Breast cancer identification using a hybrid machine learning system

Toni Arifin[1,2], Ignatius Wiseto Prasetyo Agung[1,2], Erfian Junianto[1,2], Dari Dianata Agustin[1,3], Ilham Rachmat Wibowo[1,3], Rizal Rachman[3]
[1]ARS Digital Research and Innovation, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia
[2]Informatics Study Program, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia
[3]Information System Study Program, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia

Article Info

Article history:

Received Sep 23, 2024
Revised May 31, 2025
Accepted May 24, 2025

Keywords:

Breast cancer
Feature extraction
Feature selection
Gene expression
Identification

ABSTRACT

Breast cancer remains one of the most prevalent malignancies among women and is frequently diagnosed at an advanced stage. Early detection is critical to improving patient prognosis and survival rates. Messenger ribonucleic acid (mRNA) gene expression data, which captures the molecular alterations in cancer cells, offers a promising avenue for enhancing diagnostic accuracy. The objective of this study is to develop a machine learning-based model for breast cancer detection using mRNA gene expression profiles. To achieve this, we implemented a hybrid machine learning system (HMLS) that integrates classification algorithms with feature selection and extraction techniques. This approach enables the effective handling of heterogeneous and high-dimensional genomic data, such as mRNA expression datasets, while simultaneously reducing dimensionality without sacrificing critical information. The classification algorithms applied in this study include support vector machine (SVM), random forest (RF), naïve bayes (NB), k-nearest neighbors (KNN), extra trees classifier (ETC), and logistic regression (LR). Feature selection was conducted using analysis of variance (ANOVA), mutual information (MI), ETC, LR, whereas principal component analysis (PCA) was employed for feature extraction. The performance of the proposed model was evaluated using standard metrics, including recall, F1-score, and accuracy. Experimental results demonstrate that the combination of the SVM classifier with MI feature selection outperformed other configurations and conventional machine learning approaches, achieving a classification accuracy of 99.4%.

Corresponding Author:

Toni Arifin
ADRI (ARS Digital Research and Innovation), Informatics Study Program, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University
West Java, Bandung, Indonesia
Email: toni.arifin@ars.ac.id

1. INTRODUCTION

Breast cancer is the most commonly diagnosed malignancy among women, impacting individuals across 157 countries, and is the leading cause of cancer incidence among women worldwide [1]. In 2022, approximately 2.3 million women were diagnosed with breast cancer, with the disease accounting for an estimated 670,000 deaths [2]. Projections for 2024 indicate that nearly 310,720 new cases of invasive breast cancer and 56,500 cases of ductal carcinoma in situ (DCIS) will be diagnosed [3]. As a multifaceted and heterogeneous disease, breast cancer is influenced by a variety of molecular mechanisms, including genetic mutations, epigenetic alterations, and signaling pathways, all of which contribute to its development, progression, and resistance to treatment. This molecular diversity underpins the complexity of breast cancer and highlights the need for personalized therapeutic strategies tailored to the unique biological characteristics of each patient's tumor. Messenger ribonucleic acid (mRNA) plays a crucial role in these processes, often interacting with other RNA molecules, such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs). Genomic analysis through mRNA expression profiling has proven valuable in identifying biomarkers for breast cancer diagnosis and prognosis [4]. These biomarkers help distinguishing between different molecular subtypes, predicting disease progression, and uncovering insights into the molecular mechanisms involved [5]. mRNA gene expression data provides significant opportunities to analyze complex biological patterns related to breast cancer. However, the high dimensionality of genetic data poses substantial challenges in analysis, encompassing challenges such as the risk of overfitting, elevated computational demands, and complexities in interpreting the resulting outputs. One such technique capable of handling this complexity is the machine learning approach [6].

Diagnosing breast cancer using mRNA data and machine learning has become a key research focus due to its potential for early detection and personalized treatment. Early and accurate detection is essential for improving survival rates. This overview explores various machine learning methods applied to mRNA data for breast cancer detection and diagnosis. Machine learning (ML) techniques have been increasingly utilized to optimize the accuracy and efficiency of breast cancer diagnosis by leveraging both imaging modalities and advanced data analysis. These methods significantly improve the identification of diagnostic and prognostic biomarkers and enable effective classification of breast cancer subtypes [7]. The field of ML has advanced considerably with the development of both conventional and hybrid approaches. Conventional ML methods typically rely on single algorithms, while hybrid machine learning system (HMLS) combines multiple techniques to leverage their strengths and mitigate the weaknesses of individual algorithms [8].

