Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 25 NO 04 2024, pp 468-483
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol25-iss04/553

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

The Implementation of Machine Learning Algorithms for Breast Cancer Biomarker Validation in Metabolomics Studies

Article Info

Article history :

Received December 05, 2024
Revised December 20, 2024
Accepted December 26,2024
Published December 30, 2024

Keywords :
Breast cancer, biomarker, bioinformatics, metabolomics, machine learning

Nindhyana Diwaratri Ratnaningayu[1*], Aryo Tedjo[2], Sonar Soni Panigoro[3]

[1]Master's Programme in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia
[2]Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia
[3]Surgical Oncology Division, Department of Surgery, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia

Abstract. Breast cancer is a heterogeneous disease characterized by distinct molecular and metabolic characteristics, making its diagnostics and treatment challenging. The existence of metabolic reprogramming in breast cancer underscores the potential to identify biomarkers through metabolomics studies, offering new avenues for personalized therapeutic approaches. Machine learning algorithms are now increasingly used to uncover complex patterns in metabolomics data. A comprehensive analysis of in silico metabolomics had successfully identified 24 significant metabolites after rigorous univariate and multivariate tests. Pathway analysis highlighted the apparent involvement of glycerolphosphate in glycerophospholipid and glycerolipid metabolism, indicating its potential role in breast cancer pathology. Validation of these 24 metabolites using machine learning algorithms provided superior results, with Neural Network achieving an AUC of 0.979 and a precision of 93%, Logistic Regression showing an AUC of 0.945 and a precision of 95.7%, as well as Random Forest reporting an AUC of 0.974 and a precision of 95.7% in predictive performance. These findings demonstrate the remarkable ability of machine learning to improve biomarker validation accuracy in metabolomics, facilitating better diagnostic strategies for breast cancer.

Corresponding Author :
Nindhyana Diwaratri Ratnaningayu
Master's Programme in Biomedical Science, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia
Email: nindhyana@gmail.com

1. Introduction

Breast cancer becomes the most commonly diagnosed cancer and the first most common cancer in women, surpassing lung cancer cases. The increasing incidence and mortality of breast cancer is a global problem that needs special attention [1-2]. In Indonesia, 66.271 new cases of breast cancer were recorded in 2022, making it the most prevalent type of cancer in the country [3-4]. Most breast cancer patients diagnosed in Indonesia are in the late stage, which is associated with low survival and poor prognosis [5-6]. Early detection of disease and selection of appropriate treatment can improve the prognosis of breast cancer patients [7-10].

Conventional methods of breast cancer screening are performed with imaging techniques, such as mammography, ultrasonography, and magnetic resonance imaging (MRI). Needle biopsies are commonly performed operatively to confirm and determine the histopathological classification and stage of breast cancer [11-12]. The growing emphasis on personalized precision medicine has increased the need for the development of new molecular markers. The hormone receptors estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor-2 (HER2) are used to aid in the molecular subtype classification of breast cancer. The predictive marker Ki-67 as a sign of cell proliferation is also widely used to aid classification as well as evaluation of neoadjuvant therapy [13-14]. Serological markers of serum CA-15-3 and CEA have also been utilized in the clinical setting, but have low sensitivity and specificity [8],[15].

Breast cancer is one of a complex and heterogeneous diseases. The heterogeneity of this disease becomes one of the results of changes in the regulation of cell metabolism, as proposed by Hanahan and Weinberg in the hallmark of cancer theory [16-17]. The multistep process of cancer cell occurrence can involve metabolic pathways, such as glucose, amino acids, and lipids to fulfill the needs of malignant cells, including increased proliferation, cell survival, cell differentiation, and others. These unique metabolite quantities represent a phenotype used for biomarker discovery in breast cancer [18-19].

Metabolomics is a branch of the omics approach utilized in numerous breast cancer research, involving biological samples such as cell lines, tissues, blood, urine, and saliva [7],[13],[20-21]. This approach provides insights into the dynamic interplay of endogenous metabolites in response to individual changes influenced by genetic and environmental factors [14-15],[22].Untargeted metabolomics has a broader scope and is used for exploratory analysis of the entire spectrum of metabolites present, both known and unidentified. The objective is to thoroughly characterize the metabolome in a sample, which will enable the identification of new clinically relevant biomarkers [23-24]. For instance, significant changes involving glutamic acid, lactic acid, and fructose, have been observed in breast cancer patients compared to healthy individuals, with outstanding strong discriminatory capability area under the curve (AUC). Additionally, identified metabolite panels also showed high AUC values, which helped distinguish triple-negative breast cancer (TNBC) from non-TNBC variants [25-26].

Due to its coverage of thousands of metabolic signals, the analysis process is inherently more complex and requires advanced steps to pinpoint significant biomarkers. The complexity and volume of metabolomics data present significant challenges in analysis, particularly in uncovering subtle patterns and biological insights. The use of computational methods using machine learning (ML), especially when combined with metabolomics, has emerged as a powerful tool to address these challenges. ML facilitates the processing and analysis of metabolomics data by identifying subtle patterns, relationships, and insights that traditional statistical methods might overlook [27-28].

In particular, applying computational ML strategies to metabolomics enables a more holistic and robust diagnostic approach, such as in cancer research, where precision in identifying biomarkers and pathways is critical. ML algorithms can be customized for each research project and can be improved with the availability of more data, thus increasing the precision and power of analysis in metabolomics studies. Both unsupervised and supervised methods in ML can be used for cancer classification and prediction, which has potential for classification, biomarker screening, and progression prediction in breast cancer [29-30].

2. Experimental Section

2.1. Materials

The dataset used were obtained from The Metabolomic Workbench (https://www.metabolomicsworkbench.org/). The keyword used "breast cancer", then data selection was carried out. One dataset selected with Project ID PR000284 and Study ID ST000355. The dataset was split into training and testing datasets using Data Sampler widget on Orange data mining ver. 3.37. (https://orangedatamining.com/). Data preprocessing, including data and statistical analysis were performed using Statistical Analysis modules available from MetaboAnalyst 6.0 (https://www.metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml). Meanwhile, pathway and enrichment analysis were also performed using Pathway Analysis and Enrichment Analysis modules from MetaboAnalyst 6.0. Machine learning algorithm analysis and model evaluation were conducted using Test and Score and ROC analysis from Orange data mining ver. 3.37. The research followed the flowchart described in Figure 1.

