JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org

Optimization of Breast Cancer Prediction using Optimazed Parameter on Machine Learning

Sri Nuarini[1] and Ade Rumintarsih[2]*

[1,2] *Nusa Mandiri University Jakarta, Indonesia*

ARTICLE HISTORY

Received: 13 March 23

Final Revision: 16 March 23

Accepted: 30 March 23

Online Publication: 31 March 23

KEYWORDS

RapidMiner, Machine Learning, Breast Cancer, Algorithm, Prediction

CORRESPONDING AUTHOR

14220016@nusamandiri.ac.id

DOI

10.37034/medinftech.v1i1.5

A B S T R A C T

Breast cancer is currently the most dangerous type of cancer that affects women. In order to classify the two types of cancer, this research aims to inform and educate medical professionals and cancer patients. The research project also aims to use machine learning (ML)-based data mining techniques. These algorithms include the K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Gradient Boosting Tree (Gboost), Logistic Regression (LR), and Support Vector Machine (SVM). The outcomes of these algorithms will determine the most prevalent cancer types that can be predicted. 10 characteristics from 683 patient samples with breast cancer were used in this study. Mammography and biopsy tests were used to gauge these traits. The K-Nearest Neighbor (KNN) algorithm produced the highest accuracy of 96.87% among the five algorithms, according to the results obtained using the K-Fold Validation operator. The researchers also improved the algorithm as a second comparison.

1. Introduction

In Indonesia, breast cancer is the most prevalent of the disease. Additionally, it ranks second among for cancer-related deaths [1]. Breast cancer can be diagnosed when a malignant tumor is found in the breast tissue. A malignant tumor is a type of tumor that spreads to nearby cells or even throughout the body. Breast cancer can affect both men and women, but is more common in women [2].

Cancer prediction can categories and staging of the cancer has become an important area in cancer research, as it can simplify the patient's subsequent clinical needs and determine effective treatment [3]. In breast cancer, early diagnostic can be the decisive point between life and death for patients with breast cancer.

Early diagnosis can determine what actions can be taken next to improve quality of life for breast cancer patients. Five years after diagnosis, a survival rate of 88% is estimated. While it decreases to about 80% after 10 years from diagnosis [4].

2. Research Method

2.1. Machine Learning

Machine learning techniques for classification, such as Naive Bayes, logistic regression, support vector machines, linear regression, Gaussian processes, decision trees, random forests, multi-layer perceptron's, and many others, have also been put into practice. The best results for increasing prediction accuracy come from using some feature selection techniques, such as filter methods, wrapper methods, and embedded methods [5].

With fully class-labeled data, supervised algorithms explore the connections between the data and the class. Regression or classification can be used to accomplish this. Similar to training and testing, classification involves two steps. Response variables are used to gather training data. Support Vector Machine (SVM), Discriminant Analysis, Nave Bayes, Nearest Neighbor, Neural Networks, and Logistic Regression are typical algorithms that fall under the classification category.

Unsupervised Algorithm for instruction detection, this unsupervised learning algorithm will try to find structured and hidden ways in unlabeled data. There is no training data for unsupervised learning. Where this unsupervised algorithm can be done by clustering or association analysis or dimension reduction. Clustering algorithms such as K-Means, and C-Means can be used [6].

This Wisconsin Breast Cancer dataset is originated from Kaggle.com. This dataset contains information

about breast cancer patients, including Patient ID, Clump Thickness, Cell Size, Cell shape, Adhesion, Epicell Size, Nuclei, Chromatin, Nucleoi, Mitoses, Class in Table 1.

Table 1. List of Feature

Name of Feature	Description	Domain
Clump of Thickness	Clump of thickness	1-10
Size of Cell	Uniforminty of cell size.	1-10
Shape of Cell	Uniformity of cell shape	1-10
Adhesion	Marginal adhesion	1-10
Epi Cell Size	Single epitaleal cell size.	1-10
Nuclei	Single epitaleal cell size	1-10
Chromatin	Bland chromatin	1-10
Nucleoi	Normal nucleoli	1-10
Mitoses	Mitoses	1-10
Class	Target variable	2-Benign 4-Malignant

2.2. Data and Process

The UCI Machine Learning Repository provided the data set (Original Wisconsin Breast Cancer Database). contains 683 illustrations. There are ten qualities. There are nine regular attributes and one special attribute (label), all of which have integer values, and the class value must be predicted.

