pijmath
Pattimura International Journal of Mathematics (PIJMath)

Volume 1, issue 1, pages 33-42, May 2022
E-ISSN : 2830-6791

RESEARCH ARTICLE ◙ OPEN ACCESS

DOI: https://doi.org/10.30598/pijmathvol1iss1pp33-42

Comparison of Support Vector Machine and K-Nearest Neighbors in Breast Cancer Classification

Anita Desiani[1*], Adinda Ayu Lestari[2], M. Al-Ariq[3], Ali Amran[4], Yuli Andriani[5]

[1,2,3,4,5] *Mathematics Study Program, Faculty of Math and Science, Universitas Sriwijaya*
Jl. Raya Palembang - Prabumulih Km. 32, Indralaya, 30862, Indonesia

Corresponding author's e-mail: [1] anita_desiani@unsri.ac.id*

ABSTRACT

Article History

Received : .20th March 2022
Accepted : 26th April 2022
Published : 1st Mei 2022

Keywords

Breast Cancer;
Support Vector Machine;
Data Mining;
Classification

Cancer is one of the leading causes of death, and breast cancer is the second leading cause of cancer death in women. One method to realize the level of malignancy of breast cancer from an early age is by classifying the cancer malignancy using data mining. One of the widely used data mining methods with a good level of accuracy is the Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). Evaluation techniques of percentage split and cross-validation were used to evaluate and compare the SVM and KNN classification models. The result was that the accuracy level of the SVM classification method was better than the KNN classification method when using the cross-validation technique, which is 95,7081%. Meanwhile, the KNN classification method was better than the SVM classification method when using the percentage split technique, which is 95,4220%. From the comparison results, it can be seen that the KNN and SVM methods work well in the classification of breast cancer.

How to cite this article:
Anita Desiani, *et. al.*, "COMPARISON OF SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBORS IN BREAST CANCER CLASSIFICATION", *Pattimura Int. J. Math. (PIJMath).*, vol. 1, iss. 1, pp. 33-42, May 2022.
© 2022 by the Author(s)

1. Introduction

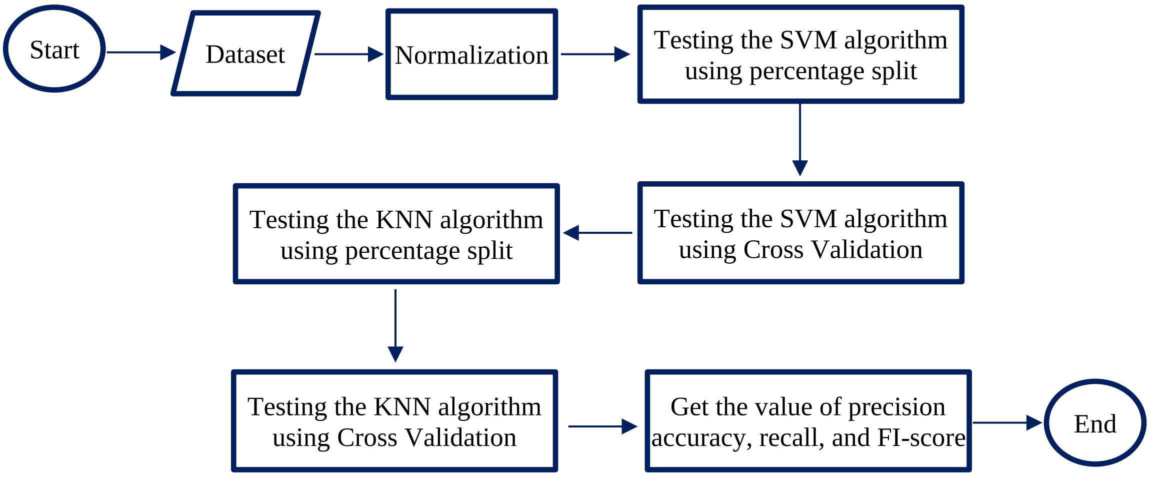
Cancer is a tumor that can invade tissues and spread to other organs. Cancer is the leading cause of death, with nearly 8 million deaths identified as malignant in 2008 and a projected 11 million cancer-related deaths by 2030 [1]. Breast cancer is the second leading cause of cancer death among women[2]. Over the last 5-year period (2012-2016), the incidence rate of breast cancer increased slightly by 0.3% per year, largely due to increasing rates of local staging and hormone receptor-positive disease [3]. Until now, one of the most common treatment methods is surgery and, if necessary, chemotherapy or radiation. However, this treatment will not significantly impact if cancer has reached an advanced stage. One way that can be done so that breast cancer can be detected early is early detection with information technology to facilitate the process of early detection of breast cancer in the community. One of the technologies used to facilitate the public in the health sector is data mining. With the advent of the information age, data mining is increasingly being used in clinical practice [4]. Data mining is the process of analyzing usable information and extracting big data from the data warehouse, which involves various patterns, intelligent methods, algorithms, and tools [5].

