JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.2 May 2025: 201-207
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i2.13409

# Breast Cancer Diagnosis Based on a Hybrid Genetic Algorithm and Neural Network Architecture

**Rifqi Alfinnur Charisma[1*], Ayu Maulina[2]**
[1,2]Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
rifqi.charisma@binus.ac.id; ayu.maulina@binus.ac.id

[*]Correspondence: rifqi.charisma@binus.ac.id

***Abstract –*** *Breast cancer is one of the diseases with a high prevalence and is a leading cause of death among women. Early detection is crucial in improving patient survival rates. However, a major challenge in diagnosis using machine learning methods is the high dimensionality of the data, which can lead to overfitting and reduced interpretability of the model. This study proposes a new approach to improve breast cancer prediction accuracy by using a combination of Genetic Algorithm + Neural Network (GA + NN). The dataset used is the Breast Cancer Wisconsin (Diagnostic) Data Set, consisting of 569 samples with 32 numerical features that describe the characteristics of tumor cells. The experimental results show that the GA + NN method achieved the highest accuracy of 99.42%, outperforming the benchmark model using PCA and logistic regression with an accuracy of 97.37%. This approach demonstrates that GA-based feature selection can improve prediction accuracy while reducing model complexity, making it more efficient for medical applications. To support clinical reliability, model evaluation was not limited to accuracy but also included precision, recall, and F1-score. The feature selection process using GA successfully identified the most relevant tumor features, such as radius, texture, and concavity measurements, contributing significantly to the model's predictive power. Moreover, the class distribution of the dataset, which consists of 357 benign and 212 malignant cases, was also considered to ensure balanced performance. These findings confirm the potential of hybrid GA + NN methods as an effective solution for high-dimensional medical classification tasks with strong applicability in real-world healthcare settings.*

***Keywords:*** *Breast Cancer; Classification; Genetic Algorithm; Neural Network; Multilayer Perceptron*

## I. INTRODUCTION

Breast cancer is one of the diseases with a high prevalence worldwide and is a leading cause of death among women (Breast Cancer, n.d.). Early detection of breast cancer is a key factor in improving patient survival rates as it allows for faster and more accurate treatment (Mangukiya & Vaghani, 2022). In an effort to improve the effectiveness of the diagnostic process, various technology-based approaches, particularly machine learning, have been widely used. Machine learning has the ability to identify patterns in complex medical data and provide accurate predictions related to the malignancy of tumors (Raj et al., 2024; Seethalakshmi B., 2024). However, high-dimensional medical datasets often include irrelevant or redundant features, which can lead to overfitting, increased computational complexity, and reduced model interpretability, ultimately hindering clinical adoption (Chen et al., 2025; Zhang & Cao, 2019).

Various studies have developed disease prediction models using feature selection methods to improve classification performance. Laghmati et al. (Laghmati et al., 2024) proposed a breast cancer prediction system combining PCA and logistic regression, achieving an accuracy of 97.37%, but PCA's transformation sacrifices feature interpretability, limiting its utility in clinical settings where understanding feature contributions is essential. Similarly, Afrin et al. (Afrin et al., 2021) used LASSO for feature selection and decision tree for liver disease prediction, with an accuracy of 94.29%, yet LASSO may discard correlated features that are biologically significant. Rustam et al. (Rustam & Kharis, 2020) compared SVM with and without feature selection using RFE for lung cancer classification, showing an accuracy of 96.79%, but RFE's iterative approach is computationally expensive, making it less feasible for large datasets.

