Engineering and Applied Science Research
https://www.tci-thaijo.org/index.php/easr/index
Published by the Faculty of Engineering, Khon Kaen University, Thailand

Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction

Lalu Ganda Rady Putra*, Khairan Marzuki and Hairani Hairani

Department of Computer Science, Faculty of Engineering, Bumigora University, Mataram, 83127, Indonesia

Received 18 July 2023
Revised 26 September 2023
Accepted 6 October 2023

Abstract

Indonesia is an archipelago with the fourth largest population in the world, with a population of 283 million. In Indonesia, breast cancer ranks first in cancer and is the highest contributor to death. Deaths caused by breast cancer can be minimized by screening and early detection to avoid the risk of more severe cancer. Early detection of breast cancer can delay the growth of cancer cells and increase the chances of recovery. This research proposed a machine learning-based application for screening and early detection of breast cancer independently based on perceived symptoms. However, developing breast cancer early detection applications requires a very high level of accuracy to minimize prediction errors. This research focused on finding the optimal accuracy of the machine learning method so that it could predict breast cancer with a very low error rate. This research aimed to improve the performance of classification methods in breast cancer disease prediction using the correlation feature selection approach and hybrid sampling Smote-Tomek Link. This research utilized Support Vector Machine (SVM) and Naive Bayes classification methods with a combination of Smote-Tomek Link hybrid sampling approach and correlation feature selection. Hybrid Sampling Smote-Tomek Link balanced the data by minimizing noise in the data created. At the same time, the correlation feature selection method was used to select relevant or influential attributes with class attributes based on a strong correlation level (≥ 0.6) between input attributes and classes. The results of this study obtained that the SVM method with hybrid sampling and correlation feature selection obtained the best performance compared to the Naive Bayes method and previous research referred to with an accuracy of 96.80%, sensitivity of 96.80%, and specificity of 96.80%. Thus, using the Smote-Tomek Link hybrid sampling approach and correlation feature selection positively impacted the performance increase in the SVM and Naive Bayes methods for breast cancer prediction.

Keywords: Breast cancer prediction, Feature selection correlation, Machine learning methods, Hybrid Smote-Tomek Link

1. Introduction

Indonesia is an archipelago with the fourth largest population in the world, with a population of 283 million. In Indonesia, breast cancer ranks first in cancer [1] and is the highest contributor to death [2]. The number of breast cancer deaths is very high, influenced by several factors such as late treatment of cancer patients and delayed early detection. In contrast, the mortality rate of breast cancer can be minimized by screening and early detection to avoid the risk of more severe cancer. Delayed treatment is caused by a lack of public knowledge about breast cancer and not early detection. Early detection is an effort to examine people who have felt the symptoms as early as possible so that action can be taken quickly and precisely. Early detection of breast cancer can delay the growth of cancer cells and increase the chances of recovery. Therefore, the solution offered by this research is to create a machine learning-based application for screening and early detection of breast cancer independently based on perceived symptoms. However, developing breast cancer early detection applications requires a very high level of accuracy to minimize prediction errors. This research focuses on finding the optimal accuracy of the machine learning method so that it can predict breast cancer with a very low error rate.

Similar research has predicted breast cancer using various approaches such as the Artificial Neural Network (ANN) method [3, 4], XGBoost [5, 6], Recurrent Neural Network (RNN) with RFE feature selection [7], SVM [8, 9], XGBoost with Pearson Correlation feature selection [10] , Random Forest and Logistic Regression with Variance Inflation Factor (VIF) feature selection [11] , Logistic Regression and SVM with feature selection [12] , Logistic Regression with Chy Square feature selection [13] , Logistic Regression [14] , Convolutional Neural Network (CNN) [15-17] , Deep Neural Network [18] , Deep Learning with feature selection [19] .

Based on the previous research above, feature selection approaches are already used to improve accuracy and speed up computation time. However, previous research does not address the problem of unbalanced data on breast cancer datasets which can cause classification methods to favor the prediction of majority classes over minority classes. Therefore, this study solves the problem of unbalanced data and selecting relevant features using feature selection to improve the accuracy of classification methods. The method of resolving unbalanced data uses the Smote-Tomek Link hybrid sampling approach because it can reduce data noise in the resulting synthetic data [20-22]. In contrast, the feature selection method used is Pearson correlation to select influential attributes because it can increase the accuracy of the classification method [23, 24] .

