frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/fonc.2022.943874

OPEN ACCESS

EDITED BY
Francesco Schettini,
Institut de Recerca Biomèdica August Pi i Sunyer (IDIBAPS), Spain

REVIEWED BY
Duraimurugan Samiayya,
St. Joseph's College of Engineering, India
Yanshuo Chu,
University of Texas MD Anderson Cancer Center, United States

*CORRESPONDENCE
Yuanjie Zheng
yjzheng@sdnu.edu.cn
Huali Pan
110096@sdnu.edu.cn

SPECIALTY SECTION
This article was submitted to Breast Cancer, a section of the journal Frontiers in Oncology

RECEIVED 14 May 2022
ACCEPTED 17 November 2022
PUBLISHED 08 December 2022

CITATION
Yang X, Zheng Y, Xing X, Sui X, Jia W and Pan H (2022) Immune subtype identification and multi-layer perceptron classifier construction for breast cancer.
Front. Oncol. 12:943874.
doi: 10.3389/fonc.2022.943874

Immune subtype identification and multi-layer perceptron classifier construction for breast cancer

Xinbo Yang[1], Yuanjie Zheng[1]*, Xianrong Xing[2], Xiaodan Sui[1], Weikuan Jia[1] and Huali Pan[1,3]*

[1]School of Information Science and Engineering, Shandong Normal University, Jinan, China, [2]Department of Pharmacy, Shandong Medical College, Jinan, China, [3]Business School, Shandong Normal University, Jinan, China

Introduction: Breast cancer is a heterogeneous tumor. Tumor microenvironment (TME) has an important effect on the proliferation, metastasis, treatment, and prognosis of breast cancer.

Methods: In this study, we calculated the relative proportion of tumor infiltrating immune cells (TIICs) in the breast cancer TME, and used the consensus clustering algorithm to cluster the breast cancer subtypes. We also developed a multi-layer perceptron (MLP) classifier based on a deep learning framework to detect breast cancer subtypes, which 70% of the breast cancer research cohort was used for the model training and 30% for validation.

Results: By performing the K-means clustering algorithm, the research cohort was clustered into two subtypes. The Kaplan-Meier survival estimate analysis showed significant differences in the overall survival (OS) between the two identified subtypes. Estimating the difference in the relative proportion of TIICs showed that the two subtypes had significant differences in multiple immune cells, such as CD8, CD4, and regulatory T cells. Further, the expression level of immune checkpoint molecules (PDL1, CTLA4, LAG3, TIGIT, CD27, IDO1, ICOS) and tumor mutational burden (TMB) also showed significant differences between the two subtypes, indicating the clinical value of the two subtypes. Finally, we identified a 38-gene signature and developed a multilayer perceptron (MLP) classifier that combined multi-gene signature to identify breast cancer subtypes. The results showed that the classifier had an accuracy rate of 93.56% and can be robustly used for the breast cancer subtype diagnosis.

Conclusion: Identification of breast cancer subtypes based on the immune signature in the tumor microenvironment can assist clinicians to effectively and accurately assess the progression of breast cancer and formulate different treatment strategies for different subtypes.

KEYWORDS

breast cancer, immune infiltration, subtype identification, tumor mutational burden, multi-layer perceptron classifier

1 Introduction

Breast cancer is a disease with high morbidity and mortality, only lower than lung cancer in women (1, 2). According to a report by the American Cancer Society in 2019, in the last 5 years (2012-2016), the incidence of breast cancer has increased slightly at a rate of 0.3% per year (3). Breast cancer is a highly heterogeneous tumor (4); the tumor tissue not only includes tumor cells, but also normal epithelial, stromal, and immune cells that are associated with tumors. The tumor microenvironment (TME) that is composed of these cells has an important impact on the tumor proliferation, metastasis, treatment, and prognosis (5–7).

Immune cells are scattered in the tumor center and infiltration margin or adjacent tertiary lymphoid tissues, and can be roughly divided into immunosuppressive and immune effector cells (8, 9). The level of immune cells infiltration reflectsthe degree of tumor development, affecting cancer progression (10). The tumor immune microenvironment (TIME) is composed of various cells that can inhibit the tumor formation (11–13) and promote tumorigenesis (14, 15).

