International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 2
ISSN: 2088-5334

Grid Search CV Implementation in Random Forest Algorithm to Improve Accuracy of Breast Cancer Data

Dimas Aryo Anggoro [a,*], Nur Aini Afdallah [b]

[a] *Informatics Department, Universitas Muhammadiyah Surakarta, Jl. Ahmad Yani, Surakarta, 57162, Indonesia*
Corresponding author: [*]dimas.a.anggoro@ums.ac.id

***Abstract*—Breast cancer is the most common cancer in women and is the second leading cause of global death. Disease diagnosis plays an important role in determining treatment strategies related to patient safety. Therefore, we need machine learning to predict disease. This paper aims to determine the best parameter values in breast cancer data using the Grid Search CV method and classify breast cancer data using the random forest algorithm. In addition, the paper aims to compare the accuracy values generated using the Grid Search CV and without the Grid Search CV. The method used to analyze breast cancer data in researchers is the Random Forest (RF) classification algorithm. In addition to using the Random Forest algorithm, this study also uses the Grid Search CV method. Grid Search CV is a method used to determine the optimal model parameters so that the classifier can predict the test data reliably. This study indicates that the highest accuracy value is obtained in the random forest algorithm using the grid search method of 0.9545. In contrast, the accuracy of the random forest algorithm without using the grid search method is 0.9480. For further research, it is suggested to develop a breast cancer dataset using the grid search cv method with other algorithms, such as Logistic Regression, Xgboost, and SVM. We can also use the same algorithm with different datasets to prove that the grid search cv method can increase accuracy.**

***Keywords*— Accuracy; breast cancer; grid search cv; random forest.**

Manuscript received 10 Jun. 2021; revised 28 Sep. 2021; accepted 30 Jan. 2022. Date of publication 30 Apr. 2022.

I. INTRODUCTION

Breast cancer is often found in women and is the second leading cause of death. In previous research, this issue is reviewed [1]. Breast cancer is a non-contagious disease. The cause of breast cancer is still unknown, yet this disease is multifactorial and interplay. A study discusses women's lifestyle risk factors that affect breast cancer at Makassar City Hospital. The results discovered that fat consumption, obesity, smoking, and stress are risk factors that influence breast cancer. Stress is the most influential risk factor that stressed individuals would be prone to cancer 2.698 times higher than those who are relaxed [2].

Breast cancer is the world's second-biggest cause of mortality, and it is forecast that it will be the first cause of mortality in 2060 (~18.63 million deaths) [3]. In 2018 as many as 18 million instances were detected, with 2.09 million cases of breast cancer obtained from information from the World Health Organization (WHO) and the American Cancer Society. A survey on the incidence of breast cancer was conducted in 187 countries in 2011 found a yearly average rise of 3.1% to 1.643.000 data in 2010 from 641.000 data in 1980 [4].

Breast cancer is a non-communicable illness that primarily affects women, has a high fatality rate, and rises every year [5]. Medical diagnosis is critical in ensuring a suitable treatment strategy for the patient's safety. Machine learning is needed to predict a disease to handle the data more efficiently. The main advantage of machine learning is that an algorithm can learn data and automatically perform its tasks [6]. This study's analysis of breast cancer data used the Random Forest (RF) classification method.

The selection of classification algorithms is used to predict the decision value of class variables for qualitative or categorical variable types with calculations with one or more independent variables or predictors. It is widely applied in computer science, medicine, botany, and psychology. Random Forest (RF) was chosen in processing breast cancer datasets because it has many advantages. The first advantage is that the algorithm can effectively handle large databases, generate input variables without deleting variables, create unbiased internal estimates of common error errors, estimate each variable for classification, and demonstrate robust and

accurate performance on complex data sets [7]. Earlier research has identified retinal anomalies using Random Forest and Naive Bayes algorithm. According to this study, the Random Forest method has a better accuracy value than Naive Bayes – 0.9358 for Random Forest and 0.8363 for Naive Bayes [8].

