 30 features to 13 features, thereby increasing the performance of both classification models. The best classification was obtained by using the Backward Elimination feature selection and linear kernel SVM with an increase in accuracy value from 96.14% to 97.02%, precision from 98.06% to 99.49%, recall from 90.48% to 92.38%, and the AUC from 0.95 to 0.96.

Keywords—*Breast Cancer Classification, Backward Elimination, Logistic Regression, SVM*

Received September 11th,2023; Revised September 21th, 2023; Accepted September 27th, 2023

1. INTRODUCTION

Breast cancer, a highly prevalent form of cancer, is a global affliction affecting women across all nations. In 2020, over 2 million new cases were reported [1]. One of the ways that can be done as a prevention to reduce the elevated fatality rate due to breast cancer is with a detection system that can determine whether a cancer is benign or malignant [2]. Accurate early diagnosis in detecting the type of cancer plays a vital role in treating patients because the prompt diagnosis of cancer allows for the timely administration of the appropriate treatment.

The cancer-type detection system can be applied through data mining algorithms with classification techniques. According to the research conducted by Siswa, it has been determined that the best classification algorithms for detecting the type of breast cancer are Logistic Regression and SVM. These two algorithms have demonstrated the highest accuracy values, equivalent to 96.8% [3]. Research by Chen et al. [4] on breast cancer classification using Logistic Regression proved that using Logistic Regression can obtain a good accuracy with a value of 94%. Other research on breast cancer diagnosis also shows that the SVM algorithm [5] is better than the other classification algorithms such as Naïve Bayes and Decision Tree. According to Ing et al. [6], the SVM algorithm, as the latest statistical algorithm, is useful to compare with traditional Logistic Regression regarding medical classification because both algorithms have the same type of data variable value, namely the response variable, which uses nominal variable values. Apart from that, SVM itself, as a modern classification algorithm, is claimed to have optimal global results compared to conventional statistical algorithms such as Logistic Regression [7]. These two classification algorithms were selected based on their demonstrated ability to produce accurate results in the context of breast cancer classification. However, their use may not always yield optimal outcomes when applied to datasets with numerous features, necessitating the inclusion of supplementary algorithms to enhance classification performance through feature selection.

Feature selection is a technique that aims to handle datasets with many features by reducing irrelevant features, leaving only the best features in a dataset [8]. Backward Elimination is an algorithm for selecting features that have been demonstrated to enhance the efficacy of classification algorithms. Backward Elimination works by selecting features backward to get the most relevant features in the classification process [9]. Research by Resmiati et al. [10] proved that combining Backward Elimination with the SVM algorithm increased the high and significant accuracy value by 30.43%. Apart from that, the addition of Backward Elimination also demonstrated the capability to enhance the accuracy of the Logistic Regression algorithm model [11]. Therefore, this research used the same additional algorithm, namely Backward Elimination, as a feature selection for Logistic Regression and SVM.

Based on the description above, the researcher proposes a comparison between two classification algorithms, which are Logistic Regression and SVM, by adding Backward Elimination feature selection. The application of feature selection in this research aims to maximize performance on breast cancer classification with a dataset that has 30 features, compared to previous research, which tends to have fewer features than this research. The researcher created 4 test models, namely two basic classification models, Logistic Regression and SVM, and two other models with the addition of Backward Elimination feature selection. The performance of each of the four models is evaluated to see which model is the best, and Backward Elimination is expected to show relevant features in the dataset so that it can be seen which features are important and which are not in breast cancer classification.

2. METHODS

This section describes the dataset used, the Backward Elimination for feature selection, the model validation using cross-validation, the theory of Logistic Regression and SVM algorithms, and the performance measurement as the evaluation algorithm using accuracy, precision, recall, and AUC. The flowchart of this research is illustrated in **Figure 1**.