In this particular example, the cancer cells are predicted to be benign if the class has a value of 2. Similarly, a value of 4 denotes the likelihood that the cancer cells are malignant. Pre-processing of data before it is used in an operation to ensure the best outcomes for a data set. The accuracy of the results may be lowered by the presence of noise, missing values or information, and unbalanced data in the data set. As a result, before running the machine learning model, we must remove these undesirable items from the dataset. Finding the original feature subset is the goal of feature selection. With various information-based methods, precision, accuracy, and prediction error [7] in Figure 1.

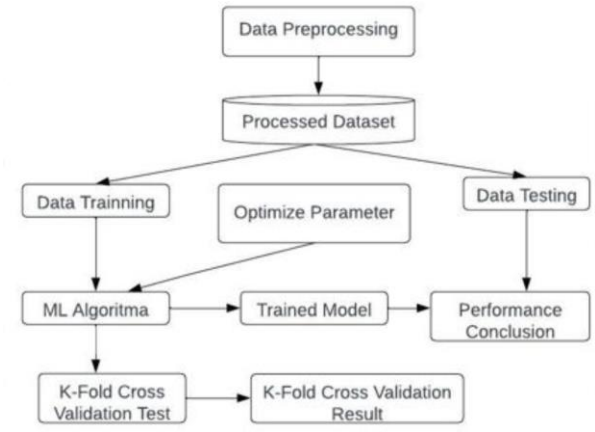

Figure. 1. Data flow Processing

Description:

a. Data collection: The first is to solving Machine Learning and collect relevant data sources.
b. Data processing: Data is cleaned using missing values, remove outliers, normalizing the data.
c. Identifying target variable: target variable determines the set of SL algorithms that can be applied.
d. Split the dataset: The dataset divided into training and testing subsets. And we have done 70:30 split, with 70% of the data for training and 30% for testing.
e. Training the model: The training subset of the dataset is applied to ML classification algorithms. We have applied nine algorithms to ML and each algorithm is trained differently.
f. Prediction: Input data to predict the results and the results are evaluated based on the model results. Include accuracy, precision, recall, f1-score, and support.

2.3. Decision Tree Algorithm

Decision trees provide a way to display the impact of any decision, through classification. Items within the data aid the mining process or facilitate the creation of forecasting models by continuously breaking down the data set into smaller, more specific groups.

The decision tree is constructed using the C4.5 algorithm. A decision tree is a structure that can be used to aggregate and divide large amounts of data into manageable groups of records by applying a set of decision-making rules [8].

The DT algorithm's goal is to create a tree data structure that can be used to forecast a class from a fresh case or record that doesn't yet have a class. C4.5 employs a divide strategy to construct a decision tree. First, the divide and conquer algorithm is used to create only the root node. By calculating and comparing the profit ratio, this algorithm can produce the best-case scenario. The divide and conquer algorithm will then be applied to the nodes created at the following level. until a leaf is formed, and so forth. A decision tree with box symbols for nodes and ellipses for leaves can be created by the C4.5 algorithm. Formula Entropy [9] in Formula (1).

$$\boldsymbol{Entropy}\ (\boldsymbol{S}) = -\sum_{i\#1}^{x} \boldsymbol{P_i} * \boldsymbol{Log_2}\ (\boldsymbol{P_1}) \qquad (1)$$

Where n is number of class S, p is Proportion of grades entered into the classroom at grade level i.

2.4. Random Forest Algorithm

ML algorithm called Random Forest aims to combine the results of various decision trees into a single output. A forest is made up of numerous trees that are acquired through a bootstrap bagging or aggregation process, as

is evident from the name alone. Only unbalanced data can be effectively handled by random forest.