Hybrid machine learning system (HMLS) provides substantial benefits by combining the strengths of different techniques. They enhance performance, accuracy, and robustness, making them well-suited for tackling complex problems. These methods are particularly effective in applications that require both data-driven insights and domain knowledge. By integrating various techniques, hybrid models not only improve accuracy but also expand the applicability of machine learning across diverse fields. They demonstrate superior performance in complex tasks, including classification, regression, and reinforcement learning. In healthcare, hybrid systems that combine physician reasoning with ML algorithms outperform traditional models by leveraging high-quality data and expert knowledge [9]. Additionally, hybrid approaches optimize feature selection, enhance generalization, and offer significant advantages over conventional methods [10].

In this study, we used three components of the HMLS: i) Machine learning classification algorithms, selected for their ability to handle complex and intricate data. These algorithms include random forest (RF) [11], naive bayes (NB) [12], k-nearest neighbors (KNN) [13], extra trees classifier (ETC) [14], and logistic regression (LR) [15]. ii) Feature selection algorithms, designed to identify the best features from large datasets. These include analysis of variance (ANOVA) [16], mutual information (MI) [17], ETC, and LR [18]. iii) Using the principal component analysis (PCA) algorithm, which enhances nonlinear dynamic process monitoring by extracting dynamic, linear, and nonlinear features from process data [19].

This research employs mRNA gene expression data to classify breast cancer and aims to identify the most optimal combination of HMLS by analyzing and comparing the results of each experiment conducted. The HMLS model, developed using the Python programming language, presents a comprehensive approach that combines feature selection, feature extraction, and classification techniques, addressing the limitations of traditional single-method pipelines. The primary objectives of this study are as follows: i) to develop a robust and precise HMLS model for identifying breast cancer using mRNA gene expression data, ii) to compare the proposed HMLS model with previous models from past research, and iii) to gain new insights into the implementation of HMLS for breast cancer identification. By employing a comparative and ensemble-based feature selection approach, the study identifies the most effective strategies that improve both the accuracy and reliability of classification, contributing to more robust and dependable outcomes. Additionally, the innovative use of the extra trees algorithm, both as a classifier and as a feature selector, offers a novel perspective on algorithm versatility and its impact on model performance. Finally, the integration of PCA with various classification algorithms provides a comprehensive evaluation framework that could serve as a benchmark for future studies involving highly complex and high-dimensional biomedical datasets. Collectively, the findings contribute to the advancement of more effective strategies, efficient, and interpretable machine learning models within the domain of cancer genomics.

2. LITERATURE REVIEW

The study by Chen *et al.* [20], datasets obtained from Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), The Cancer Genome Atlas (TCGA), and Gene Expression Omnibus (GEO) were used to predict breast cancer according to immune subtypes in triple-negative breast cancer (TNBC) patients, identifying 11 hub genes. The R F model developed in their study yielded an AUC of 0.76, a performance considered satisfactory though not remarkable, indicating that the model still has room for improvement. Subsequent research by El-Nabawy *et al.* [21] applied supervised learning to the METABRIC dataset, achieving the highest accuracy of 97.1% using linear-SVM and E-SVM algorithms. However, the model still heavily relies on manual feature pre-selection rather than utilizing more scalable automatic feature learning techniques. Further studies are needed to explain the key features driving the predictions. A study conducted by Zhao *et al.* [22] used METABRIC data with the K-Means method and machine learning classifiers, showing that Random Forest and SVM achieved an accuracy of 72.9%. However, the study did not explain how missing data in gene expression and clinical variables were handled, which could affect the accuracy and introduce bias. More advanced dimensionality reduction techniques, such as supervised PCA or autoencoders, could be integrated to improve the model's ability to preserve essential biological information, offering advantages over the conventional K-Means approach.