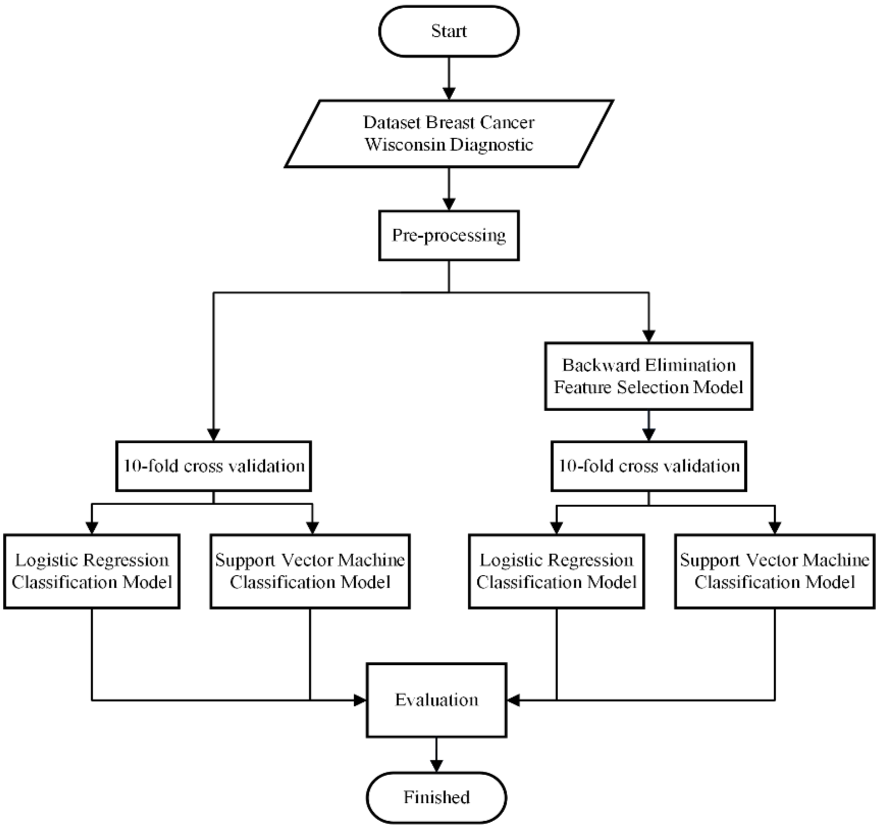
Start
Dataset Breast Cancer Wisconsin Diagnostic
Pre-processing
Backward Elimination Feature Selection Model
10-fold cross validation
10-fold cross validation
Logistic Regression Classification Model
Support Vector Machine Classification Model
Logistic Regression Classification Model
Support Vector Machine Classification Model
Evaluation
Finished

Figure 1 The Research Flowchart

2.1 Data Collection

The dataset utilized in this research is the Breast Cancer Wisconsin Diagnostic, which comprises two distinct classes: Benign (B) and Malignant (M). The dataset comprises 570 data, with 30 predictor features distributed across ten groups for each cell nucleus and a single feature designated as the target. The features extracted originate from digital images of breast masses. Among the 570 data, 357 are included as benign class, and 213 are malignant class [12], [13]. It is imperative to note that all independent variables in this dataset are of numeric data categories, while the dependent or target variable is binary.

2.2 Pre-processing

Pre-processing is the initial stage in machine learning, which involves data encoding or transforming data so that the model's algorithm can quickly analyze the data features [14]. The initial stage of data pre-processing entails identifying problematic data, encompassing data with missing values [15]. In this research, no missing values were found in the dataset, so no further action needed to be taken. The next stage involves aligning the data with the requisite data type by the classification algorithm. It is required when features on a different scale are considered [15]. The target feature used in this research is categorical data, so they must be changed first into nominal data. Converting data into nominal form is intended so that the data to be processed follows the analysis algorithm, making it easier to process the data [16]. At this stage, the value of the "Diagnosis" feature is changed, namely mapping the value of "M" which means malignant, to 1 and the value of "B" which means benign, to 0. This process is important because the Logistic Regression and SVM algorithms use nominal data.

2.3 Backward Elimination

Backward Elimination is a feature selection algorithm that uses a collection of feature combinations to find the best combination recursively. This algorithm works by testing all the features first and then gradually reducing features that are not significant based on a comparison of the evaluation of test results obtained from each combination of these features [17]. Firstly, it is imperative that a comprehensive evaluation of all features is carried out using a regression model, with a predetermined level of significance set at 0.05. Any feature with a p-value exceeding the established significance level (p > 0.05) shall be eliminated. This iterative procedure shall be repeated until only features exhibiting a p-value below 0.05 are retained [18]. The benefits of this algorithm include enhanced training time, improved performance, and reduced complexity. Backward Elimination is helpful for selecting relevant features before entering the model testing stage. It uses regression statistics to determine the closeness of each feature combination to the target. The smaller the significance level, the stricter the selection of features so that fewer features are selected in the model. The Backward Elimination algorithm operates in a way that is based on a linear regression. The steps of the Backward Elimination are illustrated in **Figure 2**.