There has been previous research on gender classification based on sound frequency using several algorithms, one of which is the random forest. In addition to using the random forest, this study also performed the grid search cv technique. The grid search cv function evaluates and optimizes the established model [9]. The grid search cv can generate new training models automatically using cross-validation to get the best parameters [10]. Moreover, the results show that the accuracy value following grid search cv 0.9691 was higher than the accuracy value without using grid search cv 0.9675 [11].

This research aims to determine the ideal parameter values for breast cancer data using the grid search cv method and classify breast cancer data through a random forest algorithm. In addition, this study aims to compare accuracy values yielded with grid search cv and without grid search cv. The output of this study implies that a random forest algorithm with grid search cv is expected to show a better accuracy value.

II. Materials and Method

A. Data Collection

Data collection is the initial stage in a study. Some variables and attributes facilitate the research process in the data mining process. Data obtained from breast cancer comprise 569 data and 32 attributes. It includes ID number, diagnosis, and ten attributes which are further divided into mean, standard error (SE), and worst or largest. For instance, the attribute radius becomes mean radius, radius SE, and worst radius. The following attributes and variables are described in Table 1.

TABLE I
VARIABLES AND DATA ATTRIBUTES OF BREAST CANCER

Variable	Attributes
X1	ID number, the ID number of data
X2	Radius_mean, the average distance between the centre and the perimeter of points
X3	Texture_mean, greyscale value standard deviation
X4	Perimeter_mean, cancer's average size
X5	Area_mean, the typical area
X6	Smoothness_mean, the mean of the local variations in radius length
X7	Compactness_mean, the mean (around^2 / large-1)
X8	Concavity_mean, the harshness of the contours concave curves
X9	Concave points_mean, the amount of contour concave sections
X10	Symmetry_mean, symmetry on average
X11	Fractal dimension_mean, approximation of the arithmetic mean coastline-1
X12	Radius_SE, the standard deviation for average distance from centre to perimeter points
X13	Texture_SE, greyscale standard error for standard deviation value
X14	Perimeter_SE, cancer size standard deviation
X15	Area_SE, error standard area
X16	Smoothness_SE, the standard deviation in radius length for local variation
X17	Compactness_SE, the standard deviation for around^2 / large-1
X18	Concave point_SE, the standard deviation in contour concave portions
X19	Symmetry_SE, symmetric standard deviation
X20	Fractal dimension_SE, approximation of the arithmetic standard deviation coastline-1
X21	Radius_worst, the average distance from the centre to the perimeter points with the lowest or highest average value
X22	Texture_worst, the lowest or highest average value for greyscale standard deviation
X23	Perimeter_worst, the lowest or highest average value for cancer size
X24	Area_worst, the lowest or highest average value for a region
X25	Smoothness_worst, the lowest or highest average value for local radius length variation
X26	Compactness_worst, the lowest or the highest average value for around^ 2 / large-1
X27	Concavity_worst, the lowest or highest average value for the concave contour sections severity
X28	Concave point_worst, for concave regions of the contour, the lowest or highest average value is used
X29	Symmetry_worst, the lowest or highest symmetry average value
X30	Fractal dimension_worst, approximation of the arithmetic lowest or highest coastline-1
Y	Diagnosis, M = Malignant, B = Benign

In this study, we are using the Wisconsin Breast Cancer Database. It was collected by Wolberg, Nick Street, and Mangasarian at the University of Wisconsin-Madi-son Hospitals [12].

B. Data Preprocessing

1) Data Oversampling: After obtaining the dataset, the next step is data oversampling. This study utilized Adaptive Synthetic (ADASYN). He *et al.* [13] proposed a new adaptive synthetic sampling approach from an imbalanced dataset. It can adaptively synthesize minority class illustrations based on their difficulty level, resulting in more information for the more difficult minority classes to understand [14]. Adaptive Synthetic (ADASYN) steps calculate the imbalance class.

$$d = m_s \div m_l \tag{1}$$

Where $d \in (0,1)$. Count how many synthetic samples there are in the minority class.

$$G = m_l - m_s \cdot \beta \tag{2}$$

Where $\beta \in (0,1)$ and G = the distinction between minority and majority classes. Count the k-neighbors for each minority class sample using the Euclidean distance, where 1 is the number of samples from the majority class among the k-neighbors. The ratio r may then be determined as follows.

