SMOTE Variants and Random Forest Method: A Comprehensive Approach to Breast Cancer Classification

Baiq Candra Herawati
Universitas Bumigora, Mataram, Indonesia

Hairani Hairani
Universitas Bumigora, Mataram, Indonesia

Juvinal Ximenes Guterres
Universidade Oriental Timur Lorosa'e, Timor Leste

Abstract: This research focused on using machine learning methods for breast cancer diagnosis, considering that breast cancer is the scariest disease for women because it can cause mortality. Not only that, but there is also an increase in breast cancer death rates in women yearly. Early prediction is the right solution to increase life expectancy and reduce mortality rates caused by breast cancer. However, breast cancer data has a problem, namely that the data is imbalanced, which harms the performance of the machine learning method itself. In the data, breast cancer had a Benign class (357 instances) more than the Malignant class (212 instances). Therefore, this study aimed to solve the problem of imbalanced data using the Smote variants and Random Forest approaches in breast cancer classification. The results of this study showed that the Smote approach with Random Forest had the best performance compared to Borderline Smote and Random Forest in the case of breast cancer data classification, where Smote with Random Forest produced an accuracy of 97.3%, sensitivity of 96.9%, and specificity of 97.8%. In comparison, Borderline Smote with Random Forest produced an accuracy of 96.4%, sensitivity of 95.6%, and specificity of 96.9%. The results of this study can contribute to predicting breast cancer using the proposed method, because it has been proven to have high accuracy.

Keywords: Breast Cancer Prediction, Machine Learning in Health, Random Forest, Smote Variants

Correspondents Author:
Hairani Hairani, Department of Computer Science, Universitas Bumigora, Indonesia
Email: Hairani@universitasbumigora.ac.id

Received October 30, 2023; Revised November 12, 2023; Accepted December 8, 2023; Published March 1, 2024

Introduction

In the current decade, machine learning in health has been widely applied. This is influenced by the increasingly massive health data (Dhillon & Singh, 2019), (Javaid et al., 2022). Examples of the application of machine learning in the field of health are the diagnosis of heart disease (Pattekari, S.A.; Parveen, 2012), hepatitis (Wang et al., 2017), ENT (Dirgantara & Hairani, 2021), diabetes (R. et al., 2022), and breast cancer (Rajendran et al., 2020). This research focuses on using machine learning methods to diagnose breast cancer, considering that breast cancer is the most frightening disease for women because it can cause mortality (Momenimovahed, 2019). Globally, there is an increase in breast cancer death rates in women every year (Azamjah et al., 2019). Early prediction is the right solution to increase life expectancy and reduce the death rate caused by breast cancer (Barrios, 2022). However, breast cancer data has a problem, namely unbalanced data (Gupta et al., 2021; Susilo & Sugiharti, 2021), so it can negatively affect the performance of the machine learning method itself (Azhar et al., 2022; Rezvani & Wang, 2023). Breast cancer data has more Benign classes (357 instances) than Malignant classes (212 instances), where the machine learning method recognizes the Benign class more than the Malignant class. In other words, the machine learning method can predict the Malignant class as the Benign class because fewer classes exist.

Some previous studies that classified breast cancer using various approaches, such as research (Jabbar et al., 2022), compare machine learning methods for breast cancer prediction. The results of their research obtained the K-Nearest Neighbors method to get the best accuracy compared to the Naïve Bayes and Decision Tree methods by 96%. Research (Achmad, 2022) conducts breast cancer predictions using the logistic linear method with training data accuracy of 76% and test data of 83%. Research (Andryan et al., 2022) compares machine learning methods, namely XGBoost and Support Vector Machine (SVM) for breast cancer prediction. The results of their research are that the XGBoost method obtained an accuracy of 95% and an SVM of 90%. Research (Muntiari & Hanif, 2022) compares several machine-learning methods for breast cancer prediction. The results of their research obtained Naïve Bayes, Decision Tree, Logistic Regression, and k-NN methods get the same accuracy of 95%.