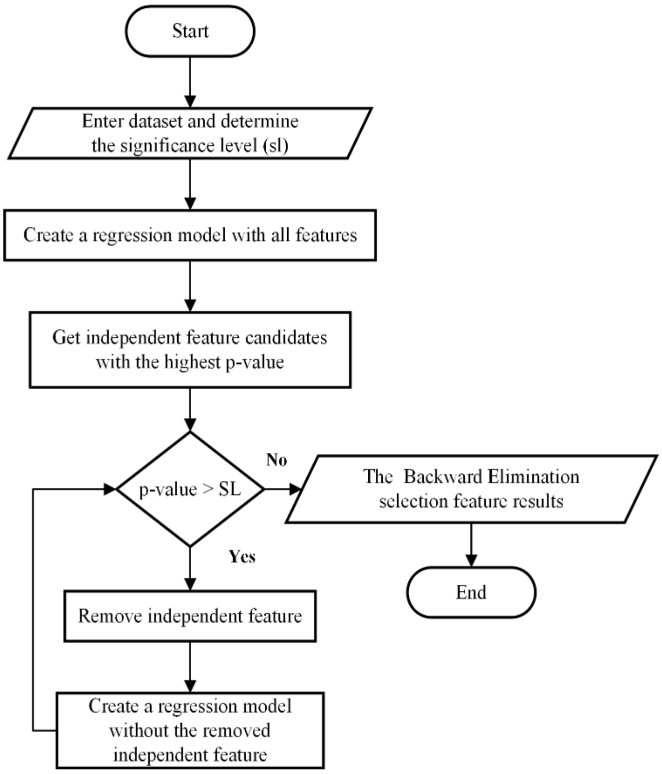
Start
Enter dataset and determine the significance level (sl)
Create a regression model with all features
Get independent feature candidates with the highest p-value
p-value > SL
No
Yes
The Backward Elimination selection feature results
End
Remove independent feature
Create a regression model without the removed independent feature

Figure 2 The Backward Elimination Flowchart

2.4 Cross-Validation

Cross-validation is a technique utilized for model validation to assess the accuracy of analysis results. The pre-processed data is cross-validated by dividing the data into training and testing data for the classification process. This technique evaluates how accurately a predictive model performs when run and is primarily used to make model predictions in practice [19]. In the cross-validation, the assessment may vary based on the level of proficiency in partitioning the training set and testing set. The k-fold cross-validation is widely regarded as one of the most frequently preferred and efficient algorithms [20]. Overall, it is recommended and universally agreed upon that a 10-fold cross-validation approach be employed. Data partitioning was carried out using the k-fold cross-validation technique, where the value of k was set at 10.

2.5 Classification

2. 5.1 Logistic Regression

Logistic Regression is a model that describes the relationship between several independent variables and the dependent variable that are dichotomous (two categories) or polychotomous (more than two categories). If the dependent variable comprises two categories, the binary Logistic Regression form can be used. The mathematical formulas for the Logistic Regression model are shown in **equation (1)** and **equation (2)** [21]:

$$\mathrm{P}\,(\mathrm{y} \mid \mathrm{x}) = \pi(x) \tag{1}$$

and

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})} \tag{2}$$

In the Logistic Regression approach, probability calculations are carried out to predict class. The determination of classification based on the binary response variable is contingent on the reference to the cut-point value. The cut point that can be used is 0.5. If the opportunity resulting from the model is less than 0.5, then the prediction result is category 0, while if the opportunity resulting from the model is greater than or equal to 0.5, then the prediction result is category 1 [21].

