

Breast cancer identification using machine learning and hyperparameter optimization

Toni Arifin^{1,2}, Ignatius Wiseto Prasetyo Agung^{1,2}, Erfian Junianto^{1,2}, Rizal Rachman³

Ilham Rachmat Wibowo^{1,3}, Dari Dianata Agustin^{1,3}

¹ARS Digital Research and Innovation (ADRI), Bandung, Indonesia

²Department of Informatics Engineering, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University (ARS University), Bandung, Indonesia

³Department of Information Systems, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University (ARS University), Bandung, Indonesia

Article Info

Article history:

Received Jun 15, 2024

Revised Aug 6, 2024

Accepted Aug 11, 2024

Keywords:

Breast cancer
Classification
Gene expression
Hyperparameter optimization
Machine learning

ABSTRACT

Breast cancer identification can be analyzed through genomic analysis using gene expression data, one type of which is mRNA. This involves analyzing gene expression patterns of breast tissue samples to distinguish breast cancer from healthy tissue or to differentiate subtypes of different breast cancers. This research developed the right computational model for breast cancer classification using machine learning and hyperparameter optimization algorithms. The primary objective of this research is to utilize various machine learning algorithms to classify breast cancer based on gene expression and enhance the models developed in previous studies. This paper provides an extensive literature review of prior breast cancer classification research and offers new theoretical perspectives. This research used a problem-solving approach with conventional machine learning techniques, most notably the decision tree. It also evaluates other machine learning algorithms for comparison, including k-nearest neighbor, naïve bayes, random forest, extra tree classifier, and support vector machine. The evaluation process used classification reports that provide insight into the precision, recall, F1-score, and accuracy of each machine learning model. The evaluation results show that the performance of the decision tree algorithm model is superior and impressive, achieving 99.73% accuracy and a score of 1 for precision, recall, and F1-score.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Toni Arifin

Department of Informatics Engineering, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University (ARS University)

Bandung, Indonesia

Email: toni.arifin@ars.ac.id

1. INTRODUCTION

Breast cancer forms when cells in the breast develop abnormally and uncontrollably, if treated too late, these cells can spread quickly to the surrounding tissue, even can spread to other organs, which can be fatal [1]. Breast cancer is the most common type of cancer suffered by women [2]. From 2014 to 2018, breast cancer cases in women worldwide increased by 0.5% per year, and in 2020, breast cancer caused 685,000 deaths globally [1].

Identification of breast cancer must be done early. With better treatment, the emergence of breast cancer cells can be addressed immediately, stopping their spread. Breast cancer screening is an important strategy for early detection and ensures a greater chance for good outcomes, one of which is using

biomarkers [3]. Biomarkers are biological indicators that can be measured and used to indicate presence, or severity of a disease condition, biological process, or response to therapy. The use of biomarkers in medical has grown rapidly due to their ability to assist in diagnosis, monitoring and assessing response to treatment. Examples of biomarkers are proteins, enzymes, genes, antigens, antibodies, and other various small molecule that can be measured or tested in biological sample [4].

Diagnosis based on gene or gene expression data is one of the main topics in cancer classification [2]. gene data analysis can help identify cancer growth, especially breast cancer [5], Gene expression data consists of DNA, RNA, protein and epigenetic levels [2], [5]. Gene data varies criteria and complexity, the difficulty in analyzing this data depends on various factors, including the type of gene data, the technique used and the data source [6]. One type of gene expression is RNA. Gene RNA data analysis differs in difficulty from other types of gene data, it is more complex due to several variations. There are three main types of RNA, one of which is messenger RNA (mRNA). The diverse mRNA expression patterns in breast cancer can be used for detailed classification of the disease. These mRNA biomarkers can aid in early detection, selecting appropriate therapy, and monitoring disease progression [7]. The mRNA has a large data volume and provides detailed information about gene expression in a specific sample [8], his detailed information is obtained because mRNA is one of the main molecules in biology, playing an important role in genetic expression, containing genetic location information, nucleotide sequences, codon sequences, exons, and introns [9], as a result, this data becomes complex and complicated [10], requiring an exact computing model for further analysis [5].

In the last few decades, computing techniques have been rapidly growing, including the use of a variety of techniques such as pattern recognition, machine learning, genetic data analysis, and artificial intelligence [11]. Identification by analyzing genes and developing machine learning models is one solution that can be applied to identify breast cancer [12]. Like research conducted by [13], research [14] and research [15] which has developed a machine learning model for breast cancer, However, the results are not optimal because gene expression data is so complicated and complex. Therefore, further research on breast cancer identification based on gene expression still needs to be carried out, and the final results need improvement through further analysis and the selection of appropriate parameters to handle such complicated and complex data types.

In this research machine learning algorithms are selected based on their ability to handle large and complex data, and to overcome the problem of overfitting. Some of these algorithms include: naïve bayes [16], [17], support vector machine [18], random forest [19], decision tree [20], k-nearest neighbor [21] and extra tree classifier [22]. In its application the machine learning algorithm has several shortcomings, These include being ineffective in handling missing data or unbalanced data and requiring the selection of appropriate parameters to produce an accurate model [23]. The solution to this problem is to apply data preprocessing to remove noise in the data and a combination of machine learning algorithms with hyperparameter optimization [19], [24]. The purpose of applying hyperparameter optimization is to strengthen the performance of the machine learning model [25] This is achieved by choosing the right combination of parameters that will be used during training [26]. With this technique machine learning algorithms can be adjusted to datasets and certain more specific problems, thus resulting in increased accuracy [27].