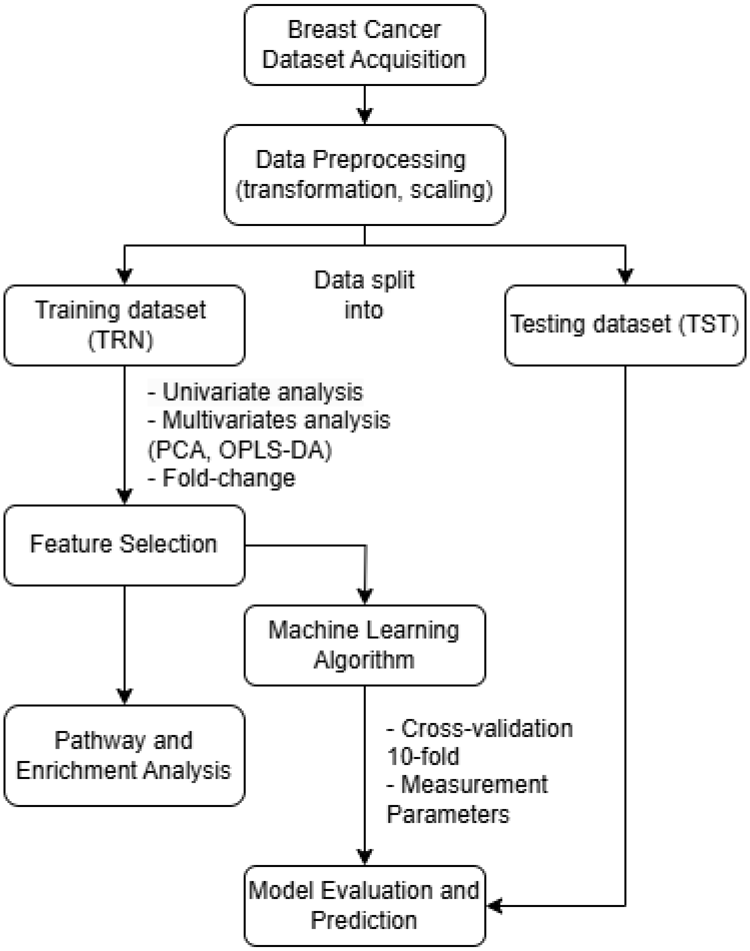

Breast Cancer Dataset Acquisition
Data Preprocessing (transformation, scaling)
Data split into
Training dataset (TRN)
Testing dataset (TST)
- Univariate analysis
- Multivariates analysis (PCA, OPLS-DA)
- Fold-change
Feature Selection
Machine Learning Algorithm
Pathway and Enrichment Analysis
- Cross-validation 10-fold
- Measurement Parameters
Model Evaluation and Prediction

Figure 1. Schematic/Flowchart of research

2.2. Tips

2.2.1 Breast Cancer Dataset Acquisition

The metabolomic dataset with Study ID ST000355 obtained from The Metabolomic Workbench. This dataset came from humans with sampling from blood plasma. The analysis instrument on the data employed mass spectrometry with a gas chromatography system (GC-MS). The dataset was then divided into two different groups, 134 breast cancer (BC) groups and 76 healthy control (HC) groups. The scope of the metabolomics study was selected to be untargeted as it can perform comprehensive profiling of metabolite compounds, both identified and unidentified, for the search of significant biomarkers. The downloaded dataset was in .TXT format. Next, these data were processed into .XLSX format first. BC and C (identifiers) were organized into columns, while metabolite data (features) were organized into rows.

2.2.2 Data Analysis and Statistics

The dataset was first subjected to data transformation and scaling to obtain uniform data distribution. Then, the dataset was divided into training (TRN) and testing (TST) data with a ratio of 80:20 randomly by the 'Data Sampler' widget on Orange Data Mining. The TRN data were subjected to univariate analysis of T-test or Mann-Whitney U statistics to compare the mean metabolite concentrations of BC and HC groups. Multivariate analysis, namely Principal Component Analysis (PCA), was used for data dimension reduction and identification of patterns and variations of principal components. Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) as a supervised method was also performed to determine the most significant variant (VIP). Feature selection was conducted to select the most relevant metabolites as breast cancer biomarker candidates. Feature selection criteria were p-value <0.05 from univariate analysis and VIP>1 from OPLS-DA analysis. Fold-change (FC) of each selected metabolite feature was calculated to examine the trend of concentration level from BC group.

2.2.3 Pathway and Enrichment Analysis

Selected metabolite features were subjected to pathway and enrichment analyses to identify the metabolite pathways and important biological processes involved in breast cancer. Each metabolite was mapped with databases from Kyoto Encyclopedia of Genes and Genomes (KEGG) and Human Metabolome Database (HMDB).

2.2.4 Machine Learning Algorithm and Validation Model

The metabolite features selected from the TRN data were trained with ML models to create classification and prediction models. There are five classification algorithms used, namely support vector machines (SVM), logistic regression (LR), neural network (NN), naive bayes (NB), and random forest (RF). These are the algorithms commonly chosen for disease classification tasks. The classification results produced several classifier measurement parameters, including area under curve (AUC), accuracy, precision, recall, and specificity. These parameters were cross-validated by 10-fold cross-validation. The ML algorithm with the best performance based on these parameters was selected for testing and validation on TST data. Model evaluation on TST data was also presented with ROC curves.

3. Results and Discussion

The initial dataset included 138 BC groups (Stage I, n= 19; Stage II, n= 49; Stage III, n= 47; Stage IV, n= 23) and 76 HC groups. After subdivision, the TRN data consisted of 110 BC groups and 61 HC groups, while the TST data had 28 BC groups and 15 HC groups (Table 1). The dataset characteristics indicated a significant difference in age, with the BC group having a higher mean age. The subtype division of breast cancer showed stage 2 and 3 predominating, making it difficult to analyze metabolite profiles by subtype. Both data subsets were stratified so that a balanced distribution of BC and HC groups was maintained.

Table 1. Dataset characteristics of training and testing data

Characteristics	Training Data (TRN)		Testing Data (TST)	
	BC (n= 110)	HC (n=61)	BC (n=28)	HC (n=15)
Age (years, mean±SD)	53.2±10.6	32.8± 5.9	53±10	34± 6
Stadium BC				
1	15 (13.6%)	N/A	4 (14.2%)	N/A
2	39 (35.4%)	N/A	10 (35.7%)	N/A
3	38 (34.5%)	N/A	9 (32.1%)	N/A
4	18 (16.3%)	N/A	5 (17.8%)	N/A

The PCA results on the TRN dataset showed the top two principal components (PC1 and PC2) at 30.1%. Although this percentage indicates that a large amount of variance remains unexplained, the separation observed between the BC group and the HC group in the PCA plot indicates a distinct metabolite profile (Figure 2). Supervised OPLS-DA analysis maximized the separation of the two groups very clearly, with the BC group predominantly on the left side and the HC group on the right side (Figure 3A). The OPLS-DA model cross-validation results showed no overfitting and could be considered accurate in prediction (R2X=0.073; R2Y=0.731; Q2=0.69). Thus, the results of important metabolites (VIP) could be considered in feature selection for significant candidates (Figure 3B).

