Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Breast Cancer Diagnosis Utilizing Artificial Neural Network (ANN) Algorithm for Integrating Multi-Omics Data and Clinical Features

Rofik[1*] , Fani Artiyani[2], Dwika Ananda Agustina Pertiwi[3]

[1,2]Department of Computer Science, Universitas Negeri Semarang, Indonesia

[3]Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Malaysia

DOI: https://doi.org/10.52465/joiser.v2i2.249

Received 22 November 2023; Accepted 22 April 2024; Available online 08 July 2024

Article Info

Keywords:

Breast cancer;
Diagnosis;
ANN;
Multi-omics data;
Clinical features

Abstracts

Breast cancer is one of the most common diseases affecting women worldwide, with a significant impact on patient's health and quality of life. Despite advances in medical technology and research, breast cancer diagnosis remains a challenge due to its complexity involving various biological and clinical factors. Several previous studies have focused on detecting this disease with optimal accuracy, but the selection of appropriate algorithms and methods is key to achieving this goal. This study aims to improve the accuracy of breast cancer diagnosis by using the ANN algorithm and data balancing method, SMOTE. This research uses Multi-Omic data and Clinical Features obtained in general from Kaggle. The research process is carried out in several stages, namely Data Collection, Preprocessing, Oversampling, Modeling, and Evaluation. This research successfully obtained an increase in accuracy, which was able to achieve an accuracy of 99.30%. This research shows that early detection of breast cancer with ANN algorithm and data balancing using SMOTE can improve accuracy performance in early detection of breast cancer. Given the use of data in this study is not too large, it is recommended for further research to use a larger dataset to validate the strength of the model that has been built on more varied data.

1. Introduction

Cancer is a complex disease that involves the relationship between genetic factors and environmental factors [1]. Cancer is categorized as one of the leading causes of death in humans worldwide [2]. According to statistics [3], in 2020 there were 19.3 million cancer patients and about 10 million patients died from cancer. Breast cancer is the disease that causes the second largest mortality

* ***Corresponding Author:***

Rofik,
Department of Computer Science,
Universitas Negeri Semarang
Semarang, Indonesia.
Email : rofikn4291@students.unnes.ac.id

in women [2], accounting for 30% of female cancers [4]. If considering breast cancer stages separately early-stage breast cancer or cancer in the early stages has a 5-year survival rate of 99% [5]. However, for breast cancer that spreads regionally, the survival rate decreases to 85% and 27% [5]. Currently, the diagnosis of breast cancer still relies on conventional methods, making it less accurate and specific to distinguish between malignant or benign lesions. Therefore, research continues to be conducted to develop more effective and accurate diagnosis methods.

Disease diagnosis is a complicated process of identifying the type of disease in the field of medicine [6], [7]. In recent years, there are various systems used for cancer diagnosis through medical imaging analysis [8] or omic data analysis [9]. Machine Learning [10], [11] is also used to integrate multi-omics data and clinical features [12], [13]. Integrating multi-omics data, whether genomic, proteomic, metabolomic, or epigenomic, can improve biomarker discovery by identifying examples of biological interactions and improving the prediction of clinical features [14], [15]. Technological advances in multi-omics approaches [16] can facilitate the diagnosis of breast cancer. Multi-omics-based Machine Learning technology has an important role in cancer diagnosis such as survival analysis, drug sensitivity response, and others [17], [18]. In recent years, several studies have been conducted to integrate omics data to predict clinical outcomes and improve cancer-related medical decisions [9], [19]. Given the limitations and urgency of breast cancer diagnosis, as well as the potential of a system capable of diagnosing a person through available data, this research focuses on further development in breast cancer diagnosis using Machine Learning.

