Machine learning techniques for breast cancer diagnosis and treatment: a narrative review

Masahiro Sugimoto[1,2,3]^, **Shiori Hikichi**[1,4], **Masahiro Takada**[3], **Masakazu Toi**[3]

[1]Research and Development Center for Minimally Invasive Therapies, Institute for Medical Science, Tokyo Medical University, Tokyo, Japan; [2]Institute for Advanced Biosciences, Keio University, Yamagata, Japan; [3]Department of Breast Surgery, Kyoto University Hospital, Kyoto, Japan; [4]Research Fellow of Japan Society for the Promotion of Science, Tokyo, Japan

Contributions: (I) Conception and design: M Toi, M Takada, S Hikichi; (II) Administrative support: M Sugimoto; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: None; (V) Data analysis and interpretation: None; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Masahiro Sugimoto, PhD. Professor, Research and Development Center for Minimally Invasive Therapies, Institute of Medical Science, Tokyo Medical University, 6-1-1, Shinjuku, Tokyo, 160-8402, Japan. Email: mshrsgmt@tokyo-med.ac.jp.

Objective: This narrative review describes the recent developments and applications of machine learning (ML), a part of artificial intelligence, concerning breast cancer.

Background: The advent of new bioinformatic approaches and artificial intelligence-based computational technologies has led to a shift in the decision-making of oncologists regarding breast cancer diagnostics and treatment processes. Various successful applications of ML on image processing, especially the use of deep neural networks and convolutional neural networks, to detect tumor and lymph nodes regions have been reported. Recent high-throughput molecular quantifications, i.e., quantitative omics techniques have enabled simultaneous monitoring of thousands of molecules to understand the molecular-level pathology. These data, including gene expression, protein, metabolite, and methylation profiling, have been analyzed via deep learning, network analysis, clustering, and dimension reductions to explore intrinsic subtypes and new biomarkers. Clinical-pathological features have been conducted by multivariable analysis to predict various outcomes, e.g., the sensitivity of adjuvant therapy and prognosis. The quantitative relationships among their variables have been visualized as nomograms. To analyze complex structures of a larger number of variables, ML combining multiple clinical-pathological features has been developed to predict the prognosis, metastasis, and treatment outcomes of breast cancer.

Methods: We provided the narrative review of ML-related topics especially in the quantitative omics data and clinical-pathological prediction models.

Conclusions: ML-based prediction methods are powerful tools and contribute to realizing personalized medicine for breast cancer.

Keywords: Artificial intelligence (AI); machine learning (ML); nomogram; omics; network analysis

Received: 11 April 2021; Accepted: 26 August 2021; Published online: 17 September 2021.
doi: 10.21037/abs-21-63
View this article at: https://dx.doi.org/10.21037/abs-21-63

^ ORCID: 0000-0003-3316-2543.

Introduction

Cancer is one of the major health problems, and 19.3 million new cases of cancer and almost 10 million deaths from cancer are estimated worldwide for 2020 (1). Among females, breast cancer is the most commonly diagnosed cancer with an estimated 2.3 million new cases, followed by colorectal and lung cancer in incidence, and vice versa in mortality (1). Among females in the United States, breast cancer is the most diagnosed cancer and the second cause of death among most of the cancer types in 2021. However, there is concern regarding the underestimation of new cases because of the coronavirus disease 2019 (COVID-19) pandemic (2). The development of new technology for automatic screening is needed for the early detection and treatment of breast cancer.

With advances in technology, the recent rapid increase in computational resources has yielded high-dimensional data, whereby accurate diagnosis and prediction of the prognosis and treatment outcome of breast cancers are possible. The capability of imaging breast lesions through techniques such as mammography (MMG), ultrasonography (US), and magnetic resonance imaging (MRI) contributes to characterizing the malignant region(s), which provides information complemental to biopsies for definitive diagnoses (3). Molecular-based profiling data, such as transcriptomics, proteomics, and metabolomics, have also been accumulated using omics technologies, and these data help decipher the molecular pathways dysregulated in breast cancer (4). To exploit the clinical utility and to understand tumorigenesis and progression of breast cancer from these types of big data, sophisticated computational analyses are necessary in addition to conventional statistical approaches.

As a computational analytical methodology, machine learning (ML), one of the branches of artificial intelligence (AI), has been implemented to analyze these data, for which there exist unsupervised and supervised methods. Unsupervised methods involve the extraction of repeatedly observed or implicit features without a pre-definition of the expected results. The clustering technique is one example thereof and is frequently employed in the analysis of quantitative omics data, such as the subtype-specific expression patterns of transcriptome data. Supervised methods, on the contrary, mimic given results by combining the observed features. Various ML methods have been used for this purpose, such as the artificial neural network (ANN), support vector machine (SVM), naïve Bayes algorithm, random forest (RF), and decision tree. Approximately 3,000 papers have been published in this area in the last five years (2015–2019) (5) owing to publicly available databases such as the Wisconsin Breast Cancer Database and Breast Cancer Coimbra Data Set, which have been repeatedly used for the development and validation of new methods (6,7). The benchmark tests to compare the prediction performance of ML methods have been reviewed using various datasets (8-10).

Image processing technologies have been improving rapidly and significantly, particularly the use of deep neural networks (DNNs) and convolutional neural network (CNN). The applications of ML for image processing have been frequently reviewed and are a topic of high research interest (11-16). A large-scale database including histopathological images of breast cancer is publicly available for these analyses (17). A combination of supervised and unsupervised CNN has been used for histopathological data (18). DNNs have been also employed in breast cancer risk assessment using whole genomic data (named polygenic risk scores), a technique that has been also well-reviewed (19,20).

Thus, image processing is currently a hot topic in breast cancer diagnosis; therefore, various review papers are already available. This manuscript provides a review of the application of ML to the other types of data, especially for clinical-pathological features and quantitative molecular profiles. We present the following article in accordance with the Narrative Review reporting checklist (available at https://abs.amegroups.com/article/view/10.21037/abs-21-63/rc).