Previous research has explored hybrid machine learning systems. For example, Al-Rajab *et al.* [23] proposed a new hybrid machine learning feature selection model to improve gene classification across multiple colon cancer datasets. The study addressed the challenges of high-dimensional and noisy gene expression data. The model combines information gain (IG) and genetic algorithm (GA) for feature extraction, and mRMR with particle swarm optimization (PSO) for gene selection. HMLFSM improved classification accuracy by identifying key genes and removing irrelevant ones, achieving up to 97% accuracy. However, the study reported only accuracy, while other metrics such as AUC, F1-score, recall, and precision are also important for diagnostic applications. It also did not assess the model's robustness to noise, missing values, or small datasets. Lastly, the feature extraction approach could be compared with other methods to better evaluate its effectiveness. Another study on hybrid machine learning was conducted by Taghizadeh *et al.* [18], combining feature selection, feature extraction, and classification to identify breast cancer. The study used RNA sequencing data from the TCGA database. The best results were obtained using the LGR feature selection method with a multilayer perceptron (MLP) classifier, achieving a balanced accuracy of 0.86 and an AUC of 0.94. However, due to the complexity and noisiness of RNA-seq data, such high accuracy raises concerns about overfitting. Additionally, the study did not employ k-fold cross-validation, making the results more susceptible to bias from data splitting and less reliable for broader application. A similar study by Nadem *et al.* [12] demonstrated that combining artificial neural networks with traditional machine learning algorithms have enhanced the accuracy of colon cancer predictions by up to 6.67% and 10.43% compared to standard methods. The highest accuracy was achieved by the RFNN model, reaching 89.81%. However, the study provided limited details about the feature selection process. It did not clearly explain which algorithms were used, how features were selected, or whether feature stability was evaluated. Furthermore, it did not mention whether k-fold cross-validation (such as 5-fold or 10-fold) was applied, which is crucial to reduce bias from random data splits.

3. METHOD

The evaluation phase of the method necessitates a clearly defined and highly accurate approach. The proposed method serves a critical function in the research process, facilitating the achievement of the desired outcomes, as illustrated in Figure 1. The design offers a thorough overview of the research process, ensuring a clear understanding of each step involved, which are outlined systematically. The steps include: i) collecting data from METABRIC, ii) data inspection, iii) preprocessing the data and performing 10-fold cross-validation, iii) feature selection and feature extraction, iv) applying machine learning algorithms and selecting the best-performing model, v) evaluating and validating the results, and vi) generating a performance report. This methodology ensures transparency and reproducibility, which are essential pillars of robust scientific inquiry.

3.1. Breast cancer mRNA dataset (METABRIC)

This study leverages the breast cancer mRNA dataset provided by the METABRIC. Recognized for its extensive validation and broad citation within the breast cancer research community, the METABRIC dataset integrates detailed clinical, pathological, and molecular profiles from a diverse array of tumor specimens, thereby offering a robust foundation for comprehensive analyses. The dataset includes a broad spectrum of genomic data, encompassing genetic mutations, gene expression profiles, and epigenetic modifications, alongside clinical variables and other pertinent risk factors. It has been widely employed in numerous studies to investigate the heterogeneity of breast cancer and to identify potential biomarkers for its early detection and diagnosis [24]. In the present study, the dataset comprises 692 attributes across 1,904 samples, providing a robust and reliable foundation for analytical modeling and predictive analysis.