One of the data mining techniques that was used is classification. The purpose of classification is to accurately predict the target class for each case in the data [6]. The classification system can help increase the accuracy and reliability of the diagnosis, minimize the possibility of errors, and make the diagnosis more time-efficient [7]. One of the classification methods used in this research is the Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). A radial basis function (RBF) kernel was used in the SVM classification model. The Euclidean function was used in the KNN classification model to calculate the distance to the nearest data. Then, the nearest neighbors were taken, as many as 3. SVM is one of the popular classification methods. SVM classifications include Huang et al. [8] adopting a fruit fly optimization (FO) algorithm which is enhanced by a levy flight (LF) strategy and building an LF and FO-based SVM (LFFO-SVM), then getting an accuracy of 93.83%, recall 91.22 %, and specificity 96.53%. Aroef et al. [9] compared the SVM classification method and the random forest classification method for breast cancer classification, obtaining an accuracy of 95.45% for the SVM classification method and an accuracy of 90.90% for the random forest classification method. Liu et al. [10] also used the SVM classification method and used the Radial Basis Function (RBF) kernel function for breast cancer classification and got an accuracy of 96.58 for the Wisconsin Breast Cancer (WBC) dataset, an accuracy of 95.91% for the Wisconsin Diagnostic Breast Cancer dataset (WDBC).

Another popular classification method is KNN. The KNN algorithm is a non-parametric method that can manage classification and regression problems as one of the simplest machine learning algorithms [11], [12]. Several studies using the KNN classification method, including Rajaguru & Sannasi Chakravarthy [13], compared the KNN algorithm and decision tree algorithm for breast cancer classification. The result was that the accuracy of the KNN algorithm is higher than the decision tree algorithm, which is 95.61%. In comparison, the accuracy of the decision tree algorithm is 91.23%. Eyupoglu [14] used 2-fold, 5-fold, and 10-fold cross-validation to test the accuracy of the KNN algorithm for breast cancer classification. As a result, he got an average accuracy rate of 97%. Mushtaq et al. [15] compared the accuracy of 3 functions used in the KNN algorithm: Euclidean, Manhattan, and Cosine. The result is that the KNN classification method using the Manhattan function gets the highest accuracy rate, 99.42%. Then, the Euclidean function gets an accuracy rate of 98.85%, while the Cosine function gets the lowest accuracy rate, 94.86%. In this study, a comparison of the level of accuracy between the classification method using SVM and KNN was carried out for the classification of breast cancer.

2. Research Methods

In this study, the Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) algorithms were used for breast cancer classification. The system schema is made as follows:


Start
Dataset
Normalization
Testing the SVM algorithm using percentage split
Testing the SVM algorithm using Cross Validation
Testing the KNN algorithm using percentage split
Testing the KNN algorithm using Cross Validation
Get the value of precision accuracy, recall, and FI-score
End

.

Figure 1. General schema of the classification process

Based on Figure 1, there are 6 processes must be carried out for breast cancer classification, namely, normalizing the data to become uniform. Then for testing the SVM and KNN classification methods, a percentage split of 80% was carried out for training data. The remaining 20% was for testing data. For testing using the cross-validation technique, 10-fold cross-validation was used for each SVM and KNN classification method.