Research (Resmiati & Arifin, 2021) proposes Backward Elimination feature selection to improve the SVM method for breast cancer classification. The results show that using the Backward Elimination feature selection significantly improves SVM performance with an accuracy of 95% on breast cancer classification. Research (Hairani et al., 2022) uses the C4.5 method for the classification of the nutritional status of toddlers with an accuracy of 95%. Research (Astuti et al., 2021) suggests forward selection feature selection to improve the Naïve

Bayes method for breast cancer classification. The results show that forward selection feature selection can significantly improve Naïve Bayes performance with an accuracy of 96% in breast cancer classification. Research (Juarto, 2023) compares machine learning methods for breast cancer prediction. The results show that the Random Forest method is more accurate than the SVM, Gradient Boosting, KNN, and Logistic Regression methods. Research (Michael Lauw et al., 2023) proposes the Smote method to improve the Random Forest method for lung cancer prediction by dividing training and testing data using 10-fold cross-validation. Based on the results of his research, the Smote method can improve the performance of the Random Forest method, such as 94.1% accuracy, 94.5% sensitivity, and 93.7% specificity.

Several previous studies have used diverse approaches to predict breast cancer, but there are still shortcomings that can be overcome. The level of accuracy needs to be increased and solve the data imbalance problem using the SMOTE variant, so that it can increase the accuracy of the classification method used. So, this research proposes an oversampling approach using Smote variants and Random Forest. The Smote variant approach is used to balance breast cancer disease data, after which classification is done using the Random Forest method. Therefore, this study aims to solve the problem of unbalanced data with the Smote variant approach and Random Forest in classifying breast cancer diseases. The Smote variant approach is expected to improve the performance of the Random Forest method based on accuracy, sensitivity, and specificity.

Research Method

This research comprises several stages, as shown in Figure 1.

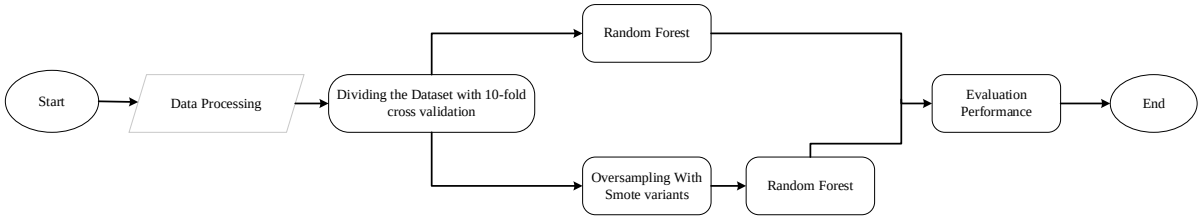

Figure 1 Research Flow

Based on Figure 1, the research started with collecting breast cancer datasets from the Kaggle website. The second process was to perform data processing. The data processing technique used was an oversampling technique to balance the data on breast cancer. The oversampling techniques used were Smote and Borderline Smote. The third process was dividing training data and testing data using 10-fold cross-validation. The fourth process was implementing the Random Forest method for breast cancer classification. After classification, the next process

was to evaluate the performance results of the Random Forest method based on accuracy, sensitivity, and specificity. The calculation formula for accuracy, sensitivity, and specificity used Equation 1, Equation 2, and Equation 3 (Anggrawan et al., 2023).

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \quad (1)$$

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

$$Specificity = \frac{True\ Negative}{True\ Negative+False\ Positive} \quad (3)$$

Result and Discussion

This section explains the research results achieved based on the research stages Figure 1. This research used breast cancer datasets obtained from Kaggle. The number of instances in the breast cancer dataset is 569 instances and 31 attributes (See Table 1). The dataset had a class imbalance issue that could affect the performance of the classification method. The imbalance in question was the number of Benign classes (357 instances) more than the number of Malignant classes (212 instances), so it was feared that the classification method would more easily predict the Benign class compared to the Malignant class. Therefore, the researcher proposed solving the problem with an oversampling approach—the oversampling technique aimed to add the Malignant class so that the number equals the Benign class.

The oversampling techniques used were Smote and Borderline Smote. Smote created artificial data in the minority class (Malignant) by linear interpolation between minority classes based on k nearest neighbors (Chawla et al., 2002). At the same time, borderline smote added the number of minority classes by creating artificial data in the borderline area or the boundary separating the majority and minority classes (Revathi & Ramyachitra, 2021). The oversampling results of the Smote and Borderline Smote methods on the breast cancer dataset can be seen in Figure 2.