Methods

Information used to write this paper was collected from PubMed and Google Scholar.

Clinical-pathological feature-based prediction

Index

Combinations of clinical-pathological features have long been used to assess diseases and treatment responses in patients. For this purpose, the Nottingham prognostic index (NPI) was developed in 1982 (21) and involved three factors, namely tumor size, stage of the disease, and tumor grade, to predict the prognosis of primary and operable breast cancers (22). This index has been frequently validated and widely employed in the prognosis of breast cancer (23). Various other factors, such as the vascular endothelial growth factor (VEGF), are prognostic factors independent of the NPI for patients with node-negative breast cancers (24,25). Lymphovascular invasion

(LVI) and progesterone receptor (PgR) were also proposed as factors independent of NPI to predict the prognosis (26,27). Thus, the NPI cannot reveal the entire clinical and survival outcome of breast cancer heterogeneity, and a new index incorporating molecular-based biomarkers was developed. The Nottingham prognostic index plus (NPI+) was developed using an ANN to combine a large number of molecular expression levels in a non-linear manner (28). This approach has improved the predictive ability of each clinical-pathological feature, considering their complex and non-linear relationship.

Nomogram

The ability to visualize the relationships among the quantitative effects of each observed feature on the outcome would aid the decision-making process when selecting the treatment mode. Adjuvant! Online was first developed for this purpose (29). This graphical web tool visualizes the risk of cancer-related mortality or relapse without therapy and the reduction in risk with therapy. However, it is limited on several counts; for example, the human epidermal growth factor receptor 2 (HER2) status was not included. As similar approaches, various nomograms have since been developed to combine the predictive effect of each feature linearly. In this regard, PREDICT was developed in the United Kingdom to predict the overall survival and treatment effect of systemic therapy in patients with primary breast cancer using the factors of age, menopausal status, disease stage, estrogen receptor (ER), HER2, and Ki-67 labeling index (Ki-67) (30). Similar to Adjuvant! Online, PREDICT has been well-validated. Subsequently, Rouzier *et al.* developed two nomograms to predict the residual tumor size and the applicable probability of conservative surgery after neoadjuvant chemotherapy (31).

In addition, nomograms for specific subtypes of breast cancer have been developed, such as for triple-negative, i.e., ER/PgR-negative and HER2-negative (32,33) as well as ER/PgR-positive and HER2-negative patients (34). Recently, a systematic review was published, evaluating a total of 58 mathematical prediction models for disease prognosis (35). Most of these models utilized Cox proportional hazards regression to predict mortality, recurrence, or both, and they were calibrated using the C-index or the area under the receiver operating curve (AUC). Moreover, nodal status, tumor size, tumor grade, age, and ER status were included as parameters. However, only 17 of the 58 models were validated in independent populations (i.e., external validation).

Some of these nomograms, validated previously, were used to aid the prognosis estimation. To enhance the generalization ability of the model, that is, the ability of the model to work accurately with independent data not included in the model development, independent minimum variables sets are usually selected (36). Thus, only a few variables can be incorporated into a single nomogram, which limits the accounts of heterogeneity as well as the accuracy of the predictions (*Figure 1A*).

Machine learning methods

ML methods can potentially enhance observable variables that predict outcomes, e.g., prognosis, diagnosis, and treatment sensitivity, by combining them in complicated structures. The objective of each method is to attempt to explore underlying data patterns and the relationship among the data, which would contribute to realizing accurate predictions. Various supervised ML methods have been used for this purpose.

The pathological complete response (pCR) of neoadjuvant chemotherapy (NAC) was predicted using various types of classification methods, such as the RF, naïve Bayes algorithm, SVM, and ANN; a cross-validation-based accuracy comparison concluded that RF yielded the best performance (37). An alternative decision tree (ADTree) was used to predict the pCR of NAC of patients with primary breast cancers (38) (*Figure 1B*). A prediction model was developed using training data collected from three institutions including Tokyo Metropolitan Cancer and Infection Diseases Centre Komagome Hospital, Osaka National Hospital, and Tsukuba University Hospital with cross-validation as a means of internal validation. The developed model was evaluated using independent validation data collected from the Organisation for Oncology and Translational Research (OOTR) N003 which is a randomized trial of patients with operable breast cancer treated with docetaxel with or without capecitabine after four cycles of NAC consisting of 5-fluorouracil, epirubicin, and cyclophosphamide (FEC) (UMIN ID: C000000322, http://www.umin.ac.jp/ctr/index.htm) (*Figure 2A*).

ADTree-based prediction model also predicted the metastasis of the axillary lymph node in patients with breast cancer who had not received prior treatment (39). The training data included the breast cancer patients from the Tokyo Metropolitan Cancer and Infectious Diseases Centre Komagome Hospital and Kyoto University Hospital, whose maximum tumor size was ≤4 cm. The independent validation