Quantification of the proportion of various cells in the TME is important to understand the occurrence and development of tumors. Yoshihara K et al. proposed a method (ESTIMATE) of using gene expression profiles to calculate the ratio of stromal to immunecells to reveal the tumor purity (16). Newman et al. utilized the gene expression data to estimate the abundance of immune cells in tumor samples, and developed the analysis tool CIBERSORT for estimating and verifying the proportion of 22 immune cells (17).

Breast cancer is very difficult to cure; however, early diagnosis and timely treatment can prolong the patients' survival. Immunotherapy is considered the most promising treatment for breast cancer currently and includes immune checkpoint blocking (ICB) therapy (18, 19), adoptive T cell immunotherapy (20), and tumor vaccine immunotherapy (21, 22). The US FDA has also approved few immunotherapy drugs mainly Keytruda (Pembrolizumab), Opdivo (Nivolumab), Tecentriq (Atezolizumab) among others.

Immunotherapy is not suitable for all breast cancer patients (23–26), and hence, it is important to accurately determine the cancer subtype in such patients so that appropriate treatment can be administered. Perou et al. distinguished the breast cancer subtypes based on the differences in mRNA expression patterns, and proposed, for the first time, that breast cancer can be divided into four subtypes (27). Subsequently, a 50-gene breast cancer classification model (PAM50) was developed based on the gene expression profile data (28), and is commonly employed in clinical practice. Further, based on the molecular subtype identification of triple-negative breast cancer, six (29), four (30), and three (31) subtypes have been proposed, while a model of six subtypes was also proposed based on the colon cancer classification method (32). Although these subtype classification methods elucidated the molecular markers, prognostic differences, and clinical significance of the subtypes; TME and the influence of TIME on the occurrence, development, and prognosis of tumors have not been evaluated. Additionally, the association of immune checkpoint molecules and breast cancer isnot comprehensively understood.

In this study, the ESTIMATE algorithm was used to determine the individual and combined scores of the immune and stromal cells of each sample in the breast cancer research cohort. Further, we used the CIBERSORT algorithm to estimate the scores of the 22 types of immune cells in the same research cohort. We propose a method to identify breast cancer subtypes by combining the estimated scores of the two immune infiltrations. Two breast cancer subtypes were identified using the consensus clustering algorithm, and the survival, immune cell differential, immune checkpoint molecules differential, tumor mutation burden correlation, differential gene enrichment, and drug sensitivity analyses were performed for these two subtypes. We showed that this classification into two subtypes has a potential for clinical application. We also developed a multi-layer perceptron (MLP) classifier based on a deep learning framework to detect two breast cancer subtypes. By using the training data to train the classifier model, the test results showed that the classifier can distinguish the two subtypes.

2 Materials and methods

2.1 Data search strategy and collection

The breast cancer data used in this study were obtained from the two public databases, The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) and Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/). The TCGA data included the transcriptome mRNA expression profile data of female patients (n=1208, 1096 cancer and 112 normal samples), clinical data (n=1085), and simple nucleotide variation (SNV) data. The research cohorts selected from the GEO database were GSE42568 (n=104) and GSE88770 (n=117), including the mRNA expression profile files of the patient cohort and probe files of the sequencing platform.

2.2 Data preprocessing

TCGA mRNA expression and clinical data were normalized through the following steps: (1) mapping of the mRNA expression data to the human genome annotation file, replacing Ensemble IDs with gene names, and deleting the genes lacking a corresponding mapping, (2) standardization of the mRNA expression data, (3) conversion of FPKM standardization data into TPM standardization; when the same sample was repeated, the average value of gene expression was used instead. Further, the normal samples were deleted, and (4) using perl language scripts to extract

the clinical information, including the sample id, overall survival (OS) in days, survival status, age, grade, and stage (T, M, and N staging). The breast cancer data of the GEO database were annotated with theprobe data of the sequencing platform GPL570. We extracted the gene expression data and clinical information separately. Finally, we consolidated and combined the TCGA and GEO expression data.

2.3 Estimate the proportion of tumor-infiltrating immune cells and tumor purity

The proportion of 22 TIICs types were estimated by using the CIBERSORT algorithm for each sample, and samples with a p-value of <0.05 were selected for the survival analysis. For tumor purity, we used the ESTIMATE algorithm for evaluation. Two non-tumor components (immune and stromal cells) was calculated by using the ESTIMATE algorithm and gene expression profiles, and obtained three tumor purity signatures (stromal, immune, and estimate scores).