Table 1. Sample Breast Cancer Disease Data

No	Radius Mean	Texture Mean	...	Symmetry Worst	Fractal Dimension Worst	diagnosis
1	17.99	10.38	...	0.4601	0.1180	Benign
...	....	.....	...	.....	.....	.....
568	20.6	29.33	...	0.4087	0.124	Benign

569	7.76	24.54	...	0.2871	0.07039	Malignant

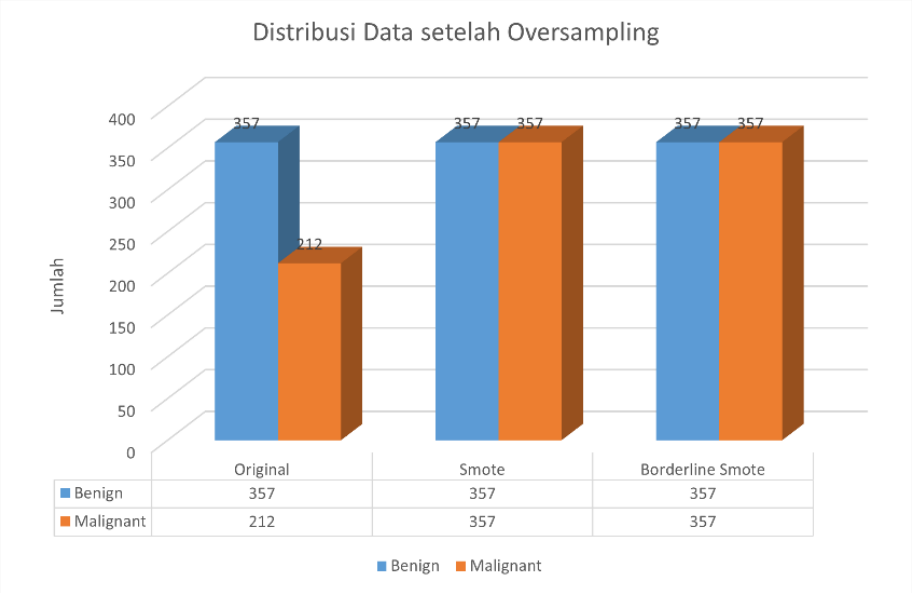

Figure 2. Distribution of Breast Cancer Disease Data

The data that had been balanced was then classified using the Random Forest method, which first divided by data training and testing. The training and testing data division used 10-fold cross-validation, where each fold or group could be used as training and testing data alternately. Datasets classified by the Random Forest method were tested based on accuracy, sensitivity, and specificity obtained from the confusion matrix table. In the classification of breast cancer data, the Random Forest method is used by tuning hyper parameters using several parameters as shown in Table 2.

Table 2. Hyper parameters of the Random Forest Method for Classifying Breast Cancer

Hyper parameter	Value
Criterion	Entropy
	Gini
Min_samples_split	2
N_estimators	100

The results of the Random Forest method confusion matrix on the original data are shown in Figure 3, the results of the Random Forest method confusion matrix on the Smote result data are shown in Figure 4, and the results of the Random Forest method confusion matrix on the Borderline Smote result data are shown in Figure 5. In Figure 4, the Random Forest method with the original data can classify the Benign class as many as 348 instances out of a total of

357 instances, while the Malignant class correctly classified as many as 198 instances out of a total of 212 instances.

In Figure 5, the Random Forest method with Smote result data can classify the Benign class as many as 349 instances out of 357 instances, while the Malignant class correctly classified as many as 346 instances out of 349 instances. In Figure 6, the Random Forest method with Smote result data can classify the Benign class as many as 346 instances out of 357 instances, while the Malignant class correctly classified as many as 342 instances out of 349 instances.