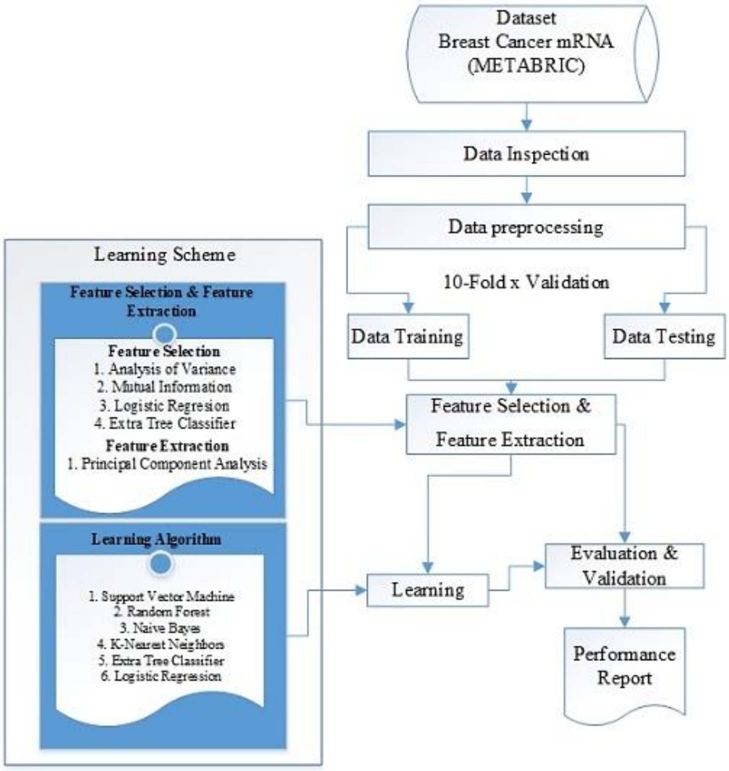


Figure 1. Proposed method hybrid machine learning system

3.2. Data inspection

Inspection and evaluation of datasets in machine learning entails a comprehensive evaluation and understanding of the dataset prior to implementing preprocessing steps or building models. This process is critical for guiding informed decisions regarding data preprocessing and feature selection, both of which are fundamental to enhancing the accuracy and performance of machine learning models. As an integral component of data analysis, data inspection involves scrutinizing the dataset to identify the most appropriate processing techniques to enhance model performance [25]. In this study, data inspection was performed at the outset, involving an in-depth analysis of the data. The objective was to evaluate the data's characteristics and quality, allowing for the identification of the appropriate preprocessing techniques to be utilized.

3.3. Preprocessing data and cross validation

Data preprocessing, particularly data transformation, is a critical step in machine learning. It involves converting raw data into a format that is more suitable for analysis and model training. This process can significantly impact the quality and efficiency of the resulting models. In this research, the first stage of the data preprocessing technique is data transformation (discretization), followed by the second stage, which involves replacing missing values. These techniques were chosen because they offer several advantages, such as converting data into a format suitable for learning algorithms and improving the accuracy and efficiency of mining algorithms. Once this stage is completed, the next step is the implementation of 10-fold cross-validation for splitting the training and testing data. An additional objective of employing cross-validation is to obtain a more accurate evaluation, ensuring that the model not only acquires knowledge from the training data but also generalizes effectively to new, unseen data [26].

3.4. Feature selection and feature extraction

Feature selection and feature extraction are essential methods in machine learning, especially when dealing with high-dimensional data. These techniques focus on reducing data dimensionality to enhance model performance, lower computational costs, and improve interpretability [25]. The feature selection algorithms used in this study include ANOVA, MI, LR, and ETC. These algorithms offer several advantages, such as reducing computation time and model complexity, improving learning accuracy, and helping to avoid overfitting. PCA is utilized for feature extraction in this study. This technique provides the benefit of reducing high-dimensional data to a lower-dimensional representation while retaining most of the original variance, thereby facilitating more efficient data analysis without significant loss of critical information [27].

3.5. Machine learning algorithm

Machine learning (ML) is a rapidly advancing discipline that lies at the intersection of computer science and statistics, focused on the development of algorithms that enable computers to learn from data and generate accurate predictions or decisions based on that information. In this study, the machine learning algorithms used for training are capable of handling large and complex datasets [12], including SVMs, RF, NB, KNN, ETC, and LR. Some of the algorithms used for feature selection, such as LR and ETC, are also applied in the classification process. These algorithms were chosen because they have made a significant impact in various fields, particularly healthcare, by allowing computers to learn from data and make data-driven decisions [28].

3.6. Evaluation and validation

This phase is dedicated to evaluating the performance of the developed model. In this study, a classification report is employed as the evaluation method, providing a detailed summary of the performance metrics associated with each applied technique. This step is fundamental for establishing the robustness, effectiveness, and reproducibility of the classification model, particularly in healthcare applications where accuracy is critical. A comprehensive set of evaluation techniques is systematically applied to rigorously assess both the predictive accuracy, generalizability, and robustness of the machine learning model. During the testing phase, quantitative metrics-including accuracy, precision, recall, and F1-score-are systematically applied to ensure a thorough and statistically sound evaluation of the model’s predictive capabilities. Accuracy reflects the overall correctness of the model’s predictions, while recall measures its ability to correctly identify all relevant positive instances. Precision evaluates the ratio of true positives to all instances predicted as positive. The F1-score, calculated as the harmonic mean of precision and recall, serves as a robust and integrative metric for assessing the classification effectiveness of the model. Taken together, these metrics provide a thorough and robust assessment of the model’s predictive performance [6].

