Journal of Computer Engineering, Electronics and Information Technology (COELITE)

Journal homepage: https://ejournal.upi.edu/index.php/COELITE

Comparison of Machine Learning and Deep Learning Algorithms for Classification of Breast Cancer

*Puji Ayuningtyas[1, *], R. Rahmawati[2], Akhmad Miftahusalam[3]*

[1]Department of Information Engineering, Institut Teknologi Telkom Purwokerto, Indonesia
[2]Department of AI Research, PT Bisa Artificial Indonesia, Indonesia
[3]Department of AI Hacker, PT Bisa Artificial Indonesia, Indonesia
Correspondence: E-mail: 20102122@ittelkom-pwt.ac.id

ABSTRACT

Statistical data from the American Cancer Society which shows that breast cancer ranks first with the highest number of cases of all types of cases of malignant tumors (cancer) worldwide. through a data mining process that is used to extract information and data analysis, a classification process can be carried out to carry out further analysis of the pattern of a data. The dataset used in this study is the Breast Cancer Wisconsin (Diagnostic) Dataset obtained from UCI Machine Learning. The purpose of this study is to compare five algorithms, namely Logistic Regression, K Neighbors Classifier (KNN), Decision Tree Classifier, Deep Neural Network, Genetic Algorithm. The results showed that deep neural network algorithms and multilayer perceptron-genetic algorithms get 96% accuracy, logistic regression algorithms have 96% accuracy, then KNN with 94%, and decision tree classifier with 92%.

ARTICLE INFO

Article History:
Submitted/Received 27 Mar 2023
First Revised 12 Apr 2023
Accepted 19 May 2023
First Available online 22 Jun 2023
Publication Date 01 Oct 2023

Keyword:
Decision Tree Classifier,
Deep Neural Network,
Genetic Algorithm,
KNN,
Logistic Regression.

1 INTRODUCTION

Cancer is a disease that occurs due to the uncontrolled growth of abnormal cells. Breast cancer is the most common cancer in women [1]. Based on data from GLOBOCAN (Global Burden of Cancer), the International Agency for Research on Cancer (IARC) it is known that in 2018 there were 18.1 million new cases of cancer and 9.6 million deaths from cancer worldwide. It is estimated that annual cancer cases will increase from 18.1 million to 22 million in the next two decades. WHO estimates that by 2030 the incidence of cancer will reach 26 million people and 17 million of them will die from cancer [2]. This makes breast cancer the most common type of cancer in women after cervical cancer [3]. This is supported by statistical data from the American Cancer Society which shows that breast cancer ranks first with the highest number of cases of all types of malignant tumors (cancer) worldwide [4].

Data mining is defined as the process of extracting or mining the required knowledge from large amounts of data. In the process, data mining will extract valuable information by analyzing the existence of certain patterns or relationships from large data. Data mining is related to other fields of science, such as Database Systems, Data Warehousing, Statistics, Machine Learning, Information Retrieval, and High Level Computing. In addition, data mining is supported by other sciences such as Neural Networks, Pattern Recognition, Spatial Data Analysis, Image Databases, Signal Processing [5].

Classification is included in supervised learning because it uses a set of data to be analyzed first, then the pattern from the results of the analysis is used to classify the test data. The data classification process consists of learning and classification. In learning training data is analyzed using a classification algorithm then in classification data testing is used to ensure the level of accuracy of the classification rules used. Classification techniques are divided into five categories based on differences in mathematical concepts, namely statistical-based, distance-based, decision tree-based, neural network-based, and rule-based [6].

As a series of processes, data mining has several data processing, one of which is classification. Based on research conducted by Pangaribuan, et al regarding a comparison between algorithms C.50, SVM, KNN, Logistic Regression, and neural networks for detecting heart disease, the accuracy rate for KNN was 86.50% while logistic regression was 83.7% [7]. Research on breast cancer using data mining and machine learning has been carried out by Higa using a Decision Tree and Artificial Neural Network, the two algorithms successfully classify more than 92% of cases correctly in 10 trials. However, the Neural Network algorithm has an average level of predictive accuracy that is better (the correct classification rate is up to 95.9%) [8]. Research conducted by Derisma also uses a Neural Network with optimization using a Genetic Algorithm to obtain an accuracy of 97.24% [9].

The purpose of this study is to compare five algorithms, namely Logistic Regression, K Neighbors Classifier (KNN), Decision Tree Classifier, Deep Neural Network, Genetic Algorithm. Based on these algorithms, it will be determined which algorithm produces a better accuracy value.

