 ISSN 2579-7298

Performance Comparison of Support Vector Machine Algorithm and Logistic Regression Algorithm

Hanny Hikmayanti [a,1,*], Anis Fitri Nurmasruriyah [a,2], Ahmad Fauzi [a,3], Nunung Nurjanah [a,4], Arphilia Nur Rani [a,5]

[a] Universitas Buana Perjuangan Karawang, Jl.HS. Ronggo Waluyo, Karawang 41361, Indonesia
[1] hanny.hikmayanti@ubpkarawang.ac.id*; [2] anis.masruriyah@ubpkarawang.ac.id; [3] afauzi@ubpkarawang.ac.id, [4] if20.nunungnurjanah@mhs.ubpkarawang.ac.id, [5] if20.arphiliarani@mhs.ubpkarawang.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 06 May 2023
Revised 13 Jul 2023
Accepted 03 Sept 2023

Keywords
Breast cancer
Regression Logistics
SVM
K-Fold cross validation

ABSTRACT

According to the World Health Organization (WHO), there are around 7 million breast cancer patients each year, with about 5 million of them dying. Based on Globocan 2018 data, the death rate from breast cancer averages 17 per 100,000 people with incidents of 2.1 per 100,000 people attacking women in Indonesia. Hence breast cancer causes spread genetic mutations in the DNA of breast epithelial cells that radiate to the ducts. The purpose of this study was to classify the type of cancer (benign or malignant) that was suffered. The difference between previous research and this research is in the algorithm testing method chosen. In this study the algorithm used is SVM and Logistic Regression by applying the SMOTE technique. The K-fold cross validation method is used in testing this research. The accuracy results obtained are 1.0, precision 1.0 and recall 1.0.While the highest evaluation results for the model without SMOTE were Accuracy 0.97, precision 1.0 and recall 0.90 with the LR method. So based on the results of the comparison, it shows that the evaluation of models using SMOTE tends to be higher than models without SMOTE.

1. Introduction

Breast cancer is the most common type of cancer affecting women, and has the second highest mortality rate after other cancers [1] [2]. Cancer is a big challenge for humans because it can affect various organs of the body. Breast cancer can strike at all ages [3] [4]. According to the World Health Organization (WHO) there are around 7 million breast cancer patients each year, with about 5 million of them dying. Based on Globocan 2018 data, the death rate from breast cancer averages 17 per 100,000 people with incidents of 2.1 per 100,000 people attacking women in Indonesia [5]. Breast cancer can cause the spread of genetic mutations in DNA from breast epithelial cells that spread to the ducts. Breast cancer can be prevented by doing breast self-examination (BSE) [6].

Breast cancer in several studies using data mining has been proven to be predictable and classifiable [7] [8] [9] [10] [11]. Data mining is used to find patterns, correlations, trends in processing large amounts of data [12]. Previously, in Athalla et al's study, the K-Nearest Neighbor method was used to classify breast cancer. In this study using the Minkowski method to calculate the closest distance to the object. This study produces an accuracy of 93% by applying the K Nearest Neighbor method [13]. Breast cancer is generally divided into two types, namely benign and malignant [14]. Benign cancer is generally characterized by a small lump, round and feels soft. While malignant cancer is generally characterized by breast shape that is asymmetrical, rough, and

causes pain [15]. Early detection of breast cancer is very important in reducing the high risk of death. The Indonesian government itself even made a program regarding early detection of breast cancer [16]. Based on the results of previous studies, this research needs to be carried out to classify the type of cancer (benign or malignant) that is suffered. This will help facilitate fast and appropriate treatment of breast cancer patients.

2. Method

In this study, to classify the types of breast cancer (benign or malignant) the methods used are Support Machine Learning and Logistic Regression. The flowchart of this research can be seen in Fig. 1.