This research uses gene expression data for breast cancer classification and improves existing machine learning models. The machine learning model was created using the Python programming language and uses a Machine Learning algorithm for classification of breast cancer types based on gene expression mRNA data. The main objectives and contributions of this research are as follows: i) analyze and apply data preprocessing to gene expression mRNA data to enable further processing, ii) design a machine learning model for breast cancer identification based on reliable and accurate gene expression mRNA data, iii) uses a machine learning algorithm for detailed classification of breast cancer types, iv) choose the right parameters and select the most suitable machine learning algorithm for breast cancer identification with the help of hyperparameter optimization, v) compare machine learning models for breast cancer identification with previous research models, and vi) gain new insights regarding the implementation of machine learning with hyperparameter optimization for breast cancer identification based on Gene expression mRNA.

2. LITERATURE REVIEW

Research by Chen *at al* [28], utilized breast cancer data from Metabric, TCGA and GEO to employ machine learning models to predict breast cancer based on the immune subtype of TNBC patients requiring ICB. This research involved the analysis of Bioinformatics techniques and resulted in the identification of 11 hub genes, utilizing the random forest method. The results showed an AUC value of 0.76. Further research was conducted by El-Nabawy *at al* [13], which utilized the Metabric dataset comprising clinical, gene

expression, CAN, CNV data, and histopathological images. Supervised learning algorithms were employed, with linear-SVM and E-SVM algorithms achieving the highest accuracy of 97.1%. Other research was conducted by Mucaki *et al* [15], utilizing the Metabric dataset and machine learning algorithms to identify precision genes based on the biochemical response to chemotherapy in breast cancer cells. The selected genes comprised 15 attributes, including ABCC10, BCL2, BCL2L1, BIRC5, BMF, FGF2, FN1, MAP4, MAPT, NKFB2, SLCO1B3, TLR6, TMEM243, TWIST1, and CSAG2. The SVM algorithm emerged as the superior machine learning algorithm in this research, achieving an accuracy of 84%. Next, the research conducted by Zhao *et al* [14], applied the K-Means method to select training data with random samples from the Metabric dataset. They then applied machine learning classification methods, with the results showing that Random Forest and SVM produced an accuracy of 72.9%.

Research by Thalor *et al* [29], utilizes breast cancer data based on the TNBC immune subtype and applies a machine learning algorithm. The data is initially processed using z-score normalization and feature selection techniques, including Pearson's correlation coefficient to reduce features and Recursive Feature Elimination (RFE) to select the most relevant features. the recursive feature elimination with random forest classifier (REFRF) method is then employed to produce 7 features from 1,150 samples. The results of this research indicate that the XGBoost algorithm using REFRF produces an AUC value of 0.99.

Khorshed *et al* [30], conducted research using a deep learning approach to diagnose various types of cancer, including breast cancer. The method utilized a convolutional neural network architecture modified and named gene expression network (GeneXNet), designed to handle complex data such as gene expression data. The data used in this research comprised 33 different types of cancer from 26 organs of the human body. The results of the study indicate that the designed classification model achieved an accuracy of 98.9%. Other research was conducted by Hussein and Al-Sarray [31], uses machine learning and deep learning approaches optimized with genetic algorithm algorithms for breast cancer classification, the results of this research show that the GA-CNN algorithm produces the highest accuracy, namely 97.76%.

3. METHOD

The testing stage for the method requires a clear and precise proposed method. The proposed method will be used as a stage in the research process to obtain the desired results, as depicted in Figure 1, this design provides a clear understanding of the research stages and is presented systematically, some of these stages include: i) the first stage starts with collecting data from METABRIC (breast cancer mRNA), followed by ii) inspection data, iii) preprocessing data, iv) implementing hyperparameter optimization, v) implementing machine learning algorithms, vi) evaluation and validation, and vii) performance report. This approach will help ensure transparency and replicability, which are fundamental aspects of any scientific investigation.