The following are the steps for classifying the Logistic Regression model [22]:

1. Determine the dependent variable, broken down into binary data scores, namely 0 and 1, where 0 is for failure events and 1 is for success events.
2. Determine the coefficient value (β) of each independent variable and intercept (a) by using **equation (3)** and **equation (4)**:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

and

$$a = \bar{\bar{y}} - \beta\bar{x} \tag{4}$$

3. Form a logistic model, because the logistic model is a non-linear form, transformation is needed to get a linear model, so the following model is obtained by using **equation (5)**:

$$g(x) = \ln\frac{\pi(x)}{1 - \pi(x)} = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots + \beta_i x_i \tag{5}$$

4. Form a probability model from the logsitic function obtained according to **equation (2)**.
5. Test the regression model that has been created by inputting data into the logistic model and probability model. If the value of π(x) < 0.5 then it is in class 0, whereas if π(x) ≥ 0.5 then it is in class 1

.

2. 5.2 Support Vector Machine (SVM)

SVM, as a prominent classification algorithm, is capable of addressing classification challenges in binary datasets. This supervised learning technique, derived from statistical theory, is widely utilized in classification and regression analyses. The primary objective of SVM is to establish an optimal high-dimensional classification hyperplane. The SVM framework endeavors to identify the most effective hyperplane that serves as a separator of two classes in the input space. The pattern is classified into two classes, the positive (+1) and the negative (-1) [23]. In the SVM algorithm transformation process, there is a mathematical function known as the kernel function. The kernel function aims to classify non-linear data. The method is to change non-linear data into linear data and then form a hyperplane. Based on previous research [24], it is known that the best SVM kernel model for dealing with breast cancer detection problems is the linear kernel, so this research will use the same kernel.

The following are the stages for the linear kernel SVM [25]:

1. Find the kernel matrix value in the training data
   To find the kernel value, use the following linear kernel function by using **equation (6)**:

$$K\ (xi, xj = xi^{T}\ . xj) \tag{6}$$

2. Looking for the alpha value (*a*)
   The alpha value can be obtained by using the **equation (7)**:

$$a_i = \frac{N}{\sum_{i=1}^{N}(K(X_iX_j)y_iy_j)} \tag{7}$$

3. Calculate the weight (*w*) and bias (*b*) values
   To find the weight and bias values, use **equation (8)** and **equation (9)**:

$$w_i\ =\ \sum_{i=1}^{l} a_i y_i\, x_i\ = 0 \tag{8}$$

$$b = -\frac{1}{2}\ (wx^{+} +\ wx^{-}) \tag{9}$$

4. Determine the test kernel value
5. Determine the value of the predicted result f(∅ (x))
   To test and classify data using the **equation (10)**:

$$f(\emptyset(x)) =\ sign(w_1x_1 + w_2x_2\ +\ w_nx_n\ + b) \tag{10}$$

2.6 Confusion Matrix

The utilization of the confusion matrix is prevalent in machine learning, particularly in the context of supervised classification models. It represents the model evaluation results using a matrix table [26]. For the binary classification, the representation of the confusion matrix is displayed as a 2*2 matrix., containing four distinct terms that signify the outcome of the classification, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [27]. The gathered data is organized in a tabular format and differentiated based on the predicted and actual values by classifying them into four categories. The confusion matrix for binary classification is indicated in Table 1.

Table 1 The Confusion Matrix

Actual Class	**Prediction Class**	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Based on Table 1, the classification performance can be evaluated by computing accuracy, precision, and recall values. Accuracy represents the percentage of correctly classified data records after testing of classification outcomes. On the other hand, precision is the proportion of classes that are predicted positively and correctly based on actual data. Recall pertains to the proportion of correctly predicted positive classes out of the actual positive ones. These following **equation (11)**, **equation (12)**, and **equation (13)** are the formula for measuring performance based on the confusion matrix in [28]:

$$\text{Accuracy} = \frac{tp + tn}{tp+tn+fp+fn} \tag{11}$$

$$\text{Precision} = \frac{tp}{tp + fp} \tag{12}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{13}$$

2.7 Area Under the ROC (Receiver Operating Characteristic) Curve

Area Under The Curve (AUC) is a widely utilized metric for computing the value beneath the Receiver Operating Characteristic (ROC) curve. The AUC value will always be in the range 0-1 because the part of the unit square area with the x-axis and y-axis has values from 0 to 1. If the result value is above 0.5, then the AUC value is considered interesting because it produces random predictions of the diagonal line between (0,0) and (1,1), with an area of 0.5. The AUC value for data mining categorization is indicated in Table 2 [29].