This study proposes a novel approach that integrates Genetic Algorithm (GA) with Multilayer Perceptron (MLP) for feature selection in breast

cancer detection, aiming to address the limitations of existing methods. This study proposes a novel approach that integrates Genetic Algorithm (GA) with Multilayer Perceptron (MLP) for feature selection in breast cancer detection, aiming to address the limitations of existing methods. By leveraging GA's global search capabilities to explore diverse feature combinations and employing MLP as a fitness function, our method optimizes both classification accuracy and feature relevance while preserving interpretability. This combination is particularly advantageous for high-dimensional medical datasets, as it reduces insignificant features, accelerates training, and enhances model efficiency. The proposed GA-MLP model is validated using standard breast cancer datasets, such as the Wisconsin Breast Cancer Dataset, with performance evaluated through metrics including accuracy, precision, recall and f1-score, ensuring robust and reliable predictions for clinical use.

As a contribution, this study introduces a hybrid GA-MLP framework that outperforms traditional feature selection methods by dynamically evaluating feature subsets, offering a balance of accuracy, efficiency, and interpretability. This approach not only improves diagnostic performance but also facilitates clinical decision-making by retaining meaningful features, making it a promising tool for medical applications.

## II. METHODS

### 2.1 Dataset

The dataset used in this study is the Breast Cancer Wisconsin (Diagnostic) Data Set, obtained from the UCI Machine Learning Repository (Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository, n.d.). This dataset was developed by Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison, and is widely used in medical classification research, particularly in detecting breast cancer based on cellular analysis. The dataset consists of 569 data samples, each representing the results of cell analysis from patients who have undergone the Fine Needle Aspiration (FNA) procedure, a technique for obtaining tissue samples from lumps in the breast using a fine needle. Each sample is digitally analyzed from microscopic images, and several numerical features are calculated to reflect the characteristics of the tumor cells. Structurally, this dataset has 32 columns, consisting of: 1 ID column (not used in modeling), 1 diagnosis column (target label), and 30 columns of numerical features. The diagnosis label is found in the "diagnosis" column, which consists of two classes: M (Malignant): indicating that the tumor is malignant, and B (Benign): indicating that the tumor is benign. The numerical features are divided into 10 main cell characteristics, which are:

- Radius: the distance from the center to the outermost boundary of the cell
- Texture: variation in gray intensity
- Perimeter: the length of the cell's perimeter
- Area: the area covered by the cell
- Smoothness: local variation in the contour length
- Compactness: the ratio between the perimeter and area
- Concavity: the degree of concavity at the cell boundary
- Concave points: the number of concave points on the cell boundary
- Symmetry: symmetry of the cell
- Fractal dimension: the complexity of the cell boundary

Each of these characteristics is measured in three statistical forms: mean, standard error (se), and worst (maximum value from the ten largest cells). Therefore, a total of 30 numerical features are obtained (10 characteristics x 3 types of measurements). The class distribution in this dataset is as follows: 357 benign samples and 212 malignant samples, as illustrated in the figure 1.

Additionally, this dataset contains no missing values, and all features are in numerical format, so minimal additional preprocessing is required, such as imputation or re-encoding. Since all features have different scales, normalization is performed before inputting them into the machine learning model.

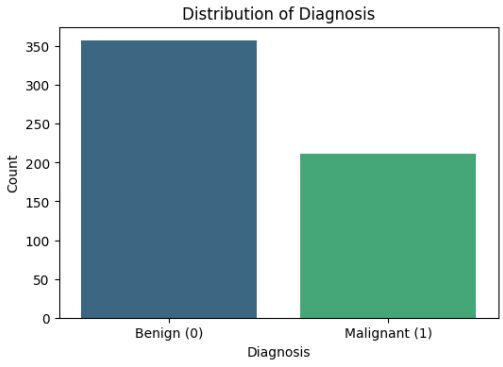


Figure 1. Data Distribution

### 2.2 Genetic Algorithm + Neural Network (GA + NN)

In this study, Genetic Algorithm (GA) is used to perform feature selection before the classification model training process using a Multilayer Perceptron (MLP) neural network. The goal is to find the best combination of

features that can improve the model's accuracy and efficiency. The architecture of the Genetic Algorithm + Neural Network is illustrated in the figure 2. The feature selection process is carried out by representing each individual in the population as a binary chromosome, where each gene represents one feature from the dataset. A value of 1 indicates that the feature is used, while a value of 0 indicates that the feature is ignored. (Pham et al., 2020).