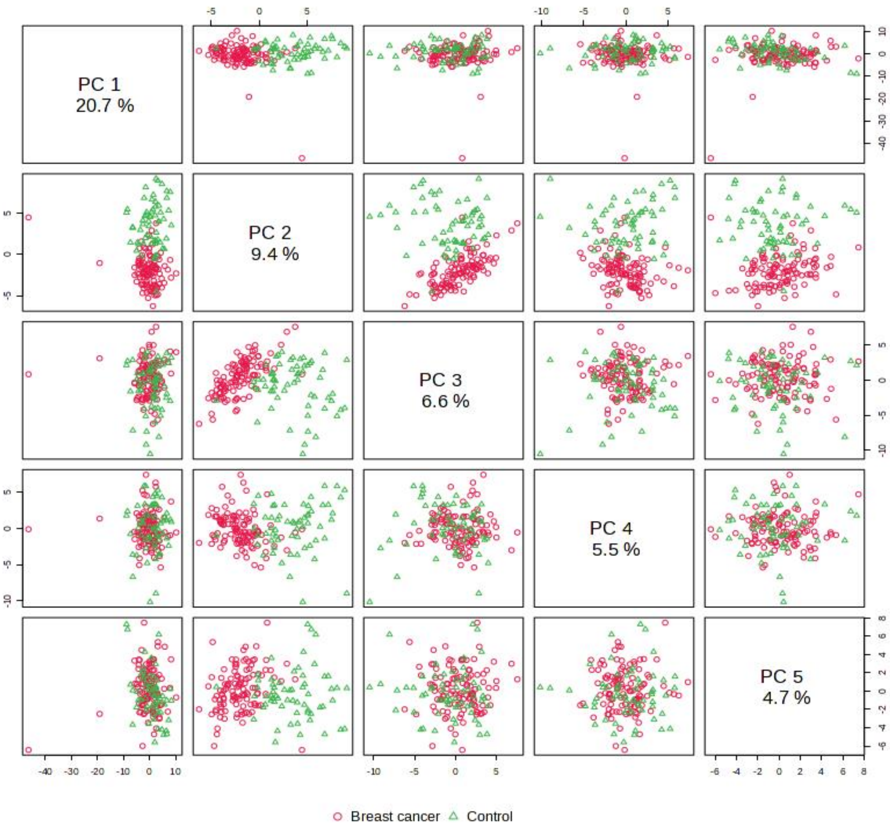
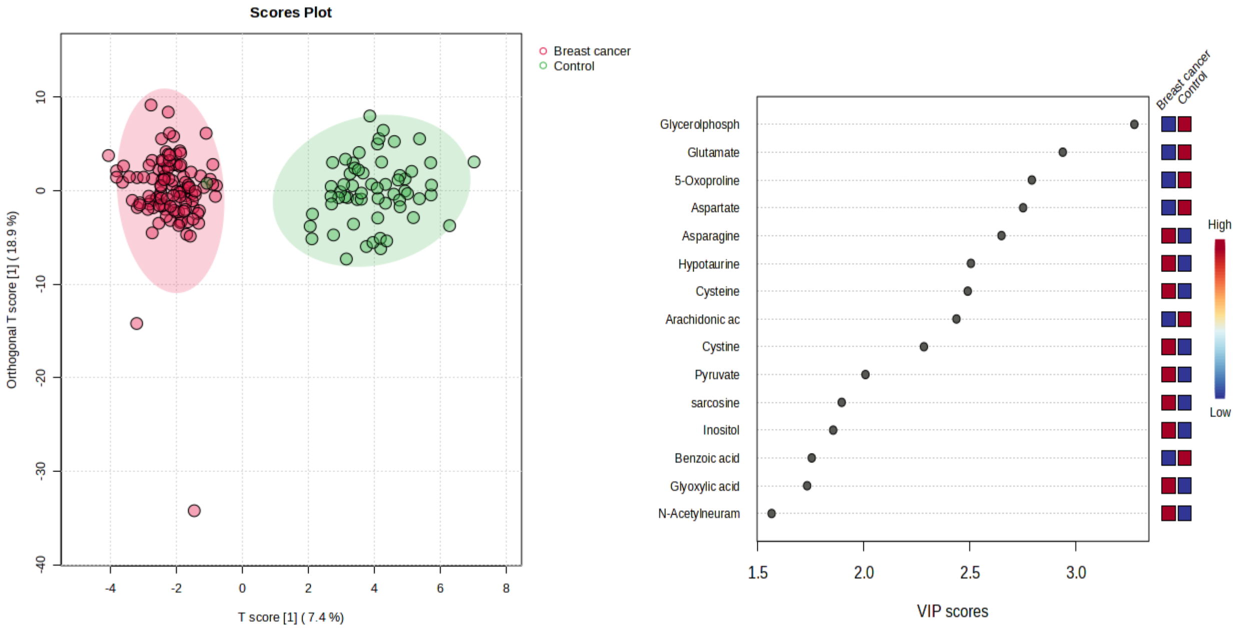


Figure 2. PCA plot of the training dataset on the BC group (red circles) and HC group (green triangles)



(a) (b)

Figure 3. Results of OPLS-DA training data set (A) Score plot. Red color indicates BC group and green color indicates HC group. (B) VIP Score. Red color indicates higher relative metabolite concentration and blue color indicates lower relative metabolite concentration.

Further data analysis on the TRN dataset revealed a distinctive set of 24 metabolite profiles that could serve as potential biomarkers for breast cancer. The identified metabolites were those that showed significant fold change (FC >2 and <0.5) and desirable statistical relevance (p-value <0.05 and VIP >1), highlighting their potential role in distinguishing the BC group from the HC group (Table 2).

Of the 24 metabolites, 22 metabolites were continued for pathway analysis and enrichment (Figure 4). The top five pathways with a significant p-value and false-discovery rate (FDR) also had an impact higher than 0 were glycerophospholipid metabolism, glycerolipid metabolism, glutathione metabolism, alanine, aspartate and glutamate metabolism, as well as arginine biosynthesis. The enrichment results confirmed the involvement of these five pathways in the occurrence of breast cancer. Of the five pathways, eight metabolites were dominantly involved, namely glycerolphosphate, glutamate, 5-oxoproline, cysteine, aspartate, asparagine, pyruvate, and succinate.

Several metabolites, including glycerolphosphate (FC=0.011), glutamate (FC=0.049), 5-oxoproline (FC=0.049), and aspartate (FC=0.054) showed exceptionally low downregulation in breast cancer patients. The exceedingly small FC, especially for glycerophosphate, coupled with the very low p-value and false discovery rates (FDR), proved the robustness of this discovery. Pathway and enrichment analysis also indicated that glycerophosphate reduction was highly involved in two pathways, namely glycerophospholipid and glycerolipid metabolism. Glycerophosphate, also known as glycerol-3-phosphate (G3P), is a lipid precursor molecule that plays a role in membrane synthesis, energy storage and signaling. The decrease in precursors for such synthesis may be due to metabolic disorders that are common in cancer to support rapid growth and proliferation. In addition, G3P with the help of the enzyme glycerol-3-phosphate acyltransferase 1 can produce lysophosphatidic acid (LPA), a component that promotes cancer cell migration. Increased invasion and migration requirements of cancer cells may lead to a state of G3P depletion. Tumor microenvironment conditions, such as the presence of hypoxia and nutrient availability can also affect the level of lipid precursors [31-32].

Table 2. Identification of metabolite differences between BC and C

Metabolites	FC	Trend	p-value	FDR	VIP
Glycerolphosphate	0.011	Down	1.92E-51	2.46E-49	3.275
Glutamate	0.049	Down	1.47E-33	9.38E-32	2.937
5-Oxoproline	0.049	Down	3.61E-29	1.30E-27	2.791
Aspartate	0.054	Down	4.05E-29	1.30E-27	2.750
Asparagine	11.274	Up	1.65E-25	4.21E-24	2.648
Hypotaurine	6.637	Up	3.25E-21	6.92E-20	2.504
Cysteine	6.098	Up	8.58E-21	1.57E-19	2.489
Arachidonic acid	0.136	Down	9.85E-20	1.58E-18	2.436
Cystine	6.098	Up	2.52E-17	3.59E-16	2.283
Pyruvate	2.382	Up	3.53E-14	4.52E-13	2.007
sarcosine	2.177	Up	2.75E-11	3.20E-10	1.895
Glyoxylic acid	4.741	Up	3.57E-11	3.81E-10	1.732
Inositol	4.073	Up	8.55E-11	8.42E-10	1.855
Benzoic acid	0.100	Down	1.48E-09	1.35E-08	1.754
Succinate	0.338	Down	3.00E-07	2.40E-06	1.454
Lactate	0.264	Down	4.75E-07	3.57E-06	1.446
Isoleucine	0.311	Down	2.43E-06	1.73E-05	1.390
3-amino-2-Piperidone	0.408	Down	5.52E-06	3.72E-05	1.274
Caproic acid	0.293	Down	8.83E-06	5.65E-05	1.333
Threonine	0.400	Down	2.51E-05	0.000135	1.222
Octadecanoic acid	0.467	Down	2.53E-05	0.000135	1.186
Malate	0.370	Down	3.61E-05	0.000178	1.225
Myristoleic acid	0.456	Down	4.59E-05	0.00021	1.129
3-hydroxyoxyisovaleric acid	0.486	Down	0.000166	0.000733	1.126

The enzyme glycerol-3-phosphate dehydrogenase, a key enzyme involved in the conversion of G3P to dihydroxyacetone phosphate, has previously been reported to be downregulated in breast tumor tissue [33]. Although no studies related to breast cancer metabolomics have directly mentioned the role of G3P, the involvement of related enzymes may emphasize its potential as a new biomarker candidate.