2.4 Identification of breast cancer subtypes

The R language "ConsensusClusterPlus" package was used to perform the consistent clustering, and to separately save the graphs of the clustering results for each K value (Integer K, $2 \leq K \leq 9$). The parameters of the unsupervised clustering weredefined, including the clustering algorithm (clusterAlg="km"), maximum number of clusters (maxK=9), number of resampling (reps=50), sampling ratio (pItem=0.8), characteristics sampling ratio (pFeature=1), and clustering distance (distance="euclidean").

2.5 Statistical analysis

The statistical analysis was performed by Rstudio software, R version 4.1.2. For clinical data, the R packages ßurvival" and ßurvminer" were used for the survival analysis, and the Kaplan-Meier survival curve was drawn. Using the "limma" package, differentially expressed genes (DEGs) between subtypes, as well as the expression differences of immune checkpoint molecules, tumor mutational burden (TMB), and drug sensitivity were statistically analyzed. "ggplot2" was used to draw the graphics and figures.

2.6 Breast cancer subtype classifier based on the neural network

Deep learning algorithms are gradually being widely used in the field of biomedicine (33–35). We designed an MLP classifier to identify the breast cancer subtypes. This classifier included an input, hidden, and output layer. The input layer contains 38 nodes, which represent 38 DEGs. The activation function of the multilayer perceptron model uses the "sigmoid", and the mathematical formula is expressed as:

$$sigmoid = \frac{1}{1 + e^{-x}} \tag{1}$$

The loss function of the model used the cross entropy loss function, and the mathematical formula is expressed as:

$$Loss = -\frac{1}{n}\sum_{x}[ylna + (1-y)ln(1-a)] \tag{2}$$

where, x represents the sample, y represents the true label, a represents the predicted output value, and n represents the total number of samples.

The optimization of the model uses the RMSProp optimization algorithm and the mathematical formula is expressed as:

$$\begin{aligned} S_{dw} &= \beta S_{dw} + (1-\beta)dw^2 \\ w &= w - \alpha \frac{dw}{\sqrt{S_{dw}}} \end{aligned} \tag{3}$$

Where dw is the gradient, S_{dw} is a value container, which stores the result of the square weighted average of all the gradients, α is the learning rate (general value: 0.001), β decay factor (general value: 0.9). In order to obtain a classifier model with robust performance and high accuracy, we also verified the impact of the number of hidden layer nodes on the classification results, ranging from 2 to 38.

3 Results

3.1 Estimate of the proportion of TIICs in the breast cancer research cohort

The proportion of immune cells and tumor purity in the research cohort were quantified by using CIBERSORT and ESTIMATE algorithms (Table S1); based on the quantitative score, the general landscape of TIICs interaction in breast cancer TME was visualized by generating the correlation coefficient heat map (Figure 1). The correlation analysis of TIICs showed that the CD8 and memory-activated CD4 T cells and M0 macrophages had the strongest positive and negative correlations, respectively. In addition, M1 macrophages and CD8, memory-activated CD4, and follicular helper T Cells showed strong positive correlations.

3.2 Subtype clustering and differential analysis of immune cells

By performing the K-means clustering algorithm, 8 cluster maps were generated (Figures 2A–H). Figure 2A shows two subtypes with the best clustering results (Table S2). We define

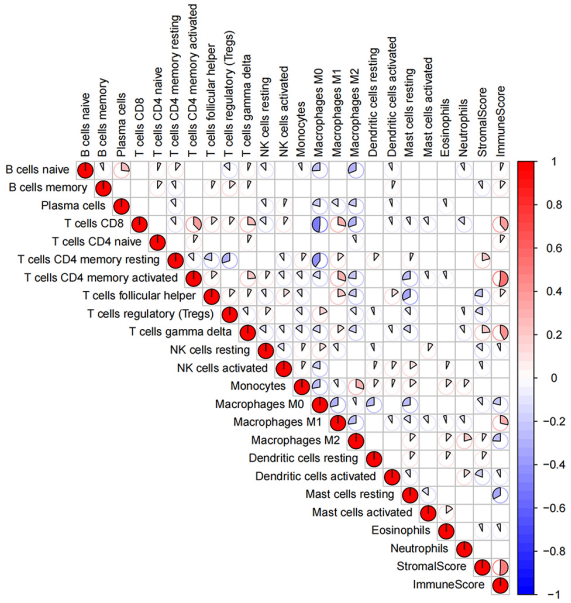

FIGURE 1
The general landscape of infiltrating immune cells interactions in breast cancer. The color value and shape size of the pie chart represent the correlation size between immune cells. The color bar shows the positive and negative values of the correlation.

