RESEARCH ARTICLE | MAY 09 2023

Comparison of Logistic Regression, K-Nearest Neighbour, and decision tree (C4.5) on parameter optimization to increase prediction of breast cancer

Harsih Rianto ✉; Omar Pahlevi; Rudianto; ... et. al

AIP Conference Proceedings 2714, 020018 (2023)
https://doi.org/10.1063/5.0128341

Comparison of Logistic Regression, K-Nearest Neighbour, and Decision Tree (C4.5) on Parameter Optimization to Increase Prediction of Breast Cancer

Harsih Rianto,[1, a)] Omar Pahlevi,[1, b)] Rudianto,[1, c)] amrin,[2, d)] and Paramita Kusumawardhani[3, e)]

[1)]*Program Studi Sistem Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia*
[2)]*Program Studi Teknologi Komputer, Universitas Bina Sarana Informatika, Jakarta, Indonesia*
[3)]*Program Studi Bahasa Inggris, Universitas Bina Sarana Informatika, Jakarta, Indonesia*

[a)]*Corresponding author: harsih.hhr@bsi.ac.id*
[b)]*omar.opi@bsi.ac.id*
[c)]*rudianto.rdt@bsi.ac.id*
[d)]*amrin.ain@bsi.ac.id*
[e)]*paramita.pmi@bsi.ac.id*

Abstract. The malignancy of breast cancer has caused many deaths in women. Breast cancer begins when a cancerous, malignant lump begins to grow from the breast cells. Over the decades machine learning has grown rapidly and is used for applications in health-related fields. In this study, we compared the Logistic Regression (LR), K-Nearest Neighbour (KNN), and C4.5 Decision Tree (DT) algorithm methods by applying parameter optimization to improve breast cancer prediction results. The dataset we use is the Wisconsin Breast Cancer Dataset (WBCD) which consists of 699 records. The results of the performance of LR, KNN and DT have increased in accuracy values after the parameter optimization method was applied. After testing using the same dataset on the three algorithms by comparing the AUC and Confusion Matrix values, the LR algorithm produces an Accuracy value of 96.9% and AUC of 99.5%. While the KNN Algorithm the resulting Accuracy value is 97.4% and AUC is 99.3%. For the DT algorithm, the test results get the Accuracy and AUC values of 96%.

INTRODUCTION

The cause of death in Indonesia after heart disease and stroke is cancer [1]. Cancer is a type of non-communicable disease. Some of the factors that cause the risk of cancer are age, gender, race or ethnicity in several countries [1]. Cancers that often occur in men are lung cancer and cholesterol. Meanwhile, in women, the most dominant cancer is breast cancer and cervical cancer. Mortality resulting from breast cancer reaches 25% in women aged 40 to 49 years [2]. In women aged 25-30 years, this type of cancer is rarely seen. From the results of the World Health Organization (WHO) statistical reports, there are 1 in 8 to 10 women suffering from breast cancer [2]. Meanwhile, mortality data in Indonesia averaged 17 cases per 100,000 populations of breast cancer patients [3]. Breast cancer is one of the most common types of cancer in the world and breast cancer causes death [4–6]. The investigators suggest that the high mortality rate from breast cancer is due to an inaccurate and acute initial diagnosis [7]. A Computer Aided Detection (CAD) system using machine learning (ML) and data mining approaches is needed to diagnose breast cancer [8].

In data mining, there are several approaches that can be used, including estimation, association and classification [9]. Classification technique is a popular approach to predict cancer [2, 6, 10, 11]. Classification algorithms include Decision Tree C4.5 (DT), Linear Regression, Logistic Regression (LR), Naïve Bayes (NB), Neural Network (NN), K-Nearest Neighbour (KNN) which is the topic of many researches [2, 5, 6, 12, 13]. There have been many applications of machine learning to predict breast cancer [2, 5, 6, 12, 13], but the machine learning algorithms used are still using the standard model. To get the results of predictions with high accuracy, it is necessary to set the correct parameters.