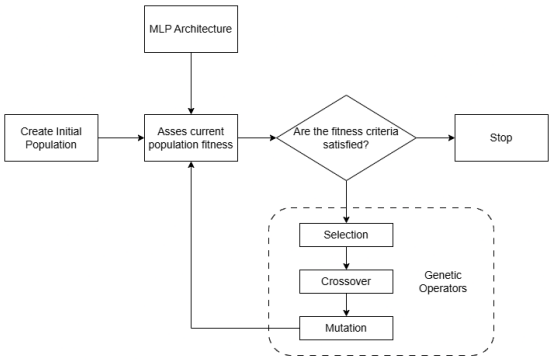


Figure 2. GA + NN Architecture

After evaluation, the evolutionary process in the genetic algorithm is carried out, including selection, crossover, and mutation. Selection is performed using the roulette wheel method to choose the best individuals as parents. The two selected parents undergo single-point crossover to form new offspring. Next, mutation is performed by randomly flipping gene values based on a certain probability (mutation rate). This process is repeated for several generations to find the feature combination that provides the highest accuracy. The best chromosome from all generations is considered the final solution. The features selected by this chromosome are then used to retrain the entire MLP model (Idrissi et al., 2016).

### 2.3 Multi-Layer Perceptron (MLP)

Multilayer Perceptron (MLP) is one of the most commonly used architectures of artificial neural networks for both classification and regression tasks (He & Chen, 2021). The MLP architecture consists of three main components: the input layer, one or more hidden layers, and the output layer, as shown in the figure 3. Each layer consists of several processing units called neurons, and each neuron in a layer is fully connected to all neurons in the subsequent layer. The input layer functions to receive input data from the features that have been processed or selected previously, such as the results from feature selection using a genetic algorithm. Each neuron in the input layer represents one feature (Kusuma et al., 2022).

After passing through the input layer, the data is processed by one or more hidden layers. The hidden layer is responsible for performing a non-linear transformation of the input using weights and biases that can be learned during the training process. Each neuron in the hidden layer calculates a linear combination of the input it receives, then applies an activation function such as ReLU (Rectified Linear Unit), sigmoid, or tanh to produce a non-linear output. This activation function enables the MLP to learn complex relationships within the data (Sudianto et al., 2022).

The output from the last hidden layer is then passed to the output layer, which produces the final prediction of the model. In binary classification tasks such as breast cancer detection, the output layer typically consists of a single neuron with a sigmoid activation function to produce a probability value between 0 and 1. This value is then used to determine the final class. The training of the MLP network is carried out through the backpropagation process, which involves calculating the error (the difference between the predicted output and the actual label), and then updating the network weights using optimization algorithms such as stochastic gradient descent (SGD) or Adam (Rashedi et al., 2024).

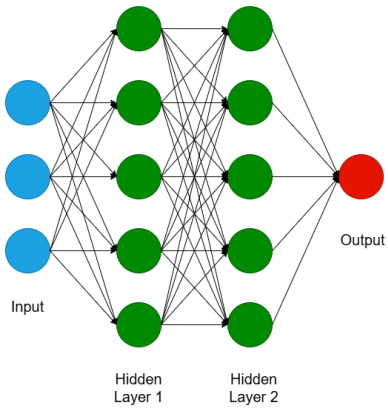


Figure 3. MLP Architecture

### 2.4 Evaluation

To assess the performance of the proposed model, several standard classification metrics are used, namely accuracy, precision, recall, and F1-score. These metrics are widely adopted in binary classification tasks and provide a comprehensive view of the model's predictive capability (Hasan et al., 2021; Rasywir et al., 2020; Sholihati et al., 2020; Wulandari et al., 2020)

The experimental results obtained from the proposed method will be compared with a benchmark model presented by Laghmati et al. in their study "An Improved Breast Cancer Disease Prediction System Using ML and PCA" (Laghmati et al., 2024). That benchmark model employed PCA for dimensionality reduction and logistic regression in a stacking ensemble setup, achieving an accuracy of 97.37% on the WBCD dataset. The comparison will highlight the effectiveness of using GA-based feature selection in optimizing relevant features and improving classification performance.