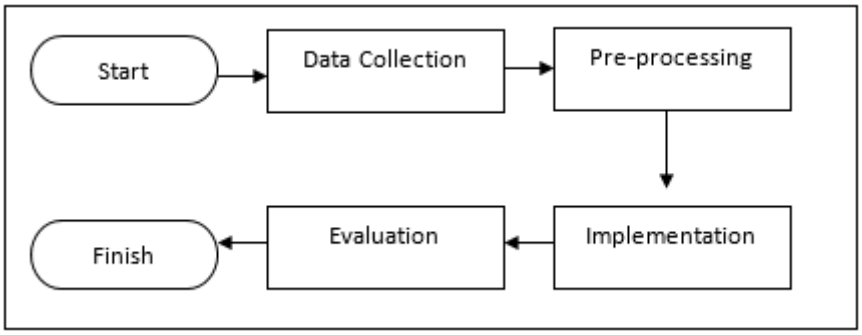

Fig. 1. Research flowchart

2.1. Data Collection

The dataset used in this study comes from Breast cancer predictions | Kagglewhich is available in .csv format, with a total of 569 data and consists of 32 attributes. This dataset contains diagnostic attributes that distinguish between benign and malignant cancers. In addition, other attributes of this dataset are divided into three features, namely Mean (average), Standard deviation (se), and worst (worst). Each column in the feature has various sizes consisting of radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensions.

2.2. Pre-Processing

As forsome of the techniques applied at this stage are described as follows;

a. Changing Data Type

The data type change phase is carried out by changing the 'object' data type to 'category' to make it easier to read the data.

b. Feature Selection

At this stage, the feature-based correlation technique is used to perform feature selection. This application of feature selection helps to find subsets of original features by taking different approaches based on information [17] [8]. Function of feature selectionused to see the relationship between one attribute and other attributes in the dataset.

c. Cleaning data

The data cleaning carried out in this study was used to overcome data from noisy, outliers, missing values and duplication of data, so that the data is ready for use in the next stage [18] [19].

d. Data Normalization

This data normalization stage is used to eliminate scale differences between attributes. The method used for data normalization is the Min-max scaling method [20].

e. Data Transformation

Data transformation in this study was carried out to change categorical data into numerical data, so that the data would be easy to process for modeling data. The technique used in this research is label encoding [21].

2.3. Implementation

a. Smote Technique

Synthetic Minority Oversampling Technique(SMOTE) is a preprocessing algorithm that is generally used to overcome data imbalance problems [22].

b. SVM

Support Vector Machine (SVM) is a method used to solve classification and regression problems. However, SVM is better known and more widely used in a classification context [23] [24]. SVM serves to find the best hyperplane (separator) that can separate the two classes and maximize the distance between the two classes [25] [26]. SVM is divided into two categories, namely linear SVM and non-linear SVM. Linear SVM is used when data can be separated linearly using hyperplanes with soft margins. Meanwhile, non-linear SVM involves using kernel functions to transform a feature space into a higher dimensional space [27] [28].

SVM	Jenis Kernel	Definisi Rumus
Linier	Linier	K (x,y) = x.y
Non linier	*Polynomial*	K (x,y) = (x.y + 1)^p
	RBF	K (x,y) = $e^{-\|x.y\|^2/2\sigma^2}$
	Sigmoid	(x,y) = tanh (Kx.y- δ)

Fig. 2. SVM Linear and Non Linear

c. Logistics Regression

Logistic Regression Algortihm is used to perform classification by predicting the possibility of an event that will occur or not. This classification process can occur if the dependent variable is a dichotomous variable. Where dichotomous variables are represented in numbers 1 and 0 [29]. Logistic Regression will identify the relationship between the dependent variable (Y) as an event that will occur or not, with the independent variable, namely categorical or continuous [30].

2.4. Evaluation

a. Confusion matrix

Confusion matrixis a table used to display the number of correct test data and wrong test data [31]. The confusion matrix is used to evaluate the performance of a model based on precision and accuracy [23].

	Aktual	
Prediksi	**A**	**B**
A	TP (*True Positive*)	FP (*False Positive*)
B	FN (*False Negative*)	TN (*True Negative*)

Fig. 3. Confusion Matrix

b. Confusion matrix

K-Fold Cross Validationis a method used to evaluate the performance of the algorithm by dividing the sample data into several groups randomly, namely K fold groups. Each fold group will be used as test data alternately. While other fold groups are used as training data [32].