2 METHODS

2.1 Dataset

The dataset used in this study is Breast Cancer Wisconsin (Diagnostic) Dataset obtained from UCI Machine Learning (see https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data). The dataset consists of 33 features and 569 data. This dataset contains patient diagnosis information and the value of breast cancer image computation results. Diagnosis consists of two classes, namely benign and malignant. A total of 357 patients were

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

diagnosed as having benign breast cancer and 212 patients were diagnosed as having malignant breast cancer. It can be seen if the target has imbalanced data.

2.2 Classification Algoritma

2.2.1 Logistic Regression

Logistic Regression is often used in medical research, especially since this procedure is readily available in statistical software packages. Logistic Regression is a statistical model that describes the relationship between the qualitative dependent variable and the independent variable [10][11][12]. Where μ is the location of the parameter (the midpoint of the curve where $p(\mu) = \frac{1}{2}$) and s is the scale of the parameter, see Equation [1].

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}} \quad (1)$$

2.2.2 K Neighbors Classifier (KNN)

KNN clusters each test sample based on the nearest neighbor k value. The cluster whose center has the minimum distance from the test sample is selected as the appropriate reduced training dataset [13][14]. The quality of the KNN classification depends on how well the closest neighbors are found. The chances of finding the exact k nearest neighbors also depend on how well the large dataset has been trimmed. There are two functions in determining the closest neighbor distance between them, see Equation [2] and Equation [3].

Euclidean:

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (2)$$

Manhattan:

$$d(x,y) \sum_{i=1}^{m}|x_i - y_i| \quad (3)$$

2.2.3 Decision Tree Classifier

Decision Tree is a machine learning algorithm where each decision is described as a node [15][16]. Each decision tree is built on several branches and nodes. Each node represents a feature in the category to be classified and each subset defines a value that can be taken by the node. The way this algorithm works is by comparing the numerical features with the threshold values in each test. Roots, branches, and nodes in the decision tree are determined by the entropy value and the gini index value, see Equation [4] and Equation [5].

$$Entropy(t) = -\sum_{i=1}^{c-1} p(i|t)\ log_2 p(i|t) \quad (4)$$

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t]^2 \quad (5)$$

2.2.4 Deep Neural Network

Deep Neural Network (DNN) is one of the Deep Learning algorithms which is inspired by the human nervous system. Like the human brain, the DNN has neurons that are

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

interconnected with each other. Deep Learning consists of several layers such as input layer, hidden layer, and output layer [17][18]. This algorithm consists of 2 stages, namely Forward Pass and Backward Pass. Forward pass is the stage where the input data will pass through each neuron at each layer, starting from the input layer to the output layer. At this stage, the error value will be calculated, see Equation (6) and Equation (7).

$$dot_j = \sum_{i}^{3} w_{ji} x_i + b_j \tag{6}$$

$$h_j = \sigma(dot_j) = max(0, dot_j) \tag{7}$$

Second, there is the Backward Pass, which is the stage where the weight and bias values will be updated using the error values that have been obtained at the Forward pass stage.

2.2.5 Genetic Algorithm

Genetic Algorithm (GA) or in Indonesian it is called Genetic Algorithm is one of the algorithms whose way of working is inspired by Darwin's Theory of Evolution. GA population-based algorithm. Each solution corresponds to a chromosome and each parameter represents a gene [19][20].

2.3 Confusion Matrix

Measurement of the classification model can be seen through the confusion matrix. The confusion matrix is a special table that visualizes the performance of the classification algorithm. The confusion matrix table can be seen in **Table 1**.

Table 1. Confusion Matrix.

Prediction	Actual		
	Positive	**Negative**	**Total**
Positive	TP	FP	TP + FP
Negative	FN	FN	FN + TN
Total	TP + FN	FP + TN	TP + FP + FN + TN

From the confusion matrix, model accuracy can be obtained. Accuracy is the most frequently used indicator to measure the classification performance of a model [21]. Model accuracy can be obtained through the Equation [8]:

$$accuration = \frac{TP + TN}{(TP + FP + TN + FN)} \tag{8}$$

Where,

TP: True Positive, positive is classified as positive
TN: True Negative, negative is classified as negative
FP: False Positive, negative is classified as positive
FN: False Negative, positive is classified as negative

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

3 RESULTS AND DISCUSSION

Comparison of the accuracy of each of the algorithms, as follows in **Table 2.**:

Table 2. Accuracy comparison

Model	Accuracy
Logistic Regression	0.96
K Neighbors Classifier (KNN)	0.94
Decision Tree Classifier	0.92
Deep Neural Network	0.96
Genetic Algorithm	0.96