## III. RESULTS AND DISCUSSION

### 3.1 Experimental Environment and Parameter Settings

All experiments in this study were conducted using Google Colaboratory (Colab), a cloud-based Python notebook environment that provides free access to GPUs and pre-installed scientific computing libraries. The default runtime environment of Google Colab was used without any hardware customization. This environment is equipped with Python 3.x and supports major machine learning libraries such as Scikit-learn, NumPy, and Pandas.

The core of the proposed method involves a hybrid model that integrates Genetic Algorithm (GA) for feature selection and Multilayer Perceptron (MLP) as the fitness function. The dataset is divided into training and testing sets in the proportion of 80:20, to evaluate model performance.

Parameter settings for the GA and MLP were selected based on preliminary experiments to ensure a balance between performance and computational efficiency. For the Genetic Algorithm (GA) used for feature selection, the following parameters were configured:

- Population size: 10 chromosomes (pop_size = 10),
- Number of generations: 10 (n_gen = 10),
- Mutation rate: 0.1 (mutation_rate = 0.1).
- Crossover rate: Configured to perform single-point crossover at a randomly chosen point between 1 and the number of features minus 1,
- Selection method: Probability-based selection, where each chromosome's selection probability is proportional to its fitness score normalized by the sum of all scores.

Each chromosome in the GA represents a binary vector corresponding to the inclusion (1) or exclusion (0) of each feature in the dataset. The fitness function is evaluated based on the classification accuracy of an MLP model trained on the selected subset of features. For this purpose, an MLP with a single hidden layer of 50 neurons, max_iter=500, and random_state=42 for reproducibility was used as the fitness evaluator. This configuration was chosen based on preliminary experiments showing that a single hidden layer with 50 neurons provides sufficient predictive performance, while max_iter=500 ensures convergence within a reasonable time. To assess computational efficiency, the GA-MLP model was evaluated, with an average runtime of approximately 120 seconds per experiment on the default Colab GPU environment, demonstrating its suitability for practical applications.

### 3.2 Breast Cancer Prediction

In this study, Genetic Algorithm (GA) is combined with Neural Network (NN), using Multilayer Perceptron (MLP) as the fitness function, for feature selection in the breast cancer dataset. The experimental results show that the combination of GA + NN achieves excellent accuracy in the classification model. The feature selection performed automatically using GA demonstrates a significant improvement in the model's accuracy as the generations progress.

The table 1 shows the feature selection results and the best accuracy achieved in each generation:

Table 1. Genetic Algorithm + NN Training Results

| Generation | Best Accuracy | Selected Features |
|---|---|---|
| 1 | 0.9824561403508771 | 2,4,8,9,11,12,13,14,16,19,20,21,22,23,24,26,27 |
| 2 | 0.9883040935672515 | 1,4,8,9,11,12,13,16,19,20,21,22,23,24,26,27 |
| 3 | 0.9883040935672515 | 1,3,5,6,7,8,10,11,13,16,19,20,21,23,24,26,27 |
| 4 | 0.9883040935672515 | 1,2,3,4,8,9,12,13,14,16,17,21,23,26 |
| **5** | **0.9941520467836257** | **0,1,4,5,8,9,10,13,14,16,17,21,26** |
| 6 | 0.9883040935672515 | 0,1,2,3,4,8,12,13,14,16,17,23,26,27 |
| 7 | 0.9883040935672515 | 3,4,7,8,9,12,13,14,15,16,17,21,23,24,26,29 |
| 8 | 0.9883040935672515 | 1,4,5,8,9,11,13,14,15,19,22,23,24,27,29 |
| 9 | 0.9883040935672515 | 0,1,2,4,5,10,11,12,13,14,15,19,20,21,26,27,29 |
| **10** | **0.9941520467836257** | **1,2,4,5,10,11,12,15,17,20,21,22,23,24,26,27** |