Table 2 Classification Quality Based on AUC Value

AUC Value	**Category**
0.90 — 1.00	Excellent Classification
0.80 — 0.90	Good Classification
0.70 — 0.80	Fair Classification
0.60 — 0.70	Poor Classification
0.50 — 0.60	Failure

AUC measures of quality between negative and positive values with a single value. AUC size was calculated as the area of the ROC curve by using **equation (14)** [30]:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{14}$$

3. RESULTS AND DISCUSSION

3.1 The Result of Backward Elimination

The feature selection algorithm used in this research is Backward Elimination, where all the features will be tested first, then gradually reduce the features that are not significant based on the evaluation of the test results obtained. The determination of the significant level is contingent on a prescribed threshold of 5% or 0.05. This refers to research [18],[31], using a significant or p-value of 0.05. Features with a p-value greater than the predetermined significant value will be eliminated. This stage continues to be repeated and will stop until the remaining features have a significant level of no more than 0.05.

Based on the results of tests on 30 features, 17 repetitions were obtained until the remaining features with a p-value below the predetermined value were obtained. Of the 30 features, only 13 remain, which are used as dataset to be tested at the classification stage. The final result of the Backward Elimination selection with 13 features remaining is indicated in Table 3.

Table 3 The Final Result of Backward Elimination

No.	Features	p-value
1.	radius_mean	0.0158
2.	compactness_mean	0.0000
3.	concave points_mean	0.0467
4.	radius_se	0.0126
5.	smoothness_se	0.0000
6.	concavity_se	0.0000
7.	concave points_se	0.0038
8.	radius_worst	0.0000
9.	texture_worst	0.0000
10.	area_worst	0.0000
11.	concavity_worst	0.0031
12.	symmetry_worst	0.0002
13.	fractal_dimension_worst	0.0019

3.2 Logistic Regression

In the testing process with the Logistic Regression algorithm, two models were created: the basic Logistic Regression model and the combination model with Backward Elimination. The basic Logistic Regression model used the dataset with 30 features, while the proposed algorithm with Backward Elimination used only 13 features. Based on the test results, accuracy, precision, recall, and AUC values are obtained. The performance of both models is indicated in Table 4.

Table 4 The Performance of Logistic Regression Model

	Accuracy	Precision	Recall	AUC
Logistic Regression	94.04%	95.36%	88.1%	0.93
Backward Elimination with Logistic Regression	95.43%	96.46%	90.95%	0.95

From Table 4, it is evident that the performance of the Logistic Regression algorithm has increased after adding the Backward Elimination algorithm. The accuracy value increased by 1.39%, the precision value increased by 1.1%, and the recall value increased by 2.85%. The AUC value also improved well and was classified as the Excellent Classification. Based on the results, it is evident that Backward Elimination can enhance the performance of the Logistic Regression algorithm, even with fewer features, without reducing previous performance.

3.3 The Linear Kernel SVM

In the testing process with the linear kernel SVM algorithm, two models were created: the basic SVM model and the combination model with Backward Elimination. The basic linear kernel SVM model used the dataset with 30 features, while the proposed algorithm with Backward

Elimination used only 13 features. Based on the test results, accuracy, precision, recall, and AUC values are obtained. The performance of both models is indicated in Table 5.

Table 5 The Performance of Linear Kernel SVM Model

	Accuracy	**Precision**	**Recall**	**AUC**
Linear Kernel SVM	96.14%	98.06%	90.48%	0.95
Backward Elimination with the Linear Kernel SVM	97.02%	99.49%	92.38%	0.96

From Table 5, it is evident that the performance of the SVM algorithm also has increased after adding the Backward Elimination algorithm. The accuracy value increased by 0.88%, the precision value increased by 1.43%, and the recall value increased by 1.9%. The AUC value also improved well and was classified as the Excellent Classification. Based on the results, it shows that Backward Elimination is also evident that Backward Elimination can enhance the performance of the SVM algorithm, even with fewer features, without reducing previous performance.

3.4 Comparison Results of Logistic Regression and SVM Algorithms

Based on the four models that have been made, the combination model of Backward Elimination with linear kernel SVM shows the highest performance, so it can be concluded that this model is the best model to use among all. The comparison results from all of the models are illustrated in **Figure 3**.