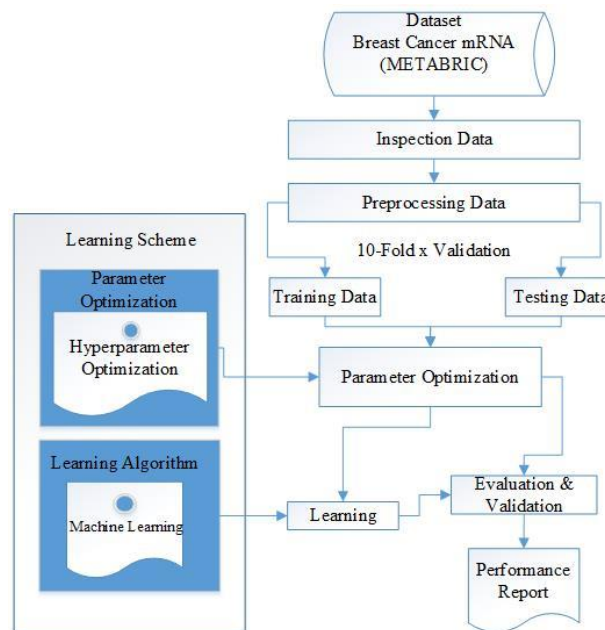


Figure 1. Proposed method

3.1. Breast cancer messenger RNA (mRNA) dataset

In this research, the dataset used was the breast cancer messenger rna dataset from the molecular taxonomy of breast cancer international consortium (METABRIC). The METABRIC dataset is a collection of data utilized in breast cancer research, combining clinical, pathologic, and molecular information from thousands of breast cancer tumor samples. This dataset comprises genomic data, such as genetic mutations, gene expression, and epigenetic changes, as well as clinical information, including age, gender, tumor size, and other risk factors. These data have been utilized in numerous studies to comprehend the diversity of breast cancer and identify potential biomarkers for breast cancer diagnosis [14], [32]. The data used in this research are 692 attributes and 1,904 data.

3.2. Inspection data

Data inspection in machine learning involves examining and understanding the data before applying any preprocessing or building models. This technique helps make informed decisions about data preprocessing and feature optimization, ultimately leading to better-performing machine learning models. Inspection data is a stage of data analysis used to observe data and determine the appropriate data processing stage that can be applied to the data. By understanding the characteristics and quality of the data, we can ensure that the analysis carried out will be based on accurate and relevant information [5], [12], [25]. Figure 2 is an example of inspection data implementation using the Python programming language.

```
dg = pd.read_csv('/content/drive/MyDrive/Dataset/METABRIC_RNA_Mutation_Edit 27 Februari 2024.csv')

dg.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1904 entries, 0 to 1903
Columns: 693 entries, patient_id to siah1_mut
dtypes: float64(496), int64(179), object(18)
memory usage: 10.1+ MB

print("data:", dg.shape)

data: (1904, 693)
```

Figure 2. Inspection data

3.3. Data preprocessing

Data preprocessing in machine learning involves a series of steps to prepare raw data for building and training models. It is a crucial step that can significantly impact the performance and accuracy of the model. One of the most difficult aspects of developing a learning model for the healthcare field is data preprocessing. High-quality data is essential for a machine learning model to achieve effective training and optimal performance in terms of accuracy. The preprocessing data stage is essential before entering the classification stage, as it involves processing the data to facilitate further analysis. In general, data preprocessing is a process aimed at addressing issues such as noisy data, missing values, duplicate data, and contradictory data [23]. Figures 3 and 4 as shown explain data preprocessing.

```
[ ] dg = dg.fillna(0)

[ ] dg.head()
```

Figure 3. Preprocessing missing value data

```

kolom_kategorikal = dg.select_dtypes(include=['object']).columns

print("Kolom-kolom kategorikal:")
print(kolom_kategorikal)

for col_name in dg.columns:
    if(dg[col_name].dtype == 'object'):
        dg[col_name] = dg[col_name].astype('category')
        dg[col_name] = dg[col_name].cat.codes
    if(dg[col_name].dtype == 'object'):
        dg[col_name] = dg[col_name].astype('category')
        dg[col_name] = dg[col_name].cat.codes
    if(dg[col_name].dtype == 'object'):
        dg[col_name] = dg[col_name].astype('category')
        dg[col_name] = dg[col_name].cat.codes
    if(dg[col_name].dtype == 'object'):
        dg[col_name] = dg[col_name].astype('category')
        dg[col_name] = dg[col_name].cat.codes

```

Figure 4. Preprocessing data

3.4. Training data and testing data

The next stage involves dividing the data into training and testing sets, aiming to develop the model and objectively measure the performance of the designed machine learning model. The technique used in this research is cross validation with K-Fold 10. Another purpose of implementing cross validation is to evaluate more accurately and ensure that the model not only learns from the training data but can also generalize well to unseen data, as demonstrated in research [23]. Figure 5 as shows the implementation of cross validation.

```

kf =KFold(n_splits=10, shuffle=True, random_state=42)
cnt = 1
# split() method generate indices to split data into training and test set.
for train_index, test_index in kf.split(x, y):
    print(f'Fold:{cnt}, Train set: {len(train_index)}, Test set:{len(test_index)}')
    cnt += 1

```

Figure 5. Cross validation

3.5. Machine learning model

Machine learning is a technology that involves designing computational algorithms to emulate human intelligence and learn from the surrounding environment. In machine learning, systems are developed and trained using large datasets to handle highly complex tasks. The machine learning model analyzes that data, and based on that trained model, we can make predictions about the future [28]. In this stage we discuss several machine learning algorithms that will be used, these algorithms are selected based on their ability to handle large and complex data including, naïve bayes, support vector machine, random forest, decision tree, k-nearest neighbor and extra tree classifier.

3.5.1. Naïve bayes

The naïve bayes classifier is a classification algorithm based on Bayes theorem. This algorithm forecasts future outcomes by leveraging past experiences. Its key feature is the strong assumption of independence between conditions or events. Naïve bayes performs exceptionally well compared to other classification algorithms and requires only a small amount of training data to estimate the parameters needed for classification. Since it assumes independence among variables, it only requires the variance of a variable within a class for classification, rather than the entire covariance matrix [16], [17].