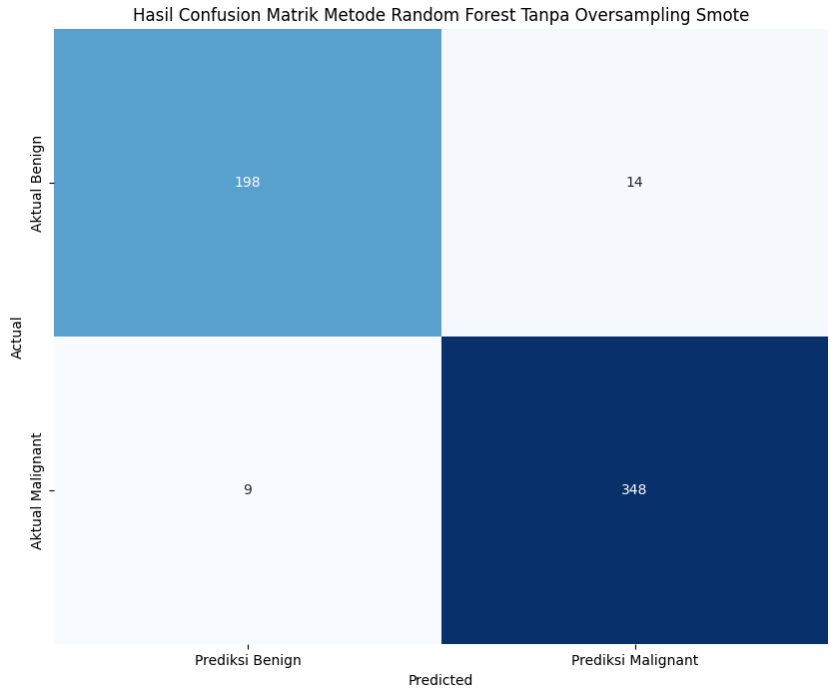

Figure 3 Classification Results of Random Forest Method on Data without Oversampling