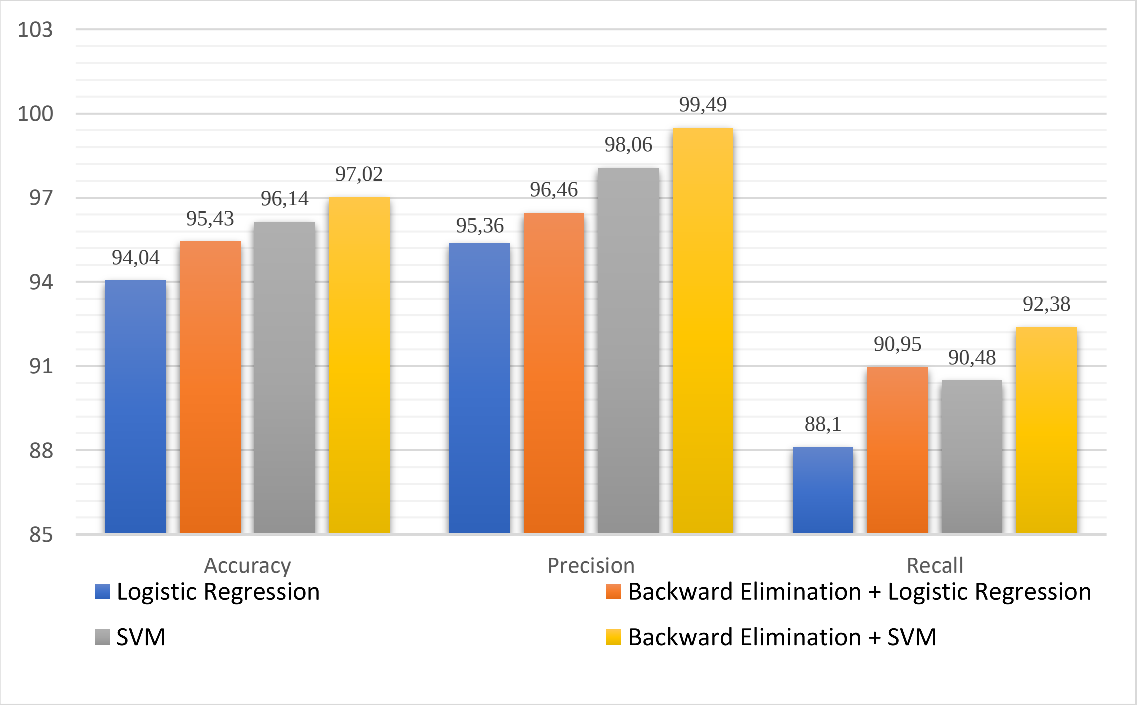

Figure 3 The Comparison Results

The AUC value was also taken into consideration in this research. The AUC results in the Logistic Regression and linear kernel SVM algorithms increased after adding the Backward Elimination algorithm. Both algorithms are included in the Excellent Classification category based on Table 2. The comparison results of the AUC value can be seen in **Figure 4**.