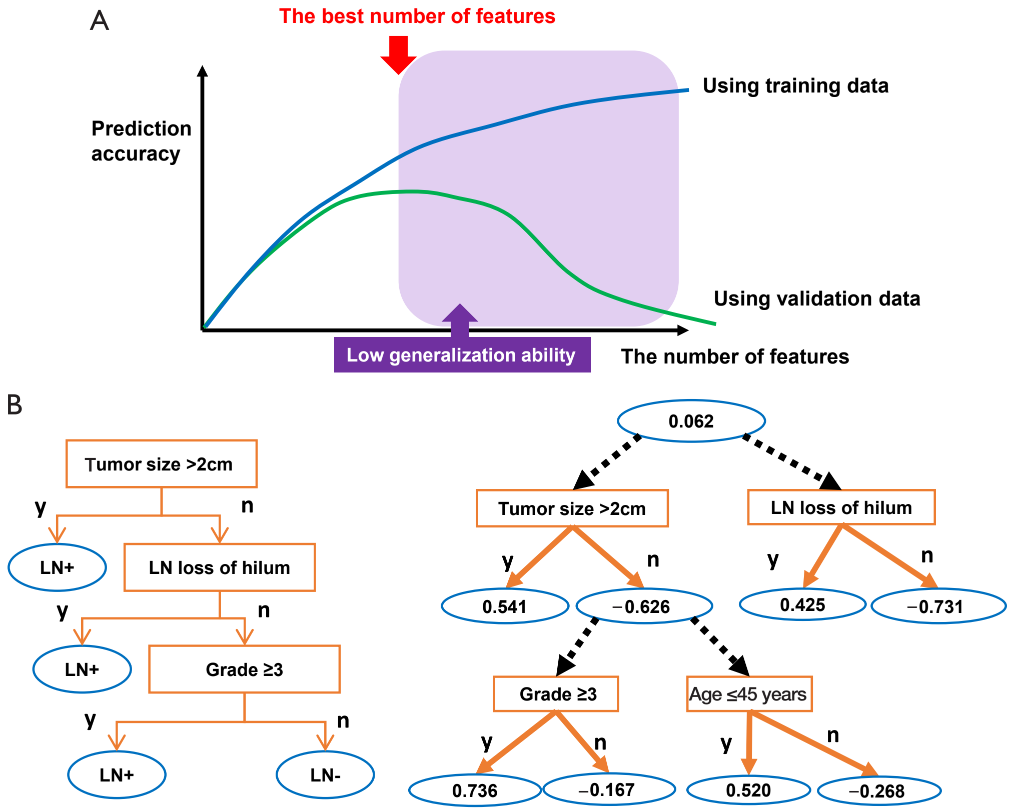


Figure 1 The machine learning method using clinical-pathological features. (A) The relationship between the prediction accuracies and the number of features used for the prediction model. The prediction accuracies using training data increase along with the number of features uses, whereas those using validation data show a peak. The number of features at this peak should be selected to enhance the generalization ability of the model. (B) Examples of tree-type algorithms. Conventional decision tree (left) and alternative decision tree (right). LN, lymph node.

data was collected from Seoul National University Hospital, Korea, and consisted of the patients who underwent sentinel lymph node (SLN) biopsy and met the same eligibility criteria as the modeling dataset (*Figure 2B*).

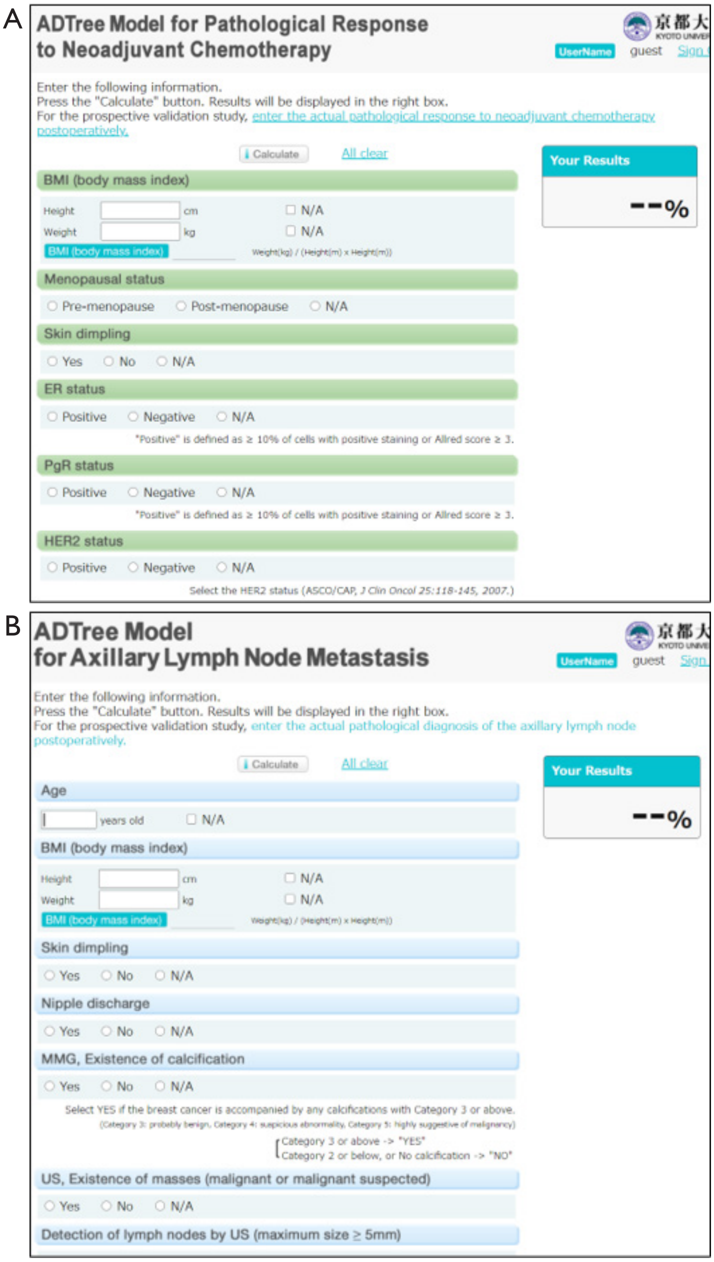
ADTree model was also used to predict the disease-free survival (DFS) and brain metastasis in patients with HER2-positive breast cancer who had received NAC plus trastuzumab in the Japan Breast Cancer Research Group (JBCRG)-03 study (40,41). These models involve an ensemble technique, whereby multiple prediction models can be developed and their predictions can be integrated to enhance the prediction accuracies. Notably, this method also showed robustness against missing values (42,43).

Quantitative omics data analysis

Commercially available gene expression analyses

Traditionally, transcriptional data based on gene expression have been analyzed. However, in recent times, other various omics data, such as proteomics and metabolites, have also been simultaneously analyzed to understand the molecular-based heterogeneity of cancers (44). The clinical utilities of several gene expression profiling tools such as Prediction Analysis of Microarray 50 (PAM50)® (45), MammaTyper® (46), MammaPrint® (47), Oncotype DX® (48), Endopredict® (49), and Genomic Grade Index® (50) have already been well-validated, and these tests are commercially available and widely used in clinical practice (51).