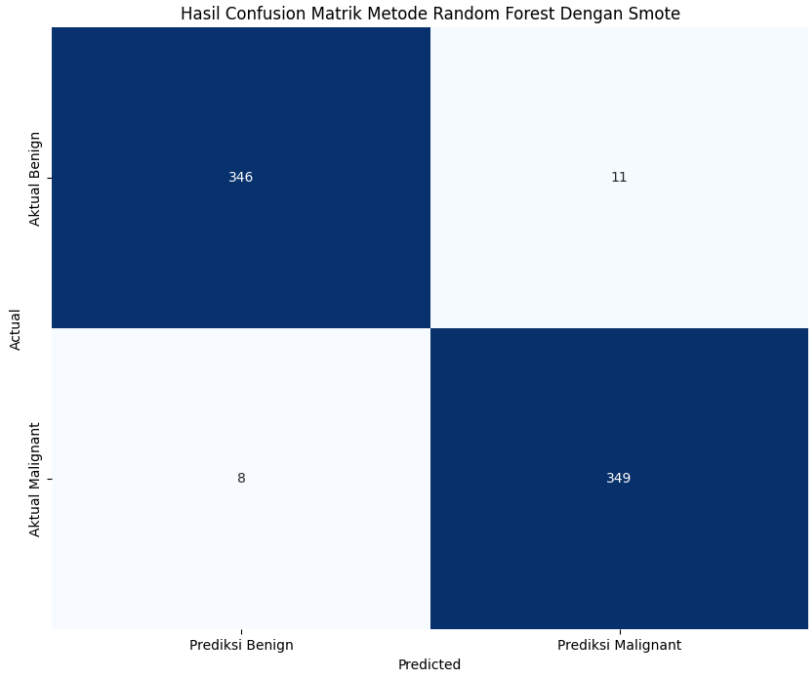

Figure 4 Classification Results of Random Forest Method with Smote Data

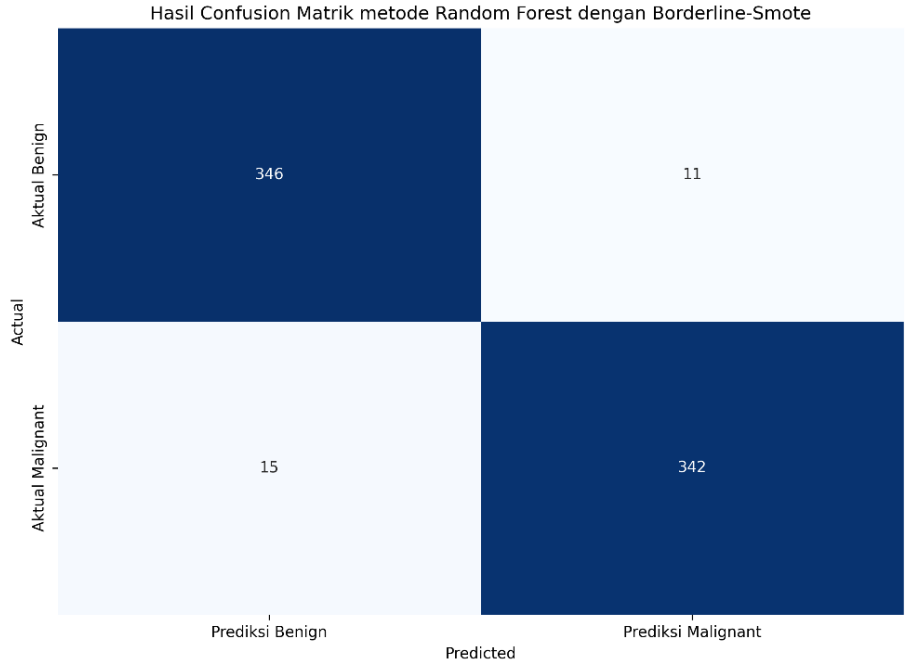

Figure 5. Classification Results of Random Forest Method with Borderline Smote Data

Figure 6 shows the performance of the Random Forest method on breast cancer classification with Smote variant oversampling and without oversampling. The Random Forest method with no oversampling resulted in an accuracy of 95.9%, sensitivity of 93.4%, and specificity of 97.5%. The Random Forest method with Smote resulted in an accuracy of 97.3%, sensitivity of

96.9%, and specificity of 97.8%. In comparison, the Random Forest method with Borderline Smote produced an accuracy of 96.4%, a sensitivity of 95.6%, and a specificity of 96.9%.

The test results showed that using Smote variants in data balancing positively impacted the performance of the Random Forest method in breast cancer classification. The Smote method produced the best performance with Random Forest compared to Borderline Smote. On average, the Smote variant with Random Forest performed better than the data without oversampling. This is in line with research, which states that the use of oversampling to solve unbalanced data problems can improve the performance of the method used (Hairani et al., 2020)(Hairani & Priyanto, 2023)(Hairani et al., 2023).

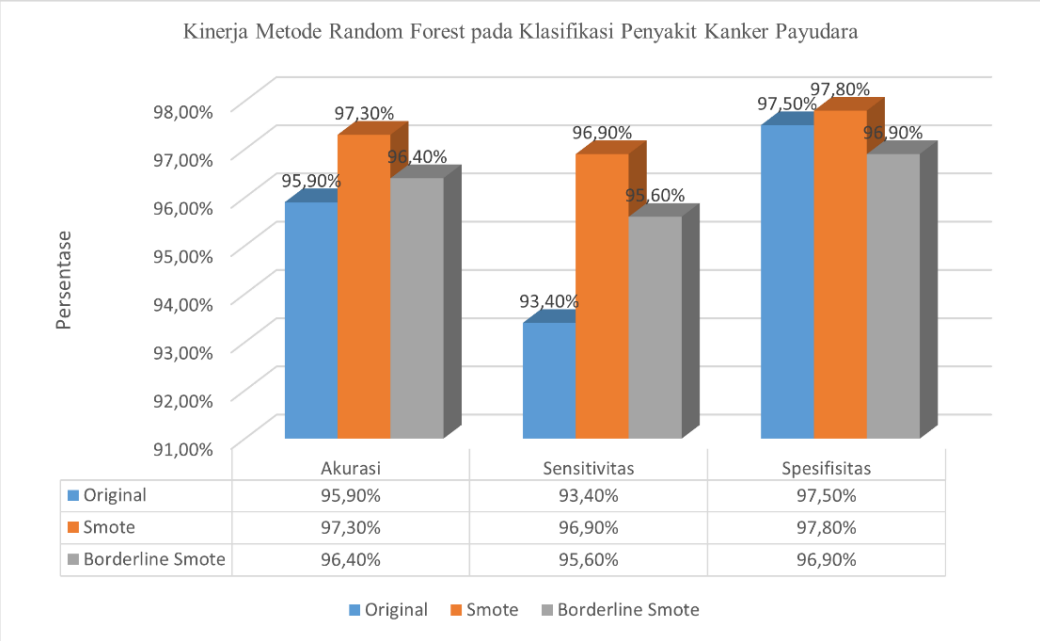

Figure 6. Performance Results of Random Forest Method on Breast Cancer Clarification