2.1 Description of data

This dataset was obtained from research conducted by doctors from the United States. This study was conducted in 1995. The data obtained were 699 from the site https://www.openml.org/search?type=data&sort=runs&id=15. In the Class attribute, a classification attribute, the benign variable representing benign tumors is 458 data. In comparison, the malignant variable representing malignant tumors is 241 data. This dataset contains nine attributes, namely Clump_Thickness, Cell_Size_Uniformity, Cell_Shape_Uniformity Marginal_Adhesion, Single_Epi_Cell_Size, Bland_Chromatin, Normal_Nucleoli, Mitoses, and Class.

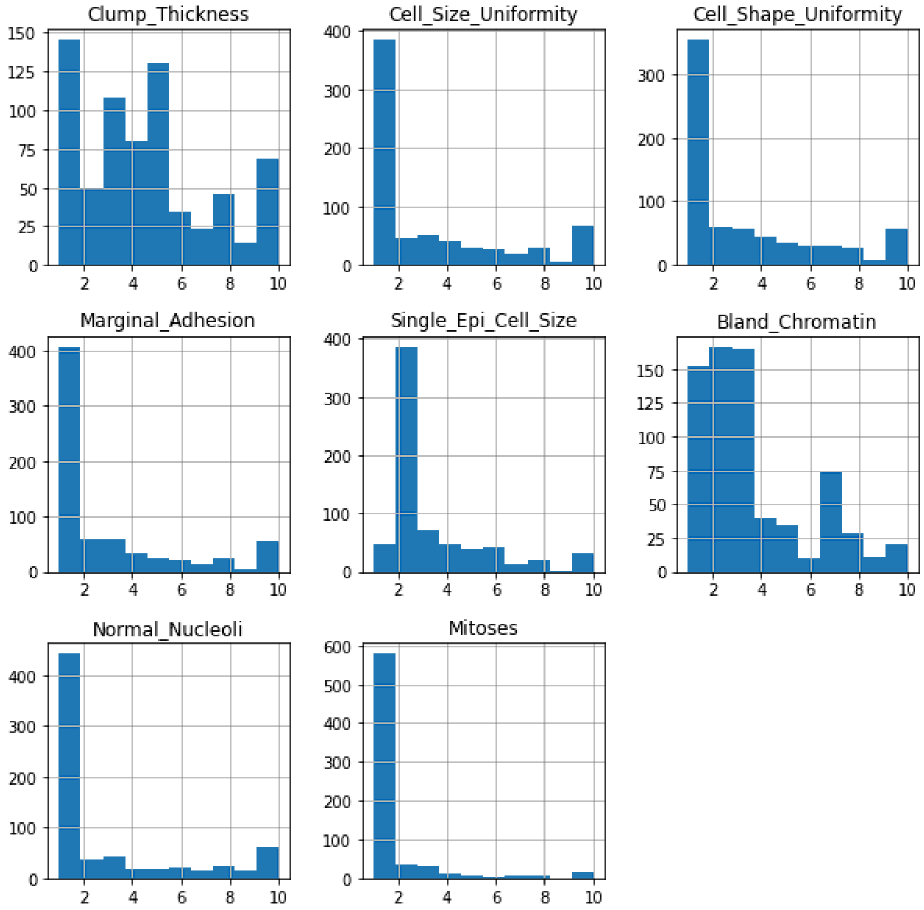


Figure 2. Visualization of each data variable

In Figure 2, the horizontal numbers from 1-10 in each attribute histogram show the variables in each attribute. Then, the vertical numbers in each histogram indicate the amount of data. The following Table 1 explains the description for each attribute:

Table 1. Attribute description

Attribute	Description
Clump_Thickness	Estimating whether it is single-celled or multilayered
Cell_Size_Uniformity	Consistency in cell size
Cell_Shape_Uniformity	Estimation of cell shape quality and identification of cell edge differences
Marginal_Adhesion	The number of epithelial cells that tend to stick together
Single_Epi_Cell_Size	Determine whether the epithelial cells are significantly enlarged
Bland_Chromatin	The value of the uniformity of the nucleus texture is in the fine or coarse range
Normal_Nucleoli	Define small and invisible or large visible ucleolus
Mitoses	Describe the level of reproductive activity
Class	Determine the type of cancer

Source: https://www.openml.org/search?type=data&sort=runs&id=15.