4. RESULTS AND DISCUSSION

Upon completion of all experimental phases, the model was rigorously evaluated using 10-fold cross-validation and further analyzed through a confusion matrix, which reports key performance metrics, including accuracy, precision, recall, and F1-score. The results demonstrate that the SVM consistently outperforms other algorithms, exhibiting superior classification performance. As presented in Table 1, Figure 2 and Figure 3(a)-3(d), the SVM combined with MI-based feature selection achieves the highest performance, with an accuracy of 99.4% and identical precision, recall, and F1-score values of 0.9940. This strong performance can be primarily attributed to the comprehensive data preprocessing conducted in the early stages, which effectively refined the input data and enabled the SVM to manage and simplify an otherwise complex classification task. Furthermore, the application of feature selection contributed to a more efficient model by reducing dimensionality and retaining the most relevant features, thereby enhancing the algorithm’s predictive capability and computational efficiency. The second and third positions are also held by SVM models but with different feature selection methods: LR and ETC, both achieving an accuracy of 99.22%. Their precision, recall, and accuracy values are closely matched at 0.9921 and 0.9922, although the F1-score differs slightly, with the ETC yielding 0.9221. In fourth place, the SVM with ANOVA achieves a slightly lower accuracy of 99.17%, a precision of 0.9918, a recall of 0.992, and an F1-score of 0.9919. Conversely, the SVM with feature extraction using PCA exhibits the lowest performance, with an accuracy of 29.68%, a precision of 0.2204, a recall of 0.3014, and an F1-score of 0.2263. These results highlight the significantly poor performance of SVM + PCA compared to other algorithms, a trend consistent with other machine learning models tested with PCA, all of which produced similarly low values, as illustrated in Figure 2.

This study includes comparisons with similar research. Some of the studies referenced are Al-Rajab *et al.* [23], who proposed a HMLS by integrating IG with GA and coupling minimum redundancy maximum relevance (mRMR) with PSO, as well as Taghizadeh *et al.* [18], who implemented HMLS using the LGR feature selection method coupled with an MLP classifier to achieve high accuracy. The results of these comparisons are presented in Table 2.

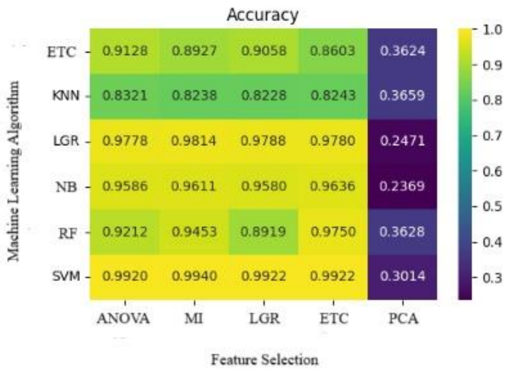
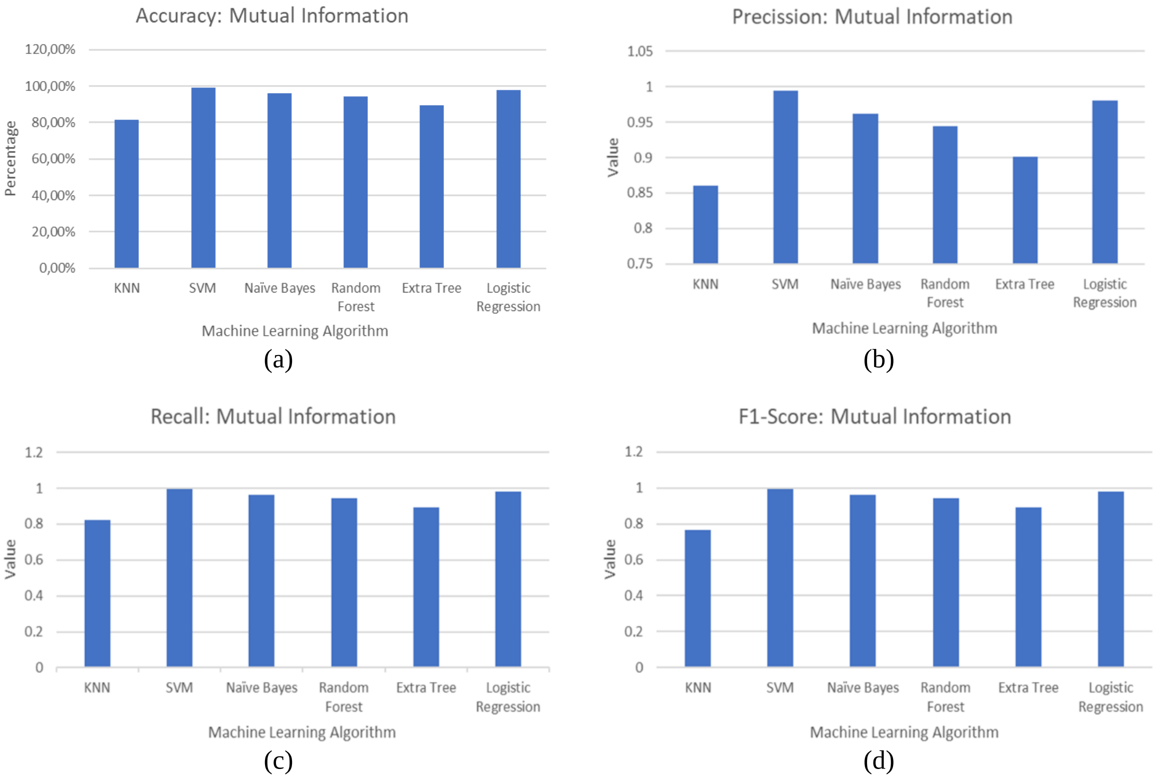