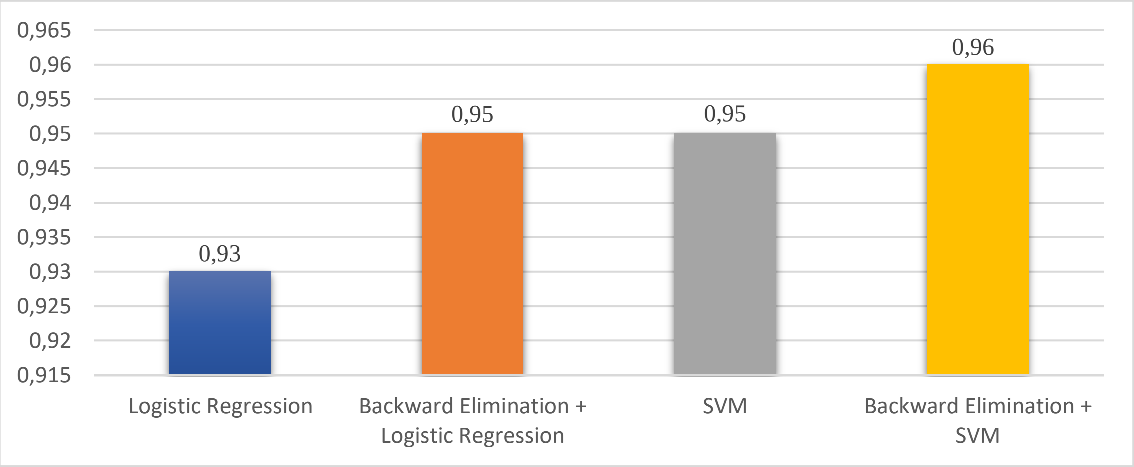

Figure 4 The Comparison Results of the AUC

Comparing the outcomes of this research with prior studies showed that the Logistic Regression and SVM algorithms combined with the Backward Elimination feature selection have great performance in classifying breast cancer. This matters because combining these algorithms increases the accuracy, precision, recall values, and AUC value compared to previous studies that only used classification algorithms or without Backward Elimination feature selection. Compared to previous research, which only used classification algorithms [5], [32], the algorithm proposed in this research exhibits enhanced performance in the classification of breast cancer dataset. In addition, this comparative analysis also facilitates a better understanding of the potential of the feature selection algorithm and the algorithm used in this research to produce better results than previous approaches.

4. CONCLUSIONS

Based on the findings of this research, it has been demonstrated that the utilization of Backward Elimination in both Logistic Regression and SVM classifications yields an impact on improving performance. Backward Elimination effectively reduced the number of features from 30 features to 13 features without compromising the accuracy of each classification model. Besides that, reducing the number of features processed by algorithms can reduce computing power and make algorithm calculations lighter.

From the various models employed to predict breast cancer, it can be concluded that the combination of Backward Elimination with the linear kernel SVM yields the best performance with an accuracy value of 97.02%, precision value of 99.49%, recall value of 92.38%, and AUC value is 0.96. The model is included in the excellent classification category.

For further research, it is highly advised to employ larger and more varied datasets and consider alternative methods of feature selection and classification algorithms to provide increased performance in the classification of breast cancer.

REFERENCES

[1] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, “Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review,” *Cancers*, vol. 13, no. 17. MDPI, Sep. 01, 2021. doi: 10.3390/cancers13174287.

[2] A. Al Bataineh, “A comparative analysis of nonlinear machine learning algorithms for breast cancer detection,” *Int J Mach Learn Comput*, vol. 9, no. 3, pp. 248–254, Jun. 2019, doi: 10.18178/ijmlc.2019.9.3.794.

[3] T. Azhima Yoga Siswa, “Perbandingan Kinerja Algoritma C4.5, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, dan Support Vector Machines Untuk Mendeteksi Penyakit Kanker Payudara,” 2018. [Online]. Available: http://archive.ics.uci.edu.

[4] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, “Classification Prediction of Breast Cancer Based on Machine Learning,” *Comput Intell Neurosci*, vol. 2023, pp. 1–9, Jan. 2023, doi: 10.1155/2023/6530719.

[5] L. Indah Prahartiwi, W. Dari, and S. Nusa Mandiri, “Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine untuk Prediksi Penyakit Kanker Payudara,” *Jurnal Teknik Komputer AMIK BSI*, vol. 7, no. 1, 2021, doi: 10.31294/jtk.v4i2.

[6] E. Ing, W. Su, M. Schonlau, and N. Torun, “Support Vector Machines and logistic regression to predict temporal artery biopsy outcomes,” *Canadian Journal of Ophthalmology*, vol. 54, no. 1, pp. 116–118, Feb. 2019, doi: 10.1016/j.jcjo.2018.05.006.

[7] D. PEMBIMBING Ir Joko Lianto Buliali and D. Manajemen Teknologi Bidang Keahlian Manajemen Teknologi Informasi Fakultas Bisnis Dan Manajemen Teknologi, “Tesis-Pm 147501 Prediksi Kinerja Mahasiswa Menggunakan Support Vector Machine Untuk Pengelola Program Studi di Perguruan Tinggi (Studi Kasus: Program Studi Magister Statistika ITS) Fathin Hilmiyah 9115 205 311,” 2017.

[8] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, “A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, May 2020, doi: 10.38094/jastt1224.

[9] K. Dissanayake and M. G. M. Johar, “Comparative study on heart disease prediction using feature selection techniques on classification algorithms,” *Applied Computational Intelligence and Soft Computing*, vol. 2021, 2021, doi: 10.1155/2021/5581806.