Previous research that utilizes data to learn and then diagnose someone as a breast cancer patient or not includes research conducted by [20]. The research focuses on finding the best data composition for training and testing. The research used the holdout method and k-fold cross-validation. By using the SVM algorithm and with a holdout validation scheme with a ratio of 75%: to 25%, the study managed to achieve the greatest accuracy of 98.89%. However, this research was conducted using rapid-miner tools, so the sustainability of model implementation and development may be a concern in the context of wider and integrated system use. Research [21] was also conducted with a focus on classifying breast cancer diseases. This research focuses on comparing the performance of the results of the implementation of FFNN in machine learning and RBM in deep learning. Using 683 data, this research successfully showed the superiority of deep learning in classifying breast cancer diseases with an accuracy of 98.5401% (higher than machine learning). By looking at the potential use of deep learning, this research aims to develop breast cancer classification with deep learning algorithms and utilize more explored (different) data..

2. Method

The research was conducted sequentially, starting from the data collection stage which involved gathering information regarding multi-omics data and relevant clinical features from available sources. After that, a preprocessing stage was conducted where the data was prepared and processed to make it suitable for further processing. Next, an oversampling stage is performed to handle the class imbalance in the dataset, thus ensuring that the built model can learn well from each existing class.
Once the data is ready, a modeling stage is performed where the Machine Learning model is built and trained using the pre-processed data. Then the testing and evaluation of the model that has been built is carried out. The series of stages carried out in this research can be seen in Figure 1.

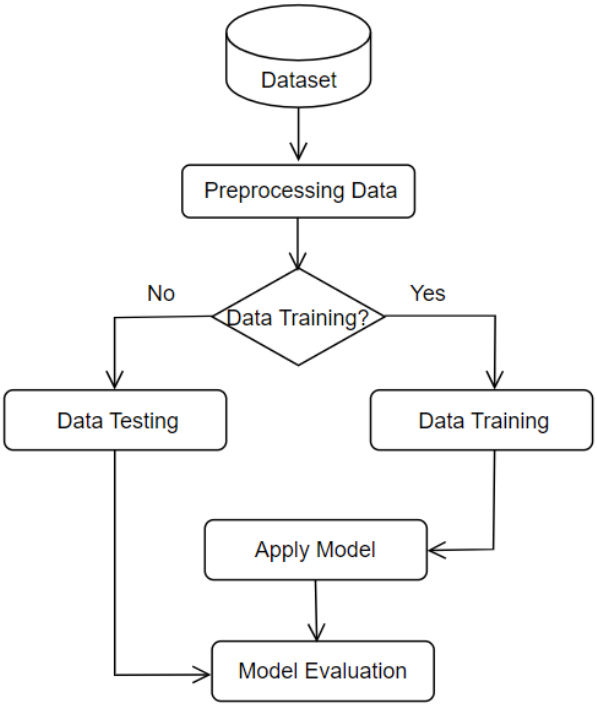


Figure 1. The methodology used for evaluation experimental

2.1 Data collection

The data collection stage is carried out by downloading it through a publicly accessible platform, namely Kaggle. The dataset can be accessed via the URL: https://www.kaggle.com/code/thebrownviking20/intro-to-keras-with-breast-cancer-data-ann/input. This dataset consists of 569 data records with 32 features. The features in the dataset include id, diagnosis, 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.

2.2 Data preprocessing

Data preprocessing is done to clean the data so that the data is ready when implemented in the modeling stage. At this stage, missing values and duplicate data were checked. However, no missing values or duplicate data were found. This stage also identified that there was a column that was not needed, namely id. So this feature was removed.

In checking the target, which is between the data class of individuals who are not breast cancer patients, and data of individuals with breast cancer, it was found that the data was not balanced between these classes. The amount of data from individuals who are not breast cancer patients is more than data from individuals with breast cancer. Which are 357 and 212 respectively. Data balancing is done to prevent the model from only favoring the majority class. This can result in the resulting model only being good at classifying the majority class, while the minority class is not well represented [22]. Data balancing is done using the Synthetic Minority Over-sampling Technique (SMOTE) method. SMOTE is one of the commonly used techniques to balance the dataset by creating a synthetic sample of the minority class so that the number is balanced with the majority class. This effort is made so that the model to be developed can be more effective in classifying both classes well and does not tend to favor one particular class. Therefore, this method also supports improving the quality and accuracy of the model in predicting whether someone has breast cancer or not.