*Corresponding author.
Email address: Laluganda@universitasbumigora.ac.id

doi: 10.14456/easr.2023.59

This research aims to improve the performance of classification methods in breast cancer disease prediction using the correlation feature selection approach and hybrid sampling Smote-Tomek Link.

2. Materials and methods

2.1 Materials

This study uses a breast cancer dataset obtained from Kaggle. The number of instances in the breast cancer dataset is 569, with 30 input attributes and 1 output attribute with two class categories, namely Benign and Malignant (See Table 1). The attributes owned by the breast cancer dataset are radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst, Class.

Table 1 Breast cancer disease sample data

No	Radius_Mean	Texture_Mean	…..	Symmetry Worst	Fractal Dimension Worst	Diagnosis
1	17.99	10.38	….	0.4601	0.1180	Benign
2	20.57	17.77	….	0.275	0.08902	Benign
3	19.69	21.25	….	0.3612	0.08758	Benign
4	11.42	20.38	….	0.6638	0.173	Benign
5	20.29	14.34	….	0.2364	0.07678	Benign
6	12.45	15.7	….	0.3985	0.1244	Benign
7	18.25	19.98	….	0.3063	0.08368	Benign
8	13.71	20.83	….	0.3196	0.1151	Benign
9	13	21.82	….	0.4378	0.1072	Benign
10	12.46	24.04	….	0.4366	0.2075	Benign
	….	……	…..	…..	……	…..
560	11,51	23.93	….	0.2112	0.08732	Malignant
561	14.05	27.15	….	0.225	0.08321	Malignant
562	11.2	29.37	….	0.1566	0.05901	Malignant
563	15.22	30.62	….	0.4089	0.1409	Benign
564	20.92	25.09	….	0.2929	0.09873	Benign
565	21.56	22.39	….	0.206	0.07115	Benign
566	20.13	28.25	….	0.2572	0.06637	Benign
567	16.6	28.08	….	0.2218	0.0782	Benign
568	20.6	29.33	….	0.4087	0.124	Benign
569	7.76	24.54	….	0.2871	0.07039	Malignant

2.2 Data preprocessing

The breast cancer dataset has the problem of unbalanced data and a large number of attributes. The number of Benign classes (355 instances) is more than the Malignant class (214 instances). This can cause machine learning methods to ignore the minority class and prioritize the majority class so that the prediction results are biased. Therefore, data preprocessing is needed to improve data quality and machine learning methods' performance. This research uses data preprocessing techniques to balance classes on breast cancer datasets using a data sampling approach. The data sampling used is the Smote-Tomek Link method. SMOTE-Tomek Link is a method that combines SMOTE (Synthetic Minority Over-sampling Technique) and Tomek Link techniques to handle data imbalance in classification problems. SMOTE is used to oversample the minority class by creating synthetic samples. In contrast, Tomek Link removes samples close to each other between the minority and majority classes (noise). SMOTE-Tomek Link aims to reduce data imbalance by reducing noise samples and increasing the number of samples in the minority class with synthetic samples. Combining SMOTE and Tomek Link can improve the performance of machine learning models on imbalanced datasets.

The Smote method was proposed by [25] aims to create artificial data between minority classes based on the distance of nearest neighbors by means of linear interpolation until the data is balanced. The Smote method creates artificial data in two steps, namely first, calculate the distance of k nearest neighbors to the Euclidean distance between the considered sample and the identified sample. After that, multiply by a randomly generated number between 0 and 1, then add the resulting artificial sample. Meanwhile, the Tomek Link method is an under sampling method developed by [26] to delete majority data samples that are close to minority data using nearest neighbors to select the majority instances to be deleted. Tomek Link aims to remove noise data at the majority and minority class decision boundaries which can complicate the classification process.