The results of identification in previous studies have built many machine learning models to predict breast cancer [2,4-13]. But the model built still uses the standard parameters of the machine learning algorithm used. The challenge of building a machine learning model using Optimization Parameters is when the computation is carried out, but it produces an optimal prediction model. In this study, we will use the tuning grid search parameters to be developed in the Logistic Regression (LR), K-Nearest Neighbour (KNN) and Decision Tree C4.5 (DT) algorithm for prediction of breast cancer. Furthermore, the focus of this research is devoted to building machine learning models using the Logistic Regression (LR) algorithm, K-Nearest Neighbour (KNN) and Decision Tree C4.5 (DT). After the development of the model is generated, a comparison between the models is carried out to get the best model by looking at the evaluation results of AUC and Accuracy.

2nd International Conference on Advanced Information Scientific Development (ICAISD) 2021
AIP Conf. Proc. 2714, 020018-1–020018-5; https://doi.org/10.1063/5.0128341
Published by AIP Publishing. 978-0-7354-4520-8/$30.00

PROPOSED METHODS

The dataset used in this study uses the Wisconsin breast cancer dataset (WBCD) from the UCI Machine Learning Repository. The dataset consists of 699 rows and 10 independent attributes with an interval value between 1-10 and the class attribute for malignant cancer with a value of 2 and a benign cancer with a value of 4. There were distributed data with 241 lines of observations for malignant cancer patients and 458 observation data for patients with benign cancer categories [2].

Initial Data Processing

The attributes of the dataset are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. In the Bare Nuclei attribute there is still missing data, so it is necessary to clean the data before it is used for training and testing the model that was formed. The method we use in this study is different from previous research [14] to obtain a machine learning model with Optimized parameters in the KNN algorithm.

Method

In the first step, we did a cleanup of the dataset by changing the missing value in the Bare Nuclei attribute with the average value. We used the 10-fold cross validation method to share 90% of the training data and 10% of the testing data. Then the data was trained and tested 10 times using the LR, KNN and C4.5 algorithms with manual optimization parameters. The second step, LR, KNN and C4.5 parameters will be optimized using the tuning grid optimized parameter 10 times. So the accuracy, AUC and ROC values can be compared and the mean of all iterations between LR, KNN and C4.5 can be compared with the t-test. The final step is to compare the Accuracy value using the t-test to test whether h1 is accepted or rejected.

The proposed method can be seen in Figure 1. In the proposed method, the preprocessing process produces a new data set with the missing value that has been changed to the average value. Then the new dataset will be transferred to training data and testing data using 10-fold cross validation. Then perform parameter optimization in each model to produce the Accuracy and AUC values. The results from Accuracy and AUC will be compared with previous research and other machine learning.

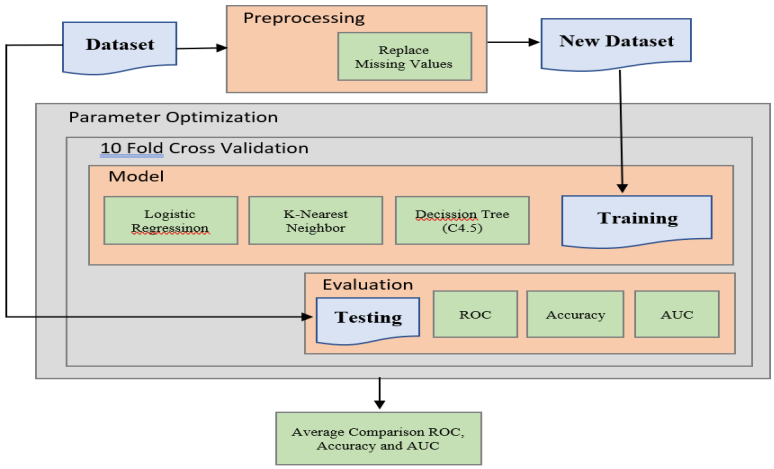

FIGURE 1. Proposed Methods

RESULTS AND DISCUSSION

Based on experiments, the results of parameter optimization manually can be seen in Table 1.And the results of parameter optimization by grid search can also be seen in Table 1.