The computing process for each model uses Google Collab. Tests are carried out by comparing simple machine learning algorithms with deep learning algorithms. Simple machine learning algorithms include Logistic Regression, KNN, and Decision Tree. The test results show that the Logistic Regression model has a training accuracy value of 98% and a testing accuracy value of 95% as shown in **Figure 1**. Matrix evaluation shows that this model has an accuracy of 96% with a precision value of 97%, a recall of 92%, and an f1-score of 94% for patients diagnosed with breast cancer.

	precision	recall	f1-score	support
0	0.95	0.98	0.97	108
1	0.97	0.92	0.94	63
accuracy			0.96	171
macro avg	0.96	0.95	0.96	171
weighted avg	0.96	0.96	0.96	171

Figure 1. Logistic Regression model result.

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

The K Neighbors Classifier (KNN) model has a training accuracy value of 96% and a testing accuracy value of 93% as shown in **Figure 2**. Matrix evaluation shows that this model has an accuracy of 94% with a precision value of 95%, a recall of 87%, and an f1-score of 91% for patients diagnosed with breast cancer.

	precision	recall	f1-score	support
0	0.93	0.97	0.95	108
1	0.95	0.87	0.91	63
accuracy			0.94	171
macro avg	0.94	0.92	0.93	171
weighted avg	0.94	0.94	0.94	171

Figure 2. K Neighbors Classifier (KNN) model result.

The Decision Tree Classifier model has a training accuracy value of 96% and a testing accuracy value of 92% as shown in **Figure 3**. Matrix evaluation shows that this model has an accuracy of 92% with a precision value of 87%, a recall of 94%, and an f1-score of 90% for patients diagnosed with breast cancer.

	precision	recall	f1-score	support
0	0.96	0.92	0.94	108
1	0.87	0.94	0.90	63
accuracy			0.92	171
macro avg	0.91	0.93	0.92	171
weighted avg	0.93	0.92	0.92	171

Figure 3. Decision Tree Classifier model result.

Furthermore, for algorithms based on deep learning, there are Deep Neural Networks (DNN) and Genetic Algorithm (GA). The DNN algorithm uses 5 layers where 2 of them are input and output layers and the other 3 are hidden layers. ReLU is used for the activation function in each hidden layer. In addition, the output layer uses the sigmoid activation function because the dataset is a binary class. The comparison between the train size and the test size is 7:3. The training process was carried out for 100 epochs with Adam as the optimizer.

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

The results show that the DNN model has an accuracy value of 96% with a precision value of 90%, 98% recall, and 94% f1-score for patients diagnosed with breast cancer as shown in **Figure 4.**, **Figure 5.**, and **Figure 6**.

	precision	recall	f1-score	support
0	0.99	0.95	0.97	113
1	0.90	0.98	0.94	58
accuracy			0.96	171
macro avg	0.95	0.96	0.96	171
weighted avg	0.96	0.96	0.96	171

Figure 4. Deep Neural Networks (DNN) model result.