From the table 1, it can be seen that in the fifth and tenth generations, the model achieved the best accuracy of 99.42%. The feature selection results in these generations show that the most frequently selected features include:

- radius_mean
- texture_mean
- smoothness_mean
- compactness_mean
- symmetry_mean
- fractal_dimension_mean
- radius_se
- area_se
- smoothness_se
- concavity_se
- concave points_se
- texture_worst
- concavity_worst

These features demonstrate high relevance in the breast cancer prediction process based on the Breast Cancer Wisconsin dataset.

### 3.3 Comparative Study

The results obtained from the GA + NN model, compared to the benchmark model proposed by Laghmati et al. (Laghmati et al., 2024), which uses PCA and logistic regression, show that the model developed in this study has an advantage in terms of accuracy. The benchmark model by Laghmati et al. achieved an accuracy of 97.37%, while the GA + NN model in this study achieved the highest accuracy of 99.42%. This demonstrates the effectiveness of using the genetic algorithm for feature selection in improving model accuracy, as well as reducing model complexity through the selection of more relevant features..

## IV. CONCLUSION

This study proposes a combination of Genetic Algorithm (GA) + Neural Network (NN) for breast cancer prediction. The experimental results show that this model achieves the highest accuracy of 99.42%, outperforming the benchmark model using PCA and logistic regression, which achieved 97.37%. This approach proves to be effective in improving the accuracy and efficiency of the classification model.

Future work should address these limitations by testing the GA-MLP framework on diverse medical datasets with varying imbalance levels to validate its applicability beyond breast cancer. Exploring alternative strategies for handling class imbalance, such as cost-sensitive learning or ensemble methods, could further enhance the model's robustness, making it a more versatile and reliable tool for clinical diagnostics with improved interpretability and performance across a broader range of medical applications.

## REFERENCES

Afrin, S., Javed Mehedi Shamrat, F. M., Nibir, T. I., Muntasim, M. F., Moharram, M. S., Imran, M. M., & Abdulla, M. (2021). Supervised machine learning based liver disease prediction approach with LASSO feature selection. *Bulletin of Electrical Engineering and Informatics*, *10*(6), 3369–3376. https://doi.org/10.11591/EEI.V10I6.3242

*Breast cancer*. (n.d.). Retrieved April 21, 2025, from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

*Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository*. (n.d.). Retrieved April 21, 2025, from https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Chen, Y., Ding, W., Huang, J., Zhang, W., & Zhou, T. (2025). Multigranularity Fuzzy Autoencoder for Discriminative Feature Selection in High-Dimensional Data. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2025.3569893

Hasan, Moh. A., Riyanto, Y., & Riana, D. (2021). Grape leaf image disease classification using CNN-VGG16 model. *Jurnal Teknologi Dan Sistem Komputer*, *9*(4), 218–223. https://doi.org/10.14710/jtsiskom.2021.14013

He, X., & Chen, Y. (2021). Modifications of the Multi-Layer Perceptron for Hyperspectral Image Classification. *Remote Sensing 2021, Vol. 13, Page 3547*, *13*(17), 3547. https://doi.org/10.3390/RS13173547

Idrissi, M. A. J., Ramchoun, H., Ghanou, Y., & Ettaouil, M. (2016). Genetic algorithm for neural network architecture optimization. *Proceedings of the 3rd IEEE International Conference on Logistics Operations Management, GOL 2016*. https://doi.org/10.1109/GOL.2016.7731699