[10] R. Resmiati and T. Arifin, “SISTEMASI: Jurnal Sistem Informasi Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination.” [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[11] R. Sari Wardani, “Model Pengambilan Keputusan dalam Prediksi Kasus Tuberkulosis Menggunakan Regresi Logistik Berbasis Backward Elimination,” 2014. Accessed: Sep. 05, 2023. [Online]. Available: https://jurnal.unimus.ac.id/index.php/psn12012010/article/view/1226

[12] M. Tech, “Breast Cancer Remnant Impact During Covid-19 Using to Machine Learning DEVANAND,” *Journal of Tianjin University Science and Technology*, vol. 55, no. 01, pp. 149–159, 2022, doi: 10.17605/OSF.IO/68GZU.

[13] M. Samieinasab, S. A. Torabzadeh, A. Behnam, A. Aghsami, and F. Jolai, “Meta-Health Stack: A new approach for breast cancer prediction,” *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2021.100010.

[14] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[15] S. Roy, P. Sharma, K. Nath, D. K. Bhattacharyya, and J. K. Kalita, “Pre-processing: A data preparation step,” in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, Elsevier, 2018, pp. 463–471. doi: 10.1016/B978-0-12-809633-8.20457-3.

[16] P. Meilina, “Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi,” Jakarta, Jan. 2015.

[17] S. Muthukumaran, P. Geetha, and E. Ramaraj, “A Rule Based Recommender System to Improve the Yield of Groundnut Crop Using Decision Tree with Backward Elimination, Principal Component Analysis,” 2021.

[18] F. Maulidina, Z. Rustam, S. Hartini, V. V. P. Wibowo, I. Wirasati, and W. Sadewo, “Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Mar. 2021. doi: 10.1088/1742-6596/1821/1/012006.

[19] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, “Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification,” *Remote Sens (Basel)*, vol. 11, no. 2, Jan. 2019, doi: 10.3390/rs11020185.

[20] S. Prusty, S. Patnaik, and S. K. Dash, “SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer,” *Frontiers in Nanotechnology*, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.

[21] D. Kartikasari, “Analysis of Factors Affecting Air Pollution Levels Using The Binary Logistic Regression Method,” 2020. doi: https://doi.org/10.26740/mathunesa.v8n1.p55-59.

[22] A. S. Arsya, “Comparison of Oversampling Methods On Balanced Data Using Logistic Regression Algorithm In Stroke Disease Classification,” 2023.

[23] S. Abdollahi, H. R. Pourghasemi, G. A. Ghanbarian, and R. Safaeian, “Prioritization of effective factors in the occurrence of land subsidence and its susceptibility mapping using an SVM model and their different kernel functions,” *Bulletin of Engineering Geology and the Environment*, vol. 78, no. 6, pp. 4017–4034, Sep. 2019, doi: 10.1007/s10064-018-1403-6.

[24] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, W. Zhou, and R. Lacuesta, “Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection,” *Wirel Commun Mob Comput*, vol. 2019, 2019, doi: 10.1155/2019/5176705.

[25] A. P. Lahagu, “Implementasi Data Mining Untuk Memprediksi Pemesanan Barang Ekspor Pada PT. Musim Mas Dengan Menggunakan Metode Support Vector Machine (Study Kasus : PT. Musim Mas),” *Pelita Informatika : Informasi dan Informatika*, vol. 9, no. 1, 2020.

[26] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, “Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking,” *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[27] R. Rajalakshmi and C. Aravindan, “A Naive Bayes approach for URL classification with supervised feature selection and rejection framework,” *Comput Intell*, vol. 34, no. 1, pp. 363–396, Feb. 2018, doi: 10.1111/coin.12158.

[28] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[29] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. in Intelligent Systems Reference Library. Springer Berlin Heidelberg, 2011. [Online]. Available: https://books.google.co.id/books?id=yJvKY-sB6zkC

[30] D. Rodriguez, I. Herraiz, R. Harrison, J. Dolado, and J. C. Riquelme, “Preliminary comparison of techniques for dealing with imbalance in software defect prediction,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2014. doi: 10.1145/2601248.2601294.

[31] S. Lonang and D. Normawati, “Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination,” *Jurnal Media Informatika Budidarma*, vol. 6, no. 1, p. 49, Jan. 2022, doi: 10.30865/mib.v6i1.3312.

[32] P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, “Comparative study of machine learning algorithms for breast cancer prediction,” in *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 796–801. doi: 10.1109/ICSSIT48917.2020.9214267.