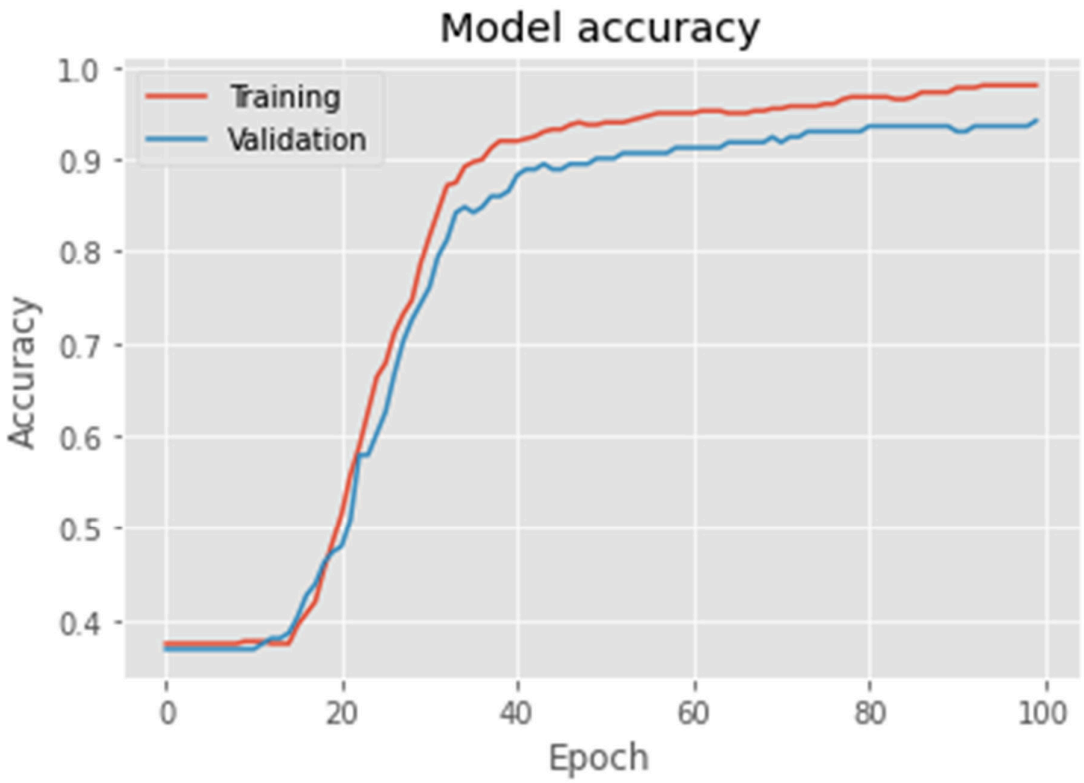

Figure 5. Graph of the accuracy of the Deep Neural Networks (DNN) model.

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

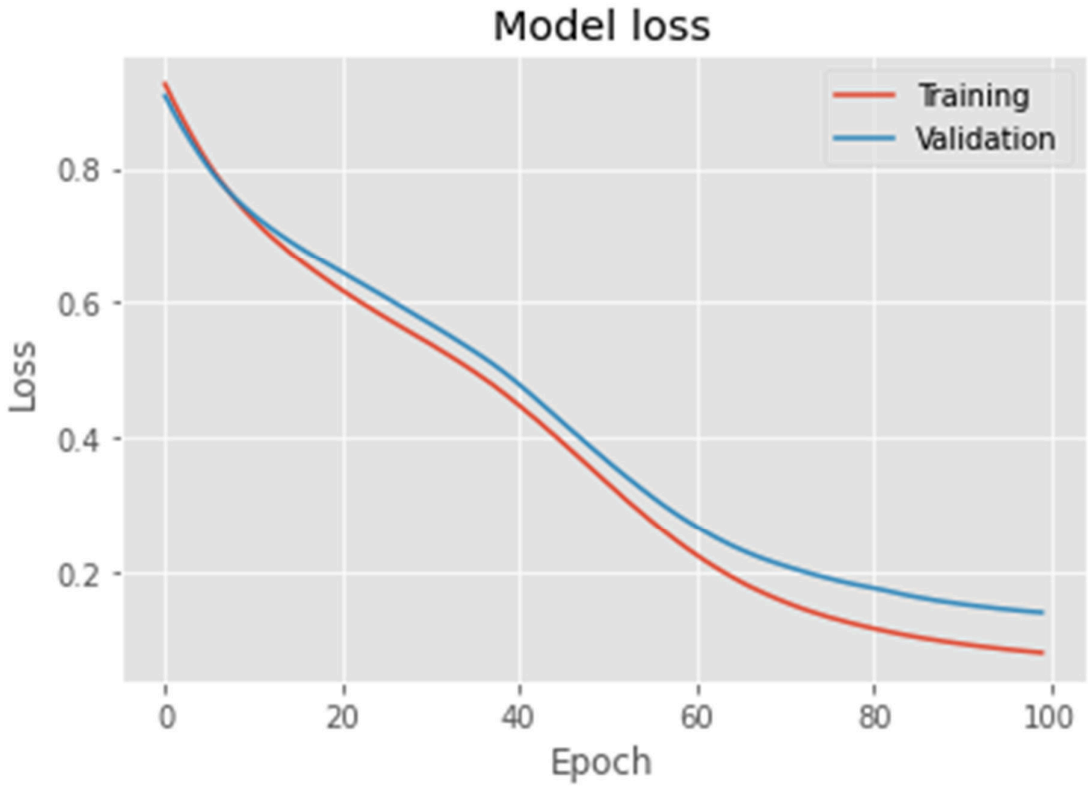

Figure 6. Graph of the loss of the Deep Neural Networks (DNN) model.

In the Genetic Algorithm, the existing data will first be classified using the MLP Classifier and Decision Tree algorithms. The classification results show that in the MLP classifier, the training process has an accuracy value of 89% and 90% in model testing. Whereas in the Decision Tree algorithm, the training process has an accuracy value of 100% and 94% when testing the model. After doing the classification, then implement the GA. The specified population size is 100, meaning that the generation to be used is half of the population, namely 50. After retraining, it was found that when adding GA as the optimizer of MLP and Decision Tree, the accuracy rate will increase to 96%.