3.5.2. Support vector machine

Support vector machine is a machine learning algorithm used for classification, estimation, and prediction. This algorithm works based on structural risk minimization, which processes data into

hyperplanes to classify the input space into two classes. The theory of the support vector machine (SVM) starts with grouping linear cases that can be separated by hyperplanes and divided according to their classes. This algorithm is often applied to large datasets to solve classification problems in research [18].

3.5.3. Decision tree

The decision tree is a structured algorithm that resembles a tree, with a root node, internal nodes representing features of the dataset, branches indicating decision rules, and leaf nodes representing outcomes. It is capable of analyzing data to reveal hidden relationships between input variables and target variables. Additionally, the decision tree simplifies complex decision-making processes into more manageable steps, enabling decision-makers to interpret solutions more effectively. Another term for the decision tree is classification and regression tree (CART), which combines two types of trees: the classification tree and the regression tree [20]. It is an appropriate and effective algorithm used in medical decision-making due to its ease of implementation and ability to produce high accuracy, even when dealing with incomplete or noisy data [33].

3.5.4. Random forest

Random forest is one of the most popular machine learning algorithms. It comprises a collection of decision trees that are combined into a single model. In random forest, decision trees are recursively divided based on data within the same class. Using a large number of trees in this algorithm tends to improve the accuracy obtained, making it more optimal. Additionally, random forest has the ability to handle complex data, many features without overfitting, and can handle imbalanced data. These characteristics make random forest a reliable and effective algorithm for various prediction and classification problems [19].

3.5.5. K-nearest neighbor

K-nearest neighbor is a machine learning algorithm that classifies objects based on learning data whose distance is closest to the object's distance, which can be calculated using Euclidean, Manhattan, or Minkowski distance. Essentially, this algorithm searches for training data that is most similar to the test data that will be classified and assigns it the same class label as the training data. The advantage of this algorithm is its effectiveness in handling large datasets, noisy data, and flexibility in handling data with various types of features [21].

3.5.6. Extra tree classifier

The extra tree classifier is a variation of random forest used in machine learning for estimation, prediction, and classification problems. This algorithm is a type of ensemble learning, resulting from a combination of several decision trees to produce more accurate output. It requires setting parameters such as the number of trees, maximum tree depth, and the size of the considered feature subset, all of which can affect algorithm performance [22].

3.6. Hyperparameter model

Hyperparameter tuning is one of the techniques used in developing machine learning models to achieve optimal performance. It involves adjusting parameters that influence the model's learning from data. By finding the optimal combination of hyperparameters, the model can easily achieve higher accuracy and better overall results. In this research, the hyperparameter tuning technique used is grid search combined with cross-validation (CV). This approach combines the grid search method for tuning hyperparameters with the use of cross-validation to objectively evaluate model performance [34]. Grid search CV is crucial for optimizing hyperparameters in each machine learning algorithm. It enhances the achievement of optimal final results by optimizing the parameters for each algorithm used, as illustrated in the Table 1, which depicts the parameters using grid search CV.

Table 1. Machine learning algorithm parameter using grid search CV

Algorithm	Parameter
Naive Bayes	{ 'var_smoothing': np.logspace (0,-9, num=100) }
Support vector machine	{ 'C' :1,'gamma':1, 'kernel': ('linear'), 'random_state': 42 }
Random forest	{ 'n_estimators':200,'min_samples_leaf':1,'min_samples_split': 2,'random_state': 42 }
Decision tree	{ "max_depth": (None),"min_samples_split": 10,"min_samples_leaf": 1,"random_state": 42 }
K-nearest neighbor	{ 'n_neighbors': 20,'weights': ('uniform'),'algorithm': ('auto') }
Extra tree classifier	{ 'n_estimators': 100, 'max_depth': 70, 'min_samples_split': 10,'min_samples_leaf': 1,'random_state': 4 }

3.7. Evaluation model

This stage involves evaluating the performance of the model being designed. In this research, the evaluation model is a classification report, which describes the performance of each method used. This stage is crucial for ensuring the reliability and capability of the classification algorithm, especially in the health sector. Various methods are employed to assess the performance and robustness of the machine learning model. During testing, evaluation metrics such as accuracy, recall, precision, and F1-score are utilized. Accuracy measures the quality of the training data used in forming the machine learning model. Recall assesses the ability of the model to find all positive instances, while precision measures the accuracy and reliability of the classification model. The F1-score represents the harmonic average of weighted precision and recall, providing a balance between the two. precision, recall, F1-score, and accuracy calculations can be determined using the [35]:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

$$\text{F1 - Score} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (4)$$

4. RESULTS AND DISCUSSION

After completing all stages using a previously designed model, the decision tree algorithm proves to be more accurate in identifying breast cancer based on gene expression messenger RNA data. In this experiment, the decision tree algorithm demonstrates its advantages as a simple classification algorithm with fast classification speed, ease of extraction of explicit rules, and high accuracy classification for large-scale data processing. Moreover, parameter optimization with the hyperparameter optimization technique significantly enhances the performance of machine learning models by improving generalization and robustness. This is evident from the combination of decision tree with hyperparameter optimization, resulting in much better and more impressive results. These findings were obtained through experiments comparing the outcomes of experiments that solely utilized machine learning algorithms with those that incorporated hyperparameter optimization. Importantly, these results were based on tests conducted on each algorithm using separate test data, distinct from the training data.