these two independent subtypes as ICS-A and ICS-B. In subtype ICS-A, the scores of regulatory T cells and M0 and M2 macrophages were significantly higher than that of the subtype IC-B (Figure 3). Further, in subtype ICS-B, the proportion of B cells, CD8 T cells, memory activated CD4 T cells, memory resting CD4 T cells, NK cells, and M1 macrophages was significantly higher than that of the subtype ICS-A. On the other hand, no significant difference was observed in the proportion of native CD4 T cells, follicular helper T cells, eosinophils, and neutrophils between the two subtypes.

3.3 Kaplan–-Meier survival analysis

In order to investigate the clinical significance of the subtype identification, we performed the Kaplan-Meier survival analysis on the OS of the two subtypes. The survival curve showed that the two subtypes had a significant difference in the OS, and the median survival time of the subtype ICS-B was 8 years longer than that of the subtype ICS-A (Figure 4).

3.4 Differential expression and drug sensitivity analysis of immune checkpoint molecules in the breast cancer

ICB therapy is currently the most promising immunotherapy for the treatment of breast cancer. We revealed differences in the expression levels of several key immune-modulatory molecules, including the co-stimulatory (CD27, ICOS, CD28, CD80, CD86, CD40, and CD276) and co-suppressive molecules (PDL1, CTLA4, LAG3, TIGIT, and IDO1) in the two subtypes. The expression levels of immuno-modulators (PDL1, CTLA4, LAG3, CD27, ICOS, CD28, CD86, CD40, TIGIT, and IDO1) in subtype ICS-B were significantly higher, while the CD276 levels were significantly lower than that

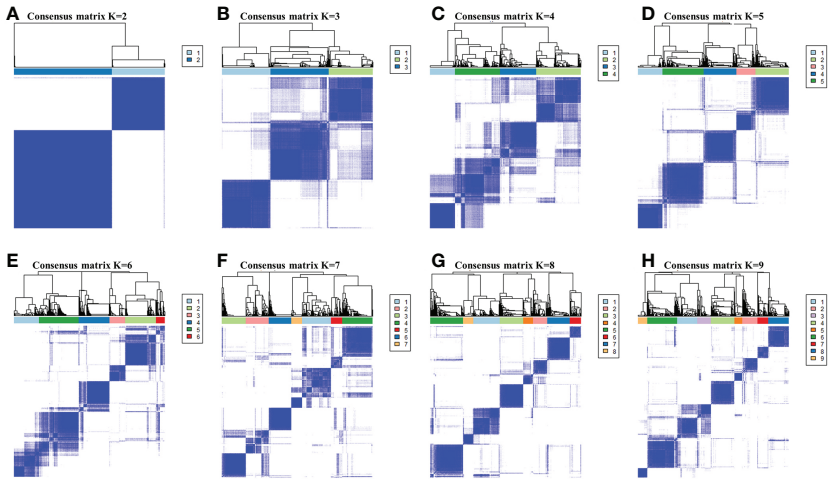

FIGURE 2
(A–H) respectively showed the consensus matrices of all breast cancer samples in the research cohort for each k ($2 \leq k \leq 9$), displaying the clustering stability by performing 100 iterations of hierarchical clustering. Perform subtype clustering using the K-means clustering algorithm, with k in the range of 2 to 9.

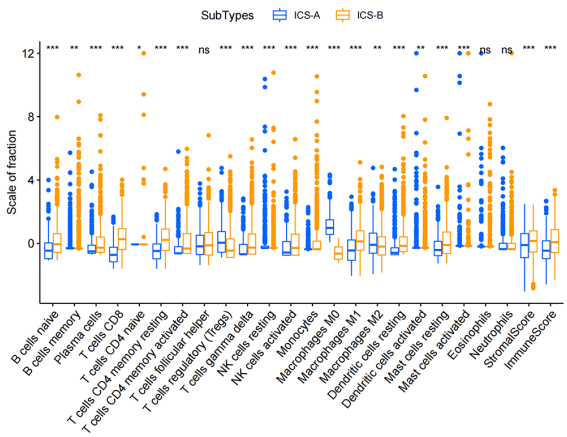

FIGURE 3
Boxplots showing statistical differences in immune cells in the two immune subtypes ICS-A and ICS-B. Comparison was performed using the Wilcoxon signed-rank test. *p < 0.05, **p < 0.01, ***p < 0.001 (p-values were adjusted using FDR correction).