Moreover, the number of attributes in the breast cancer dataset is so large that it can reduce the performance of the classification method. Therefore, this study uses a correlation-based feature selection approach to select influential features in the breast cancer dataset. The correlation-based feature selection (CFS) method selects features on the dataset based on the level of correlation between input and output attributes. The formula for calculating correlation uses Equation (1), while the level of correlation strength is shown in Table 2.

$$r = \frac{n.(\sum XY)-(\sum X).(\sum Y)}{\sqrt{(n.\sum X^2-(\sum X)^2)}\sqrt{(n.\sum Y^2-(\sum Y)^2)}} \quad (1)$$

r is the correlation coefficient value, while ***n*** is the amount of data. $\sum$***XY*** is the total of the multiplication of variable X with variable Y. $\sum$**X** is the total of variable X, $\sum$***Y*** is the total of variable Y. $\sum$***X²*** is the total of the X variables squared, **(**$\sum$***X*)²** is the square of the total X variable. $\sum$***Y²*** is the total of the squared Y variable, while **(**$\sum$***Y*)²** is the square of the total Y variable.

Table 2 Correlation level

Coefficient Range	Correlation Strength
0 – 0.19	Very Low
0.2 – 0.29	Low
0.4 – 0.59	Adequate
0.6 – 0.79	Strong
0.8 – 1	Very Strong

2.3 Machine learning implementation

This research uses machine learning methods for breast cancer prediction, namely SVM and Naive Bayes methods. The Naive Bayes method is one of the classification methods based on Bayes' Theorem, assuming that all features or attributes used are conditionally independent. The Naive Bayes method is naive because the values of the features are independent of each other if the class is known. The Naive Bayes method calculates the probability of a class (label) based on the associated features. The Naive Bayes method is relatively simple and efficient in data processing and classification. Despite its naive assumptions, this method often gives good results in many applications. While the SVM (Support Vector Machine) method is a machine learning algorithm used for classification. The working concept of the SVM method involves separating data into feature spaces by finding the best hyperplane that separates two classes with maximum distance. The SVM method can work on both linear and non-linear data. If the data is not linearly separable in the feature space, SVM uses kernel techniques to project the data into a high-dimensional feature space. The kernel trick allows SVM to tackle the linear non-separable problem without explicitly projecting the data into a higher feature space. This saves computational time and avoids very high dimensionality. Some of the commonly used kernels in SVM include linear kernel, polynomial kernel, Gaussian kernel (RBF), sigmoid kernel, and radial basis function (RBF) kernel. The SVM method has the advantage of handling complex data and has limitations in large amounts of data. In addition, SVM can also use different kernels to adjust to the characteristics of the data. However, SVM also has disadvantages in computational performance when faced with very large datasets.

2.4 Performance testing

This research uses model performance testing with a confusion matrix table. The confusion matrix table describes the model's performance by comparing the model's prediction with the actual value of the observed data. The confusion matrix contains four values, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). This research uses several performance measurement metrics on the implemented model: accuracy, sensitivity, and specificity. Accuracy is used to see the ratio between the number of correct predictions (true positives and true negatives) and the total number of samples. Sensitivity measures the accuracy of predicting the positive class (true positives). This metric is important when we want to minimize the number of false negatives (positive cases that are misclassified as negative). While specificity measures the accuracy of predicting the negative class (true negatives). The calculation formulas for accuracy, sensitivity or recall, specificity, precision, F1 score, and AUC are shown in Formulas (2), (3), (4), (5), (6),and (7) [27, 28].

$$Accuracy \ = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \quad (2)$$

$$Sensitivity/Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \quad (3)$$

$$Specificity = \frac{True\ Negative}{True\ Negative+False\ Positive} \quad (4)$$

$$Precision = \frac{True\ Positive}{True\ True\ Positive+False\ Positive} \quad (5)$$

$$F1\ score \ = \frac{2*(Precision*Recall)}{(Precision+\ Recall)} \quad (6)$$

$$AUC \qquad = \frac{Recall\ \ +Specificity}{2} \quad (7)$$

3. Results

This section describes the research results obtained according to the previous research stages. Breast cancer disease data has several shortcomings, namely unbalanced data and a large number of attributes. Unbalanced data and many attributes in the data can reduce the performance of SVM and Naïve Bayes methods. Therefore, this research uses data sampling, and feature selection approaches on breast cancer datasets to improve the performance of classification methods. Resolving unbalanced data in breast cancer data using the Smote-Tomek Link method. The results of breast cancer data balancing are shown in Figure 1.