In this test, the evaluation results are obtained in the form of a classification report for each algorithm used, along with graphs depicting the evaluation results for each algorithm. Table 2 explains the classification report using a machine learning algorithm without hyperparameter optimization. This report indicates that the extra tree classifier algorithm performs better at identifying breast cancer. Additionally, Figure 6, which consists of sub-images (a) accuracy, (b) precision, (c) recall, and (d) F1-score, presents graphs showing the evaluation and validation results for each algorithm. Meanwhile, Table 3 and Figure 7, which also consists of sub-images (a) accuracy, (b) precision, (c) recall, and (d) F1-score, present a classification report and graph, respectively. These depict the results of classification using machine learning algorithms with hyperparameter optimization. The report and graph show that the decision tree algorithm outperforms other algorithms. Furthermore, the extra tree classifier also demonstrates improved evaluation results compared to previous experiments.

Table 3 and Figures 7(a)-(d) demonstrate that the evaluation results of the decision tree + hyperparameter optimization algorithm are superior and highly favorable compared to other algorithms. This is evident from the Accuracy, precision, recall, and F1-Score values, which exhibit maximum precision and recall results, leading to optimal F1-score results. The effectiveness of the proposed model can be attributed to several factors. Firstly, data preprocessing is applied at the beginning of the process, simplifying complex decision-making processes for the decision tree algorithm. Additionally, hyperparameters are optimized to maximize the parameters of the decision tree, enabling better handling of the overfitting problem. To further illustrate the effectiveness of the proposed model, this research conducted a comparison with similar previous studies, some of these research studies include Zhao *at al* [14] applying K-means to select training data with random samples and SVM, El-Nabawy *at al* [13] linear-SVM and E-SVM with the most value and Mucaki *at al* [15] with 15 selected attributes and SVM became the best algorithm. The comparison results can be seen in Table 4.

Table 2. Results of the classification report evaluation without hyperparameter optimization

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision tree	90.66%	0.82	0.89	0.84
Random forest	79.63%	0.61	0.78	0.68
Extra tree	91.39%	0.87	0.89	0.85
K-nearest neighbor	66.81%	0.63	0.64	0.64
Support vector machine	78.00%	0.61	0.78	0.68
Naïve Bayes	79.63%	0.82	0.89	0.84

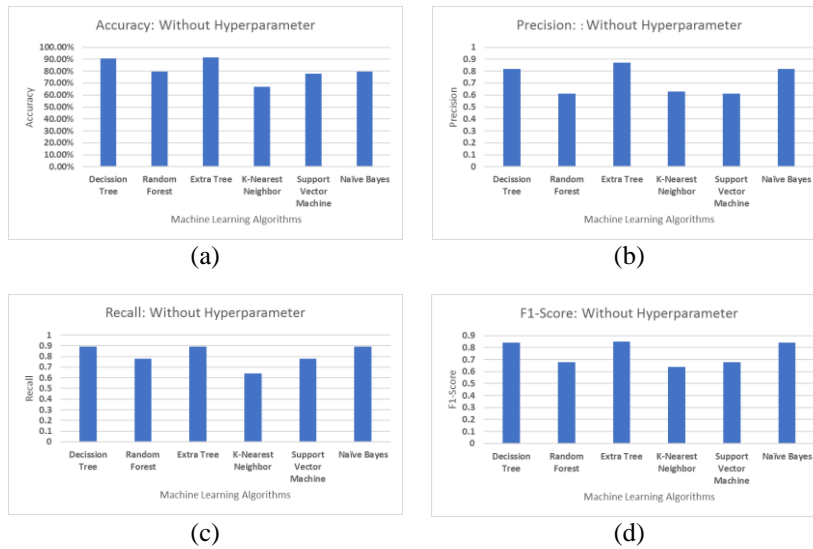


Figure 6. The evaluation results without hyperparameter optimization are presented in (a) accuracy, (b) precision, (c) recall, and (d) F1-score

Table 3. Evaluation results of the classification report with hyperparameter

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision tree	99.73%	1.00	1.00	1.00
Random forest	96.43%	0.98	0.96	0.94
Extra tree	92.17%	0.88	0.90	0.87
K-nearest neighbor	79.63%	0.61	0.79	0.72
Support vector machine	93.17%	0.91	0.92	0.91
Naïve Bayes	98.01%	0.98	0.98	0.98