3. Results and Discussion

3.1. Pre-Processing

In the preprocessing stage, changes are made to the data types to improve the correspondence with the data in the data set. Changes to data types are shown in the following figure :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      569 non-null    int64
 1   diagnosis               569 non-null    object
 2   radius_mean             569 non-null    float64
 3   texture_mean            569 non-null    float64
```

Fig. 4. Before Change Data Type

```
[ ]
df['diagnosis'] = df['diagnosis'].astype('category')
df.dtypes

id                          int64
diagnosis                category
radius_mean               float64
texture_mean              float64
```

Fig. 5. After Change Data Type

Then feature selection is carried out to determine the attributes that have a large influence on the data [33]. Also, the data is cleaned so there is no noise and no outliers. Cleaned data can be shown in the following figure:

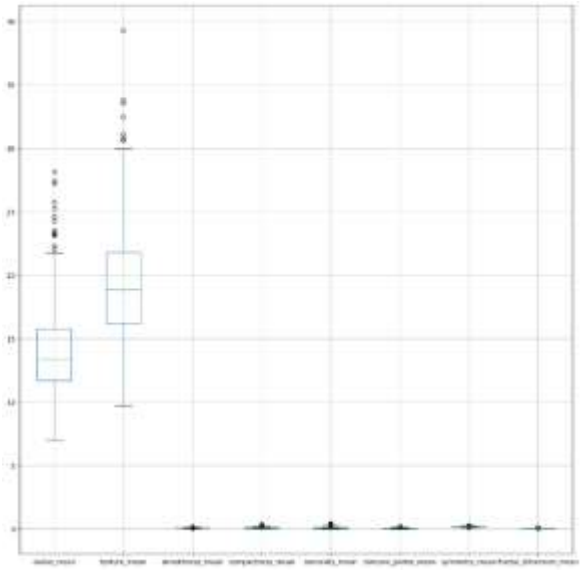

Fig. 6. Before Cleaning Data

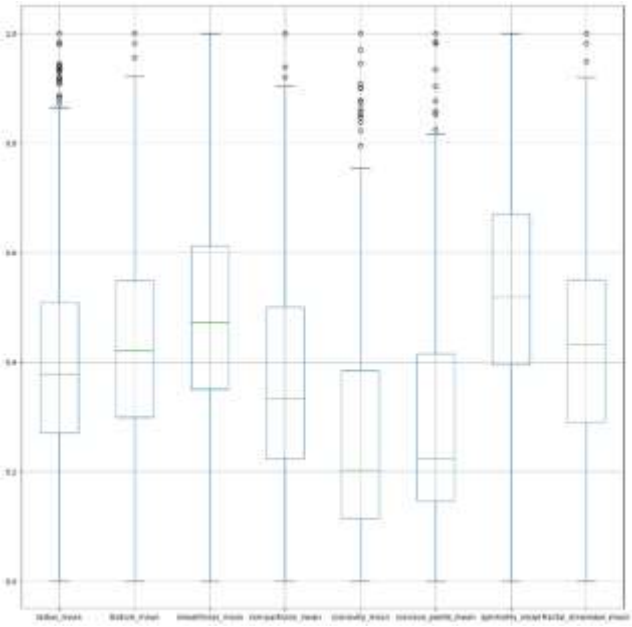

Fig. 7. After Cleaning Data

After the data is cleaned, the numeric data is normalized to make it easier to scale the data. Meanwhile, categorical data is converted into data to facilitate inspection [23]. The results of data normalization can be seen in the following figure:

	diagnosis	radius_mean	texture_mean	smoothness_mean	compactness_mean
1	M	20.57	17.77	0.08474	0.07864
2	M	19.69	21.25	0.10960	0.15990
4	M	20.29	14.34	0.10030	0.13280
6	M	18.25	19.98	0.09463	0.10900
7	M	13.71	20.83	0.11890	0.16450

Fig. 8. Before Data Normalization

diagnosis	radius_mean	texture_mean	smoothness_mean	compactness_mean
M	0.939274	0.400995	0.351796	0.331766
M	0.872476	0.574129	0.745213	0.786698
M	0.918020	0.230348	0.598038	0.634979
M	0.763170	0.510945	0.508308	0.501736
M	0.418552	0.553234	0.892388	0.812451

Fig. 9. After Data Normalization

Because breast cancer data from Kaggle has an imbalance, it is necessary to use the oversampling method to achieve data balance. Thus, breast cancer data from Kaggle has an imbalance, so it is necessary to use the oversampling method to achieve data balance. The technique used is SMOTE. The results of applying the oversampling method are shown in the figure below:

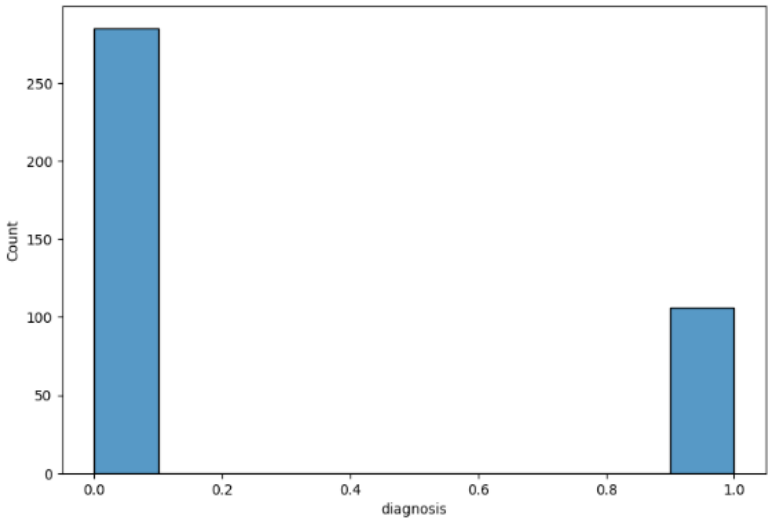

Fig. 10. Before SMOTE

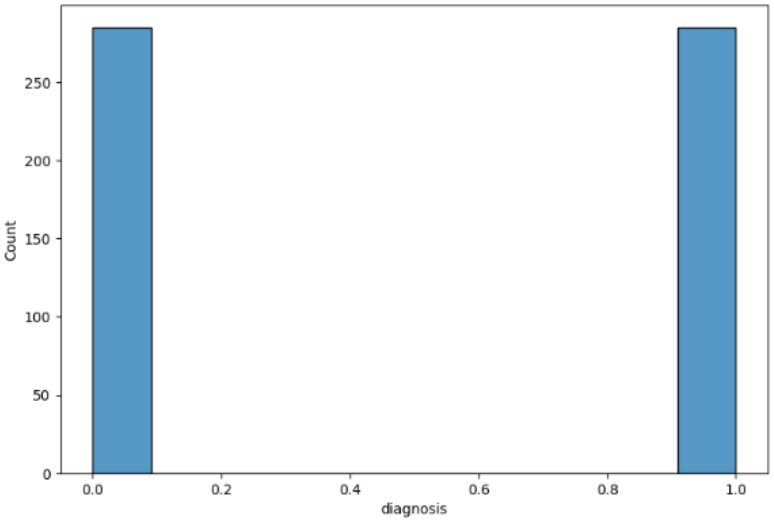

Fig. 11. After SMOTE

3.2. Modelling

In this modeling phase, data classification is carried out. Moreover, the pre-processed data is initially partitioned with the help of K-fold cross validation. After partitioning, modeling is performed using the SVM algorithm with a polynomial kernel and logistic regression using the SMOTE technique. SMOTE is an oversampling procedure. Furthermore, the performance of the SVM algorithm is then evaluated using SMOTE or oversampling techniques.

3.3. Evaluation and Comparison