2.3 Split Data

After the data between classes is balanced, split data is performed. Split data is done to divide the data into 2 main parts, namely training data (i.e. data used for the training process using deep learning models) and testing data, to test the models that have been built. The division of data between training data and testing data is 80% and 20%, respectively.

2.4 Apply Model

This research uses the Artificial Neural Network (ANN) algorithm, which is one of the deep learning algorithms. ANN, also known as a neural network, is a mathematical model consisting of a series of interconnected processing units (neurons). These neurons are organized in layers and are capable of learning complex patterns in data. The selection of the algorithm is based on seeing the potential for development from previous studies.

The way the ANN algorithm works is based on the principle that the network consists of interconnected neurons, organized in layers [23]. Each neuron receives an input, multiplies it by the appropriate weight, adds a bias, and applies it to the activation function. This process produces an output that is passed on to the next layer. During training, the network learns from the data by adjusting its weights and biases using the backpropagation algorithm, where the prediction error is compared to the expected value, and the gradient of the loss function over the weights and biases is calculated. The weights and biases are updated iteratively using optimization methods such as Stochastic Gradient Descent (SGD) to minimize the prediction error. This iterative process allows the network to learn from the data and improve its ability to learn complex patterns in the data, making it a powerful tool for predicting or classifying data.

2.5 Evaluation

Model evaluation is done to test the model that has been built with the dataset that has been prepared. This research uses a confusion matrix to evaluate the performance of breast cancer classification models using ANN. The confusion matrix consists of four main cells, including as in Table 1.

Table 1. Confusion matrix.

Classification		Class Classification Prediction Result	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Description:
True Positive (TP) : Positive cases classified as positive.
False Positive (FP) : Negative cases classified as positive.
True Negative (TN : Negative cases classified as negative.
False Negative (FN) : Positive cases classified as negative.

This research focuses on developing an accuracy performance metric. Accuracy is an evaluation metric that measures how well the model performs overall classification. Accuracy is expressed as a percentage of the number of correct predictions (TP and TN) compared to the total amount of data. The formula is:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

3. Results And Discussion

The research was carried out sequentially starting from the data collection stage from the Kaggle platform. Perform preprocessing, which includes separating features (X) and target variables (Y). Perform data standardization using StandardScaler. Then oversampling the data using SMOTE to overcome data imbalance. Divide the data into training data and testing data. Modeling is done using the ANN algorithm. An evaluation using confusion matrix, where this research focuses on improving the accuracy metric.

In the preprocessing stage, no missing values or duplicate data were found. Therefore, there is no need for data reduction or data addition, which causes the amount of data at this stage to remain intact. However, in checking the data distribution between the class of breast cancer patients and healthy individuals is shown in Figure 2.

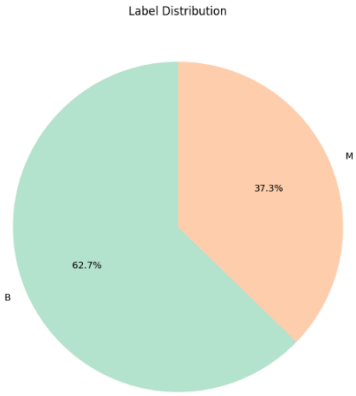
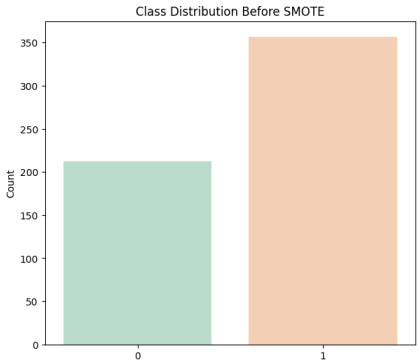
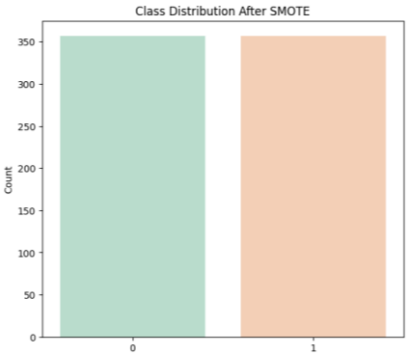