It is resistant to anomalous data. is effective with non-linear data. less chance of overfitting. operates effectively with large data sets. superior to other classification algorithms in terms of accuracy. Oncologists in the field could thus use the Random Forest model to validate their diagnosis. To spread knowledge of and access to breast cancer treatment in rural areas, it may also be used. With timely treatment being a rarity in many developing countries, this will undoubtedly save many lives [10] in Figure 2.

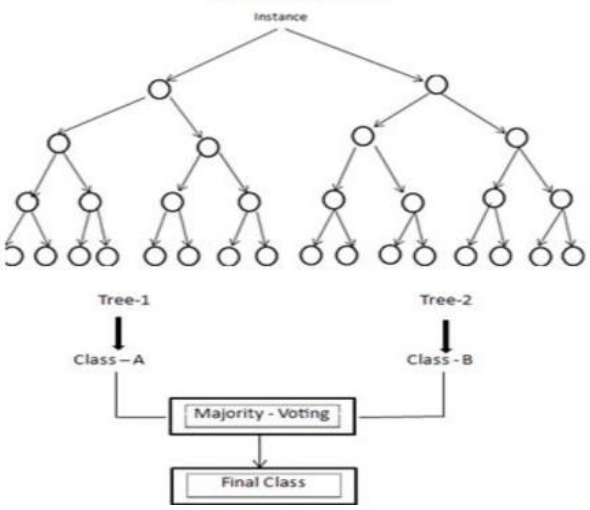

Figure 2. Random Forest

The figure above illustrates that the Random Forest algorithm has a difference with the Decision Tree algorithm. Where in Random Forest will produce smaller and more specific branches.

2.5. K-Nearest Neighbor

K-Nearest Neighbor Algorithm is a non-parametric algorithm. K-Nearest Neighbor is selected based on the Euclidean distance calculated between vector x and vector y given in equation. The result of KNN varies for different K. A large value of K will lead to overlapping classes, while a small value of K will increase computation. This algorithm is semi-supervised learning in nature which requires training data and a predefined k value. In KNN, k is the number of neighbors taken in making a decision. The KNN algorithm has a working principle that is looking for the nearest neighbor distance between the data to be evaluated and the training data. The following is an equation in determining the Euclidean distance in K-NearestNeighbor:

Formula Euclidien [11] in Formula (2).

$$Euclidien = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \qquad (2)$$

Where p is data dimension, X1 is Data train, and X2 is Data Set.

2.6. Naïve Bayes

Given two pieces of information—evidence E and a hypothesis H—Naive Bayes is an algorithm for probabilistic classification based on rules developed by a man by the name of Bayes. P(H|E) = P(H) P(E|H) / P(E) (1), where P(H|E) is the belief about the hypothesis after learning about E (called the posterior) and P(H) is the belief about H before learning about E (called the probability)

P(E|H) is the likelihood of H (L(H|E)) given E is a measure of how well H explains E.

2.7 Support Vector Machine

Support Vector Machines (SVM) is supervised learning algorithm to classify, predict, and discover outliers. SVM works effectively in high-dimensional spaces. To use an SVM to create and predictions for sparse data, however, it must be fitted on sparse data, for correct prediction [12].

2.8. Logistic Regression

Regression is a method of analytical modeling where a set of explanatory variables are linked to a likelihood that exists at some level. When analyzing a data set with one or more independent variables that affect the result, regression analysis is used. A binary variable (there are only two possible outcomes) is used to measure the result. It is used to forecast a binary result on a collection of independent variables (True/False, 1/0, Yes/No). The LR model is represented by the equations below [13] in Equation (3).

$$x = C_0 + \sum_{i=1}^{n} C_i X_i \qquad (3)$$

$$P_{(x)} = \frac{e^2}{1+ e^x}$$

Where Ci is the regression coefficient that can only be reached with the highest probability in relation to the usual error, and x is the percentage of the example variable Xi (i = 1,..., n) that is included in the analysis. A specific value of a variable, P(x), will be used to describe the probability of excitation in the outcome. In this study, the threshold is taken to be equal to or greater than 0.5, i.e., P(x) 0.5, leading to the classification of a record as an excitement [14].