of the subtype ICS-A (Figure 5). Furthermore, we investigated the expression of the immunomodulators in the 60 human cancer cell lines (NCI-60), and systematically tested the correlation between their expression levels in the NCI-60 cell lines with drug sensitivity of 218 FDA-approved chemotherapy drugs (Table S3). Figure 6 shows the association between expression of immunomodulatory molecules (PDL1 and CTLA4) and drug sensitivity. We noticed that increased PDL1 expression was associated with increased cellular resistance to chemotherapy drugs such as Tamoxifen and Nilotinib; we also observed inverse associations of multiple genes to these drugs, Furthermore, PDL1 was associated with increased sensitivity of cells to Dasatinib (treatment for mantle cell lymphoma and

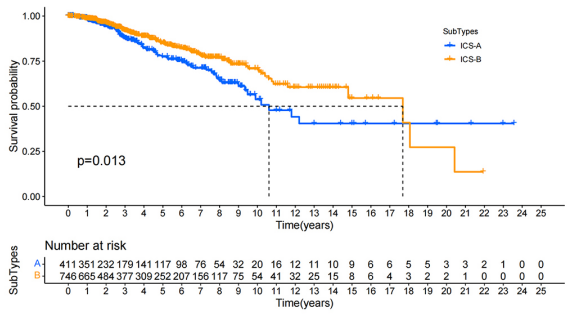

FIGURE 4
Kaplan-Meier survival curves for the overall survival of all breast cancer patients in the research cohort. Log rank test showed that the overall survival of ICS-A and ICS-B subtypes were significantly different (p-value = 0.013, p-values were adjusted using FDR correction).

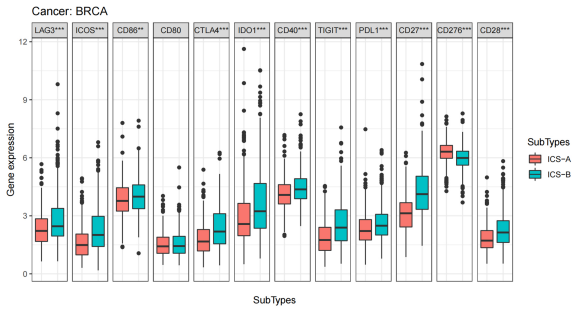

FIGURE 5
Boxplots of differential expression of immune checkpoint molecules. The analysis was performed using the Wilcoxon signed-rank test, *p < 0.05, **p < 0.01, ***p < 0.001 (p-values were adjusted using FDR correction).

chronic lymphocytic leukemia), while CTLA4 was associated with increased resistance of cells to Dasatinib.

3.5 Analysis of TMB in two subtypes

Considering the impact of TMB in tumor development, we further explored and revealed the correlation of TMB with OS and Age, respectively. We first counted the SNV of each sample in the TCGA cohort, and the frequency (number of samples) of the mutated genes in the research cohort (Table S4). The results showed that the number of samples with PIK3CA mutation was the largest, followed by TP53, TTN, CDH1, and GATA3. Further, the TMB subtype ICS-A was significantly higher than that of subtype ICS-B (Wilcoxon test p <0.001) (Figure 7). Furthermore, TMB showed significant negative and positive associations with the OS (Spearman coefficient: R = -0.12, p = 0.00043) and age (Spearman coefficient: R =0.14, p = 1.8e-05) (Figures 8, 9).

3.6 Differentially expressed genes in the two subtypes

By using the Bayesian estimation test, more than 5000 DEGs were found between the subtypes ICS-A and ICS-B (Table S5). Further, under the conditions of p value <0.05,| *logFC*| >1, and 95% confidence interval (CI), a signature of 38 DEGs was used to identify the two subtypes. Table 1 shows the list of genes identified.