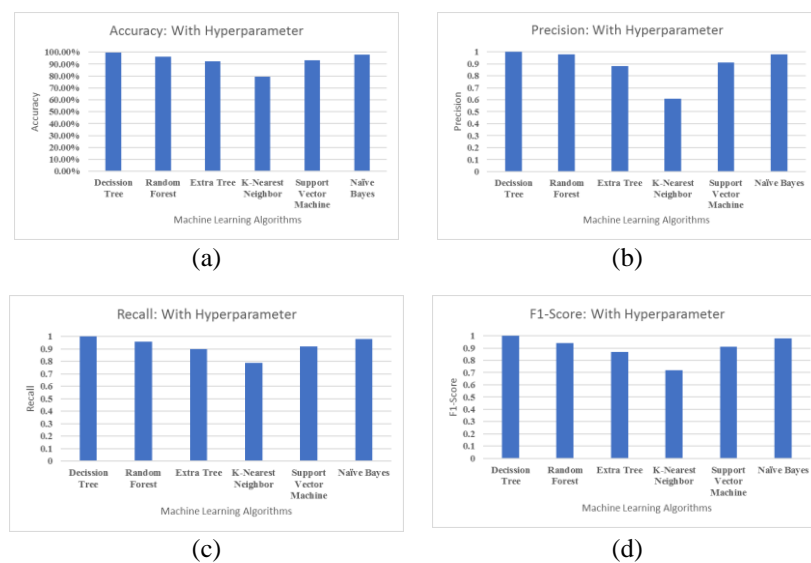


Figure 7. Evaluation results with hyperparameter are shown in (a) accuracy, (b) precision, (c) recall, and (d) F1-Score

Table 4. Comparison of the machine learning model with previous research

Research	Machine learning algorithm	Accuracy
Zhao <i>et al</i> [14]	Random forest & SVM	97.1%
Nabawy <i>et al</i> [13]	Linear SVM & E-SVM	84%
Mucaki <i>et al</i> [15]	SVM	72.9%
This research	Decision tree	99.73%

5. CONCLUSION

Our research on breast cancer classification using gene expression data, machine learning, and hyperparameter optimization algorithms has successfully created an accurate model to help health experts identify breast cancer types. This study achieved a remarkable accuracy rate of 99.73%, along with precision, recall, and F1-score values of 1, demonstrating the effectiveness of the decision tree algorithm. The results indicate that the decision tree, combined with hyperparameter optimization, surpasses other machine learning algorithms and demonstrates its potential in classifying breast cancer based on gene expression data, as observed in other studies. Nevertheless, there are still opportunities for further improvements that can be explored in future research. The following points outline potential areas for further investigation: i) increasing the quantity and diversity of data: by expanding the dataset, researchers can assess the model's performance and its ability to handle various types of breast cancer data, ii) comparing with other hyperparameter optimization techniques: analyzing alternative methods such as Bayes search CV and Random Search CV can help determine which parameter optimization technique yields the best performance, and iii) properly tuning model hyperparameters can result in better generalization and robustness, leading to increased predictive accuracy.

ACKNOWLEDGEMENTS

The author wishes to express gratitude to the Directorate of Research, Technology, and Community Service of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for funding this research.




REFERENCES

- [1] WHO, "Breast Cancer," 2023. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed Feb. 07, 2024).
- [2] Y. Zhang, Q. Deng, W. Liang, and X. Zou, "An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data," *Biomed Research International*, vol. 2018, 2018, doi: 10.1155/2018/7538204.
- [3] E. Tarighati, H. Keivan, and H. Mahani, "A Review of Prognostic and Predictive Biomarkers In Breast Cancer," *Clinical and Experimental Medicine*, vol. 23, no. 1, pp. 1–16, 2023, doi: 10.1007/s10238-021-00781-1.
- [4] C. J. Li, H. M. Chen, and J. C. Lai, "Diagnostic, Prognostic, and Predictive Biomarkers in Breast Cancer," *Journal Oncology*, vol. 2020, 2020, doi: 10.1155/2020/1835691.
- [5] L. Venkataramana, S. G. Jacob, S. Saraswathi, and D. V. V. Prasad, "Identification of Common and Dissimilar Biomarkers for Different Cancer Types from Gene Expressions of RNA-sequencing Data," *Gene Reports*, vol. 19, 2020, doi: 10.1016/j.genrep.2020.100654.
- [6] J. C. Silva, D. S. Domingues, D. Menotti, M. Hungria, and F. M. Lopes, "Temporal Progress of Gene Expression Analysis with RNA-Seq Data: A Review n The Relationship Between Computational Methods," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 86–98, 2023, doi: 10.1016/j.csbj.2022.11.051.
- [7] M. L. Zheng *et al.*, "Circulating Exosomal Long Non-Coding RNAs in Patients With Acute Myocardial Infarction," *Journal of Cellular and Molecular Medicine*, vol. 24, no. 16, pp. 9388–9396, 2020, doi: 10.1111/jcmm.15589.
- [8] A. S. Thind *et al.*, "Demystifying Emerging Bulk RNA-Seq Applications: The Application And Utility Of Bioinformatic Methodology," *Briefings in Bioinformatics*, vol. 22, no. 6, pp. 1–16, 2021, doi: 10.1093/bib/bbab259.
- [9] S. Das, M. Vera, V. Gandin, R. H. Singer, and E. Tutucci, "Intracellular mRNA Transport And Localized Translation," *Nature Reviews Molecular Cell Biology*, vol. 22, no. 7, pp. 483–504, 2021, doi: 10.1038/s41580-021-00356-8.
- [10] M. Khalsan *et al.*, "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," *IEEE Access*, vol. 10, pp. 27522–27534, 2022, doi: 10.1109/ACCESS.2022.3146312.
- [11] A. Sebastian and L. Jacob, "Breast Cancer Survival Prediction using Gene Expression Data," *13th Int. Conf. Adv. Comput. Control. Telecommun. Technol. ACT 2022*, vol. 8, pp. 352–359, 2022, doi: 10.1101/2022.01.22.22269470.
- [12] T. Takeshita, H. Iwase, R. Wu, D. Ziazadeh, L. Yan, and K. Takabe, "Development of a Machine Learning-Based Prognostic Model for Hormone Receptor-Positive Breast Cancer Using Nine-Gene Expression Signature," *World Journal of Oncology*, vol. 14, no. 5, pp. 406–422, 2023, doi: 10.14740/wjon1700.
- [13] A. El-Nabawy, N. El-Bendary, and N. A. Belal, "A feature-fusion framework of clinical, genomics, and histopathological data for METABRIC breast cancer subtype classification," *Applied Soft Computing*, vol. 91, 2020, doi: 10.1016/j.asoc.2020.106238.
- [14] M. Zhao, Y. Tang, H. Kim, and K. Hasegawa, "Machine learning with K-means dimensional reduction for predicting survival outcomes in patients with breast cancer," *Cancer Informatics*, vol. 17, 2018, doi: 10.1177/1176935118810215.
- [15] P. K. Rogan *et al.*, "Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning," *F1000Research*, vol. 5, p. 2124, 2017, doi: 10.12688/f1000research.9417.3.
- [16] A. Khajenezhad, M. A. Bashiri, and H. Beigy, "A Distributed Density Estimation Algorithm And Its Application To Naive Bayes Classification," *Applied Soft Computing*, vol. 98, p. 106837, 2021, doi: 10.1016/j.asoc.2020.106837.
- [17] D. P. Alamsyah, Y. Ramdhani, T. Arifin, F. Febrilla, and S. Setiawan, "Prediction Of Immunotherapy Success Rate: Particle Swarm Optimization Approach," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, 2022, pp. 1–5.