The results of this study are in line with research (Jabbar et al., 2022; (Andryan et al., 2022); Muntiari & Hanif, 2022) which on average obtained 95% accuracy in breast cancer classification. However, the level of accuracy can be increased by resolving data imbalances using the Smote variant to increase the accuracy of the Random Forest method. The Smote method with Random Forest can increase accuracy by 1.4% and sensitivity by 3.5% compared to without sampling. The increase in accuracy and sensitivity is not significant because Smote results in creating samples for minority classes that still contain noise. In conclusion, future research should consider using hybrid sampling methods in dealing with data imbalance in breast cancer to reduce data noise, so that the Random Forest method can produce significant performance improvements (Hairani & Priyanto, 2023; Khushi et al., 2021; Swana et al., 2022)

Conclusions

This study proposed a Smote variant approach and Random Forest method for breast cancer classification. Smote variants used were Smote and Borderline Smote methods to balance breast cancer classes. Before oversampling, classes in breast cancer data had an unbalanced data issue: the Benign class, as many as 357 instances, and the Malignant class, as many as 212 instances. It became balanced after oversampling with the Smote variant; the Benign and Malignant classes each had 357 instances. Using the Smote variant in data balancing positively impacted the performance of the Random Forest method in breast cancer classification, where the Smote variant with Random Forest performed better than the data without oversampling. The Smote with Random Forest method produced the best performance compared to Borderline Smote, where the accuracy was 97.30%, sensitivity was 96.90%, and specificity was 97.8%. Future research can use other Smote variants, such as k-means Smote and Adasyin, to solve the unbalanced breast cancer data problem.

References

Abdul Jabbar, M., Hasmin, E., Susanto, C., Musu, W., & Artikel, I. (2022). Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara. *CSRID Journal*, *14*(3), 258–270. https://www.doi.org/10.22303/csrid.14.3.2022.258-270

Achmad, A. D. (2022). Klasifikasi Breast Cancer Menggunakan Metode Logistic Regression. *Jtriste*, *9*(1), 143–148.

Andryan, M. R., Fajri, M., & Sulistyowati, N. (2022). Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosis Penyakit Kanker Payudara. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(1), 1–5. https://doi.org/10.26798/jiko.v6i1.500

Anggrawan, A., Hairani, H., & Satria, C. (2023). Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE. *International Journal of Information and Education Technology*, *13*(2), 289–295. https://doi.org/10.18178/ijiet.2023.13.2.1806

Astuti, L. W., Saluza, I., Faradilla, F., & Alie, M. F. (2021). Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes. *Jurnal Ilmiah Informatika Global*, *11*(2), 63–67. https://doi.org/10.36982/jiig.v11i2.1235

Azamjah, N., Soltan-Zadeh, Y., & Zayeri, F. (2019). Global trend of breast cancer mortality

rate: A 25-year study. *Asian Pacific Journal of Cancer Prevention*, *20*(7), 2015–2020. https://doi.org/10.31557/APJCP.2019.20.7.2015

Azhar, N. A., Mohd Pozi, M. S., Mohamed Din, A., & Jatowt, A. (2022). An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2022.3179381

Barrios, C. H. (2022). Global challenges in breast cancer detection and treatment. *The Breast*, *62*(S1), S3–S6. https://doi.org/10.1016/j.breast.2022.02.003

Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling TEchnique. *Journal of Artificial Intelligence Research*, *16*, 341–378.