3.7 Gene ontology and KEGG pathway enrichment analysis

GO enrichment analysis covers three domains: cellular component (CC), molecular function (MF), and biological process (BP). Figures 10, 11 show the results of GO terms

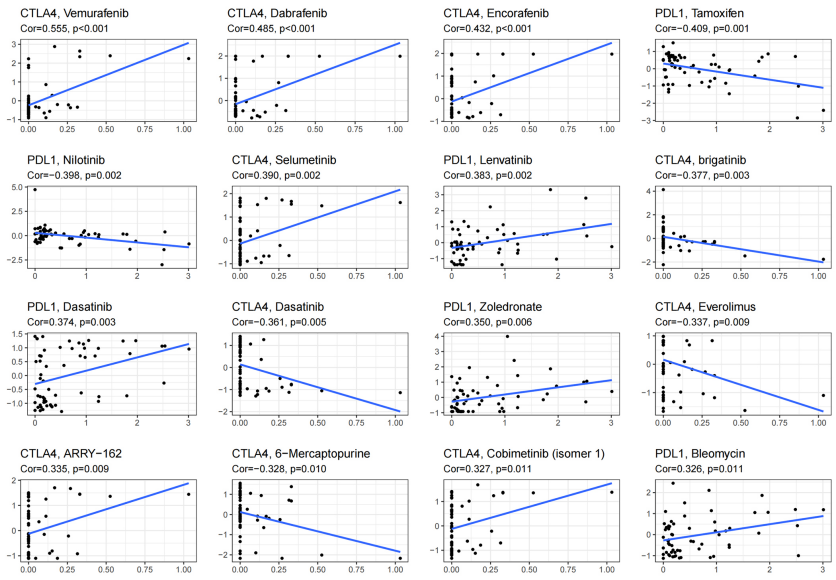

FIGURE 6
Scatter plot showing the association of expression of immune checkpoint molecules with drug sensitivity of NCI-60 cell line. Top 16 molecules are listed; PDL1 and CTLA4 were significantly associated with the drug sensitivity.

and KEGG pathways enrichment analysis, respectively. The top 5 BPs were significantly enriched in the T cell activation, leukocyte mediated immunity, positive regulation of cell activation, mononuclear cell differentiation, and positive regulation of leukocyte activation. The CC analysis revealed that DEGs were mainly enriched in the external side of the plasma membrane, membrane raft, and membrane micro domain, while the MF significantly enriched in immune receptor activity, cytokine receptor binding, cytokine activity, and carbohydrate binding. KEGG pathway analysis showed that cytokine-cytokine receptor interaction was the most significant pathway for the DEGs enrichment, followed by cell adhesion molecules, chemokine signaling pathway, and hematopoietic cell lineage.

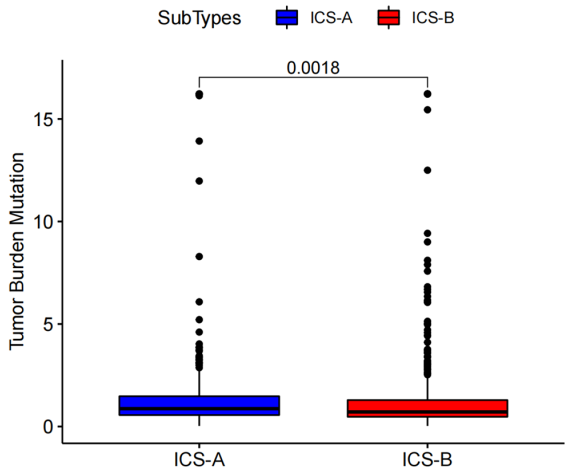

FIGURE 7
Tumor mutational burden (TMB) differences between the ICS-A and ICS-B subtypes. p-values were adjusted using FDR correction.

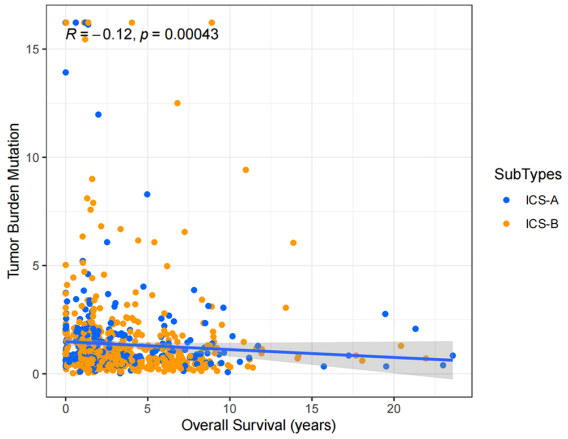

FIGURE 8
Tumor mutational burden (TMB) correlation analysis with the overall survival.