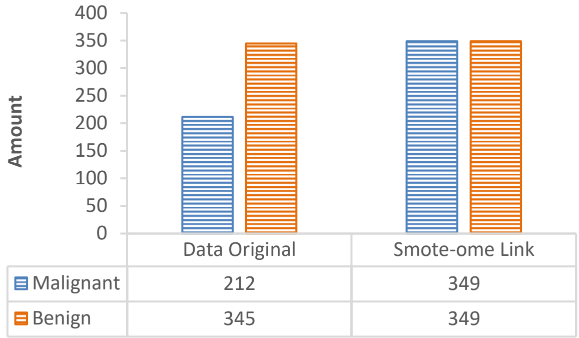

Figure 1 Breast cancer data distribution before and after Smote-Tomek Link

Breast cancer data has many attributes, namely 30 attributes. So it needs to select influential features to improve the performance of the method used. This research uses a correlation-based feature selection approach to select relevant attributes. The selected attribute is the attribute that has a higher correlation value of 0.6, because it is in the strong correlation range [29]. Based on Figure 2, the number of input the attribute that has a higher correlation value of 0.6 is 11 attributes such as the radius mean, perimeter mean, area mean, compactness mean, concavity mean, concave points mean, radius worst perimeter worst, area worst, concavity worst, and concave points worst.