Figure 2. Accuracy heatmap of the feature selection, feature extraction, and classification procedures

Table 1. The evaluation results with the SVM, feature selection, and feature extraction

	Support vector machine			
Algorithm	Accuracy	Precision	Recall	F1-score
Mutual information	99.40%	0.9940	0.9940	0,9940
Logistic regression	99.20%	0.9921	0.9922	0.9921
Extra tree classifier	99.20%	0.9921	0.9922	0.9221
Analysis of variance	99.17%	0.9918	0.992	0.9919
Principal component analysis	29.68%	0.2204	0.3014	0.2263



(a) (b)

(c) (d)

Figure 3. Depicts the evaluation outcomes based on mutual information, including (a) accuracy, (b) precision, (c) recall, and (d) F1-score

Table 2 shows that the proposed model, particularly the SVM + MI algorithm, outperforms previous studies that also utilized hybrid machine learning techniques and complex datasets. Another notable finding in this study is that the implementation of feature extraction with PCA, as explained in Table 1, yielded the lowest results compared to the feature selection algorithms applied. This highlights a limitation of our study:

we only used mRNA data. Future research could incorporate miRNA and lncRNA data and apply feature extraction algorithms to assess their effectiveness when dealing with more complex datasets.

Table 2. Comparative analysis of the hybrid machine learning system against previous studies

Research	Hybrid machine learning system	Accuracy
Al-Rajab *et al.* [23]	IG-GA and mRMR-PSO	97%
Taghizadeh *et al.* [18]	LGR + MLP	86%
Malik *et al.* [12]	RF + NN	89.81%
This research	SVM + MI	99.4%

Previous research on hybrid machine learning systems has focused on improving cancer gene classification and has also demonstrated significant improvements in prediction performance. Based on the comparison of previous studies, the implementation of hybrid machine learning systems is important because it can improve the accuracy and efficiency of detecting and classifying cancer, especially breast cancer. By combining various machine learning algorithms, such as feature selection and feature extraction, this system can handle complex, high-dimensional data. HMLS also help reduce noise in the data and select the most relevant features. Other HMLS techniques should be applied to enable the model to more accurately distinguish between healthy and cancerous conditions. Therefore, HMLS research has the potential to provide better and faster diagnostic tools for breast cancer treatment.