TABLE I. Classifier Optimization Experiment Comparizon

No	Classifier	Accuracy		AUC	
		Manually	Optimization Parameter	Manually	Optimization Parameter
1	OpDecission Tree C4.5 (DT)	95.60	96.00	95.30	96.00
2	K-Nearest Neighbour (KNN)	96.90	97.40	98.90	99.30
3	Logistic Regression (LR)	96.40	96.90	99.40	99.50

Table 1 shows the comparison between manual parameter setting and parameter optimization using grid search which results in different values for Accuracy and AUC. Where the smallest accuracy value is in the DT algorithm and the largest accuracy value is in the KNN algorithm. There is an increase in the accuracy value of 0.50 in the KNN algorithm by implementing parameter optimization using grid search. The increase in the AUC value is also seen in the KNN algorithm, which has increased by 0.40.

TABLE II. Comparison Means of Accuracy

Method	Accuracy
KNN[14]	94.35
KNN[7]	95.80
DT[7]	95.80
LR[7]	95.80
LR[15]	95.26
DT[15]	93.25
KNN Proposed Method	97.40

Table 2 shows that optimization of machine learning parameters using grid search has a positive effect on the Accuracy value even on previous work [7, 14, 15]. The Accuracy value was obtained at 97.40 on the KNN algorithm. The comparison of average accuracy is also shown in Figure 2.

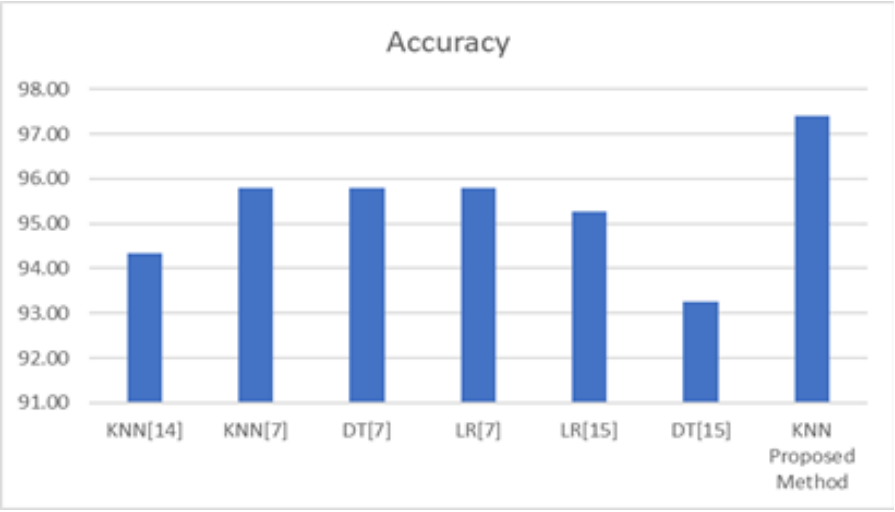

FIGURE 2. Comparison of Average Accuracy

Based on the two-sample paired t-test that has been carried out, the results can be seen in table 3. In table 3, the Accuracy value is compared with the manual parameter method and the application of optimization using the grid

search. From the comparison results obtained t count value of 0.002532 and t table of 4.302653 which means it can be concluded that H0 is rejected and H1 is accepted.