- [18] P. Du *et al.*, “Advances of Four Machine Learning Methods for Spatial Data Handling: a Review,” *Journal of Geovisualization and Spatial Analysis*, vol. 4, no. 1, 2020, doi: 10.1007/s41651-020-00048-5.
- [19] M. Daviran, M. Shamekhi, R. Ghezlbash, and A. Maghsoudi, “Landslide Susceptibility Prediction Using Artificial Neural Networks, SVMs and random Forest: Hyperparameters Tuning by Genetic Optimization Algorithm,” *International Journal of Environmental Science and Technology*, vol. 20, no. 1, pp. 259–276, 2023, doi: 10.1007/s13762-022-04491-3.
- [20] V. G. Costa and C. E. Pedreira, “Recent Advances In Decision Trees: An Updated Survey,” *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4765–4800, 2023, doi: 10.1007/s10462-022-10275-5.
- [21] B. K. Singh, “Determining Relevant Biomarkers for Prediction of Breast Cancer Using Anthropometric and Clinical Features: A Comparative Investigation In Machine Learning Paradigm,” *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 393–409, 2019, doi: 10.1016/j.bbe.2019.03.001.
- [22] H. Zheng, A. Mahmoudzadeh, B. Amiri-Ramsheh, and A. Hemmati-Sarapardeh, “Modeling Viscosity of CO(2)-N(2) Gaseous Mixtures Using Robust Tree-Based Techniques: Extra Tree, Random Forest, GBoost, and LightGBM,” *ACS omega*, vol. 8, no. 15, pp. 13863–13875, Apr. 2023, doi: 10.1021/acsomega.3c00228.
- [23] A. M. Rahmani *et al.*, “Machine learning (ML) in Medicine: Review, Applications, and Challenges,” *Mathematics*, vol. 9, no. 22, pp. 1–52, 2021, doi: 10.3390/math9222970.
- [24] M. S. A. Basha, K. Desai, S. Christina, M. M. Sucharitha, and A. Maheshwari, “Enhancing Red Wine Quality Prediction Through Machine Learning Approaches With Hyperparameters Optimization Technique,” in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2023, pp. 1–8. doi: 10.1109/ICEEICT56924.2023.10157719.
- [25] A. Wibowo, “Forecasting Water Quality Through Machine Learning And Hyperparameter Optimization,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 496–506, 2024, doi: 10.11591/ijeecs.v33.i1.pp496-506.
- [26] B. Bischl *et al.*, “Hyperparameter Optimization: Foundations, Algorithms, Best Practices, And Open Challenges,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, 2023, doi: 10.1002/widm.1484.
- [27] A. Morales-Hernández, I. V. Nieuwenhuys, and S. Rojas Gonzalez, “A Survey on Multi-Objective Hyperparameter Optimization Algorithms For Machine Learning,” *Artificial Intelligence Review*, vol. 56, no. 8, 2023. doi: 10.1007/s10462-022-10359-2.
- [28] Z. Chen *et al.*, “A Machine Learning Model to Predict the Triple Negative Breast Cancer Immune Subtype,” *Front. Immunol.*, vol. 12, pp. 1–14, 2021, doi: 10.3389/fimmu.2021.749459.
- [29] A. Thalor, H. Kumar Joon, G. Singh, S. Roy, and D. Gupta, “Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 1618–1631, 2022, doi: 10.1016/j.csbj.2022.03.019.
- [30] T. Khorshed, M. N. Moustafa, and A. Rafea, “Deep Learning for Multi-Tissue Cancer Classification of Gene Expressions (GeneXNet),” *IEEE Access*, vol. 8, pp. 90615–90629, 2020, doi: 10.1109/ACCESS.2020.2992907.
- [31] N. A. K. Hussein and B. Al-Sarray, “Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data,” *Iraqi Journal of Science.*, vol. 63, no. 7, pp. 3153–3168, 2022, doi: 10.24996/ijs.2022.63.7.36.
- [32] METABRIC, “Breast Cancer Gene Expression Profiles (METABRIC),” *Kaggle*, 2016. <https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data> (accessed Nov. 02, 2023).
- [33] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, “Decision Trees: An Overview and Their Use in Medicine,” *Journal of Medical System*, vol. 26, pp. 445-463, 2002, doi: 10.1023/A:1016409317640.
- [34] B. Bischl *et al.*, “Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, pp. 1–43, 2023, doi: 10.1002/widm.1484.
- [35] N. Al Mudawi and A. Alazeb, “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms,” *Sensors*, vol. 22, no. 11, 2022, doi: 10.3390/s22114132.