For example, MammaPrint® is an FDA-approved test for *in vitro* diagnostic multivariate index assays. This test previously involved the use of frozen samples, whereas it currently involves formalin-fixed, paraffin-embedded (FFPE) blocks for the analysis of 70 gene signatures using a microarray (52). Notably, this test showed better prognosis prediction ability than that of Adjuvant! Online (53). In MINDACT (a prospective study), patients with ER-



ADTree Model for Pathological Response to Neoadjuvant Chemotherapy

Enter the following information.
Press the "Calculate" button. Results will be displayed in the right box.
For the prospective validation study, enter the actual pathological response to neoadjuvant chemotherapy postoperatively.

BMI (body mass index)

Height ___ cm □ N/A
Weight ___ kg □ N/A
BMI (body mass index) ___ Weight(kg) / (Height(m) x Height(m))

Menopausal status

○ Pre-menopause ○ Post-menopause ○ N/A

Skin dimpling

○ Yes ○ No ○ N/A

ER status

○ Positive ○ Negative ○ N/A

"Positive" is defined as ≥ 10% of cells with positive staining or Allred score ≥ 3.

PgR status

○ Positive ○ Negative ○ N/A

"Positive" is defined as ≥ 10% of cells with positive staining or Allred score ≥ 3.

HER2 status

○ Positive ○ Negative ○ N/A

Select the HER2 status (ASCO/CAP, *J Clin Oncol 25:118-145, 2007.*)

ADTree Model for Axillary Lymph Node Metastasis

Enter the following information.
Press the "Calculate" button. Results will be displayed in the right box.
For the prospective validation study, enter the actual pathological diagnosis of the axillary lymph node postoperatively.

Age

___ years old □ N/A

BMI (body mass index)

Height ___ cm □ N/A
Weight ___ kg □ N/A
BMI (body mass index) ___ Weight(kg) / (Height(m) x Height(m))

Skin dimpling

○ Yes ○ No ○ N/A

Nipple discharge

○ Yes ○ No ○ N/A

MMG, Existence of calcification

○ Yes ○ No ○ N/A

Select YES if the breast cancer is accompanied by any calcifications with Category 3 or above.
(Category 3: probably benign, Category 4: suspicious abnormality, Category 5: highly suggestive of malignancy)
[Category 3 or above -> "YES"
Category 2 or below, or No calcification -> "NO"

US, Existence of masses (malignant or malignant suspected)

○ Yes ○ No ○ N/A

Detection of lymph nodes by US (maximum size ≥ 5mm)

Figure 2 Web site to predict the response of neoadjuvant chemotherapy (A) and axillary lymph node metastasis (B) of primary breast cancer. http://www.brca.jp/index.html. The account is issued upon request to the corresponding author. ADTree, alternative decision tree.

positive disease, who were at high clinical risk as defined by Adjuvant! Online and low genomic risk as defined by MammaPrint, showed an excellent prognosis without having received chemotherapy (54). Oncotype DX^{TM} extracts RNA from FFPE blocks and quantifies 21 gene profiles via RT-PCR (55). The 21 genes include 16 tumor-related and 5 reference genes for the prediction of the recurrence score (RS). This test was validated via tamoxifen-treated breast cancer patients with ER+/node-negative and showed a significantly different prognosis among high- and low-risk groups (48). Validation via these patients also showed prognostic value (56), and the test showed the predictive ability for the case of adjuvant therapy of tamoxifen + chemotherapy for breast cancer patients with ER+/node-

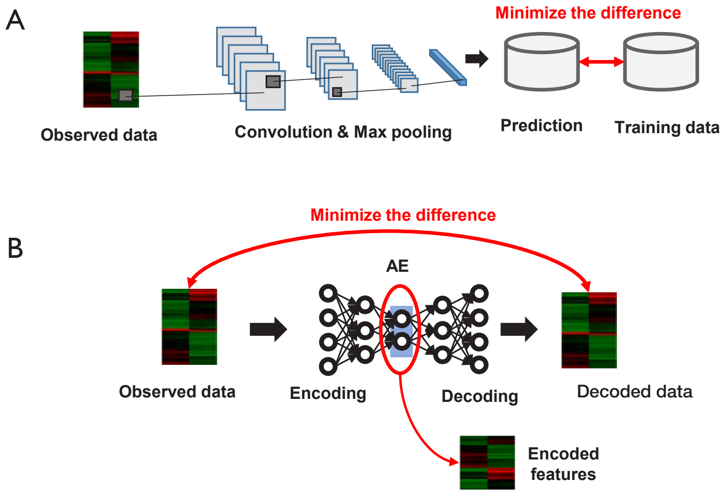


Figure 3 ANN-based data analysis of high-dimensional data, e.g., microarray gene expression. (A) A typical supervised learning method. CNN tries to predict a phenotype, e.g., good/poor prognosis, from observed data. The CNN evaluates the difference between the prediction and training data and tries to minimize this difference. (B) A typical unsupervised method. AE tries to produce lower-dimensional features from the observed data (encoding) and also reproduces the original data from the encoded features (decoding). AE tries to minimize the difference between the observed and decoded data. AE, autoencoder; ANN, artificial neural network; CNN, convolutional neural network.

negative (57). The clinical utility of Oncotype DX^{TM} was evaluated in large prospective trials. In the TAILORx trial, adjuvant endocrine therapy and chemoendocrine therapy showed similar efficacies in patients with RS of 11–25 (58). In another study, PAM50 was used to analyze 50 gene expression profiles to classify breast cancer into intrinsic subtypes (45). MammaPrint involved the use of profiles of 70 gene expressions for prediction of recurrence (59) and PAM50 involved clustering to calculate the distance of gene expression patterns for evaluation of the distances among the given and intrinsic subtypes (60). All these assays utilized relatively simple data analytical methods, such as multiple logistic regression (MLR) and clustering.