Dhillon, A., & Singh, A. (2019). Biology and Today's World Machine Learning in Healthcare Data Analysis: A Survey. *J. Biol. Today's World*, *8*(2), 1–10. https://doi.org/10.15412/J.JBTW.01070206

Dirgantara, B., & Hairani, H. (2021). Sistem Pakar Diagnosa Penyakit THT Menggunakan Inferensi Forward Chaining dan Metode Certainty Factor. *Jurnal Bumigora Information Technology (BITe)*, *3*(1), 1–8. https://doi.org/10.30812/bite.v3i1.1241

Gupta, R., Bhargava, R., & Jayabalan, M. (2021). Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models. *2021 14th International Conference on Developments in ESystems Engineering (DeSE)*, 162–167. https://doi.org/10.1109/DeSE54285.2021.9719398

Hairani, H., Anggrawan, A., & Priyanto, D. (2023). Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. *International Journal on Informatics Visualization*, *7*(1), 258–264.

Hairani, H., Nurhayati, L., & Innuddin, M. (2022). Web-Based Application for Toddler Nutrition Classification Using C4.5 Algorithm. *International Journal of Engineering and Computer Science Applications (IJECSA)*, *1*(2), 77–82. https://doi.org/10.30812/ijecsa.v1i2.2387

Hairani, H., & Priyanto, D. (2023). A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data. *International Journal of Advanced Computer Science and Application*, *14*(8), 585–590.

Hairani, H., Suweleh, A. S., & Susilowaty, D. (2020). Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *20*(1), 109–116.

https://doi.org/10.30812/matrik.v20i1.846

Javaid, M., Haleem, A., Pratap Singh, R., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*(June), 58–73. https://doi.org/10.1016/j.ijin.2022.05.002

Juarto, B. (2023). Breast Cancer Classification Using Outlier Detection and Variance Inflation Factor. *Engineering, MAthematics and Computer Science (EMACS) Journal*, *5*(1), 17–23. https://doi.org/10.21512/emacsjournal.v5i1.9223

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, *9*, 109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399

Michael Lauw, C., Hairani, H., Saifuddin, I., Ximenes Guterres, J., Maariful Huda, M., & Mayadi, M. (2023). Combination of Smote and Random Forest Methods for Lung Cancer Classification. *International Journal of Engineering and Computer Science Applications (IJECSA)*, *2*(2), 59–64. https://doi.org/10.30812/ijecsa.v2i2.3333

Momenimovahed, Z. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. *Dovepress*, *11*(April), 151–164.

Muntiari, N. R., & Hanif, K. H. (2022). Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Ilmu Komputer Dan Teknologi*, *3*(1), 1–6. https://doi.org/10.35960/ikomti.v3i1.766

Pattekari, S.A.; Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*, *3*(3), 290–294.

R., S., M., S., Hasan, M. K., Saeed, R. A., Alsuhibany, S. A., & Abdel-Khalek, S. (2022). An Empirical Model to Predict the Diabetic Positive Using Stacked Ensemble Approach. *Frontiers in Public Health*, *9*(January), 1–13. https://doi.org/10.3389/fpubh.2021.792124

Rajendran, K., Jayabalan, M., & Thiruchelvam, V. (2020). Predicting breast cancer via supervised machine learning methods on class imbalanced data. *International Journal of Advanced Computer Science and Applications*, *11*(8), 54–63. https://doi.org/10.14569/IJACSA.2020.0110808

Resmiati, R., & Arifin, T. (2021). Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination. *SISTEMASI*, *10*(2), 381–393. https://doi.org/10.32520/stmsi.v10i2.1238

Revathi, M., & Ramyachitra, D. (2021). A Modified Borderline Smote with Noise Reduction in Imbalanced Datasets. *Wireless Personal Communications*, *121*(3), 1659–1680. https://doi.org/10.1007/s11277-021-08690-y

Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, *143*, 110415. https://doi.org/10.1016/j.asoc.2023.110415

Susilo, A. B., & Sugiharti, E. (2021). Accuracy Enhancement in Early Detection of Breast Cancer on Mammogram Images with Convolutional Neural Network (CNN) Methods using Data Augmentation and Transfer Learning. *Journal of Advances in Information System and Technology*, *3*(1), 9–16.

Swana, E. F., Doorsamy, W., & Bokoro, P. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors*, *22*(9), 1–21. https://doi.org/10.3390/s22093246

Wang, H., Liu, Y., & Huang, W. (2017). The application of feature selection in Hepatitis B virus reactivation. *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, 893–896. https://doi.org/10.1109/ICBDA.2017.8078767