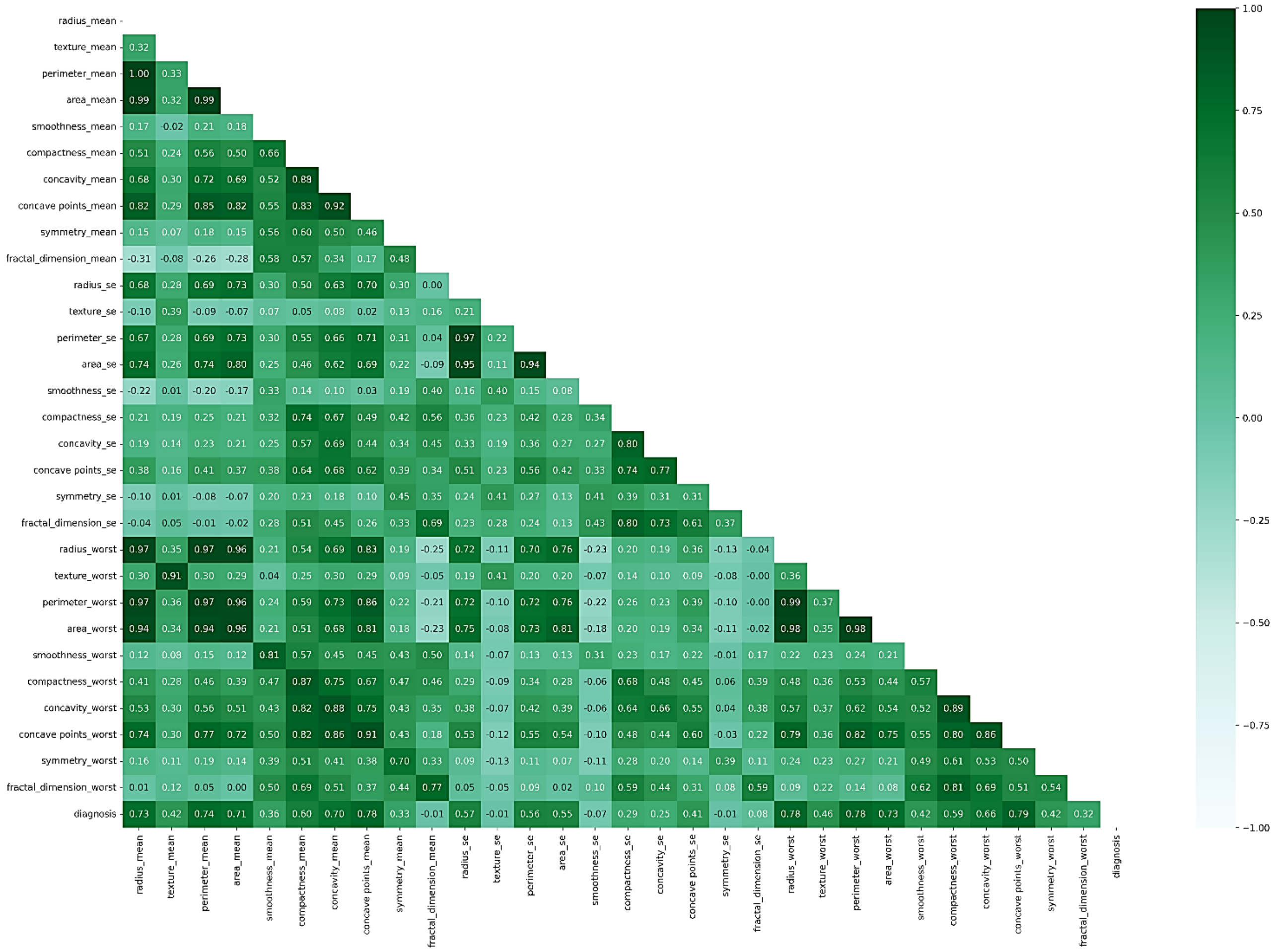

Figure 2 Correlation between attributes on breast cancer dataset

This research uses SVM and Naive Bayes methods for breast cancer prediction. The data is first divided into training and testing data using 10-Fold Cross-Validation. The prediction results are evaluated using a confusion table based on accuracy, sensitivity, and specificity. The test results of SVM and Naive Bayes methods on breast cancer prediction are shown in Table 3. Based on Table 3, the SVM method with Smote-Tomek Link and correlation feature selection correctly predicted the benign class in 338 out of 349 instances. In comparison, the Malignant class correctly predicted as many as 338 out of 349 instances. While the Naive Bayes method with Smote-Tomek Link and correlation feature selection managed to predict the benign class correctly in as many as 335 out of 349 instances, while the Malignant class was correctly predicted in as many as 328 out of 349 instances. Table 3 shows that the SVM method with Smote-Tomek Link and correlation feature selection predicts Benign and Malignant classes best compared to Naive Bayes with Smote-Tomek Link and correlation feature selection.

From the confusion matrix results of SVM and Naive methods, performance metrics such as accuracy, sensitivity, and specificity can be calculated, as shown in Table 4. In Table 4, the highest performance is obtained by the SVM method with Smote-Tomek Link and correlation feature selection with an accuracy of 96.80%, sensitivity of 96.80%, and specificity of 96.80%. While the Naive Bayes method with Smote-Tomek Link and correlation feature selection with an accuracy of 95.00%, sensitivity of 94.00%, and specificity of 96.00%. Using the Smote-Tomek Link hybrid sampling approach and feature selection can improve accuracy, sensitivity, and specificity in the SVM and Naive Bayes methods. However, the sensitivity metric with the highest rate of increase is the sensitivity part [30], where the SVM method with Smote-Tomek Link and correlation feature selection obtained an increased rate of 6.80%, while the Naive Bayes method with Smote-Tomek Link and correlation feature selection obtained an increased rate of 7.20%. In the accuracy metric, there is also an increase, but not too high, as the SVM method with Smote-Tomek Link and correlation feature selection obtained 2.80%, while the Naive Bayes method with Smote-Tomek Link and correlation feature selection obtained 2.40%.

Table 3 Confusion metric results of SVM and Naïve Bayes methods

Methods	Data Sampling	Feature Selection	Benign	Malignant	
SVM	Original	-	344	13	Benign
			21	191	Malignant
		Correlation	344	13	Benign
			16	196	Malignant
	Smote-Tomek Link	-	335	14	Benign
			12	337	Malignant
		Correlation	338	11	Benign
			11	338	Malignant
Naïve Bayes	Original	-	343	14	Benign
			28	184	Malignant
		Correlation	343	14	Benign
			20	192	Malignant
	Smote-Tomek Link	-	339	10	Benign
			26	323	Malignant
		Correlation	335	14	Benign
			21	328	Malignant