Multi-gene assay with ML

To consider more complex relationships among genes, ML methods have been used instead of the aforementioned established methods. The genes considered for PAM50 on RNA-sequence data were also analyzed via ML methods for evaluating the similarity among the given data and pre-defined subtypes. Moreover, CNN (*Figure 3A*) was used to analyze gene expression profiles collected from The Cancer Genome Atlas (TCGA) database (61); DNNs were used to explore the subtype-specific expression pattern, whereby six new subtypes of triple-negative breast cancer were found (62).

It should be noted that there exist several difficulties in analyzing high-dimensional gene expression data. For example, various genes show positive correlations (multi co-linearity among the observed features), which prevent important genes with higher effects on the other molecules from being found. Instead, network analyses are used to visualize their relationship using quantitative criteria between two genes, including the correlation of expression profiles (in most cases) to explore hub genes (63). The edges of the gene co-expression network have been used as features to predict the metastasis of breast cancer (64). Through co-expression networks and clustering, the hypoxia-related dysregulation of mRNA and microRNAs were identified using the TCGA database to predict the breast cancer prognosis (65). Moreover, intra-chromosomal and inter-chromosomal interactions among genes were analyzed for each subtype to understand the subtype-specific imbalance of these relationships (66).

These networks are frequently categorized as scale-free networks that sometimes do not yield independent multiple clusters (67). Another drawback is that the constructed network is sensitive to the threshold pre-defined by analysts, and more robust methods needed to be developed. To solve these problems, the weighted gene correlation network analysis (WGCNA) that implements the soft-threshold has been developed, and this method yielded independent clusters and robust analytical results (68,69). The WGCNA was then explored to find hub genes that possess different biological functions to predict the breast cancer prognosis (70), and it was also used to find independent hub genes to predict breast cancer of the triple-negative subtype (71). Although these results should be validated using various datasets during the comprehensive molecule profiling, these analytical methods do contribute to the determination of independent features to characterize breast cancers.

Another approach for microarray data analysis is dimensional reduction, involved in techniques such as principal component analysis (PCA). This method does not identify a single molecule but instead enables the extraction of gene patterns that contribute to discriminating the defined phenotypes. For example, the combination of PCA and PAM50 was used for subtyping breast cancers, and consistent subtype-specific gene clusters were yielded in multiple datasets (72). The autoencoder (AE) (*Figure 3B*), a substream of an ANN, is another dimension reduction method part of which prominent lower-dimensional features are extracted from high-dimensional datasets (73). DNA methylation data were analyzed using the AE to identify the CpG site enabling the prognosis prediction for breast cancer (74). The AE was also used to extract gene features enabling the prognosis prediction of the luminal A subtype of breast cancer (75). Denoising the AE, which results in more robustness against data noise than AE, was developed and used to extract gene patterns to predict the breast cancer prognosis (76). These unsupervised methods have displayed the potential to extract multiple features showing independent biological functions and prediction of the breast cancer prognosis.

More recently, different molecular data were simultaneously analyzed to further explore the understanding of breast cancer biology. For example, gene expression and copy number alternations were analyzed via a modified DNN algorithm, and six subgroups in HER2-positive groups were found (77). Another customized DNN was used to analyze gene expression data, and the identified biomarkers for each subtype were visualized using PCA to access their heterogeneity (78). Additionally, a hybrid method involving both clustering and DNN was used to analyze RNA-seq data (79). As conveyed in this section, these high-dimensional data were analyzed using AI-based computational approaches, which enable the extraction of biological knowledge.

Other quantitative omics data

Through proteomics, hundreds of proteins can be profiled; therefore, this technique has become popular upon the advent of high-throughput mass spectrometry. Protein-based biomarkers have been identified to detect breast cancers at early stages (80). A supervised ML method was used to integrate quantified levels of 16 proteins to distinguish breast cancer tissues with and without metastasis (81).

Metabolomics is the newest approach in omics science for the identification and quantification of hundreds of metabolites in biological samples. The concentration pattern of the observed metabolites is named the metabolomic profile and it has been analyzed via multivariate analysis methods, such as PCA and partial least squares-discriminant analysis (PLS-DA), and pathway analysis (82). The metabolomic profiles of breast cancer tissues with ER+ and ER− revealed significant differences in terms of the beta-alanine and glutamine pathways (83). A DNN was used for the analysis of this dataset and new findings were obtained, such as the difference in adenosine triphosphate (ATP)-binding cassette transporter pathways, based on the estrogen receptors (84). Biofluid samples have been frequently analyzed, and classification methods have been developed to utilize their molecular patterns. We previously quantified salivary metabolites and developed an ADTree-based model to distinguish patients with breast cancer from healthy controls (85). The dimension of the observed data produced via a single measurement instrument is less than that of gene expression data; however, the combination of these data and AI showed the potential to enable the combination of the prediction ability using multiple markers and also for extraction of biological information in various pathways.

Notably, most ML methods require feature selection, wherein only the observed variables relevant to the outcome are selected upon elimination of the lower informative variables. Feature ranking (86), Markov blanket filtering (87), and the variance inflation factor (VIF) are commonly used to assess the dependence of features to reduce multicollinearity among the observed features (88). We used them in combination to select the minimum feature sets to predict the breast cancer prognosis (89). This combination of feature selection and prediction models significantly improves the accuracy, generalization ability, and robustness against noise frequently observed in quantitative omics data.

Recently, the concept of sparseness is introduced to identify only a few features showing predictivities among a large number of observed features. Gene expression data were analyzed using sparse k-means to subgroups analyses of breast cancer data (90). The least absolute shrinkage selector operator (LASSO) and Ridge regression techniques were used to analyze gene expression data of breast cancer (91). The improvement in omics technologies enables the observation of a larger number of features, whereas these analytical technologies would contribute to reducing the problem of multicollinearity.