In contrast, some metabolites were found to be highly upregulated in breast cancer, namely asparagine (FC=11.274) and cysteine (FC=6.098). These two metabolites play an important role in breast cancer for supporting glutathione metabolism and the regulation of amino acid metabolism of alanine, aspartate, and glutamate. An increase in asparagine may act as a nitrogen reservoir used to the synthesis of other amino acids and nucleotides, such as alanine, aspartate, and glutamate. This is supported by the decrease in the amount of aspartate and glutamate in the data analysis results. Aspartate and glutamate have been widely mentioned in previous breast cancer studies and have potential as markers for early detection and diagnosis [34-35].

The downregulation of aspartate and glutamate reflects the metabolic adaptation of cancer as tumors often modify the availability of amino acids as essential ingredients for protein, energy formation, and signaling molecules [19],[36]. In addition to these changes, cancer cells undergo amino acid metabolism reprogramming to meet their high demands for growth and survival. For instance, the conversion of threonine, sarcosine, and glyoxylic acid into glycine is essential for the rapid proliferation of cancer cells and contributes to nucleotide biosynthesis [36-37].

Branched-chain amino acids (BCAAs) like isoleucine and its intermediate, 3-hydroxyisovaleric acid, both downregulated in this study, also reflect the high demand from tumor and play a pivotal role in protein synthesis [36],[38]. There are several metabolomic studies on breast cancer have highlighted the potential of amino acid as biomarkers [26],[39-40]. However, the results on specific amino acids differ across studies, requiring further standardized research and validation to establish consistent and clinically reliable biomarker profiles.

Cysteine, along with glutamate and 5-oxoproline, plays a role in the synthesis of glutathione, one of the body's important antioxidants. Increased cysteine in breast cancer may increase the production of glutathione, which helps cancer cells deal with increased reactive oxygen species (ROS). This study also shows an increase in cystine, an amino acid that serves as a precursor for glutathione. Glutathione is used in neutralizing these molecules to protect cancer cells in an oxidative stress environment so that they can survive longer in the tumor microenvironment [35],[41]. Previous metabolomic studies have suggested triple negative breast cancer (TNBC) dependence on glutathione for survival, making it a promising potential therapeutic target [42].

The upregulation of pyruvate, together with the downregulation of lactate, shows the classic picture of the Warburg effect where cancer cells prefer to convert more glucose to lactate despite the availability of oxygen [43-44]. Pyruvate, along with succinate and malate, are also intermediate metabolites in the TCA cycle, which contribute to the progression of breast cancer [45].Succinate also affects immune system modulation and is a potential therapeutic target in cancer. This metabolite shift could be useful for future diagnostic and prognostic needs in breast cancer [46].

The dysregulation of lipid metabolism and signaling in breast cancer is reflected by the decrease in fatty acid metabolites and their derivatives found in the results of this study, namely arachidonic acid, octadecanoic acid, myristoleic acid, and caproic acid. Because of the lipid reprogramming, there are changes in lipid availability and composition, hence affecting the signaling and proliferation [47-49]. This is also supported by the rising of inositol concentration that influences the behaviour of cancer cells. Past studies reported involvement of inositol monophosphatase 1 (IMPA1) in TNBC progression [50].

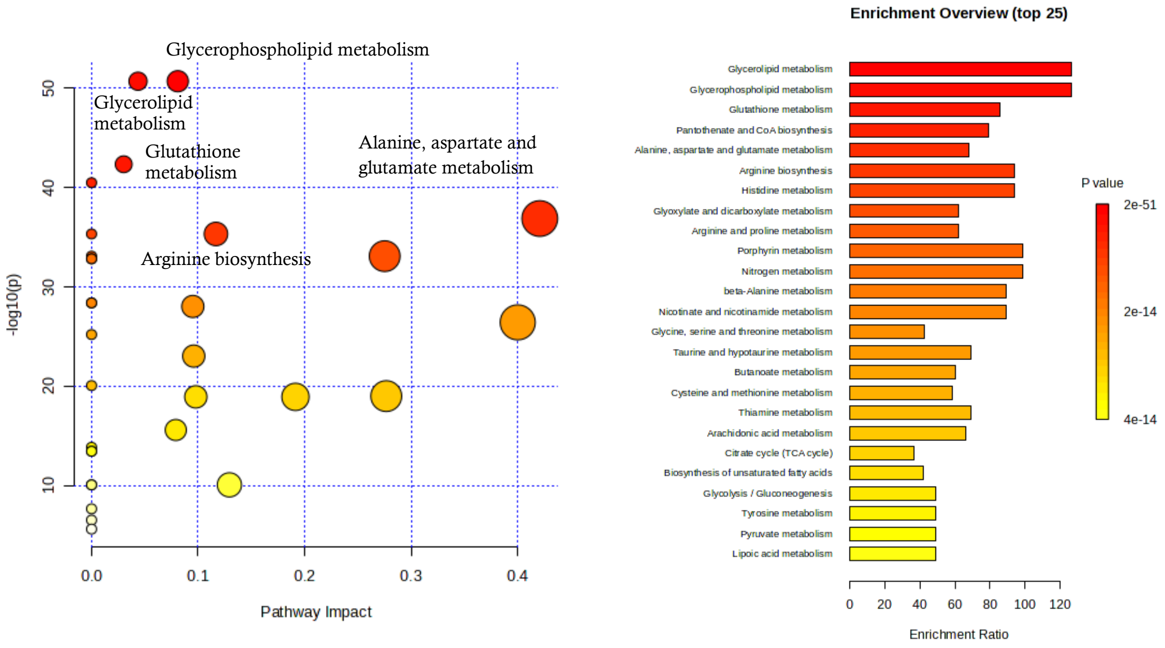


Figure 4. Pathway and Enrichment between BC and HC groups. Red color indicates more significant p-value; then circle size indicates significance of pathway impact and enrichment ratio.

After metabolite profiling between BC and HC groups, validation was done using several ML algorithm methods, namely SVM, RF, NN, NB, and LR with 10-fold cross validation based on 24 metabolites. NN, LR, and RF have advantages in all parameter metrics. In order to test the TST data, the three ML models could be used and seen based on AUC-ROC (Figure 5,6,7). Table 3 presents the validation differences between TRN data and TST data based on the three algorithms.

Table 3. Machine Learning Prediction Model Results on Biomarker candidates

Metric Parameter	Training Data			Testing Data		
	NN	LR	RF	NN	LR	RF
AUC	0.987	0.984	0.984	0.979	0.945	0.974
Precision	0.988	0.982	0.983	0.930	0.957	0.957
Recall	0.988	0.982	0.982	0.930	0.953	0.953
Specificity	0.986	0.976	0.968	0.901	0.913	0.913

NN, Neural Network; LR, Logistic Regression; RF, Random Forest; AUC, Area Under Curve;