K-fold cross-validation is used to reduce the distortion that can occur with random data [23]. In this study, a 10-fold cross-validation method was used, in which the dataset was divided into 10 different datasets of the same size. Each fold is used as a test set, while the other folds serve as a training set. This procedure was repeated 10 times to comprehensively test the model [34]. Then the classification is done using the SVM algorithm and logistic regression using the SMOTE technique and without SMOTE. The classification results in terms of accuracy, precision and recognition value are shown in Table 1 and 2 :

Table 1. Model Evaluation Results with SMOTE

K-Fold	SVM			LR		
	Acuracy	**Precision**	**recall**	**Acuracy**	**Precision**	**recall**
1	0.96	0.96	0.96	0.92	0.96	0.90
2	0.98	1.0	0.96	0.96	1.0	0.92
3	1.0	1.0	1.0	0.94	0.92	0.92
4	1.0	1.0	1.0	0.98	1.0	0.97
5	0.98	0.95	1.0	0.92	0.90	0.90
6.	0.91	0.88	0.91	0.91	0.85	0.95
7.	0.92	0.92	0.92	0.96	0.96	0.96
8.	0.94	0.93	0.96	0.98	0.96	1.0
9.	0.98	1.0	0.96	0.96	1.0	0.93
10.	1.0	1.0	1.0	0.91	1.0	0.84
Average	**0.97**	**0.96**	**0.97**	**0.94**	**0.96**	**0.93**

Using Table 1, it was found that the highest accuracy, precision, and recognition were found in K-Fold 10 using the SVM method. The evaluation results achieved were Accuracy 1.0, Precision 1.0 and Recall 1.0. In addition, the classification is carried out using the SVM and Logistic Regression methods, without using the SMOTE technique. The results of the evaluation of the model without SMOTE are presented in Table 2 :

Table 2. Model Evaluation Results Without SMOTE

K-Fold	SVM			LR		
	Acuracy	**Precision**	**recall**	**Acuracy**	**Precision**	**recall**
1	0.97	1.0	0.91	0.95	1.0	0.83
2	0.94	0.90	0.90	0.92	0.9	0.81
3	0.92	0.81	0.9	0.94	1.0	0.8
4	1.0	1.0	1.0	0.97	1.0	0.83
5	0.92	0.66	0.8	1.0	1.0	1.0
6.	0.94	0.8	1.0	0.94	1.0	0.75
7.	0.94	0.90	0.90	0.97	1.0	0.90
8.	0.92	0.93	0.87	0.89	1.0	0.75
9.	0.97	1.0	0.93	0.94	1.0	0.87
10.	0.94	0.90	0.90	0.97	1.0	0.90
Average	**0.95**	**0.89**	**0.91**	**0.95**	**0.99**	**0.84**

Based on Table 2, the highest accuracy, precision and recognition can be found in K-Fold 9. The accuracy is 0.97, the precision is 1.0 and the recognition is 0.93. Sourced from previous research [23] [35] mentioned that K-Fold10 is the best choice to get an accurate estimate. In addition, a comparison of the results of the model evaluation is carried out to determine the performance of the model that will be implemented into the system. The results of the model comparison are shown in Table 3:

Table 3. Comparison of Models

Method	Acuracy	Precision	recall
SVM	0.94	0.90	0.90
LR	**0.97**	**1.0**	**0.90**
SVM + SMOTE	**1.0**	**1.0**	**1.0**
LR + SMOTE	0.91	1.0	0.84

The results of the comparison of model evaluation with SMOTE in this study were the highest accuracy 1.0, precision 1.0 and recall 1.0 with the SVM method. While the highest evaluation results for the model without SMOTE were Accuracy 0.97, precision 1.0 and recall 0.90 with the LR method. So the results of comparison evaluation of models using SMOTE tend to be higher than models without SMOTE. As seen from the results of the comparison of models, it shows that there is a significant increase in SVM with SMOTE compared to SVM without SMOTE.