Figure 2. Class distribution in the dataset

It can be seen in Figure 2, that it turns out that the number of data of breast cancer sufferers is less than the data of people with conditions free from breast cancer. This imbalance of data classes needs to be addressed to prevent the model from classifying only the majority class. Therefore, SMOTE is applied to create new synthesized data from the minority class. Here in Figure 3, is an overview of the data from before and after SMOTE.



a. Class distribution before SMOTE

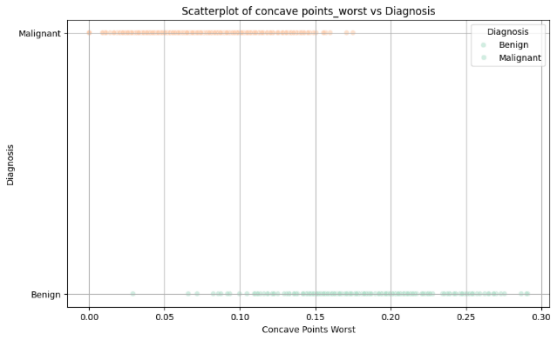


b. Class distribution after SMOTE

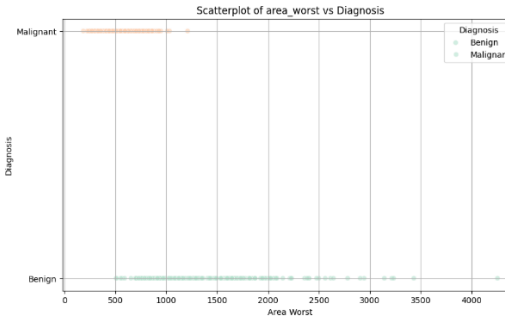
Figure 3a, 3b. Class distribution

The previous data between the classes of healthy individuals and breast cancer patients are 357 and 212 respectively, after SMOTE, each class is 357.

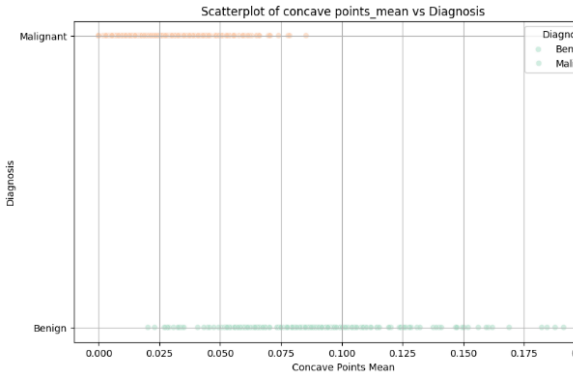
This research tries to see the relationship of several features in the dataset, as shown in Figure 4.



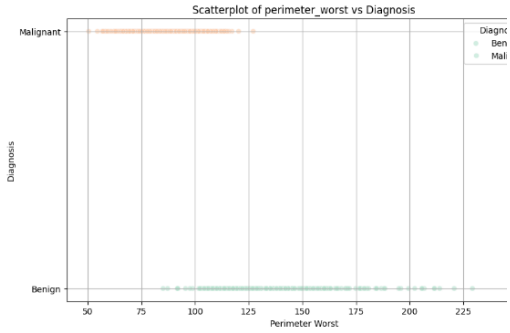
a. Relationship between concave points worst feature and diagnosis



b. Relationship between area worst feature and diagnosis



c. Relationship between concave points mean feature and diagnosis



d. Relationship between perimeter worst feature and diagnosis

Figure 4. a, b, c, d The relationship between features and their diagnoses.