$$r = \Delta \div k \tag{3}$$

In step 3, generate r_i from each minority class sample and utilize the nearby majority class sample to show the situation for each minority class sample.

$$r_i' = r_i / \sum_{i=1}^{m_s} r_i \tag{4}$$

Count how many synthetic samples there are for each minority group sample.

$$g_i = r_i' \cdot G \tag{5}$$

Choose one sample of a minority group based on the samples k-neighbors of the synthesized minority class. Synthesize the data using the formula below where s_i is synthetic data, x_{zi} is a randomly selected k-neighbors minority class sample x_i.

$$s_i = x_i + (x_{zi} - x_i) \cdot \alpha \quad (6)$$

2) Principal Component Analysis: PCA summarized high-dimensional data but maintained trends and patterns. The PCA technique may efficiently remove feature correlation and achieve feature matrix dimension reduction [15]. PCA extracts its features through eigenvector and eigenvalue [16]. Steps in dimensional reduction using PCA are to enter X for PCA, where X is training data composed of n-vectors with data dimensions m. Calculate the average of each dimension (X') in equation 7.

$$X' = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (7)$$

Where n = amount of data samples and X_i = observation data. Calculate the covariance matrix (C_x) using equation 8.

$$C_x = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X')(X_i - X')^T \quad (8)$$

Where X' = average data. Calculate the eigenvector (v_m) and eigenvalue (λ_m) in equation 9.

$$C_x v_m = \lambda_m v_m \quad (9)$$

The eigenvalues are then sorted in descending order. Principal Component (PC) is a collection of eigenvectors corresponding to previously sorted eigenvalues. PC dimension is reduced based on the eigenvalue. A way to reduce the PC dimension according to eigenvalue is using the accumulated variance value in the eigenvector [16].

3) Data Splitting: Data splitting is a study design widely used in high-dimensional datasets, and it is possible to divide the originally available datasets into training and testing data. The training dataset is a subset of the original dataset used to estimate and study the required machine learning algorithm parameters. The testing data is a subset of the original dataset used to evaluate the model's performance. The dataset is divided into 70% training data and 30% testing data [17].

C. Data Processing

1) Random Forest: Breiman defines random forest as a classifier built of a series of structured tree classifiers $\{h(x, \theta k), k = 1, \ldots\}$ where $\{\theta k\}$ is an independently and equally distributed random vector, with each tree forming a unit and picking the most popular class from the x input [18].

Given a set of classifiers $h_1(x), h_2(x), \ldots, h_k(x)$ and with a training set drawn at random from the random vector distribution Y, X, determine the margin function as:

$$mg(X,Y) = av_k I(h_x(X) = Y) - max_{j \neq Y} av_k I(h_k(X) = j) \quad (10)$$

In this case, I is the role of indication. The margin is the difference between the average number of votes in X, Y for the correct class surpasses the average number of votes in the other classes. Here are some instances of generalization errors:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \quad (11)$$

The subscript X, Y indicates that the probability exceeds space X, Y. In a random forest: $h_k(X) = h(X, \theta k)$.

In this algorithm, several decision trees are constructed as they operate together. The decision tree acts as a pillar in this algorithm. Random forest is a decision tree group whose nodes are determined in the preprocessing step. After generating numerous trees, the best features are chosen from a random subset of features. Random forest algorithm has several features. The former can handle several input variables without deleting the variable, showing important variables in classification. Large databases also run efficiently. Also, the resulting trees or forests can be saved for future use.

The Random Forest (RF) algorithm consists of several steps. Step 1 is to choose point K from the random data based on the training data. Step 2 is to create a decision tree using the K data point. Step 3 enters the testing data through the rules that have been created using a tree to predict the classification output of the data. Count the votes of each predicted target. Moreover, step 4 considers prediction targets by selecting the most predicted target class, which is the random forest algorithm final prediction result is utilized.