In this study, we scaled the data to suit the model training. Therefore, a normalization process was carried out. Normalization is an operation on raw data that rescales or transforms it so that each feature has a uniform contribution [16]. There are many ways to carry out the normalization process, one of which is the min-max normalization technique. Min-max normalization is usually used to train machine learning models because it scales back values to a range between 0 and 1 while maintaining the characteristics of the data [17]. Researchers have used the normalization technique to improve classification performance in various application areas [16]. The equation used to perform the min-max normalization technique is (1):

$$X_{Normalization} = \frac{X_{old}-X_{minimum}}{X_{maximum}-X_{minimum}} \tag{1}$$

2.2 Implementation of the Support Vector Machine (SVM) algorithm

The SVM concept can be explained simply as an effort to find the best hyperplane that functions as a separator of two classes. Available data were denoted as $x_i \in R^d$ then each label was denoted by $y_i \in \{-1,+1\}$ for $i = 1,2,\dots,l$ where l is the number of data. It is assumed that there were two classes, namely positive sample (+1) and negative sample (-1) can be separated completely by a d-dimensional hyperplane, which is defined in the equation below (2):

$$\vec{w_i}.\vec{x_i} + b = 0 \tag{2}$$

Pattern *i* which includes classes -1 and +1 (negative sample) can be formulated as a pattern that satisfies inequality (3):

$$\vec{w_i}.\vec{x_i} + b \leq -1 \tag{3}$$

while pattern i which belongs to class +1 (positive sample) satisfies inequality (4):

$$\vec{w_i}.\vec{x_i} + b \geq +1 \tag{4}$$

Where $\vec{w_i}$ is a vector, which is normal for the hyperplane. The perpendicular distance from the hyperplane to the origin is given by $\frac{b}{||\vec{w}||}$ where $||\vec{w}||$ is the Euclidean norm of the vector $\vec{w_i}$ (3) and (4) can be combined as (5) [18]:

$$y_i(\vec{w}.\vec{x_i} + b) \geq 1 \tag{5}$$

Furthermore, this hyperplane was the decision function for the classification problem of the two classes above (6):

$$f\big(\phi(x)\big) = sign\big(w.\phi(x)\big) + b \tag{6}$$

$$f\big(\phi(x)\big) = sign\left(\textstyle\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)^T.\phi(x) + b\right) \tag{7}$$

The formula used to calculate the prediction results with a single hyperplane and according to b and w to be obtained is defined in equations (3) and (4) as follows (8):

$$K(x, x_i) = \emptyset(x).\emptyset(x_i) \tag{8}$$

Next, the width of the margin y or the distance from x^+ (data located in class y = +1) to hyperplane or distance from x^- (data located in class y = -1) to *hyperplane* was found by maximizing $||w||$ with the conditions defined in equations (9) and (10) as follows:

$$w = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) \tag{9}$$

$$b = -\frac{1}{2}(w.x^+ + w.x^-) \tag{10}$$

With obstacles $y_i(< w, x_i > +b \geq 1), i = 1,2, \dots n$ is a hyperplane with maximum margin. The equation used to find the alpha value is defined in equation (11) below:

$$\alpha_1 = \alpha_2 = \cdots = \alpha_n = \frac{n}{\sum K(n*n)} \tag{11}$$

2.3 Application of the K-Nearest Neighbors (KNN) Algorithm

In 1968 Cover and Hart [11] proposed the K-Nearest Neighbor (KNN) algorithm, which was completed later. KNN is the procedure of choice in many scenarios, especially when the underlying model is complex due to its simplicity and flexibility [19]. Here is a visualization of how the simple KNN algorithm works:

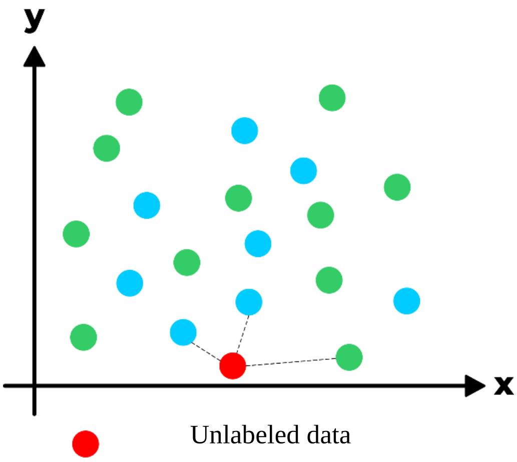


Figure 3. Visualization of how K-Nearest Neighbor works

In Figure 1, it can be seen that the red circle is a test sample that has not been classified or has no label. The other colored circles were test samples that have been classified or have a label. The KNN algorithm classifies unlabeled test samples based on the most similar samples between the KNNs closest to the test sample. A certain distance measure determines the distance between the test sample and each training data sample [20]. One of the distance functions that was used in this study was the Euclidean distance. Paredes et al. [21] analyzed 5 distance functions, namely euclidean distance, manhattan distance, canberra distance, chebychev distance, and minkowsky distance, based on an instance-based learning algorithm using the 1-nearest neighbor classification method and incremental hypersphere. They concluded that the euclidean distance function and the manhattan distance significantly yielded good results in various problems. The following is the Euclidean distance equation that was used in this study (12):

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_a - y_a)^2} \tag{12}$$

Where $x = (x_1, x_2, \dots, x_n)$ and $a = (a_1, a_2, \dots, a_n)$ are an attribute vector used for breast cancer classification. In this study, the nearest neighbor was taken as much as 3.

2.4 Result Evaluation

Evaluation of the results was carried out with two techniques, namely cross-validation and percentage split. The cross-validation technique works by alternating the data that would become training data and testing data. The percentage split works by dividing the data used for training and the data used for testing. A confusion matrix was used to calculate the results of the two methods. A confusion matrix is a matrix that displays a visualization of the performance of the classification algorithm using the data in the matrix, then compares the predicted classification against the actual classification in the form of False Positive (FP), True Positive (TP), False Negative (FN), and True Negative (TN) of information. The confusion matrix is represented by a matrix in which each row represents an example in the predicted class. In contrast, each column represents the actual class [22]. The confusion matrix for the two-class classification system is as follows:

Table 2. Confusion Matrix

Class	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Postive)	TN (True Negative)

When the result is in the TP column, the result is true and is identified as positive. When the result is in the FP column, the result is false and identified as positive. When the result is in the FN column, the result is false and is identified as negative. When the result is in the TN column, the result is true and is identified as negative.

Accuracy (13) or error rate is the number of correct predictions made by the model through the collection of data. Accuracy is usually calculated using independent tests which are not always used in the learning process.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (13)$$

Precision is the ratio of positive correct predictions to overall positive predicted outcomes (14):

$$Precision = \frac{TP}{(TP+FP)} \qquad (14)$$

Recall is the ratio of true positive predictions compared to all true positive data (15):

$$Recall = \frac{TP}{(TP+FN)} \qquad (15)$$

F1-score is a confusion matrix that takes into account the ratio of precision and recall defined as follows (16):

$$F1 - score = 2\frac{presisi\ X\ recall}{presisi+recall} \qquad (16)$$

3. RESULTS AND DISCUSSION

This study used Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) for classifying malignant and benign breast cancer patients. The dataset used in this study amounted to 699 data. This dataset contained nine attributes, namely Clump_Thickness, Cell_Size_Uniformity, Cell_Shape_Uniformity, Marginal_Adhesion, Single_Epi_Cell_Size, Bland_Chromatin, Normal_Nucleoli, Mitoses, and *Class*. In the Class attribute, a classification attribute, the benign variable representing benign tumors is 458 data. In comparison, the malignant variable representing malignant tumors is 241 data. The evaluation technique is a percentage split of 80% for training data and the remaining 20% for testing data on both classification methods, namely SVM and KNN. The second evaluation technique used was cross-validation. 10-fold cross-validation was used for each SVM and KNN classification method.