Discussion

This paper reviewed the recent application of ML-based technologies in clinicopathological prediction and the analyses of quantitative omics data. ML provides an evident advantage for the development of classification models: classification and regression via statistical methods are limited on several counts. Colinearity limits the number of variables that can be used in multiple logistic regression and multiple linear regressions. Moreover, these methods linearly integrate the predictable features. ML methods, however, have the potential to capture the non-linear structure of the observed data. Several classification methods, such as ANNs, also suffer from collinearity; therefore, appropriate feature selection before the use of these classification methods is necessary.

Notably, several technical issues have been raised during rigorous evaluations of the training of supervised methods. Although various performance comparisons have been reviewed, the performance ultimately depends on the parameters considered for each method, such as the number of layers and nodes in an ANN model. To optimize these parameters, the enhancement of the generalization ability should be a primary goal. Combinations of the best parameters must also be searched in large parameter space with global parameter optimization, such as via a genetic algorithm (92). However, the performance of such algorithms also depends on their parameters, and empirical optimizations are currently practical, despite the induction of subjective training risk in the model. Therefore, more rigorous validation of the developed model should be employed as compared to that in the case of conventional statistical analysis methods. In conclusion, the improvement of automatic training and validation algorithm of each ML method is still necessary but enough validated ML models have shown the potential to contribute to the diagnosis and decision of treatment of breast cancers.

This review focuses on the application of ML to clinical-pathological features and quantitative molecular profiles. However, wider applications of AI, such as natural language processing, have been developed to help the decision-making of oncologists for diagnosis and treatment. A recent review in this field pointed out the need for standardization of benchmark tests and a variety of processed reporting before the use of developed methods in the clinical setting (15). Another review to introduce the application of ML on chronic diseases also requires the standardization of the evaluation metrics to determine ML performance and also data governance which realizes unbiased and objective data storage from the patients (13). A review of DL-based image-processing also claimed the need for reporting standardization even though the accuracy performance is equivalent to that of healthcare professionals (14). These problems are common for the topics dealt with in this review; accuracy and validation of ML algorithms, data management, objective evaluation criteria, and informative reporting, are necessary to assist the clinicians.

As a limitation, we did not cover all AI topics for breast cancer diagnosis and prognosis prediction in this review. The topics were limited to the application of ML on clinical-pathological features and quantitative omics. Especially, multi-gene assays have been already evaluated in the prospective clinical trials. Oncotype DX was validated in TAILORx, RxPONDER, and MINDACT clinical trials and analyzed to identify the beneficial patients to adjuvant chemotherapy (58,93-95). For example, the integrated use of clinical features based on tumor size and histologic grade to Oncotype DX enabled more precise identification of beneficial patients to adjuvant chemotherapy (96). Accumulation of these analytical results enables these tools to influence clinical decision-making in breast cancer treatment (97). However, the recent AI-based electronic health records found the need for standardizing the eligibility criteria on cancer trial population and there are still difficulties in obtaining robust analytical results for non-cancer diseases (98). Currently, real-world populations are unable to participate in clinical trials because of stringent exclusion criteria, but many patients receive treatment (98). Thus, the development of robust and versatile ML-based analytical tools is still necessary to fill in the gap so the wider patient base would receive the benefit of diagnosis and treatment therapy.

Acknowledgments

Funding: This work was supported by a Grant-in-Aid for Scientific Research (Grant No. 20B205).

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Annals of Breast Surgery* for the series "Cutting-edge Surgical Research". The article has undergone external peer review.

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://abs.amegroups.com/article/view/10.21037/abs-21-63/rc

Peer Review File: Available at https://abs.amegroups.com/article/view/10.21037/abs-21-63/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://abs.amegroups.com/article/view/10.21037/abs-21-63/coif). The series "Cutting-edge Surgical Research" was commissioned by the editorial office without any funding or sponsorship. MS reports that this work was supported by a Grant-in-Aid for Scientific Research (Grant No. 20B205). MT served as the unpaid Guest Editor of the series and serves as an unpaid editorial board member of *Annals of Breast Surgery* from July 2018 to June 2024. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
2. Siegel RL, Miller KD, Fuchs HE, et al. Cancer Statistics, 2021. CA Cancer J Clin 2021;71:7-33.
3. Goh T, Dao K, Rives AF, et al. Systemic diseases affecting the breast: Imaging, diagnosis, and management. Clin Imaging 2021;77:76-85.
4. Kalita-de Croft P, Al-Ejeh F, McCart Reed AE, et al. 'Omics Approaches in Breast Cancer Research and Clinical Practice. Adv Anat Pathol 2016;23:356-67.
5. Salod Z, Singh Y. A five-year (2015 to 2019) analysis of studies focused on breast cancer prediction using machine learning: A systematic review and bibliometric analysis. J Public Health Res 2020;9:1792.
6. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci U S A 1990;87:9193-6.
7. Patrício M, Pereira J, Crisóstomo J, et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 2018;18:29.
8. Zerouaoui H, Idri A. Reviewing Machine Learning and Image Processing Based Decision-Making Systems for Breast Cancer Imaging. J Med Syst 2021;45:8.
9. Salod Z, Singh Y. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. J Public Health Res 2019;8:1677.
10. Thomas T, Pradhan N, Dhaka VS, editors. Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey. 2020 International Conference on Inventive Computation Technologies (ICICT); 2020: IEEE.
11. Malherbe K. Tumor microenvironment and the role of artificial intelligence in breast cancer detection and prognosis. Am J Pathol 2021;191:1364-73.
12. Yassin NIR, Omran S, El Houby EMF, et al. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Comput Methods Programs Biomed 2018;156:25-45.
13. Choudhury A, Renjilian E, Asan O. Use of machine learning in geriatric clinical care for chronic diseases: a systematic literature review. JAMIA Open 2020;3:459-71.
14. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1:e271-97.
15. Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform 2020;8:e18599.
16. Choudhury A, Perumalla S. Detecting breast cancer using artificial intelligence: Convolutional neural network.