The method used in this research is the *RapidMiner Machine Learning* method with six algorithms as a comparison. Starting with data collection, understanding the data, then continuing the process one by one. So as to produce an accuracy that can be used as knowledge.

2.9 . Performance Evaluation Matrix

Accuracy, precision, recall, and F-Score are a few examples of metrics that can be used to gauge how effective ML algorithms are. A few metrics are TP, TN, FP, and FN. Model Assessment Confusion matrix is a method that is frequently used to evaluate a model's performance. Confusion matrix generates four binary classifications: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The outcomes of this confusion matrix can then be used as a guide to assess the model's accuracy [15] in Equation (4).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (4)$$

$$\text{F} - \text{Score} = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

2.10. Replace Missing Value

Data mining processing is carried out on this algorithm using ML RapidMiner. Replace Missing Value is required. Because there are still some data that has not been filled incompletely.

2.11. Split Data

Split data is used to divide the data into two parts. That isa ratio of 70:30. Split the data into two parts 70% for trainingdata and 30% for testing data.

2.12 Cross Validation

The Cross Validation operator is used to evaluate the algorithm in separating two data subsets, namely training data, testing data so as to obtain maximum accuracy results.

2.13 Optimized Parameter

To improve the ML model, we first need to figure out what can improve the main hyper-parameter. The operator with the name optimized parameter must adjust the ML.

Model to a particular problem or dataset. Generally, ML models can be classified as supervised and non-passed learning algorithms, based on whether they are built for. Data obtained or unobtained from supervised learning research results. The algorithm itself is a series of machine-learning algorithms that maps the input. Identify targets through training on marked data, and in particular include Linear Model, K-Nearest Neighbors (KNN), Vector Engine Support (SVM), Bayes naive (NB), tree-based model of decision, anduse of learning algorithms.

3. Result and Discussion

The Wisconsin Breast Cancer Data Set as discussed in this research is using six kinds of algorithms on ML RapidMiner. These six algorithms are used as a comparison and to see which algorithm has a high accuracy value to predict the outcome.

The six algorithms are: Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosted Trees (XGBoost). And each algorithm has different accuracy results.

The six algorithms are: Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosted Trees (XGBoost). And each algorithm has different accuracy results.

3.1 K -Fold Validation

Results from earlier research are shown in Table 2. Additionally, it displays the comparison outcomes of the six algorithms. The LR algorithm has an accuracy score of 96,32, followed by KNN at 94,11, RF at 95,22, DT at 91,19, SVM at 93,46, and GBoost at 97,81 in Table 2.

Table 2. Previous Research Accuracy Table K-Fold

Algorithm	Accuracy
LR	96,32
KNN	94,11
RF	95,22
DT	91,19
SVM	93,46
XGBOOST	97,81

The comparison outcomes for the six algorithms are also shown in table 3. The accuracy of the LR algorithm is 96,86, which is higher than before, while the accuracy of the KNN is 96,87. This is also more accurate than the previous. RF is 96.03% accurate. In comparison to the previous algorithm, this one is more accurate. DT is 94.75% accurate. DT is also more accurate. Accuracy for SVM is 96,23. also greater accuracy. And the accuracy of XGBoost is 95,82. Only this algorithm, then, has a lower accuracy rate when compared to the earlier studies. Additionally, we can state that this study's findings are more accurate overall than those of earlier research in Table 3.