4. Conclusion

This study uses breast cancer data from Kaggle. The various preprocessing stages are data type modification, characteristic selection, data cleaning, data normalization and data transformation. The SMOTE technique is then applied to process unbalanced data. The model is then implemented using the SVM algorithm and logistic regression with K-fold cross validation. The results of the comparative evaluation of the model that has the best value in this study is the SVM method using SMOTE. The result is 1.0 for accuracy, 1.0 for precision and 1.0 for recall. While the highest evaluation results for the model without SMOTE were Accuracy 0.97, precision 1.0 and recall 0.90 with the LR method. So based on the results of the comparison, it shows that the evaluation of models using SMOTE tends to be higher than models without SMOTE.

References

[1] J. W. Zhu, P. Charkhchi, S. Adekunte, and M. R. Akbari, "What Is Known about Breast Cancer in Young Women?," *Cancers (Basel).*, vol. 15, no. 6, p. 1917, Mar. 2023, doi: 10.3390/cancers15061917.

[2] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, p. 736, Feb. 2023, doi: 10.11591/ijece.v13i1.pp736-745.

[3] S. Rana, D. Kaushik, A. Singh, D. Gautam, J. Rai, and J. S. Rathore, "Aptamer: A theranostic approach towards breast cancer," *Clin. Immunol. Commun.*, vol. 3, pp. 61–73, Dec. 2023, doi: 10.1016/j.clicom.2023.06.002.

[4] M. M. Rahman, Z. Ferdousi, P. Saha, and R. A. Mayuri, "A Machine Learning Approach to Predict Breast Cancer Using Boosting Classifiers," *Indian J. Comput. Sci. Eng.*, vol. 14, no. 3, pp. 409–415, Jun. 2023, doi: 10.21817/indjcse/2023/v14i3/231403009.

[5] M. G. Fazliddinovna, "Oncopsychology of Patients with Breast Cancer after Treatment," *J. Nat. Med. Educ.*, vol. 2, no. 2, pp. 111–116, 2023.

[6] B. Sain *et al.*, "Clinico-Pathological Factors Determining Recurrence of Phyllodes Tumors of the Breast: The 25-Year Experience at a Tertiary Cancer Centre," *J. Pers. Med.*, vol. 13, no. 5, p. 866, May 2023, doi: 10.3390/jpm13050866.

[7] L. Yang, S. Peng, R. O. Yahya, and L. Qian, "Cancer detection in breast cells using a hybrid method based on deep complex neural network and data mining," *J. Cancer Res. Clin. Oncol.*, Jul. 2023, doi: 10.1007/s00432-023-05191-2.

[8] P. Dikshit, B. Dey, A. Shukla, A. Singh, T. Chadha, and V. K. Sehgal, "Prediction of Breast Cancer

using Machine Learning Techniques," *ACM Int. Conf. Proceeding Ser.*, no. April 2019, pp. 382–387, 2022, doi: 10.1145/3549206.3549274.

[9] Dr. Nikhat Akhtar, Dr. Hemlata Pant, Apoorva Dwivedi, Vivek Jain, and Dr. Yusuf Perwej, "A Breast Cancer Diagnosis Framework Based on Machine Learning," *Int. J. Sci. Res. Sci. Eng. Technol.*, pp. 118–132, May 2023, doi: 10.32628/IJSRSET2310375.