Technol Health Care 2021;29:33-43.
17. Yan R, Ren F, Wang Z, et al. Breast cancer histopathological image classification using a hybrid deep neural network. Methods 2020;173:52-60.
18. Xie J, Liu R, Luttrell Jt, et al. Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Front Genet 2019;10:80.
19. Allman R, Spaeth E, Lai J, et al. A streamlined model for use in clinical breast cancer risk assessment maintains predictive power and is further improved with inclusion of a polygenic risk score. PLoS One 2021;16:e0245375.
20. Jurj MA, Buse M, Zimta AA, et al. Critical Analysis of Genome-Wide Association Studies: Triple Negative Breast Cancer Quae Exempli Causa. Int J Mol Sci 2020;21:5835.
21. Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. Br J Cancer 1982;45:361-6.
22. Galea MH, Blamey RW, Elston CE, et al. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat 1992;22:207-19.
23. Balslev I, Axelsson CK, Zedeler K, et al. The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). Breast Cancer Res Treat 1994;32:281-90.
24. Heimann R, Ferguson D, Powers C, et al. Angiogenesis as a predictor of long-term survival for patients with node-negative breast cancer. J Natl Cancer Inst 1996;88:1764-9.
25. Coradini D, Boracchi P, Daidone MG, et al. Contribution of vascular endothelial growth factor to the Nottingham prognostic index in node-negative breast cancer. Br J Cancer 2001;85:795-7.
26. Rakha EA, Martin S, Lee AH, et al. The prognostic significance of lymphovascular invasion in invasive breast carcinoma. Cancer 2012;118:3670-80.
27. Prat A, Cheang MCU, Martín M, et al. Prognostic significance of progesterone receptor–positive tumor cells within immunohistochemically defined luminal A breast cancer. J Clin Oncol 2013;31:203.
28. Rakha EA, Soria D, Green AR, et al. Nottingham Prognostic Index Plus (NPI+): a modern clinical decision making tool in breast cancer. Br J Cancer 2014;110:1688-97.
29. Ravdin PM, Siminoff LA, Davis GJ, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001;19:980-91.
30. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res 2010;12:R1.
31. Rouzier R, Pusztai L, Garbay JR, et al. Development and validation of nomograms for predicting residual tumor size and the probability of successful conservative surgery with neoadjuvant chemotherapy for breast cancer. Cancer 2006;107:1459-66.
32. Xie N, Xu Y, Zhong Y, et al. Clinicopathological Characteristics and Treatment Strategies of Triple-Negative Breast Cancer Patients With a Survival Longer than 5 Years. Front Oncol 2021;10:617593.
33. Cui X, Song D, Li X. Construction and Validation of Nomograms Predicting Survival in Triple-Negative Breast Cancer Patients of Childbearing Age. Front Oncol 2021;10:636549.
34. Yu J, Wu J, Huang O, et al. A nomogram to predict the high-risk RS in HR+/HER2-breast cancer patients older than 50 years of age. J Transl Med 2021;19:75.
35. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. BMC Cancer 2019;19:230.
36. Weisberg S. Applied Linear Regression. Wiley, New York 1980.
37. Meti N, Saednia K, Lagree A, et al. Machine Learning Frameworks to Predict Neoadjuvant Chemotherapy Response in Breast Cancer Using Clinical and Pathological Features. JCO Clin Cancer Inform 2021;5:66-80.
38. Takada M, Sugimoto M, Ohno S, et al. Predictions of the pathological response to neoadjuvant chemotherapy in patients with primary breast cancer using a data mining technique. Breast Cancer Res Treat 2012;134:661-70.
39. Takada M, Sugimoto M, Naito Y, et al. Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model. BMC Med Inform Decis Mak 2012;12:54.
40. Takada M, Sugimoto M, Masuda N, et al. Prediction of postoperative disease-free survival and brain metastasis for HER2-positive breast cancer patients treated with neoadjuvant chemotherapy plus trastuzumab using a machine learning algorithm. Breast Cancer Res Treat 2018;172:611-8.
41. Takada M, Ishiguro H, Nagai S, et al. Survival of HER2-positive primary breast cancer patients treated by neoadjuvant chemotherapy plus trastuzumab: a multicenter retrospective observational study (JBCRG-C03 study). Breast Cancer Res Treat 2014;145:143-53.
42. Sugimoto M, Takada M, Toi M. Development of Web tools to predict axillary lymph node metastasis and pathological response to neoadjuvant chemotherapy in breast cancer patients. Int J Biol Markers 2014;29:e372-9.

43. Sugimoto M, Takada M, Toi M. Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:3054-7.
44. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46:10546-62.
45. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160-7.
46. Varga Z, Lebeau A, Bu H, et al. An international reproducibility study validating quantitative determination of ERBB2, ESR1, PGR, and MKI67 mRNA in breast cancer using MammaTyper®. Breast Cancer Res 2017;19:55.
47. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999-2009.
48. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817-26.
49. Filipits M, Rudas M, Jakesz R, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res 2011;17:6012-20.
50. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006;98:262-72.
51. Sinn P, Aulmann S, Wirtz R, et al. Multigene Assays for Classification, Prognosis, and Prediction in Breast Cancer: a Critical Review on the Background and Clinical Utility. Geburtshilfe Frauenheilkd 2013;73:932-40.
52. Brandão M, Pondé N, Piccart-Gebhart M. Mammaprint™: a comprehensive review. Future Oncol 2019;15:207-24.
53. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183-92.
54. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med 2016;375:717-29.
55. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol 2008;26:721-8.
56. Albain KS, Barlow WE, Shak S, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. Lancet Oncol 2010;11:55-65.
57. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 2006;24:3726-34.
58. Sparano JA, Gray RJ, Makower DF, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N Engl J Med 2018;379:111-21.
59. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.
60. Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567-72.
61. Mostavi M, Chiu YC, Huang Y, et al. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 2020;13:44.
62. Liu J, Su R, Zhang J, et al. Classification and gene selection of triple-negative breast cancer subtype embedding gene connectivity matrix in deep neural network. Brief Bioinform 2021. [Epub ahead of rint]. doi: 10.1093/bib/bbaa395.
63. Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci U S A 2002;99:12783-8.
64. Adnan N, Lei C, Ruan J. Robust edge-based biomarker discovery improves prediction of breast cancer metastasis. BMC Bioinformatics 2020;21:359.
65. Gong PJ, Shao YC, Huang SR, et al. Hypoxia-Associated Prognostic Markers and Competing Endogenous RNA Co-Expression Networks in Breast Cancer. Front Oncol 2020;10:579868.
66. García-Cortés D, de Anda-Jáuregui G, Fresno C, et al. Gene Co-expression Is Distance-Dependent in Breast Cancer. Front Oncol 2020;10:1232.
67. Albert R. Scale-free networks in cell biology. J Cell Sci 2005;118:4947-57.
68. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4:Article17.
69. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
70. Tian Z, He W, Tang J, et al. Identification of Important