Figure 4a. shows the relationship between the worst concave points feature and the diagnosis, where healthy people (not affected by breast cancer) tend to have larger worst concave points values. Whereas breast cancer patients have concave points with worst values mostly between 0 to 0.15 only. The same applies to the area worst (Figure 4b.) and perimeter worst (Figure 4d.) features, where healthy people (not affected by breast cancer) tend to have larger area worst and perimeter worst values, from 80 to 225. Whereas people with breast cancer have the worst area and worst parameter values mostly between 50 to 125 only. Figure 4d. also shows the same condition, where large values of concave points mean features tend to be owned by healthy people (not breast cancer patients) while breast cancer patients have concave points mean feature values mostly only between 0 to 0.075.

Evaluation of breast cancer classification models on datasets that have been prepared and cleaned is done using a confusion matrix. Where this method is also widely used to evaluate cases of disease classification in previous research studies. The confusion matrix table for breast cancer classification in this study can be seen in Figure 5.

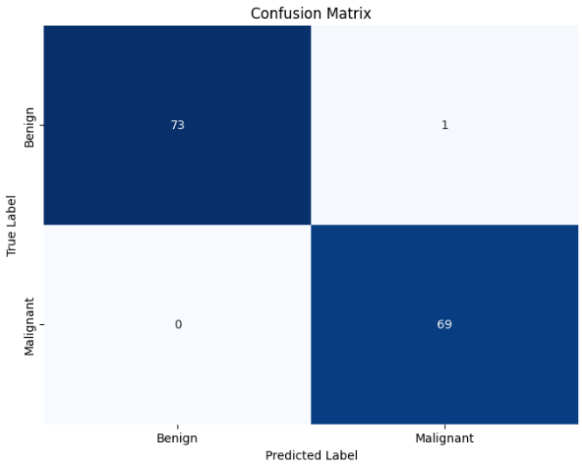


Figure 5. Testing with confusion matrix

Figure 5. Shows the number of correctly classified data from the data with the correct class. Incorrectly classified data, which is the correct case. Correctly classified data, which is the wrong case. And wrongly classified data, which is the wrong case. From these data and quantities, accuracy can be calculated according to the calculation formula. By using the ANN algorithm model, the dataset used, and all treatments in the dataset, this study managed to obtain the greatest accuracy of 99.30%. The accuracy shows that this research succeeded in producing very good performance, which was able to outperform the accuracy in previous studies.

4. Conclusion

This research utilizes multi-omics data and clinical features to develop a breast cancer prediction model. The ANN algorithm was chosen and implemented due to its significant potential in disease classification contexts, as observed in previous studies. Through data collection, preprocessing, oversampling using SMOTE, modeling, and evaluation stages, this research successfully demonstrates the effectiveness of the built ANN model in classifying individuals with and without breast cancer. Evaluation results indicate that the constructed model achieves a high level of accuracy, reaching 99.30%. However, due to the limited dataset used in this study, it is recommended for future research to explore larger and more diverse datasets to ensure a stronger and more reliable model, suitable for clinical practice.

References

[1] Z. He, J. Zhang, X. Yuan, and Y. Zhang, "Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods," *Front. Genet.*, vol. 11, no. January, pp. 1–12, 2021, doi: 10.3389/fgene.2020.632901.

[2] A. Tayanloo-Beik *et al.*, "OMICS insights into cancer histology; Metabolomics and proteomics approach," *Clin. Biochem.*, vol. 84, no. April, pp. 13–20, 2020, doi: 10.1016/j.clinbiochem.2020.06.008.

[3] A. C. Gamboa, A. Gronchi, and K. Cardona, "Soft-tissue sarcoma in adults: An update on the current state of histiotype-specific management in an era of personalized medicine," *CA. Cancer J. Clin.*, vol. 70, no. 3, pp. 200–229, 2020, doi: 10.3322/caac.21605.

[4] American Cancer Society, "American Cancer Society. Cancer Facts & Figures 2021. Atlanta: American Cancer Society; 2021." pp. 1–72, 2021.

[5] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–12, 2020, doi: 10.1186/s12911-020-01225-8.