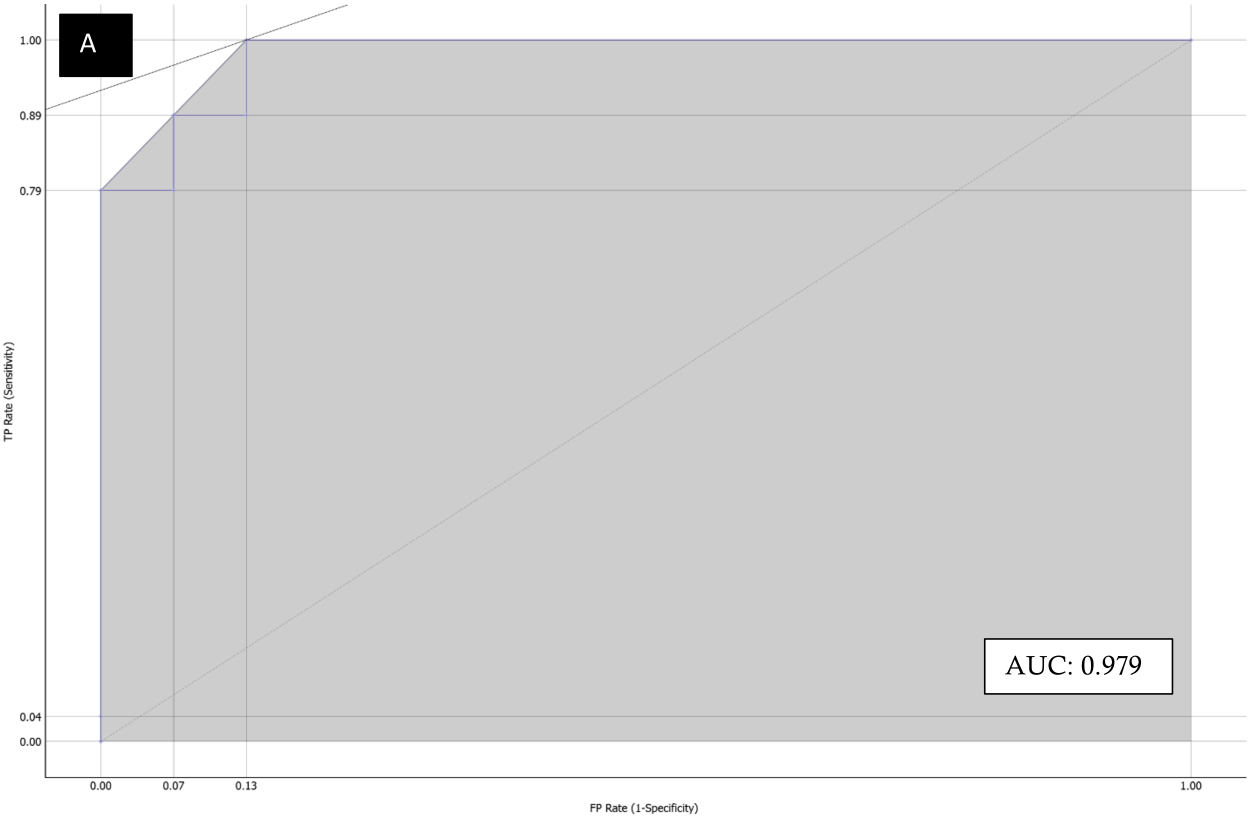


Figure 5. AUC-ROC on Testing Data based on biomarkers for BC group for Neural Network

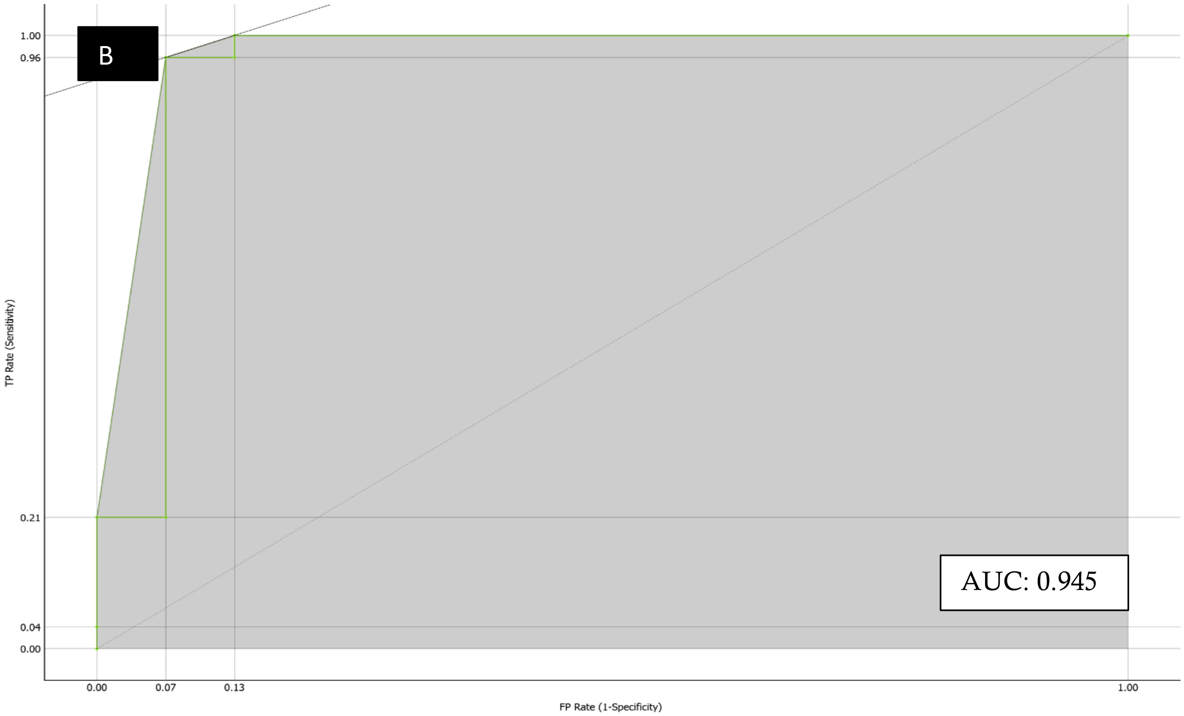


Figure 6. AUC-ROC on Testing Data based on biomarkers for BC group for Logistic Regression

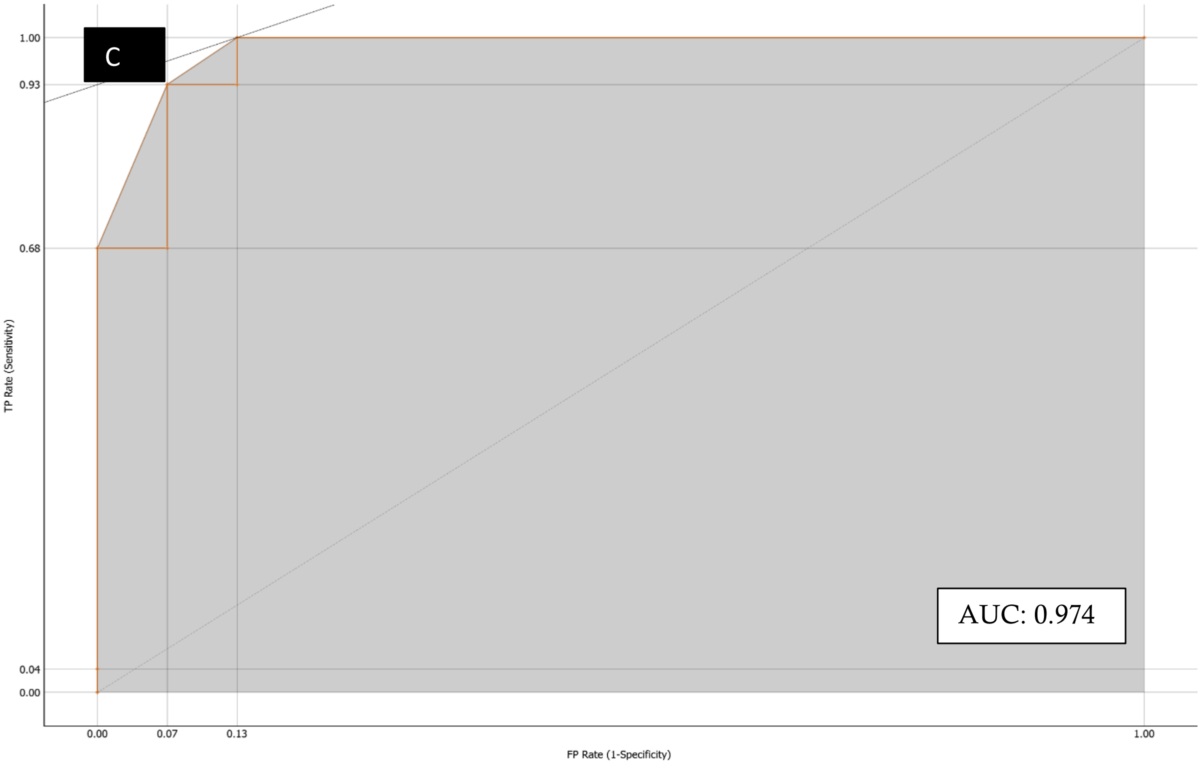


Figure 7. AUC-ROC on Testing Data based on biomarkers for BC group for Random Forest