BIOGRAPHIES OF AUTHORS






Toni Arifin    he is a member of the Faculty of Engineering, majoring in Informatics Engineering, Adhirajasa Reswara Sanjaya (ARS) University, and researcher ARS Digital Research & Innovation (ADRI). He received his Bachelor's Degree in Informatics Engineering from Bina Sarana Informatika University in 2013 and graduated from the computer science master's program at Nusa Mandiri University Jakarta in 2015. He has authored or coauthored more than 68 publications: 4 proceedings and 66 journals, with 12 H-index and more than 537 citations. Research interests include machine learning, image processing and deep learning. He can be contacted at email: toni.arifin@ars.ac.id.






Ignatius Wiseto Prasetyo Agung    after retired from PT. Telkom Indonesia, he is now dedicated his time in the ARS (Adhirajasa Reswara Sanjaya) University Bandung, Indonesia, as a lecturer and Vice Rector for Collaboration & Innovation, since October 2019. In Telkom Indonesia, he worked since 1988 in various divisions e.g satellite development, network operation, R&D, and Digital Business. He received the Sarjana (Bachelor Degree) in Telecommunication from Institut Teknologi Bandung, Indonesia in 1987. He was also graduated from University of Surrey, UK and received the MSc in Telematics (1994) and PhD in Multimedia Communication (2002). He was also in charge in several professional forums, for instance the Asia Pacific Telecommunity Wireless Forum (AWF) as Convergence Working Group Chairman (2008- 2011); in ITU-D as Vice Rapporteur (2007-2009); as Chairman (2020, 2021) and Vice Chair (2018-2019) of IEEE Communications Society Indonesia Chapter; and as General Chair of several IEEE Conferences. He can be contacted at email: wiseto.agung@ars.ac.id.






Erfian Junianto    he is a member of the Faculty of Engineering, majoring in Informatics Engineering, at Adhirajasa Reswara Sanjaya (ARS) University, and a researcher at ARS Digital Research & Innovation (ADRI). He graduated from the computer science master's program at Nusa Mandiri University Jakarta in 2014. He has authored or co-authored more than 38 publications, including 2 proceedings and 36 journals, with an H-index of 10 and more than 450 citations. His research interests include text mining, artificial intelligence, and classification. He can be contacted at email: erfian.ejn@ars.ac.id.






Rizal Rachman    he studied undergraduate at Padjadjaran University from 2000 to 2005, majoring in Mathematics with a Computer Science study program. He pursued a Master's in Management at Bina Sarana Informatika University from 2013 to 2015 and a Master's in Information Systems at STMIK LIKMI Bandung from 2019 to 2021. He has authored or co-authored more than 79 publications, including 2 proceedings and 36 journals, with an H-index of 10 and more than 986 citations. His research interests include data mining, artificial intelligence, and information systems. He can be contacted via email: rizalrachman@ars.ac.id



Ilham Rachmat Wibowo    he is a bachelor student in the Faculty of Engineering, majoring in Information Systems, at Adhirajasa Reswara Sanjaya (ARS) University, and works as a research assistant at ARS Digital Research & Innovation (ADRI). He has participated in research focused on identifying cancer using machine learning methods, utilizing the Python programming language. He can be contacted via email: ihamwibowo125@gmail.com.



Dari Dianata Agustin    she is a bachelor student in the Faculty of Engineering, majoring in Information Systems, at Adhirajasa Reswara Sanjaya (ARS) University, and works as a research assistant at ARS Digital Research & Innovation (ADRI). Previously, she participated in research using machine learning methods and the Python programming language. She can be contacted via email: 16213056@ars.ac.id.