Table 3. Accuration Result K-Fold Validation

Algorithm	Accuracy
LR	96.86
KNN	98.87
RF	96.03
DT	94.75
SVM	96.23
XGBOOST	95.82

3.2 Confusion Matrix

From the Table 3, it can be observed that the classification results for benign and malignant tumors are obtained using six algorithms. The LR algorithm shows that there are 305 more instances of benign tumors compared to malignant tumors. Similarly, the KNN algorithm yields a higher numberof benign cases, with 304 compared to malignant cases. The RF algorithm also obtains a high number of benign cases, specifically 303, while the number of malignant cases is only 156. As for the DT algorithm, it produces a higher count of benign cases, namely 301, compared to 151 malignant cases. The SVM algorithm yields a higher count of benign cases as well, with 303 compared to 157 malignant cases. Finally, theXGBoost algorithm generates a confusion matrix where the count of benign cases is higher at 299 compared to 159 malignant cases in Table 4.

Table 4. Confusion Matrix Classification

Algorithm	Malignant	Benign	Dec
LR	305	9	Benign\
	6	158	Malignant
KNN	304	8	Benign\
	7	155	Malignant
RF	303	11	Benign\
	8	158	Malignant
DT	301	16	Benign\
	10	151	Malignant
SVM	303	10	Benign\
	8	157	Malignant
XGBOOST	299	8	Benign\
	12	159	Malignant

The Table 4 explains that the machine learning tool RapidMiner can also be used to determine accuracy, recall, precision, and F1 score values. The accuracy, recall, precision, and F1 score from the prior study are displayed in the table above. For this reason, the researcher will use the performance evaluation matrix method to compare the accuracy results of the six algorithms.

In the previous research results, it can be seen that the highest accuracy is achieved by the LR algorithm at 96.32. For recall, the highest value is obtained by the Gboost algorithm, which is 100. In terms of precision, the LR algorithm produces the highest value at 95.17. As for the F1 score, the Gboost algorithm obtains the highest value of 97.06. These are the calculated values from the Performanceevaluation matrix in the previous researcher's study in Table 5.

Table 5. Previous Result of Performance Evaluation Matrix

Algorithm	Accuracy	Recall	Precision	F-Score
LR	96.32	89.14	95.17	94.27
KNN	94.11	92.36	93.42	92.08
RF	95.22	93.56	92.38	94.38
DT	91.19	88.72	90.15	92.21
SVM	93.46	94.48	89.17	92.11
XGBOOST	97.81	100.00	94.30	97.06

It is clear from table 6 of the results of the performance evaluation matrix that KNN algorithm 96,87 is more accurate than the other six algorithms. In terms of recall, the XGboost algorithm outperforms the other six algorithms with a recall value of 95.26.

Looking at the precision calculation, it can be seen that the highest result is obtained by the LR algorithm with a value of 96.34. In terms of F1 score calculation, the DT algorithm has the highest value at 100 among the other six algorithms in Table 6.

Table 6. Result of Performance Evaluation Matrix present

Algorithm	Accuracy	Recall	Precision	F-Score
LR	96.86	94.52	96.34	95.42
KNN	96.87	95.15	95.72	94.53
RF	96.03	93.42	95.49	94.43
DT	94.57	90.44	93.78	100.00
SVM	96.23	95.41	94.98	94.98
XGBOOST	95.82	93.35	94.98	94.98

In the Table 7 can be shown that the accuracy result after the calculation is done using the operator optimize parameters. The RF algorithm has the highest accuracy of 100% among the other 5 algorithms. The second algorithm is KNN with a value of 98.55. Then the LR algorithm with an accuracy of 98.54. And also, the DT algorithm with accuracy 96.57. And the last is the Gboost algorithm with a 96.35 accuracy result.

Table 7. Result Optimized Parameter

Algorithm	Accuracy	Optimized
LR	96.86	98.54
KNN	96.87	98.55
RF	96.03	100.00
DT	94.57	96.57
SVM	96.23	98.08
XGBOOST	95.82	96.35

RapidMiner uses six algorithms, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Gradient Boosting Trees (Gboost), based on calculations with machine learning. When compared to other algorithms, the K-Nearest Neighbor algorithm had the highest accuracy outcome. Six algorithms are used by RapidMiner after processing the machine learning data for breast cancer prediction. K-Fold Validation is the operator used to determine the prediction of this breast cancer. The parameter optimize operator is then used to further optimize it. As a result, the accuracy of the data used to predict breast cancer is even higher. The Decision Tree algorithm (DT) produces the most accurate result. The algorithm that to be used to predict breast cancer is the Decision Tree. So Medical professionals can predict breast cancer early.