5. CONCLUSION

This study successfully demonstrates that the HMLS holds significant potential for improving the accuracy of breast cancer identification compared to conventional methods or single algorithms. By combining the SVM classification algorithm with Mutual Information as a feature selection method, it outperforms other machine learning algorithms, achieving an exceptional accuracy rate of 99.4%, accompanied by recall, precision, and F1-score values of 0.9940. The primary contribution of this research lies in the design of the HMLS architecture, which not only enhances accuracy but also provides more reliable predictive analysis for grouping complex and intricate data. Furthermore, this approach has proven its flexibility in processing large and heterogeneous datasets, reflecting the complexity of real-world medical data. These findings reinforce the relevance of HMLS as a modern computational solution in the healthcare domain. Although significant progress has been achieved, Opportunities for further refinement persist and warrant examination in forthcoming studies. The following considerations highlight potential avenues for continued exploration: i) Expanding the dataset in terms of size and diversity-increasing both the volume and variety of data would enable a more comprehensive evaluation of the model's performance and its capacity to generalize across a wider range of breast cancer cases; and ii) Exploring other hybrid techniques, such as implementing ensemble models by combining the SVM algorithm with Gradient Boosting, or applying multi-layer models with deep learning algorithms like CNN or ANN for feature extraction, followed by the application of classification algorithms such as SVM or KNN.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	**C**	**M**	**So**	**Va**	**Fo**	**I**	**R**	**D**	**O**	**E**	**Vi**	**Su**	**P**	**Fu**
Toni Arifin	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ignatius Wiseto Prasetyo Agung		✓		✓	✓	✓		✓		✓		✓		✓
Erfian Junianto		✓	✓	✓	✓					✓		✓		✓
Dari Dianata Agustin	✓		✓	✓		✓	✓				✓		✓	
Ilham Rachmat Wibowo	✓		✓	✓		✓	✓				✓		✓	
Rizal Rachman		✓				✓				✓				

C	**C**onceptualization	I	**I**nvestigation	Vi	**Vi**sualization
M	**M**ethodology	R	**R**esources	Su	**Su**pervision
So	**So**ftware	D	**D**ata Curation	P	**P**roject administration
Va	**Va**lidation	O	Writing - **O**riginal Draft	Fu	**Fu**nding acquisition
Fo	**Fo**rmal analysis	E	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The METABRIC public dataset that support the findings of this study are openly available at https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data, reference number [24].

REFERENCES

[1] "Union for internasional cancer control," *Breast cancer*. https://www.uicc.org/what-we-do/thematic-areas/breast-cancer?gad_source=1&gclid=Cj0KCQjw_-GxBhC1ARIsADGgDjvZPrSxddgDVRw3gp_ZpjHoT_-LgeJqYk5H39BPsTbei4f4i0-D7BUaAgI7EALw_wcB (accessed Sep 23, 2024).

[2] "Breast cancer statistics and resources," *Breast Cancer Research Foundation*. https://www.bcrf.org/breast-cancer-statistics-and-resources/ (accessed Sep 23, 2024).

[3] A. N. Giaquinto *et al.*, "Breast cancer statistics 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 6, pp. 477–495, Nov. 2024, doi: 10.3322/caac.21863.

[4] D. H. Kim and K. E. Lee, "Discovering breast cancer biomarkers candidates through mRNA expression analysis based on the cancer genome atlas database," *Journal of Personalized Medicine*, vol. 12, no. 10, p. 1753, Oct. 2022, doi: 10.3390/jpm12101753.

[5] S. Zhang, H. Jiang, B. Gao, W. Yang, and G. Wang, "Identification of diagnostic markers for breast cancer based on differential gene expression and pathway network," *Frontiers in Cell and Developmental Biology*, vol. 9, Jan. 2022, doi: 10.3389/fcell.2021.811585.

[6] L. Peng, J. Yang, M. Wang, and L. Zhou, "Editorial: Machine learning-based methods for RNA data analysis," *Frontiers in Genetics*, vol. 13, May 2022, doi: 10.3389/fgene.2022.828575.

[7] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, p. 114161, Apr. 2021, doi: 10.1016/j.eswa.2020.114161.

[8] D. Machalek, T. Quah, and K. M. Powell, "A novel implicit hybrid machine learning model and its application for reinforcement learning," *Computers & Chemical Engineering*, vol. 155, p. 107496, Dec. 2021, doi: 10.1016/j.compchemeng.2021.107496.