TABLE III. T-Test result between Manually Classifier and Classifier Optimization Parameter

t-Test: Paired Two Sample for Means		
	Variable 1	Variable 2
Mean	96.3	96.76667
Variance	0.43	0.503333
Observations	3	3
Pearson Correlation	0.999519	
Hypothesized Mean Difference	0	
Df	2	
t Stat	-14	
P(T<=t) one-tail	0.002532	
t Critical one-tail	2.919986	
P(T<=t) two-tail	0.005063	
t Critical two-tail	4.302653	

CONCLUSION

In this study, the proposed method is the optimization of Logistic Regression (LR), K-Nearest Neighbour (KNN) and Decision Tree C4.5 (DT) parameters using a Grid Search inspired by research [14]. The experimental results in this study obtained the highest accuracy value of 97.40% on the WBCD dataset with the algorithm model (KNN), the DT algorithm with a value of 96.00% and the LR algorithm of 96.90%. The results of comparisons with previous researchers show that the accuracy of the KNN increases after parameter optimization is applied.

From the t-test conducted, it shows that H1 is accepted, which means that the t-test results show a significant difference between the two models being compared. The proposed method has an increasing accuracy value compared to the previous method. And also the method in previous research shows that the Accuracy value increases with the application of parameter optimization that we do.

REFERENCES

1. Kementerian Kesehatan RI, "Jenis Kanker ini Rentan Menyerang Manusia," (2019), https://www.kemkes.go.id/article/view/20011400002/jenis-kanker-ini-rentan-menyerang-manusia.html.
2. Z. Khandezamin, M. Naderan, and M. J. Rashti, "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier," Journal of Biomedical Informatics **111** (2020).
3. Kemenkes RI, "Hari Kanker Sedunia 2019," (2019), https://sehatnegeriku.kemkes.go.id/baca/fokus-utama/20190131/2329273/hari-kanker-sedunia-2019/.
4. T. A. Assegie and P. S. Nair, "The Performance Of Different Machine Learning Models On Diabetes Prediction," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH **9**, 2491–2494 (2020).
5. T. A. Assegie and S. J. Sushma, "A Support Vector Machine and Decision Tree Based Breast Cancer Prediction," International Journal of Engineering and Advanced Technology (IJEAT) , 2972–2976 (2020).
6. A. Kaur and P. Kaur, "Breast Cancer Detection and Classification using Analysis and Gene-Back Proportional Neural Network Algorithm," International Journal of Innovative Technology and Exploring Engineering (IJITEE) , 2798–2803 (2019).
7. P. Gupta and S. Garg, "Breast Cancer Prediction using varying Parameters of Machine Learning Models," Procedia Computer Science **171**, 593–601 (2020).
8. D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," IOP Conference Series: Materials Science and Engineering **495**, 0–16 (2019).
9. H. Rianto, Amrin, Rudianto, O. Pahlevi, P. Kusumawardhani, and S. S. Hadi, "Determining the Eligibility of Providing Motorized Vehicle Loans by Using the Logistic Regression, Naive Bayes and Decission Tree (C4.5)," Journal of Physics: Conference Series **1641** (2020), 10.1088/1742-6596/1641/1/012061.
10. S. Thawkar and R. Ingolikar, "Classification of masses in digital mammograms using Biogeography-based optimization technique," Journal of King Saud University - Computer and Information Sciences **32**, 1140–1148 (2018).
11. R. Chtihrakkannan, P. Kavitha, T. Mangayarkarasi, and R. Karthikeyan, "Breast Cancer Detection using Machine Learning," International Journal of Innovative Technology and Exploring Engineering (IJITEE) **8**, 3123–3126 (2019).

12. S. Sharma, A. Aggarwal, and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) , 114–118 (2018).
13. S. N. Singh and S. Thakral, "Using Data Mining Tools for Breast Cancer Prediction and Analysis," 2018 4th International Conference on Computing Communication and Automation (ICCCA) , 1–4 (2018).
14. T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," Journal of Robotics and Control (JRC) **2**, 115–118 (2021).
15. A. Athar and A. K. Ilavarasi, "A Comparative study of machine learning models for breast cancer prediction," Journal of Physics: Conference Series **1716**, 012052 (2020).