4. Conclusion

The conclusion that can be drawn is that this machine learning is very useful in predicting breast cancer patients. There are many algorithms in this ML. It can

provide the predictive value and desired accuracy. Cancer canbe classified as benign or malignant. These data can be used by oncologists in the field to confirm their diagnosis and increase the range of breast cancer awareness and treatment. This would result in many lives being saved, with timely care becoming rare in many developing countries.

References

[1] M. Shanbehzadeh, H. Kazemi-Arpanahi, M. Bolbolian Ghalibaf, and A. Orooji, "Performance evaluation of machine learning for breast cancer diagnosis: A case study," *Informatics Med. Unlocked*, vol. 31, p. 101009, 2022, doi: https://doi.org/10.1016/j.imu.2022.101009.

[2] G. Kaur, R. Gupta, N. Hooda, and N. R. Gupta, "Machine Learning Techniques and Breast Cancer Prediction: A Review," *Wirel. Pers. Commun.*, vol. 125, no. 3, pp. 2537–2564, 2022, doi: 10.1007/s11277-022-09673-3.

[3] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," *Procedia Comput. Sci.*, vol. 171, pp. 1251–1260, 2020, doi: https://doi.org/10.1016/j.procs.2020.04.133.

[4] N. Arya and S. Saha, "Multi-modal advanced deep learning architectures for breast cancer survival prediction," *Knowledge-Based Syst.*, vol. 221, p. 106965, 2021, doi: https://doi.org/10.1016/j.knosys.2021.106965.

[5] V. N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning," *Measurement*, vol. 178, p. 109442, 2021, doi: https://doi.org/10.1016/j.measurement.2021.109442.

[6] P. Kaur, G. Singh, and P. Kaur, "Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification," *Informatics Med. Unlocked*, vol. 16, p. 100151, 2019, doi: https://doi.org/10.1016/j.imu.2019.01.001.

[7] A. Conti, A. Duggento, I. Indovina, M. Guerrisi, and N. Toschi, "Radiomics in breast cancer classification and prediction," *Semin. Cancer Biol.*, vol. 72, pp. 238–250, 2021, doi: https://doi.org/10.1016/j.semcancer.2020.04.002.

[8] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," *IEEE Access*, vol. 9, pp. 71194–71209, 2021, doi: 10.1109/ACCESS.2021.3079204.

[9] A. Solichin, "Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation," in *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2019, pp. 217–222. doi: 10.23919/EECSI48112.2019.8977081.

[10] S. Kabiraj *et al.*, "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–4. doi: 10.1109/ICCCNT49239.2020.9225451.

[11] A. Budiyantara, I. Irwansyah, E. Prengki, P. Pratama, and N. Wiliani, "KOMPARASI ALGORITMA DECISION TREE, NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK MEMPREDIKSI MAHASISWA LULUS TEPAT WAKTU," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 265–270, 2020, doi: 10.33480/jitk.v5i2.1214.

[12] M. T R, A. C. Kaladevi, B. J M, V. Vivek, M. Prabu, and V. Muthukumaran, "An Efficient Ensemble Method Using K-Fold Cross Validation for the Early Detection of Benign and Malignant Breast Cancer ," *Int. J. Integr. Eng.*, vol. 14, no. 7 SE-Articles, pp. 204–216, Dec. 2022, [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/ijie/article/view/10810

[13] 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, p. 290, 2020, doi: 10.1007/s42979-020-00305-w.

[14] T. Thomas, N. Pradhan, and V. S. Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 192–196. doi: 10.1109/ICICT48043.2020.9112464.

[15] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020, doi: https://doi.org/10.1016/j.neucom.2020.07.061.