D. Evaluation Model

1) Grid Search Cross-Validation: Grid Search CV is a technique for determining which model parameter is the best can accurately predict data. There are two reasons the grid search method is performed. Firstly, it is unsafe to use a method that avoids searching for complete parameters with a heuristic approach. Secondly, because there are just two parameters, the calculation time necessary to locate appropriate parameters using grid search is no longer an advanced technique [11]. Optimization of grid search conducts cross-validation as a performance metric. The aim is to find a decent set of hyperparameters that allow classifiers to predict unknown data reliably [19]. Grid search cv optimizes parameters C, γ, degrees, etc., to select C and γ using k-fold cross-validation. Begin by dividing the data into k-subset. One subset is used as training data, while the remaining k-1 testing subsets are used to assess it. Various hyperparameter combinations of the best accuracy values are selected. There is just one crucial parameter in the linear kernel that has to be optimized, and that is C. There are two parameters in the RBF and sigmoid kernels; C and γ, whereas the polynomial kernel has three parameters; C, γ, and degree [19].

2) Confusion Matrix: The confusion matrix is a mechanism for assessing categorization performance. Accuracy is the ratio between the correct data and the entire data [20].

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (12)$$

Precision is calculated by dividing the true positive by the true positive plus the false positive.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

Recall is calculated of correct guesses divided by the total number of instances [21].

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

F1_score is the average accuracy and recall value. Where f1 score of 1 represents the highest score and f1 score of 0 represents the poorest score.

$$\text{f1}_{\text{score}} = \frac{\text{TP}}{\text{TP} + 0.5(\text{FP} + \text{FN})} \tag{15}$$

III. Results and Discussion

A. Data Preprocessing

The dataset used is breast cancer data with a total of 569 data with 31 variables and one target. The initial stage carried out in this research is the process of existing data collection and processing.

In this study, data preprocessing is divided into three steps. The first is the oversampling process of ADASYN by balancing the breast cancer dataset. The third is the data splitting process, in which the breast cancer data will be categorized into training data and testing data. Both PCA simplifies the complexity of high-dimensional data. Breast cancer data research utilized Python programming language in the Anaconda Navigator application.

1) Data Oversampling Results: After getting the data, the next step is data oversampling. Breast cancer diagnosis data are divided into two; benign and malignant. Benign data are denoted by 0, and malignant data is denoted by 1. Data for benign breast cancer is 357 data, and data for malignant breast cancer is 212 data, presented in Figure 1. There is a significant difference between malignant and benign data (145 data). It requires an oversampling process to increase the sample size, and you may balance the dataset. The results after the oversampling process using ADASYN can be seen in Figure 2, where 0 is 357 and 1 is 358. ADASYN technique in balancing data utilizes the combination of minority class samples according to their difficulty level, producing more data for minority classes that are difficult to learn.