The AUC on training data in all three ML models indicates near-perfect performance in effectively predicting breast cancer based on metabolite profiles. Although on testing data, the AUC value slightly decreased in performance, the three ML models could still produce >0.9. Comparison of prediction model results with ML algorithm on training and testing data for precision, recall, and specificity parameters shows the same trend as AUC. All experienced a slight decrease in performance on testing data, but still in the >0.9 range. AUC is a well-known metric to assess diagnostic tests because it can measure the model's ability to distinguish between diseased and healthy individuals with a consistent measurement, regardless of the prevalence of disease. However, AUC can be misleading as a sole metric due to its inability to reflect the model’s performance in real clinical settings, particularly when there are biases in the dataset. The inclusion of other classification metrics such as precision, recall, and specificity in evaluation can provide a more comprehensive understanding of model performance, particularly when there is an imbalance in class or when positive and negative predictions have different clinical significance [51-52].

Precision, also known as positive predictive value, measured the proportion of correct positive results among all positive predictions made by the model, so it is crucial in medical diagnostics. Recall, or sensitivity, assesses the ability of the model to identify true positive cases out of all true positive cases, which is also very important for early disease detection. Specificity evaluates the proportion of true negatives correctly identified out of all true negatives, ensuring that individuals without disease are accurately identified, thus reducing false positives and preventing unnecessary interventions, especially in high-risk scenarios, such as cancer detection, where misdiagnosis can have a major impact [53-54].

ML models have an important role in validating biomarkers in breast cancer metabolomics studies by analyzing complex data sets to identify patterns and predict clinical outcomes. These models utilized high-dimensional data from metabolomics, lipidomics, and other omics technologies to discover biomarkers that could aid in diagnosis, prognosis, and treatment response prediction. Integration with metabolomics enables the development of predictive models that can categorize patients based on their likelihood of responding to certain therapies or risk of disease progression.

In previous breast cancer metabolomics studies, the use of ML algorithms with NN, LR, and RF have been utilized. An Artificial Neural network (ANN) model achieved 97% accuracy in classifying breast cancer, demonstrating its effectiveness in processing complex data sets and identifying relevant biomarkers. ANN is particularly useful in handling high-dimensional data, such as metabolomics, it can study the non-linear relationship between metabolites and clinical outcomes, thus aiding in the identification of potential biomarkers [55-56]. Besides biomarkers, the utilization of deep neural networks is also used for the prediction of clinical response in anti-cancer drugs, which has the potential to integrate precision oncology in the future [57-58].

LR is a widely used ML algorithm for predicting breast cancer diagnosis from metabolomics data. It has simplicity and ease of interpretation, making it a popular choice in medical diagnosis. Breast cancer prediction using the LR algorithm is widely available and often achieves high accuracy when combined with optimized techniques. A breast cancer research study reported a prediction accuracy of 98.83% after applying optimization methods to LR [59]. It is also used to build diagnostic models by correlating metabolomic features with clinical phenotypes, achieving high accuracy and specificity. An ensemble approach model using LR achieved 98.8% accuracy in breast cancer classification, highlighting its robustness in biomarker validation [60-61].

Then, RF was used for its ability to handle large datasets and identify important features as well as to handle complex datasets and capture intricate patterns, making it highly effective for predicting breast cancer diagnosis on metabolomics data. An accuracy of 99.12% was reported for breast cancer classification using the Breast Cancer Wisconsin dataset. RF was also found to have better accuracy and reliability performance than other ML algorithms, such as decision tree and LR for breast cancer detection [62-63]. In addition, the combination of RF with other models to validate gene signatures in breast cancer confirmed its potential as a diagnostic tool [64].

Many challenges remained although machine learning models of ANN, LR, and RF were potential approaches for biomarker validation in breast cancer metabolomics. The complexity of metabolomics data, including high dimensionality and variability, required careful model selection and validation. A combination of multiple algorithms might also be required if it aimed to improve the prediction accuracy and outcome of breast cancer patients.

In addition, this study is a simple in silico analysis with inherent limitations, including the small size of the dataset, which constrains its ability to perform detailed classification of breast cancer, such as molecular subtyping. To overcome these limitations, future studies should use larger, more diverse datasets to boost the reliability of findings, and incorporate deep learning models to improve classification and performance. The accuracy and applicability of these models in clinical settings may be enhanced by integrating multi-omics data, such as genomics and proteomics, with metabolomics. Future research should also focus on improving the interpretability of the models, ensuring the reproducibility of findings across different populations, and validating results in clinical settings.

4. Conclusion

In silico metabolomics analysis yielded 24 metabolites, including glycerolphosphate, glutamate, 5-oxoproline, cysteine, aspartate, asparagine, pyruvate, and succinate, which showed distinct differences between the breast cancer group and healthy controls. Validation of these metabolites through ML algorithm implementation resulted in highly superior metric parameters, which highlight their power as potential biomarker candidates for distinguishing cancer from non-cancer group. These findings highlight the potential of ML-based metabolomics in performing early, personalized breast cancer diagnosis, paving the way for targeted therapeutic strategies and improved patient outcomes. However, further validation in clinical cohorts remains crucial to confirm the identified biomarkers. Additionally, integrating these findings with multi-omics approaches may provide a more comprehensive understanding of breast cancer, advancing precision oncology efforts.

References

[1] Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., Siesling, S., & Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, *66*, 15–23.

[2] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249.

[3] Gautama, W. (2022). Breast Cancer in Indonesia in 2022: 30 Years of Marching in Place. *Indonesian Journal of Cancer*, *16*(1), 1.

[4] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, *149*(4), 778–789.

[5] Widiana, I. K., & Irawan, H. (2020). Clinical and Subtypes of Breast Cancer in Indonesia. *Asian Pacific Journal of Cancer Care*, *5*(4), 281–285.

[6] Oluogun, W. A., Adedokun, K. A., Oyenike, M. A., & Adeyeba, O. A. (2019). Histological classification, grading, staging, and prognostic indexing of female breast cancer in an African population: A 10-year retrospective study. *International Journal of Health Sciences*, *13*(4), 3–9.

[7] Yang, L., Wang, Y., Cai, H., Wang, S., Shen, Y., & Ke, C. (2020). Application of metabolomics in the diagnosis of breast cancer: a systematic review. *Journal of Cancer*, *11*(9).

[8] He, Z., Chen, Z., Tan, M., Elingarami, S., Liu, Y., Li, T., Deng, Y., He, N., Li, S., Fu, J., & Li, W. (2020). A review on methods for diagnosis of breast cancer cells and tissues. *Cell Proliferation*, *53*(7).