[9] N. Alromema, A. H. Syed, and T. Khan, "A hybrid machine learning approach to screen optimal predictors for the classification of primary breast tumors from gene expression microarray data," *Diagnostics*, vol. 13, no. 4, p. 708, Feb. 2023, doi: 10.3390/diagnostics13040708.

[10] M. Arashpour *et al.*, "Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization," *Computer Applications in Engineering Education*, vol. 31, no. 1, pp. 83–99, Jan. 2023, doi: 10.1002/cae.22572.

[11] M. Daviran, M. Shamekhi, R. Ghezelbash, 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, Jan. 2023, doi: 10.1007/s13762-022-04491-3.

[12] M. S. A. Nadeem, M. H. Waseem, W. Aziz, U. Habib, A. Masood, and M. Attique Khan, "Hybridizing artificial neural networks through feature selection based supervised weight initialization and traditional machine learning algorithms for improved colon cancer prediction," *IEEE Access*, vol. 12, pp. 97099–97114, 2024, doi: 10.1109/ACCESS.2024.3422317.

[13] L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm," *Building and Environment*, vol. 202, p. 108026, Sep. 2021, doi: 10.1016/j.buildenv.2021.108026.

[14] 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.

[15] A. Muslim, A. Benny, R. Refianti, C. Maisyarah, and G. Setiawan, "Comparison of accuracy between long short-term memory-deep learning and multinomial logistic regression-machine learning in sentiment analysis on twitter," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020, doi: 10.14569/IJACSA.2020.0110294.

[16] M. Islam and R. Islam, "Exploring the impact of univariate feature selection method on machine learning algorithms for heart disease prediction," in *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, Jun. 2023, pp. 1–5, doi: 10.1109/NCIM59001.2023.10212832.

[17] D. K. Rakesh and P. K. Jana, "A general framework for class label specific mutual information feature selection method," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 7996–8014, Dec. 2022, doi: 10.1109/TIT.2022.3188708.

[18] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. JafarpoorNesheli, and S. M. Rezaeijo, "Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods," *BMC Bioinformatics*, vol. 23, no. 1, p. 410, Oct. 2022, doi: 10.1186/s12859-022-04965-8.

[19] L. Guo, P. Wu, S. Lou, J. Gao, and Y. Liu, "A multi-feature extraction technique based on principal component analysis for nonlinear dynamic process monitoring," *Journal of Process Control*, vol. 85, pp. 159–172, Jan. 2020, doi: 10.1016/j.jprocont.2019.11.010.

[20] Z. Chen *et al.*, "A machine learning model to predict the triple negative breast cancer immune subtype," *Frontiers in Immunology*, vol. 12, Sep. 2021, doi: 10.3389/fimmu.2021.749459.

[21] 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, p. 106238, Jun. 2020, doi: 10.1016/j.asoc.2020.106238.

[22] 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, Jan. 2018, doi: 10.1177/1176935118810215.

[23] M. Al-Rajab *et al.*, "A hybrid machine learning feature selection model-HMLFSM to enhance gene classification applied to multiple colon cancers dataset," *PLOS ONE*, vol. 18, no. 11, p. e0286791, Nov. 2023, doi: 10.1371/journal.pone.0286791.

[24] Kaggle, "Breast cancer gene expression profiles (METABRIC)," *Kaggle*. Accessed: Nov. 02, 2023. [Online]. Available: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data.

[25] T. Buckley, B. Ghosh, and V. Pakrashi, "A feature extraction and selection benchmark for structural health monitoring," *Structural Health Monitoring*, vol. 22, no. 3, pp. 2082–2127, May 2023, doi: 10.1177/14759217221111141.

[26] R. Sharma, J. B. Sharma, and R. Maheshwari, "Comparative analysis of different texture features in breast abnormality prediction," *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3885562.

[27] H. Holubova, "A comparative analysis of the principal component method and parallel analysis in working with official statistical data," *Statistics in Transition new series*, vol. 24, no. 1, pp. 199–212, Feb. 2023, doi: 10.59170/stattrans-2023-011.

[28] S. Badillo *et al.*, "An introduction to machine learning," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 871–885, Apr. 2020, doi: 10.1002/cpt.1796.

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 73 publications: 4 proceedings and 69 journals, with 14 h-index and more than 630 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's degree) in Telecommunication from Institut Teknologi Bandung, Indonesia in 1987. He also graduated from the 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.

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.

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.

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.