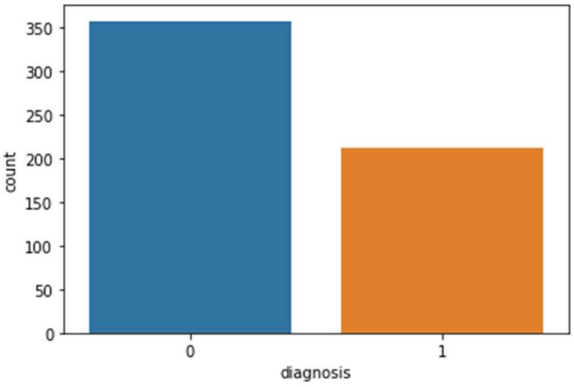

Fig. 1 Data before oversampling

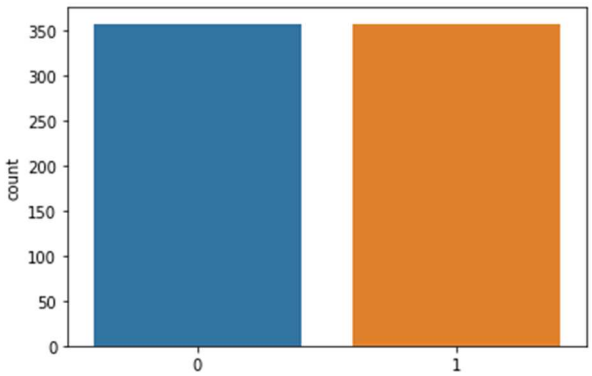

Fig. 2 Data after oversampling

2) Principal Component Analysis Results: The dataset was administered with dimensional reduction using PCA after the oversampling process. The dimensional reduction process in PCA is based on the eigenvalue and eigenvector obtained from the covariance matrix. The number of eigenvectors is computed when the threshold is compared to the cumulative proportion of variation (PPV). Therefore, the threshold plays an important role in determining PPV [22]. Two target classes, dark is benign data and light is malignant data, as shown in Figure 3. The breast cancer data contains 30 dimensions, and then it is subtracted by creating seven major components to observe if the variables can be separated into clusters.

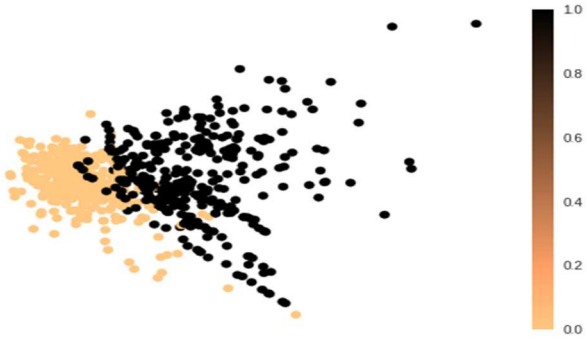

Fig. 3 Principal Component Analysis

3) Data Splitting Results: The next step is data splitting after the data oversampling and PCA process. Breast cancer data were separated into 70 percent training data and 30 percent testing data. Due to the oversampling process, the breast cancer data from 569 data became 715 data and resulted in 500 training data and 215 testing data.

B. Data Processing

This study implemented the random forest algorithm on the breast cancer dataset. The classification accuracy of the results from the random forest algorithm is presented in Table 2. This study calculates the random forest algorithms classification accuracy is calculated using 5-fold cross-validation in this study. The use of the cross-validation method benefits to obtain maximum accuracy results in breast cancer research performed five trials. Cross-validation divides the training data into separate parts of approximately the same size. Each section is selected sequentially as testing data, while the other sections are used as training data. The prediction model built on the training data is then applied to predict the class label of the testing data. This procedure was

continued until all sections were closed once, at which point the prediction accuracy across all tests was aggregated to offer an estimate of total performance [23].

TABLE II
ACCURACY OF RANDOM FOREST (RF) ALGORITHM

NO	Accuracy (5-fold cross-validation)
1	0.97
2	0.94
3	0.90
4	0.95
5	0.98

The results from Table 2 show that the average value of the 5-fold cross-validation research results from the random forest algorithm was 0.9480. This accuracy was obtained through random forest algorithm testing by considering point K randomly on the training data. Next, the testing data using a tree to predict the classification output. This study exploited 100 decision trees.

It is tuning the parameters of the random forest algorithm using a grid search cv to increase the accuracy of breast cancer data. The random forest algorithm has several parameters that are adjusted to obtain an optimal classification. This study executed four parameters. The first parameter is max_depth – the largest tree share or maximum tree depth for all trees in the forest. Max_features is the maximum number of characteristics utilized in node splitting. Criterion is the metric used to assess the termination criterion for the decision tree. There are two metrics in criterion, they are gini, and entropy. Gini measures the frequency of each element of the dataset, while entropy measures information that shows a feature interference with the target. Min_samples_split is the least number of samples required to separate the internal nodes is specified.

Grid Search CV takes into account all parameter combinations to find the best parameter values. All potential parameter value combinations are assessed in this approach, and the best mixture is prevented from offering the best classifier. The results of tweaking the random forest algorithm settings on the breast cancer dataset are shown in Table 3. The accuracy worth in Table 3 represents the classification accuracy calculated using the 5-fold cross-validation method. Tuning parameters in this study comprises 'criterion': ('gini', 'entropy'), 'max_depth': (3,5,7,9,10), 'max_features': ('auto', 'sqrt'), and 'min_samples_split': (2,4,6).