Kusuma, J., Hayadi, B. H., Wanayumini, W., & Rosnelly, R. (2022). Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara. *MIND (Multimedia Artificial Intelligent Networking Database) Journal*, *7*(1), 51–60. https://doi.org/10.26760/MINDJOURNAL.V7I1.51-60

Laghmati, S., Hamida, S., Hicham, K., Cherradi, B., & Tmiri, A. (2024). An improved breast cancer disease prediction system using ML and PCA. *Multimedia Tools and Applications*, *83*(11), 33785–33821. https://doi.org/10.1007/S11042-023-16874-W/METRICS

Mangukiya, M., & Vaghani, A. (2022). *Breast Cancer Detection with Machine Learning*. *10*. https://doi.org/10.22214/ijraset.2022.40204

Pham, T. A., Tran, V. Q., Vu, H. L. T., & Ly, H. B. (2020). Design deep neural network architecture using a genetic algorithm for estimation of pile bearing capacity. *PLOS ONE*, *15*(12), e0243030. https://doi.org/10.1371/JOURNAL.PONE.0243030

Raj, R., Jyoti, Singh, A., & Kumar, K. (2024). Machine Learning in Medical Diagnosis of Cancer. *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024*. https://doi.org/10.1109/I2CT61223.2024.10544354

Rashedi, K. A., Ismail, M. T., Al Wadi, S., Serroukh, A., Alshammari, T. S., & Jaber, J. J. (2024). Multi-Layer Perceptron-Based Classification with Application to Outlier Detection in Saudi Arabia Stock Returns. *Journal of Risk and Financial Management 2024, Vol. 17, Page 69*, *17*(2), 69. https://doi.org/10.3390/JRFM17020069

Rasywir, E., Sinaga, R., & Pratama, Y. (2020). Analisis dan Implementasi Diagnosis Penyakit Sawit dengan Metode Convolutional Neural Network (CNN). *Paradigma - Jurnal Komputer Dan Informatika*, *22*(2). https://doi.org/10.31294/p.v22i2.8907

Rustam, Z., & Kharis, S. A. A. (2020). Comparison of Support Vector Machine

Recursive Feature Elimination and Kernel Function as feature selection using Support Vector Machine for lung cancer classification. *Journal of Physics: Conference Series*, *1442*(1), 012027. https://doi.org/10.1088/1742-6596/1442/1/012027

Seethalakshmi B. (2024). Brain Tumor Malignancy Prediction Using Machine Learning Techniques. *Irish Interdisciplinary Journal of Science & Research*, *08*(02), 86–93. https://doi.org/10.46759/iijsr.2024.8210

Sholihati, R. A., Sulistijono, I. A., Risnumawan, A., & Kusumawati, E. (2020). Potato Leaf Disease Classification Using Deep Learning Approach. *IES 2020 - International Electronics Symposium: The Role of Autonomous and Intelligent Systems for Human Life and Comfort*, 392–397. https://doi.org/10.1109/IES50839.2020.9231784

Sudianto, S., Sripamuji, A. D., Ramadhanti, I. R., Amalia, R. R., Saputra, J., & Prihatnowo, B. (2022). Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasifikasi Topik Berita. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, *11*(2), 84–91. https://doi.org/10.23887/JANAPATI.V11I2.44151

Wulandari, I., Yasin, H., & Widiharih, T. (2020). Klasifikasi Citra Digital Bumbu Dan Rempah Dengan Algoritma Convolutional Neural Network (CNN). *Jurnal Gaussian*, *9*(3). https://doi.org/10.14710/j.gauss.v9i3.27416

Zhang, B., & Cao, P. (2019). Classification of high dimensional biomedical data based on feature selection using redundant removal. *PLoS ONE*, *14*(4), e0214406. https://doi.org/10.1371/JOURNAL.PONE.0214406