[6] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 5, pp. 1–14, 2020, doi: 10.1007/s42979-020-00305-w.

[7] S. H. Park and K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," *Radiology*, vol. 286, no. 3, pp. 800–809, 2018, doi: 10.1148/radiol.2017171920.

[8] K. T. Chui *et al.*, "Transfer Learning-Based Multi-Scale Denoising Convolutional Neural Network for Prostate Cancer Detection," *Cancers (Basel).*, vol. 14, no. 15, 2022, doi: 10.3390/cancers14153687.

[9] I. Zenbout, A. Bouramoul, S. Meshoul, and M. Amrane, "Efficient Bioinspired Feature Selection and Machine Learning Based Framework Using Omics Data and Biological Knowledge Data Bases in Cancer Clinical Endpoint Prediction," *IEEE Access*, vol. 11, no. October 2022, pp. 2674–2699, 2023, doi: 10.1109/ACCESS.2023.3234294.

[10] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," *IETE J. Res.*, vol. 68, no. 4, pp. 2488–2507, 2022, doi: 10.1080/03772063.2020.1713916.

[11] M. R. Haque, M. M. Islam, H. Iqbal, M. S. Reza, and M. K. Hasan, "Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder," *Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018*, pp. 1–5, 2018, doi: 10.1109/IC4ME2.2018.8465658.

[12] C. Boeri *et al.*, "Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation," *Cancer Med.*, vol. 9, no. 9, pp. 3234–3243, 2020, doi: 10.1002/cam4.2811.

[13] S. Wang, S. Wang, and Z. Wang, "A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer," *Front. Med.*, vol. 9, 2023, doi: 10.3389/fmed.2022.1109365.

[14] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, "Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools," *Front. Oncol.*, vol. 10, no. June, 2020, doi: 10.3389/fonc.2020.01030.

[15] Y. H. Chuang *et al.*, "Convolutional neural network for human cancer types prediction by

integrating protein interaction networks and omics data," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021, doi: 10.1038/s41598-021-98814-y.

[16] E. Lin and H. Y. Lane, "Machine learning and systems genomics approaches for multi-omics data," *Biomark. Res.*, vol. 5, no. 1, pp. 1–6, 2017, doi: 10.1186/s40364-017-0082-y.

[17] P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, "Using machine learning approaches for multi-omics data analysis: A review," *Biotechnol. Adv.*, vol. 49, no. December 2020, p. 107739, 2021, doi: 10.1016/j.biotechadv.2021.107739.

[18] D. Leng *et al.*, "A benchmark study of deep learning-based multi-omics data fusion methods for cancer," *Genome Biol.*, vol. 23, no. 1, pp. 1–32, 2022, doi: 10.1186/s13059-022-02739-2.

[19] N. Biswas and S. Chakrabarti, "Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer," *Front. Oncol.*, vol. 10, no. October, pp. 1–13, 2020, doi: 10.3389/fonc.2020.588221.

[20] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *J. Sains dan Inform.*, vol. 9, no. April, pp. 19–28, 2023, doi: 10.34128/jsi.v9i1.622.

[21] Y. Amelia, P. Eosina, and F. A. Setiawan, "Perbandingan Metode Deep Learning Dan Machine Learning Untuk Klasifikasi (Ujicoba Pada Data Penyakit Kanker Payudara)," *Inova-Tif*, vol. 1, no. 2, p. 109, 2018, doi: 10.32832/inova-tif.v1i2.5449.

[22] Jumanto *et al.*, "Optimizing Support Vector Machine Performance for Parkinson's Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 1, pp. 38–50, 2024, doi: 10.20473/jisebi.10.1.38-50.

[23] M. Ridwan, I. Sembiring, A. Setiawan, and I. Setyawan, "Analysis of Attack Detection on Log Access Servers Using Machine Learning Classification: Integrating Expert Labeling and Optimal Model Selection," *Sci. J. Informatics*, vol. 11, no. 1, pp. 119–126, 2024, doi: 10.15294/sji.v11i1.49424.