TABLE III
THE RESULTS OF THE BEST PARAMETER USING GRID SEARCH CV

NO	Accuracy (5-fold cross-validation)	Best Parameter			
		criterion	max depth	max features	min sample split
1	0.9475	'entropy'	10	'sqrt'	2
2	0.9550	'gini'	9	'sqrt'	4
3	0.9575	'entropy'	9	'sqrt'	4
4	0.9600	'gini'	9	'auto'	4
5	0.9525	'entropy'	9	'auto'	4

Based on Table 3, the highest classification accuracy was 0.9600 with the parameters' criterion': 'gini', 'max_depth': 9, 'max_features': 'auto', and 'min_samples_split': 4. The average accuracy value of the random forest algorithm using the grid search cv was 0.9545. The explanation in Table 3 grid search worked through the optimization of the criterion, max_depth, max_features, and min_samples_split parameters using cross-validation by dividing the data into k subset. One subset of training data that was evaluated employed testing data. Then, the accuracy value and the best parameter were selected, as presented in Table 3.

C. Evaluation Model

The evaluation was done using the confusion matrix method and grid search cv method to calculate the accuracy, recall, precision, and f1_score values – and compare the accuracy, recall, precision, and f1_score between the two methods. The results of the accuracy comparison are shown in Table 4.

TABLE IV
ACCURACY OF RANDOM FOREST (RF) ALGORITHM

	Random Forest without Grid Search CV	Random Forest with Grid Search CV
accuracy	0.9480	0.9545
precision	0.9455	0.9512
recall	0.9437	0.9497
f1_score	0.9438	0.9499

The outcomes of the assessment model using the confusion matrix, namely accuracy, precision, recall, and f1_score from breast cancer data, are shown in Table 4. The highest accuracy value produced from the random forest method utilizing a grid search yielded 0.9545. In comparison, the random forest accuracy that does not employ a grid search has 0.9480. Previous research indicated that the greater the precision and recall values, the higher the accuracy values created. Conversely, the lower the precision and recall values, the lower the accuracy values generated and confirmed in this study [24]. Table 4 compares the random forest algorithm's accuracy, precision, recall, and f1 score with and without the grid search cv technique. These results were found because the grid search cv method generates the best model parameters to predict the data more accurately, affecting the results. The better the resulting parameters, the higher the accuracy value is. It can be determined that the grid search cv method can improve the accuracy of breast cancer data.

IV. CONCLUSION

According to the investigation findings, the random forest algorithm can be used to diagnose malignant and benign breast cancer using previous data. It is possible to say that diagnosis using random forest with grid search cv is more accurate than diagnosis without grid search cv. And grid search cv yields an accuracy of 0.9545, whereas the prediction model without grid search cv yields 0.9480. The grid search cv increased the accuracy value by 0.0065. Apart from higher accuracy values, the precision, recall, and f1_score values of the random forest algorithm using the grid search cv method were also higher.

For further research, it is suggested to develop a breast cancer dataset using the grid search cv method with other algorithms, such as Logistic Regression, Xgboost, SVM, and other algorithms or equivalent algorithms with different datasets to prove that the grid search cv method can increase the accuracy value.

REFERENCES

[1] M. F. Ullah, *Breast Cancer: Current Perspectives on the Disease Status*. 2019.

[2] I. L. Maria, A. A. Sainal, and M. Nyorong, "Risiko Gaya Hidup Terhadap Kejadian Kanker Payudara Pada Wanita," *Media Kesehat. Masy. Indones.*, vol. 13, no. 2, p. 157, 2017, doi: 10.30597/mkmi.v13i2.1988.

[3] C. Mattiuzzi and G. Lippi, "Current Cancer Epidemiology glossary," *J. Epidemiol. Glob. Health*, vol. 9, no. 4, pp. 217–222, 2019, doi: DOI: https://doi.org/10.2991/jegh.k.191008.001.

[4] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, 2018, doi: 10.1016/j.ejor.2017.12.001.

[5] C. E. DeSantis *et al.*, "Breast cancer statistics, 2019," *CA. Cancer J. Clin.*, vol. 69, no. 6, pp. 438–451, 2019, doi: 10.3322/caac.21583.