3.1. Classification results using the Support Vector Machine (SVM)

The results of the evaluation using the percentage split technique, the breast cancer classification method using SVM, was shown in Table 3, below:

Table 3. Percentage split results for Support Vector Machine (SVM)

	Benign	Malignant	Total
Positive	95	4	99
Malignant	0	41	41
Total	**95**	**45**	**140**

The value of precision, recall, and f1-score used the percentage split technique for the SVM classification method as shown in Table 4, below:

Table 4. Values of precision, recall, and f1-score for SVM using the percentage split technique

	Precision	Recall	F1-score
Benign	100%	96%	98%
Malignant	91%	100%	95%

Based on the results of the evaluation using a percentage split, the accuracy rate for the SVM classification method was 97,1428%. The results of the evaluation used the cross validation technique, the breast cancer classification method using SVM as shown in Table 5, below:

Table 5. Cross validation results for Support Vector Machine (SVM)

	Benign	Malignant	Total
Benign	438	20	458
Malignant	10	231	241
Total	**448**	**251**	**699**

The value of precision, recall, and f1-score using the cross validation technique for the SVM classification method was shown in Table 6, below:

Table 6. Precision values, recall, and f1-score for SVM using cross validation technique

	Precision	Recall	F1-score
Benign	98%	96%	97%
Malignant	92%	96%	94%

Based on the results of the evaluation using cross validation, the accuracy rate for the SVM classification method was 95,7081%. It can be concluded that the percentage split evaluation technique provides a higher level of accuracy than the cross validation evaluation technique for the Support Vector Machine (SVM) classification method, which is 97.1428%.

3.2. Classification results using K-Nearest Neighbor (KNN)

The results of the evaluation using the percentage split technique, the method of classifying breast cancer using SVM were shown in Table 7, below:

Table 7. The results of the percentage split for K-Nearest Neighbor (KNN)

	Benign	Malignant	Total
Benign	97	2	99
Malignant	1	40	41
Total	**98**	**42**	**140**

The values of precision, recall, and f1-score using the percentage split technique for the KNN classification method were shown in Table 8, below:

Table 8. The values of precision, recall, and f1-score for KNN using the percentage split technique

	Precision	Recall	F1-score
Benign	**99%**	**98%**	**98%**
Malignant	**95%**	**98%**	**96%**

Based on the results of the evaluation using a percentage split, the accuracy rate for the KNN classification method was 97.8571%. The results of the evaluation using the cross validation technique, the breast cancer classification method using KNN were shown in Table 9, below:

Table 9. Cross validation results for K-Nearest Neighbor (KNN)

	Benign	Malignant	Total
Benign	**443**	**15**	**458**
Malignant	**17**	**224**	**241**
Total	**460**	**239**	**699**

The values of precision, recall, and f1-score using the cross validation technique for the SVM classification method were shown in Table 10, below:

Table 10. The values of precision, recall, and f1-score for KNN using cross validation technique

	Precision	Recall	F1-score
Benign	**96%**	**97%**	**97%**
Malignant	**94%**	**93%**	**93%**

Based on the evaluation results using cross-validation, the accuracy rate for the SVM classification method was 95.4220%. Therefore, the percentage split evaluation technique provides a higher level of accuracy than the cross-validation evaluation technique for the Support Vector Machine (SVM) classification method, which is 97.1428%.

4. Conclusions

Based on the comparison of the results between 2 methods and 2 training techniques, namely cross-validation and percentage split, it can be concluded that the Support Vector Machine (SVM) and KNN classification models work well in the classification of breast cancer. The performance results show that SVM gets a better accuracy, 95,7081% when using the cross-validation evaluation technique. The K-Nearest Neighbor (KNN) classification model gets a better accuracy rate than the Support Vector Machine (SVM), which is 97.8571% when using the percentage split evaluation technique. Both methods' precision, recall, and F1-score performance measures are above 90%. It shows that the SVM and KNN methods are strong and excellent in the classification of breast cancer.