[10] S. R. Mary, R. M. Prasad, and R. Suguna, "A feature selection using improved dragonfly algorithm with support vector machine for breast cancer prediction," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, pp. 1–11, May 2023, doi: 10.1080/21681163.2023.2212086.

[11] A. A. Sukmandhani, Lukas, Y. Heryadi, W. Suparta, and A. Wibowo, "Classification Algorithm Analysis for Breast Cancer," *E3S Web Conf.*, vol. 388, p. 02012, May 2023, doi: 10.1051/e3sconf/202338802012.

[12] M. Kirola, M. Memoria, M. Shuaib, K. Joshi, S. Alam, and F. Alshanketi, "A Referenced Framework on New Challenges and Cutting-Edge Research Trends for Big-Data Processing Using Machine Learning Approaches," in *2023 International Conference on Smart Computing and Application (ICSCA)*, Feb. 2023, pp. 1–5. doi: 10.1109/ICSCA57840.2023.10087686.

[13] K. Khadijah and R. Kusumaningrum, "Ensemble Classifier untuk Klasifikasi Kanker Payudara," *It J. Res. Dev.*, vol. 4, no. 1, pp. 61–71, 2019, doi: 10.25299/itjrd.2019.vol4(1).3540.

[14] H. Kaur, "Dense Convolutional Neural Network Based Deep Learning Framework for the Diagnosis of Breast Cancer," *Wirel. Pers. Commun.*, Jul. 2023, doi: 10.1007/s11277-023-10678-9.

[15] J. Y. You, S. Park, E.-G. Lee, and E. S. Lee, "Detection and Diagnosis of Breast Cancer," in *A Practical Guide to Breast Cancer Treatment*, Singapore: Springer Nature Singapore, 2023, pp. 1–17. doi: 10.1007/978-981-19-9044-1_1.

[16] F. T. Liza, M. C. Das, P. P. Pandit, A. Farjana, A. M. Islam, and F. Tabassum, "Machine Learning-Based Relative Performance Analysis for Breast Cancer Prediction," in *2023 IEEE World AI IoT Congress (AIIoT)*, Jun. 2023, pp. 0007–0012. doi: 10.1109/AIIoT58121.2023.10174469.

[17] D. P. M. Abellana and D. M. Lao, "A new univariate feature selection algorithm based on the best–worst multi-attribute decision-making method," *Decis. Anal. J.*, vol. 7, p. 100240, Jun. 2023, doi: 10.1016/j.dajour.2023.100240.

[18] X. Li, M. Liu, K. Wang, Z. Liu, and G. Li, "Data cleaning method for the process of acid production with flue gas based on improved random forest," *Chinese J. Chem. Eng.*, vol. 59, pp. 72–84, Jul. 2023, doi: 10.1016/j.cjche.2022.12.013.

[19] A. Lia Hananto *et al.*, "Analysis of Drug Data Mining with Clustering Technique Using K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1908, no. 1, 2021, doi: 10.1088/1742-6596/1908/1/012024.

[20] A. L. Hananto, A. P. Nardilasari, A. Fauzi, A. Hananto, B. Priyatna, and A. Y. Rahman, "Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter," *Orig. Res. Pap. Int. J. Intell. Syst. Appl. Eng. IJISAE*, vol. 2023, no. 6s, pp. 473–481, 2023, [Online]. Available: www.ijisae.org

[21] F. Aryanto, A. Fauzi, A. Fitri Nur Masruriyah, A. Lia Hananto, and Darmansyah, "Sentiment Analysis Of Vaccination Using The K-Nearest Neighbor Algorithm," *Edutran Comput. Sci. Inf. Technol.*, vol. 1, no. 1, pp. 34–41, 2023, doi: 10.59805/ecsit.v1i1.6.

[22] N. Anđelić and S. Baressi Šegota, "Development of Symbolic Expressions Ensemble for Breast Cancer Type Classification Using Genetic Programming Symbolic Classifier and Decision Tree Classifier," *Cancers (Basel).*, vol. 15, no. 13, p. 3411, Jun. 2023, doi: 10.3390/cancers15133411.