[6] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *J. Pet. Sci. Eng.*, vol. 174, pp. 776–789, 2019, doi: 10.1016/j.petrol.2018.11.067.

[7] J. Dou *et al.*, "Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan," *Sci. Total Environ.*, vol. 662, no. January, pp. 332–346, 2019, doi: 10.1016/j.scitotenv.2019.01.221.

[8] A. R. Chowdhury, T. Chatterjee, and S. Banerjee, "A Random Forest classifier-based approach in the detection of abnormalities in the retina," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 193–203, 2019, doi: 10.1007/s11517-018-1878-0.

[9] W. Dong, Y. Huang, B. Lehane, and G. Ma, "XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring," *Autom. Constr.*, vol. 114, no. March, p. 103155, 2020, doi: 10.1016/j.autcon.2020.103155.

[10] Y. Shuai, Y. Zheng, and H. Huang, "Hybrid Software Obsolescence Evaluation Model Based on PCA-SVM-GridSearchCV," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2018, pp. 449–453, 2019, doi: 10.1109/ICSESS.2018.8663753.

[11] M. M. Ramadhan, I. S. Sitanggang, F. R. Nasution, and A. Ghifari, "Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency," *DEStech Trans. Comput. Sci. Eng.*, pp. 625–629, 2017, doi: 10.12783/dtcse/cece2017/14611.

[12] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 23, pp. 9193–9196, 1990, doi: 10.1073/pnas.87.23.9193.

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proc. Int. Jt. Conf. Neural Networks*, no. 3, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.

[14] B. Tan, J. Yang, Y. Tang, S. Jiang, P. Xie, and W. Yuan, "A Deep Imbalanced Learning Framework for Transient Stability Assessment of Power System," *IEEE Access*, vol. 7, pp. 81759–81769, 2019, doi: 10.1109/ACCESS.2019.2923799.

[15] H. Zhao, J. Zheng, J. Xu, and W. Deng, "Fault diagnosis method based on principal component analysis and broad learning system," *IEEE Access*, vol. 7, pp. 99263–99272, 2019, doi: 10.1109/ACCESS.2019.2929094.

[16] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1521–1530, 2018, doi: 10.3844/jcssp.2018.1521.1530.

[17] A. N. Zuda Pradana Putra, "Pebandingan Performa Naïve Bayes dan KNN pada Klasifikasi Teks Sentimen Jasa Ekspedisi," vol. 3, no. 1, pp. 145–152, 2022.

[18] S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," *2019 Int. Conf. Comput. Inf. Sci. ICCIS 2019*, pp. 1–4, 2019, doi: 10.1109/ICCISci.2019.8716405.

[19] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, pp. 1502–1509, 2016, doi: 10.12928/TELKOMNIKA.v14i4.3956.

[20] H. Zhang, H. Zhang, S. Pirbhulal, W. Wu, and V. H. C. D. Albuquerque, "Active balancing mechanism for imbalanced medical data in deep learning-based classification models," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 16, pp. 1–15, 2020, doi: 10.1145/3357253.

[21] R. Maglietta *et al.*, "Convolutional Neural Networks for Risso's Dolphins Identification," *IEEE Access*, vol. 8, pp. 80195–80206, 2020, doi: 10.1109/ACCESS.2020.2990427.

[22] D. A. Anggoro and P. I. Rahmatullah, "The implementation of subspace outlier detection in K-nearest neighbors to improve accuracy in bank marketing data," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 2, pp. 545–550, 2020, doi: 10.30534/ijeter/2020/44822020.

[23] T. H. Kerbaa, A. Mezache, and H. Oudira, "Model Selection of Sea Clutter Using Cross Validation Method," *Procedia Comput. Sci.*, vol. 158, pp. 394–400, 2019, doi: 10.1016/j.procs.2019.09.067.

[24] G. A. Buntoro, "Analisi Sentimen Hatespeech Pada Twitter dengan Metode Naive Bayes Classifier dan Support Vector Machine," *Resma*, vol. 3, no. 2, pp. 13–22, 2016.