Table 4 Experimental results of SVM and Naive Bayes methods

Methods	Data Sampling	Feature Selection	Accuracy	Sensitivity	Specificity	AUC
SVM	Original	-	94.00%	90.00%	96.40%	93.20%
		Correlation	95.00%	92.50%	96.40%	94.40%
	Smote-Tomek Link	-	96.30%	96.60%	95.90%	96.30%
		Correlation	**96.80%**	**96.80%**	**96.80%**	**96.80%**
Naïve Bayes	Original	-	92.60%	86.80%	96.10%	91.40%
		Correlation	94.00%	90.60%	96.10%	93.30%
	Smote-Tomek Link	-	94.80%	92.60%	97.10%	94.80%
		Correlation	95.00%	94.00%	96.00%	95.00%

4. Discussion

On average, the SVM method is better than the Naïve Bayes method in the aspect of performance metrics used in breast cancer disease prediction based on test results. The results of this study need to be compared with several previous studies referred to in order to see the level of contribution of the proposed method (See Table 5). Based on Table 5, the proposed method gets a better accuracy rate than previous research, with an accuracy of 96.80%. In general, the use of data sampling and feature selection approaches can improve the accuracy of the classification method [31-34].

Table 5 Comparison of the results of this research with previous research

No	Author	Methods	Scope of Study	Accuracy	Recall	F1 score	Precision
1	Assegie [35]	K-Nearest Neighbor	Breast Cancer	94.35%	-	-	-
2	Kurian and Jyothi [36]	Decision Tree		94.03%	94.00%	94.00%	94.00%
3	Alfian et al. [37]	SVM		80.23%	78.57%	-	82.71%
4	Iparraguirre-Villanueva et al. [9]	SVM		93.00%	93.00%	93.00%	93.00%
5	Imran et al. [38]	Random Forest		96.00%	96.00%	96.00%	96.00%
6	Enriko et al. [39]	KNN		77.98%	-	-	-
7	Anklesaria [40]	SVM		95.80%	-	-	-
8	Telsang and Hegde [41]	SVM		96.20%	-	-	-
9	Rabiei et al. [42]	Random Forest		80.00%	95.00%	-	-
10	**The Proposed Research**	**SVM + Smote-Tomek Link + Correlation**		**96.80%**	**96.80%**	**96.80%**	**96.80%**

This research will also be evaluated using a statistical approach using R square, R adjusted value, and SE to evaluate the performance of classification method used (See in Table 6). Based on Table 6, the use of the Smote-Tomek Link approach and correlation-based feature selection makes the machine learning method's performance more stable and prevents overfitting. Where the difference in values between R Square and R Adjusted in the SVM and Naive Bayes methods is exceedingly small using the Smote-Tomek Link data sampling approach and correlation-based feature selection. The SVM method obtained an R Square of 83.50% and an Adjusted R of 80.40% with a difference of 3.10%. Meanwhile, the Naïve Bayes method obtained an R Square of 75.10% and Adjusted R of 70.30% with a difference of 4.80%.

Table 6 Performance evaluation of classification methods based on R square, R adjusted, and SE

Methods	Data Sampling	Feature Selection	R square	R adjusted	SE
SVM	Original	-	81.60%	60.30%	19.00%
		Correlation	79.00%	74.00%	20.00%
	Smote-Tomek Link	-	86.10%	75.30%	17.80%
		Correlation	83.50%	80.40%	18.30%
Naïve Bayes	Original	-	73.00%	41.50%	23.30%
		Correlation	73.00%	66.60%	24.00%
	Smote-Tomek Link	-	72.70%	51.50%	23.90%
		Correlation	75.10%	70.30%	24.10%

5. Conclusions

This research proposes a framework for classification of imbalanced data in breast cancer data with Smote-Tomek Link and correlation-based feature selection. The Smote-Tomek Link algorithm is used to balance data and Correlation is used to select relevant attributes in breast cancer data. The proposed approach for resolving imbalanced data and selecting attributes in breast cancer data uses accuracy, sensitivity and AUC testing. The results showed that the performance of the classification method increased. The SVM method achieved an accuracy of 96.80%, sensitivity 96.80%, and AUC 96.80%. Meanwhile, the Naïve Bayes method achieved an accuracy of 95.00%, sensitivity of 94.00%, and AUC of 95.00%. The future research can use an ensemble learning approach to improve the performance of classification methods for breast cancer disease prediction. Not only that, but it is also necessary to conduct attribute correlation analysis using canonical correlation analysis.