[23] C. B. Sonjaya, A. Fitri, N. Masruriyah, and D. Sulistya, "The Performance Comparison of Classification Algorithm in Order to Detecting Heart Disease," vol. 5, no. 2, pp. 166–175, 2022.

[24] I. Candradewi, A. Harjoko, and B. A. A. Sumbodo, "Intelligent Traffic Monitoring Systems: Vehicle Type Classification Using Support Vector Machine," *Int. J. Artif. Intell. Res.*, vol. 5, no. 1, pp. 78–90, 2021, doi: 10.29099/ijair.v5i1.201.

[25] M. Wati, R. Alfred, A. Ery, and A. Ardi, "An extraction of shapes and support vector machine

methods for identification of decorative wall ‘ Lamin ’ motifs of the Dayak Kenyah Pampang tribe,” vol. 7, no. 1, 2023.

[26] I. Kurniawan *et al.*, “Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 10, no. 1, pp. 731–740, 2023, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/3582

[27] F. G. Altin, İ. Budak, and F. Özcan, “Predicting the amount of medical waste using kernel-based SVM and deep learning methods for a private hospital in Turkey,” *Sustain. Chem. Pharm.*, vol. 33, p. 101060, Jun. 2023, doi: 10.1016/j.scp.2023.101060.

[28] A. P. Nardilasari, A. L. Hananto, S. S. Hilabi, and B. Priyatna, “Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM,” vol. 7, no. 1, pp. 11–18, 2024.

[29] T. Meenakshi, “Automatic Detection of Diseases in Leaves of Medicinal Plants Using Modified Logistic Regression Algorithm,” *Wirel. Pers. Commun.*, vol. 131, no. 4, pp. 2573–2597, Aug. 2023, doi: 10.1007/s11277-023-10555-5.

[30] Y. Song *et al.*, “Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: A retrospective study,” *CNS Neurosci. Ther.*, vol. 29, no. 1, pp. 158–167, Jan. 2023, doi: 10.1111/cns.13991.

[31] J. B. B. Bell, A. Rajkumar, S. M. C. Vigila, M. G. A. Selvan, and J. S. Binoj, “Development of novel methodology for gene identification-based classification of leukaemia disorder,” *Res. Biomed. Eng.*, Jun. 2023, doi: 10.1007/s42600-023-00289-5.

[32] X. Tang *et al.*, “Explainable multi-task learning for multi-modality biological data analysis,” *Nat. Commun.*, vol. 14, no. 1, 2023, doi: 10.1038/s41467-023-37477-x.

[33] K. Makimoto *et al.*, “Comparison of Feature Selection Methods and Machine Learning Classifiers for Predicting Chronic Obstructive Pulmonary Disease Using Texture-Based CT Lung Radiomic Features,” *Acad. Radiol.*, vol. 30, no. 5, pp. 900–910, May 2023, doi: 10.1016/j.acra.2022.07.016.

[34] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, “Threshold optimization and random undersampling for imbalanced credit card data,” *J. Big Data*, vol. 10, no. 1, p. 58, May 2023, doi: 10.1186/s40537-023-00738-z.

[35] Y. Widyaningsih, G. P. Arum, and K. Prawira, “Aplikasi K-Fold Cross Validation Dalam Penentuan Model Regresi Binomial Negatif Terbaik,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 15, no. 2, pp. 315–322, 2021, doi: 10.30598/barekengvol15iss2pp315-322.