4 CONCLUSION

The conclusions obtained in this study are that 5 (five) algorithms are used, namely logistic regression, KNN, Decision tree classifier, deep neural network, and genetic algorithm using the Breast Cancer Wisconsin (Diagnostic) Dataset obtained from UCI Machine Learning. The research results obtained that deep learning classification algorithms have a tendency to obtain high accuracy. The deep learning algorithms in question are deep neural networks and multilayer perceptron-genetic algorithms which get 96% accuracy. Meanwhile, the machine learning algorithm that has the highest accuracy is the logistic regression algorithm, which is 96%, then KNN with 94%, and decision tree classifier with 92%.

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

5 AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

6 REFERENCES

[1] Cahyanti, D., Rahmayani, A., dan Husniar, S. A. (2020). Analisis performa metode KNN pada dataset pasien pengidap kanker payudara. *Indonesian Journal of Data and Science*, *1*(2), 39-43.

[2] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209-249.

[3] Ma'arif, F., dan Arifin, T. (2017). Optimasi fitur menggunakan backward elimination dan algoritma SVM untuk klasifikasi kanker payudara. *Jurnal Informatika*, *4*(1), 46-53.

[4] Mattiuzzi, C., and Lippi, G. (2019). Current cancer epidemiology. *Journal of Epidemiology and Global Health*, *9*(4), 217-222.

[5] Ikotun, A. M., Almutari, M. S., and Ezugwu, A. E. (2021). K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: recent advances and future directions. *Applied Sciences*, *11*(23), 1-61.

[6] Sartika, D., dan Indra, D. (2017). Perbandingan algoritma klasifikasi naive bayes, nearest neighbour, dan decision tree pada studi kasus pengambilan keputusan pemilihan pola pakaian. *Jurnal Teknik Informatika dan Sistem Informasi*, *1*(2), 151–161.

[7] Pangaribuan, J. J., Tanjaya, H., dan Kenichi, K. (2021). Mendeteksi penyakit jantung menggunakan machine learning dengan algoritma logistic regression. *Journal Information System Development (ISD)*, *6*(2), 1-10.

[8] Higa, A. (2018). Diagnosis of breast cancer using decision tree and artificial neural network algorithms. *International Journal of Computer Applications Technology and Research*, *7*(1), 23–27.

[9] Rane, N., Sunny, J., Kanade, R., and Devi, S. (2020). Breast cancer classification and prediction using machine learning. *International Journal of Engineering Research and Technology*, *9*(2), 576-580.

[10] Wang, Q. Q., S. C. Yu, Xiao Qi, Y. H. Hu, W. J. Zheng, J. X. Shi, and H. Y. Yao. (2019). Overview of logistic regression model analysis and application. *Chinese Journal of Preventive Medicine, 53*(9), 955-960.

[11] Silva, J. C. F., Teixeira, R. M., Silva, F. F., Brommonschenkel, S. H., and Fontes, E. P. (2019). Machine learning approaches and their current application in plant molecular biology: a systematic review. *Plant Science*, *284*(1), 37-47.

[12] Senaviratna, N. A. M. R., and A Cooray, T. M. J. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, *5*(2), 1-9.

[13] Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., and Shamshirband, S. (2020). A new K-nearest neighbors classifier for big data based on efficient data pruning. *Mathematics*, *8*(2), 1-12.

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149

[14] Alfeilat, H. A. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Salman, H. S. E., and Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big Data*, *7*(4), 221-248.

[15] Charbuty, B., and Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends, 2*(01), 20-28.

[16] Cichy, R. M., and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305-317.

[17] Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*(1), 1-15.

[18] Patra, T. K., Meenakshisundaram, V., Hung, J. H., and Simmons, D. S. (2017). Neural-network-biased genetic algorithms for materials design: evolutionary algorithms that learn. *ACS Combinatorial Science*, *19*(2), 96-107.

[19] Katoch, S., Chauhan, S. S., and Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, *80(1)*, 8091-8126.

[20] Purwa, T. (2019). Perbandingan metode regresi logistik dan random forest untuk klasifikasi data imbalanced (studi kasus: klasifikasi rumah tangga miskin di kabupaten Karangasem, Bali tahun 2017). *Jurnal Matematika, Statistika dan Komputasi*, *16*(1), 58-73.

DOI: https://doi.org/10.17509/coelite.v2i2.59717
p- ISSN 2829-4157 e- ISSN 2829-4149