6. Acknowledgements

Thank you to the Directorate of Research and Community Service of the Ministry of Education and Culture of the Republic of Indonesia for funding the Beginner Lecturer Research Grant Year 2023 so that this article can be published.

7. References

[1] Gautama W. Breast cancer in Indonesia in 2022 : 30 years of marching in place. Indones J Cancer. 2022;16(1):1-2.

[2] Marfianti E. Peningkatan Pengetahuan Kanker Payudara dan Ketrampilan Periksa Payudara Sendiri (SADARI) untuk Deteksi Dini Kanker Payudara di Semutan Jatimulyo Dlingo. J Abdimas Madani dan Lestari. 2021;3(1):25-31. (In Indonesian)

[3] Han L, Yin Z. A hybrid breast cancer classification algorithm based on meta-learning and artificial neural networks. Front Oncol. 2022;12:1-9.

[4] Alafeef M, Srivastava I, Pan D. Machine learning for precision breast cancer diagnosis and prediction of the nanoparticle cellular internalization. ACS Sens. 2020;5(6):1689-98.

[5] Behravan H, Hartikainen JM, Tengström M, Kosma VM, Mannermaa A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. Sci Rep. 2020;10(1):1-16.

[6] Nemade V, Fegade V. Machine learning techniques for breast cancer prediction. Procedia Comput Sci. 2023;218:1314-20.

[7] Saleh H, Abd-el ghany SF, Alyami H, Alosaimi W. Predicting breast cancer based on optimized deep learning approach. Comput Intell Neurosci. 2022;2022:1-11.

[8] Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Comput Sci. 2021;191:487-92.

[9] Iparraguirre-Villanueva O, Epifanía-Huerta A, Torres-Ceclén C, Ruiz-Alvarado J, Cabanillas-Carbonell M. Breast cancer prediction using machine learning models. Int J Adv Comput Sci Appl. 2023;14(2):610-20.

[10] Chen H, Wang N, Du X, Mei K, Zhou Y, Cai G. Classification prediction of breast cancer based on machine learning. Comput Intell Neurosci. 2023;2023:1-9.

[11] Juarto B. Breast cancer classification using outlier detection and variance inflation factor. Eng Math Comput Sci J. 2023;5(1):17-23.

[12] Hasan R, Shafi ASM. Feature selection based breast cancer prediction. Int J Image Graph Signal Process. 2023;15(2):13-23.

[13] Dehdar S, Salimifard K, Mohammadi R, Marzban M, Saadatmand S, Fararouei M, et al. Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. Front Oncol. 2023;13:1-10.

[14] Singh AP, Agrawal S. Accuracy prediction on detection of breast cancer using machine learning classifiers. 14th International Conference on Computational Intelligence and Communication Networks (CICN); 2022 Dec 4-6; Al-Khobar, Saudi Arabia. USA: IEEE; 2022. p. 401-5.

[15] Sunardi S, Yudhana A, Windra Putri AR. Mass classification of breast cancer using CNN and faster R-CNN model comparison. KINETIK. 2022;7(3):243-50.

[16] Chowanda A. Exploring the best parameters of deep learning for breast cancer classification. CommIT J. 2022;16(2):143-8.

[17] Aslam MA, Aslam, Cui D. Breast cancer classification using deep convolutional neural network. J Phys Conf Ser. 2020;1584:1-10.

[18] Nurtiyasari D, Abdurakhman A, Hilmi MR. The application of deep neural network for breast cancer classification. J Sains Dasar. 2018;7(1):1-4.

[19] Jabeen K, Khan MA, Balili J, Alhaisoni M, Almujally NA, Alrashidi H, et al. BC^{2}NetRF: Breast cancer classification from mammogram images using enhanced deep learning features and features selection. Diagnostics. 2023;13(7):1-22.