References

[1] J. R. Benson and I. Jatoi, “The global breast cancer burden,” *Futur. Oncol.*, vol. 8, no. 6, pp. 697–702, 2012, doi: 10.2217/fon.12.61.

[2] Y.-S. Sun *et al.*, “Risk Factors and Preventions of Breast Cancer,” *Int. J. Biol. Sci.*, vol. 13, no. 11, pp. 1387–1397, Nov. 2017, doi: 10.7150/ijbs.21635.

[3] C. E. Desantis *et al.*, “Breast cancer statistics, 2019,” *CA. Cancer J. Clin.*, vol. 69, no. 6, pp. 438–451, Nov. 2019, doi: https://doi.org/10.3322/caac.21583.

[4] J. Yang *et al.*, “Brief introduction of medical database and data mining technology in big data era,” *J. Evid. Based. Med.*, vol. 13, no. 1, pp. 57–69, Feb. 2020, doi: https://doi.org/10.1111/jebm.12373.

[5] M. J. H. Mughal, “Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview,” 2018.

[6] G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013, pp. 1–7. doi: 10.1109/ICCCNT.2013.6726842.

[7] R. Geetha, R. Professor, & Head, and G. Sivagami, “Parkinson Disease Classification using Data Mining Algorithms,” 2011.

[8] H. Huang *et al.*, “A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features,” *BMC Bioinformatics*, vol. 20, no. Suppl 8, pp. 1–15, 2019, doi: 10.1186/s12859-019-2771-z.

[9] C. Aroef, Y. Rivan, and Z. Rustam, “Comparing random forest and support vector machines for breast cancer classification,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 815–821, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14785.

[10] N. Liu, J. Shen, M. Xu, D. Gan, E. S. Qi, and B. Gao, “Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis,” *Math. Probl. Eng.*, vol. 2018, 2018, doi: 10.1155/2018/3875082.

[11] T. M. Cover and P. E. Hart, “Approximate Formulas for the Information Transmitted by a Discrete Communication Channel,” *IEEE Manhattan, NY, USA*, vol. 24, 1952.

[12] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, “Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data.” Wiley-Blackwell, Mar. 01, 2019.

[13] H. Rajaguru and S. C. S R, “Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer,” *Asian Pac. J. Cancer Prev.*, vol. 20, no. 12, pp. 3777–3781, Dec. 2019, doi: 10.31557/APJCP.2019.20.12.3777.

[14] C. Eyupoglu, “Breast cancer classification using k-nearest neighbors algorithm,” *Online J. Sci. Technol.*, vol. 8, no. 3, pp. 29–34, 2018.

[15] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, “Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets,” Jan. 2020.

[16] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[17] H. Kwon, K. C. Oh, Y. Choi, Y. G. Chung, and J. Kim, “Development and application of machine learning-based prediction model for distillation column,” *Int. J. Intell. Syst.*, vol. 36, no. 5, pp. 1970–1997, May 2021, doi: https://doi.org/10.1002/int.22368.

[18] A. Rizwan, N. Iqbal, R. Ahmad, and D.-H. Kim, “WR-SVM model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification,” *Appl. Sci.*, vol. 11, no. 10, p. 4657, 2021.

[19] O. Anava, K. Y. Levy, and E. Zurich, “K-Nearest Neighbors: From Global to Local,” 2016.

[20] H. A. Abu Alfeilat *et al.*, “Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review,” *Big Data*, vol. 7, no. 4. Mary Ann Liebert Inc., pp. 221–248, Dec. 01, 2019. doi: 10.1089/big.2018.0175.

[21] R. Paredes, J. S. Cardoso, and X. M. Pardo, “Pattern recognition and image analysis: 7th Iberian conference, IbPRIA 2015 Santiago de Compostela, Spain, june 17–19, 2015 proceedings,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9117. doi: 10.1007/978-3-319-19390-8.

[22] A. Krouska, C. Troussas, and M. Virvou, “Comparative evaluation of algorithms for sentiment analysis over social networking services.,” *J. Univers. Comput. Sci.*, vol. 23, no. 8, pp. 755–768, 2017.