Modules and Biomarkers in Breast Cancer Based on WGCNA. Onco Targets Ther 2020;13:6805-17.

71. Wang Y, Li H, Ma J, et al. Integrated Bioinformatics Data Analysis Reveals Prognostic Significance Of SIDT1 In Triple-Negative Breast Cancer. Onco Targets Ther 2019;12:8401-10.
72. Raj-Kumar PK, Liu J, Hooke JA, et al. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. Sci Rep 2019;9:7956.
73. Tan J, Hammond JH, Hogan DA, et al. Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. MSystems 2016;1:e00025-15.
74. Macías-García L, Martínez-Ballesteros M, Luna-Romera JM, et al. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. Artificial Intelligence in Medicine 2020;110:101976.
75. Wang S, Lee D, editors. Identifying prognostic subgroups of luminal-A breast cancer using a deep autoencoder. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2020: IEEE.
76. Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 2010;11:3371-408.
77. Mohaiminul Islam M, Huang S, Ajwad R, et al. An integrative deep learning framework for classifying molecular subtypes of breast cancer. Comput Struct Biotechnol J 2020;18:2185-99.
78. Beykikhoshk A, Quinn TP, Lee SC, et al. DeepTRIAGE: interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types. BMC Med Genomics 2020;13:20.
79. Srinivasan S, Leshchyk A, Johnson NT, et al. A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. RNA 2020;26:1303-19.
80. Yao F, Yan C, Zhang Y, et al. Identification of blood protein biomarkers for breast cancer staging by integrative transcriptome and proteome analyses. J Proteomics 2021;230:103991.
81. Shapanis A, Lai C, Sommerlad M, et al. Proteomic Profiling of Archived Tissue of Primary Melanoma Identifies Proteins Associated with Metastasis. Int J Mol Sci 2020;21:8160.
82. Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. Curr Protoc Bioinformatics 2019;68:e86.
83. Budczies J, Brockmöller SF, Müller BM, et al. Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism. J Proteomics 2013;94:279-88.
84. Alakwaa FM, Chaudhary K, Garmire LX. Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. J Proteome Res 2018;17:337-47.
85. Murata T, Yanagisawa T, Kurihara T, et al. Salivary metabolomics with alternative decision tree-based machine learning methods for breast cancer discrimination. Breast Cancer Res Treat 2019;177:591-601.
86. Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinformatics 2015;2015:198363.
87. Xing EP, Jordan MI, Karp RM, editors. Feature selection for high-dimensional genomic microarray data. Icml; 2001: Citeseer.
88. Lin D, Foster DP, Ungar LH. VIF regression: a fast regression algorithm for large data. Journal of the American Statistical Association 2011;106:232-47.
89. Hikichi S, Sugimoto M, Tomita M. Correlation-centred variable selection of a gene expression signature to predict breast cancer metastasis. Sci Rep 2020;10:7923.
90. Prado-Vázquez G, Gámez-Pozo A, Trilla-Fuertes L, et al. A novel approach to triple-negative breast cancer molecular classification reveals a luminal immune-positive subgroup with good prognoses. Sci Rep 2019;9:1538.
91. Gupta M, Gupta B. A novel gene expression test method of minimizing breast cancer risk in reduced cost and time by improving SVM-RFE gene selection method combined with LASSO. J Integr Bioinform 2020;18:139-53.
92. Mishra AK, Roy P, Bandyopadhyay S, editors. Genetic Algorithm Based Selection of Appropriate Biomarkers for Improved Breast Cancer Prediction. Proceedings of SAI Intelligent Systems Conference; 2019: Springer.
93. Sparano JA, Gray RJ, Makower DF, et al. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N Engl J Med 2015;373:2005-14.
94. Ramsey SD, Barlow WE, Gonzalez-Angulo AM, et al. Integrating comparative effectiveness design elements and endpoints into a phase III, randomized clinical trial (SWOG S1007) evaluating oncotypeDX-guided management for women with breast cancer involving lymph nodes. Contemp Clin Trials 2013;34:1-9.
95. Piccart M, van 't Veer LJ, Poncet C, et al. 70-gene

signature as an aid for treatment decisions in early breast cancer: updated results of the phase 3 randomised MINDACT trial with an exploratory analysis by age. Lancet Oncol 2021;22:476-88.

96. Sparano JA, Gray RJ, Ravdin PM, et al. Clinical and Genomic Risk to Guide the Use of Adjuvant Therapy for Breast Cancer. N Engl J Med 2019;380:2395-405.

97. Jayasekera J, Sparano JA, O'Neill S, et al. Development and Validation of a Simulation Model-Based Clinical Decision Tool: Identifying Patients Where 21-Gene Recurrence Score Testing May Change Decisions. J Clin Oncol 2021:Jco2100651. doi: 10.1200/JCO.21.00651.

98. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 2021;592:629-33.

doi: 10.21037/abs-21-63

Cite this article as: Sugimoto M, Hikichi S, Takada M, Toi M. Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. Ann Breast Surg 2023;7:7.