[20] Hairani H, Anggrawan A, Priyanto D. Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link. Int J Informatics Vis. 2023;7(1):258-64.

[21] Swana EF, Doorsamy W, Bokoro P. Tomek Link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors. 2022;22(9):1-21.

[22] Yang H, Li M. Software defect prediction based on SMOTE-Tomek and XGBoost. In: Pan L, Cui Z, Cai J, Li L, editors. International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2021). Communications in Computer and Information Science, vol 1566. Singapore: Springer; 2022. p. 12-31.

[23] Hairani H, Innuddin M, Rahardi M. Accuracy enhancement of correlated naive bayes method by using correlation feature selection (CFS) for health data classification. 2020 3rd International Conference on Information and Communications Technology (ICOIACT); 2020 Nov 24-25; Yogyakarta, Indonesia. USA: IEEE; 2020. p. 51-5.

[24] Tasnim F, Habiba SU. A comparative study on heart disease prediction using data mining techniques and feature selection. 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST); 2021 Jan 5-7; Dhaka, Bangladesh. USA: IEEE; 2021. p. 338-41.

[25] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE : Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.

[26] Tomek I. Two modifications of CNN. IEEE Trans Syst Man Cybern. 1976:769-72.

[27] Anggrawan A, Hairani H, Satria C. Improving SVM classification performance on unbalanced student graduation time data using SMOTE. Int J Inf Educ Technol. 2023;13(2):289-95.

[28] Rezvani S, Wang X. A broad review on class imbalance learning techniques. Appl Soft Comput. 2023;143:110415.

[29] Blessie EC, Karthikeyan E. Sigmis: A feature selection algorithm using correlation based method. J Algorithm Comput Technol. 2012;6(3):385-94.

[30] Hairani H, Priyanto D. A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data. Int J Adv Comput Sci Appl. 2023;14(8):585-90.

[31] Sun Y, Que H, Cai Q, Zhao J, Li J, Kong Z, et al. Borderline SMOTE algorithm and feature selection-based network anomalies detection strategy. Energies. 2022;15(13):1-13.

[32] Ramos-Pérez I, Arnaiz-González Á, Rodríguez JJ, García-osorio C. When is resampling beneficial for feature selection with imbalanced wide data ?. Expert Syst Appl. 2022;188:1-12.

[33] Nakkaş BN. Feature selection and SMOTE based recommendation for Parkinson's imbalanced dataset prediction problem. 2022 30th Signal Processing and Communications Applications Conference (SIU); 2022 May 15-18; Safranbolu, Turkey. USA: IEEE; 2022. p. 1-4.

[34] Sreejith S, Khanna Nehemiah H, Kannan A. Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. Comput Biol Med. 2020;126:103991.

[35] Assegie TA. An optimized K-Nearest neighbor based breast cancer detection. J Robot Control. 2021;2(3):115-8.

[36] Kurian B, Jyothi VL. Breast cancer prediction using an optimal machine learning technique for next generation sequences. Concurr Eng Res Appl. 2021;29(1):49-57.

[37] Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FTD, Widodo T, et al. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. Computers. 2022;11(9):1-14.

[38] Imran B, Hambali H, Subki A, Zaeniah Z, Yani A, Alfian MR. Data mining using random forest, naïve bayes, and adaboost models for prediction and classification of benign and malignant breast cancer. J Pilar Nusa Mandiri. 2022;18(1):37-46.

[39] Enriko IKA, Melinda M, Sulyani AC, Astawa IGB. Breast cancer recurrence prediction system using k-nearest neighbor, naïve-bayes, and support vector machine algorithm. J Infotel. 2021;13(4):185-8.

[40] Anklesaria S, Maheshwari U, Lele R, Verma P. Breast cancer prediction using optimized machine learning classifiers and data balancing techniques. 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA); 2022 Aug 26-27; Pune, India. USA: IEEE; 2022. p. 1-7.

[41] Telsang VA, Hegde K. Breast cancer prediction analysis using machine learning algorithms. 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4); 2020 Dec 17-18; Bangalore, India. USA: IEEE; 2020. p. 1-5.

[42] Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of breast cancer using machine learning approaches. J Biomed Phys Eng. 2022;12(3):297-308.