[9] Lopez-Gonzalez, L., Sanchez Cendra, A., Sanchez Cendra, C., Roberts Cervantes, E. D., Espinosa, J. C., Pekarek, T., Fraile-Martinez, O., García-Montero, C., Rodriguez-Slocker, A. M., Jiménez-Álvarez, L., Guijarro, L. G., Aguado-Henche, S., Monserrat, J., Alvarez-Mon, M., Pekarek, L., Ortega, M. A., & Diaz-Pedrero, R. (2024). Exploring Biomarkers in Breast Cancer: Hallmarks of Diagnosis, Treatment, and Follow-Up in Clinical Practice. *Medicina*, *60*(1), 168.

[10] Wei, Y., Jasbi, P., Shi, X., Turner, C., Hrovat, J., Liu, L., Rabena, Y., Porter, P., & Gu, H. (2021). Early Breast Cancer Detection Using Untargeted and Targeted Metabolomics. *Journal of Proteome Research*, *20*(6), 3124–3133

[11] Neves Rebello Alves, L., Dummer Meira, D., Poppe Merigueti, L., Correia Casotti, M., do Prado Ventorim, D., Ferreira Figueiredo Almeida, J., Pereira de Sousa, V., Cindra Sant'Ana, M., Gonçalves Coutinho da Cruz, R., Santos Louro, L., Mendonça Santana, G., Erik Santos Louro, T., Evangelista Salazar, R., Ribeiro Campos da Silva, D., Stefani Siqueira Zetum, A., Silva dos Reis Trabach, R., Imbroisi Valle Errera, F., de Paula, F., de Vargas Wolfgramm dos Santos, E., … Drumond Louro, I. (2023). Biomarkers in Breast Cancer: An Old Story with a New End. *Genes*, *14*(7), 1364.

[12] Bhushan, A., Gonsalves, A., & Menon, J. U. (2021). Current State of Breast Cancer Diagnosis, Treatment, and Theranostics. *Pharmaceutics*, *13*(5), 723.

[13] Silva, C., Perestrelo, R., Silva, P., Tomás, H., & Câmara, J. S. (2019). Breast Cancer Metabolomics: From Analytical Platforms to Multivariate Data Analysis. A Review. *Metabolites*, *9*(5), 102.

[14] Subramani, R., Poudel, S., Smith, K. D., Estrada, A., & Lakshmanaswamy, R. (2022). Metabolomics of Breast Cancer: A Review. *Metabolites*, *12*(7), 643.

[15] An, R., Yu, H., Wang, Y., Lu, J., Gao, Y., Xie, X., & Zhang, J. (2022). Integrative analysis of plasma metabolomics and proteomics reveals the metabolic landscape of breast cancer. *Cancer & Metabolism*, *10*(1), 13.

[16] Gupta, A., Siddiqui, Z., Sagar, G., Rao, K. V. S., & Saquib, N. (2023). A non-invasive method for concurrent detection of multiple early-stage cancers in women. *Scientific Reports*, *13*(1), 19083.

[17] Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46.

[18] Danzi, F., Pacchiana, R., Mafficini, A., Scupoli, M. T., Scarpa, A., Donadelli, M., & Fiore, A. (2023). To metabolomics and beyond: a technological portfolio to investigate cancer metabolism. *Signal Transduction and Targeted Therapy*, *8*(1), 137.

[19] Bel'skaya, L. V., Gundyrev, I. A., & Solomatin, D. V. (2023). The Role of Amino Acids in the Diagnosis, Risk Assessment, and Treatment of Breast Cancer: A Review. *Current Issues in Molecular Biology*, *45*(9), 7513–7537.

[20] Ruan, X., Wang, Y., Zhou, L., Zheng, Q., Hao, H., & He, D. (2022). Evaluation of Untargeted Metabolomic Strategy for the Discovery of Biomarker of Breast Cancer. *Frontiers in Pharmacology*, *13*.

[21] Díaz-Beltrán, L., González-Olmedo, C., Luque-Caro, N., Díaz, C., Martín-Blázquez, A., Fernández-Navarro, M., Ortega-Granados, A. L., Gálvez-Montosa, F., Vicente, F., Pérez del Palacio, J., & Sánchez-Rovira, P. (2021). Human Plasma Metabolomics for Biomarker Discovery: Targeting the Molecular Subtypes in Breast Cancer. *Cancers*, *13*(1), 147.

[22] Chen, Z., Li, Z., Li, H., & Jiang, Y. (2019). Metabolomics: A promising diagnostic and therapeutic implement for breast cancer. In *OncoTargets and Therapy* (Vol. 12, pp. 6797–6811). Dove Medical Press Ltd.

[23] Jacob, M., Lopata, A. L., Dasouki, M., & Abdel Rahman, A. M. (2019). Metabolomics toward personalized medicine. *Mass Spectrometry Reviews*, *38*(3), 221–238.

[24] Lelli, V., Belardo, A., & Maria Timperio, A. (2021). From Targeted Quantification to Untargeted Metabolomics. In *Metabolomics - Methodology and Applications in Medical Sciences and Life Sciences*. IntechOpen.

[25] Gong, S., Wang, Q., Huang, J., Huang, R., Chen, S., Cheng, X., Liu, L., Dai, X., Zhong, Y., Fan, C., & Liao, Z. (2024). LC-MS/MS platform-based serum untargeted screening reveals the diagnostic biomarker panel and molecular mechanism of breast cancer. *Methods*, *222*, 100–111.

[26] Zou, Y., Song, D., Cai, Y., Liang, K., Fu, J., & Zhang, H. (2024). *Comprehensive untargeted serum metabolomics identifies biomarkers and metabolic pathways in breast cancer*.

[27] Shoaib, A. S. M., Nishat, N., Raasetti, M., & Arif, I. (2024). Integrative Machine Learning Approaches For Multi-Omics Data Analysis In Cancer Research. *Global Mainstream Journal*, *1*(2), 26–39.

[28] Ngan, H.-L., Lam, K.-Y., Li, Z., Zhang, J., & Cai, Z. (2023). Machine learning facilitates the application of mass spectrometry-based metabolomics to clinical analysis: A review of early diagnosis of high mortality rate cancers. *TrAC Trends in Analytical Chemistry*, *168*.

[29] Kuang, A., Kouznetsova, V. L., Kesari, S., & Tsigelny, I. F. (2023). Diagnostics of Thyroid Cancer Using Machine Learning and Metabolomics. *Metabolites*, *14*(1), 11.

[30] Sugimoto, M., Hikichi, S., Takada, M., & Toi, M. (2023). Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. *Annals of Breast Surgery*, *7*, 7–7.

[31] Albrecht, W. (2019). Highlight report: Role of choline phospholipid metabolism in tumor progression. *EXCLI Journal*, *18*, 1097–1098.

[32] Ma, Y., Zhang, S., Jin, Z., & Shi, M. (2020). Lipid-mediated regulation of the cancer-immune crosstalk. *Pharmacological Research*, *161*, 105131.

[33] Yoneten, K. K., Kasap, M., Akpinar, G., Gunes, A., Gurel, B., & Utkan, N. Z. (2019). Comparative Proteome Analysis of Breast Cancer Tissues Highlights the Importance of Glycerol-3-phosphate Dehydrogenase 1 and Monoacylglycerol Lipase in Breast Cancer Metabolism. *Cancer Genomics - Proteomics*, *16*(5), 377–397.

[34] Lin, Y., Yang, Z., Li, J., Sun, Y., Zhang, X., Qu, Z., Luo, Y., & Zhang, L. (2022). Effects of glutamate and aspartate on prostate cancer and breast cancer: a Mendelian randomization study. *BMC Genomics*, *23*(1), 213.

[35] Ullah Khan, S., & Ullah Khan, M. (2022). The Role of Amino Acid Metabolic Reprogramming in Tumor Development and Immunotherapy. *Biochemistry and Molecular Biology*, *7*(1), 6

[36] Zhang, J., Chen, M., Yang, Y., Liu, Z., Guo, W., Xiang, P., Zeng, Z., Wang, D., & Xiong, W. (2024). Amino acid metabolic reprogramming in the tumor microenvironment and its implication for cancer therapy. *Journal of Cellular Physiology*.

[37] Sun, W., Zhao, E., & Cui, H. (2023). Target enzymes in serine-glycine-one-carbon metabolic pathway for cancer therapy. *International Journal of Cancer*, *152*(12), 2446–2463.

[38] Lieu, E. L., Nguyen, T., Rhyne, S., & Kim, J. (2020). Amino acids in cancer. *Experimental & Molecular Medicine*, *52*(1), 15–30.

[39] Da Cunha, P. A., Nitusca, D., Canto, L. M. Do, Varghese, R. S., Ressom, H. W., Willey, S., Marian, C., & Haddad, B. R. (2022). Metabolomic Analysis of Plasma from Breast Cancer Patients Using Ultra-High-Performance Liquid Chromatography Coupled with Mass Spectrometry: An Untargeted Study. *Metabolites*, *12*(5).

[40] Panigoro, S. S., Kurniawan, A., Ramadhan, R., Sukartini, N., Herqutanto, H., Paramita, R. I., & Sandra, F. (2023). Amino Acid Profile of Luminal A and B Subtypes Breast Cancer. *The Indonesian Biomedical Journal*, *15*(3), 269–276.

[41] Jaune-Pons, E., & Vasseur, S. (2020). Role of amino acids in regulation of ROS balance in cancer. *Archives of Biochemistry and Biophysics*, *689*, 108438.

[42] Beatty, A., Fink, L. S., Singh, T., Strigun, A., Peter, E., Ferrer, C. M., Nicolas, E., Cai, K. Q., Moran, T. P., Reginato, M. J., Rennefahrt, U., & Peterson, J. R. (2023). *Data from Metabolite Profiling Reveals the Glutathione Biosynthetic Pathway as a Therapeutic Target in Triple-Negative Breast Cancer*.

[43] Niepmann, M. (2024). Importance of Michaelis Constants for Cancer Cell Redox Balance and Lactate Secretion—Revisiting the Warburg Effect. *Cancers*, *16*(13), 2290.

[44] Barba, I., Carrillo-Bosch, L., & Seoane, J. (2024). Targeting the Warburg Effect in Cancer: Where Do We Stand? *International Journal of Molecular Sciences*, *25*(6), 3142.

[45] Zakic, T., Kalezic, A., Drvendzija, Z., Udicki, M., Ivkovic Kapicl, T., Srdic Galic, B., Korac, A., Jankovic, A., & Korac, B. (2024). Breast Cancer: Mitochondria-Centered Metabolic Alterations in Tumor and Associated Adipose Tissue. *Cells*, *13*(2), 155.

[46] Zhang, W., & Lang, R. (2023). Succinate metabolism: a promising therapeutic target for inflammation, ischemia/reperfusion injury and cancer. *Frontiers in Cell and Developmental Biology*, *11*.

[47] Moreira-Barbosa, C., Matos, A., Fernandes, R., Mendes-Ferreira, M., Rodrigues, R., Cruz, T., Costa, Â. M., Cardoso, A. P., Ghilardi, C., Oliveira, M. J., & Ribeiro, R. (2023). The role of fatty acids metabolism on cancer progression and therapeutics development. In *Bioactive Lipids* (pp. 101–132).

[48] He, M., Xu, S., Yan, F., Ruan, J., & Zhang, X. (2023). Fatty Acid Metabolism: A New Perspective in Breast Cancer Precision Therapy. *Frontiers in Bioscience-Landmark*, *28*(12).

[49] Du, A., Wang, Z., Huang, T., Xue, S., Jiang, C., Qiu, G., & Yuan, K. (2023). Fatty acids in cancer: Metabolic functions and potential treatment. *MedComm – Oncology*, *2*(1).

[50] Yang, S., Xie, Y., Zhang, T., Deng, L., Liao, L., Hu, S., Zhang, Y., Zhang, F., & Li, D. (2023). Inositol monophosphatase 1 (<scp>IMPA1</scp>) promotes triple-negative breast cancer progression through regulating <scp>mTOR</scp> pathway and <scp>EMT</scp> process. *Cancer Medicine*, *12*(2), 1602–1615.

[51] Kleppe, A. (2022). Area under the curve may hide poor generalisation to external datasets. *ESMO Open*, *7*(2), 100429.

[52] Parodi, S., Verda, D., Bagnasco, F., & Muselli, M. (2022). The clinical meaning of the area under a receiver operating characteristic curve for the evaluation of the performance of disease markers. *Epidemiology and Health*, *44*, e2022088.

[53] Chowdhury, M. Z. I., & Turin, T. C. (2020). Precision health through prediction modelling: factors to consider before implementing a prediction model in clinical practice. *Journal of Primary Health Care*, *12*(1), 3.

[54] Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., & Dmochowski, R. R. (2021). Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina*, *57*(5), 503.

[55] Bhuta, N., & Raut, R. (2023). Breast Cancer Classification Using ANN and ML Techniques. *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–5.

[56] Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1), 64.

[57] Liu, X., Song, C., Huang, F., Fu, H., Xiao, W., & Zhang, W. (2022). GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Briefings in Bioinformatics*, *23*(1).

[58] Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T. J., Piha-Paul, S., Zhou, H., Kardala, E., Damianidou, E., Alexopoulos, L. G., Aifantis, I., Townsend, P. A., Panayiotidis, M. I., Sfikakis, P., Bartek, J., Fitzgerald, R. C., Thanos, D., …

Gorgoulis, V. G. (2019). A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Reports*, *29*(11), 3367-3373.e4.

[59] Sun, T. (2024). Breast cancer prediction based on multiple machine learning algorithms. *Highlights in Science, Engineering and Technology*, *92*, 241–247.

[60] Vaida, M. L., Arumalla, K., Tatikonda, P., Popuri, B., Bux, R., Tappia, P. S., Huang, G., Haince, J.-F., & Ford, W. R. (2024). *Identifying Robust Biomarker Panels for Breast Cancer Screening*.

[61] Wang, J., Ma, F., Sun, X., Wang, J., Guo, F., Liu, B., Wang, W., Li, Q., & Xu, B. (2022). Peripheral lipidomics analyses with ensemble machine learning predict response to neoadjuvant therapy in breast cancer. *Journal of Clinical Oncology*, *40*(16_suppl), 582–582.

[62] Md Zahidul Islam, Md Nasiruddin, Shuvo Dutta, Rajesh Sikder, Chowdhury Badrul Huda, & Md Rasibul Islam. (2024). A Comparative Assessment of Machine Learning Algorithms for Detecting and Diagnosing Breast Cancer. *Journal of Computer Science and Technology Studies*, *6*(2), 121–135.

[63] Rahman, S., Siregar, D., Syah, R. B. Y., Setiawan, H., Maulana, A. E., & Hamsiah. (2023). The Effective Breast Cancer Classification with the Random Forest Algorithm. *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 1–5.

[64] Mirza, Z., Ansari, M. S., Iqbal, M. S., Ahmad, N., Alganmi, N., Banjar, H., Al-Qahtani, M. H., & Karim, S. (2023). Identification of Novel Diagnostic and Prognostic Gene Signature Biomarkers for Breast Cancer Using Artificial Intelligence and Machine Learning Assisted Transcriptomics Analysis. *Cancers*, *15*(12), 3237.