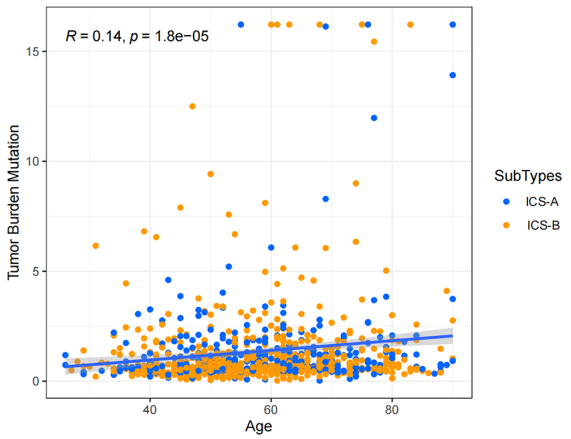

FIGURE 9
Tumor mutational burden (TMB) correlation analysis with age.

3.8 Breast cancer subtype classifier based on the neural network

The MLP classifier for identifying subtypes was defined as three layers, including the input, hidden, and output layers, while the number of nodes in each layer was respectively defined as 38, 5, and 2, according to the training and testing results of the program. Figure 12A shows the accuracy of the classifier model with different numbers of nodes in the hidden layer.

4 Discussion

We performed a detailed and comprehensive assessment of the TIICs in breast cancer using the research cohort from TCGA and GEO databases (Figure 13). Compared with previous studies (27, 29–32), this study determined the composition of TIICs in breast cancer, and the research cohort was divided into two subtypes, ICS-A and ICS-B, according to the composition of TIICs. A 38-ene signature tumor marker was identified and a classifier for subtype identification was developed using a deep learning framework. Our study confirmed that the proportion of immune cells and the expression level of immune checkpoint molecules in subtype ICS- were significantly higher than those in subtype ICS-A. Further, subtype ICS-B had better OS, suggesting that it is more suitable for immune checkpoint blockade therapy than subtype ICS-A. At the same time, we also conducted the drug sensitivity (FDA-approved chemotherapy drugs) analysis of theimmune checkpoint molecules that provided a reference for the selection of these drugs for breast cancer patients.

The TIME has an important impact on tumor diagnosis, treatment, and prognosis. The immune score has been used in renal and lung cancer studies in terms of estimating the relative proportion fraction of TIICs in tumors, and has shown its prognostic value (36–38). In this study, subtype ICS-B (with higher levels of CD8 and memory resting CD4 T cells) was found to be associated with better OS. CD8 T cells are key anti-tumor effector T cells, and CD4 T cells can be further differentiated to perform various functions (for instance, to differentiate into CD8 memory T cells to suppress tumor growth) (36, 39, 40); however the M2 macrophages can also suppress anti-tumor immune responses by secreting multiple mediators such as the inhibitory cytokines IL-10 or TGF-B, down-regulating antitumor immune response, promotion of angiogenesis, enhancement of cancer cell proliferation, invasion, intravascular penetration, and spread have been metastasized (41–44). This suggests that the immune subtype ICS-A with a lower proportion of CD8 T cells but a higher proportion of M2 macrophages may have an immunosuppressive (immune rejection) phenotype, and M2 macrophages or CD8 T cells may provide a therapeutic target for future breast cancer immunotherapy.

TMB has been recognized as a predictive marker of immunotherapy response and a prognostic marker in various tumor types (45–48). In this study, the TMB level of immune subtype ICS-A was significantly higher than that of the subtype ICS-B, indicating that patients with subtype ICS-A may produce more neo-antigens and will adversely affect the patient survival. Therefore, subtype ICS-B is indicative of better OS. This idea is supported by the correlation analysis between the TMB and OS. The level of TMB was significantly and negatively correlated with the OS in breast cancer patients. Further, since the TMB and age showed positive interaction, older patients had relatively higher TMB. This is consistent with a recent study showing that TMB increases with age, while the T cell receptor decreases (49). This provides unique insights into clinical prognostic diagnosis.

Statistical analysis of SNV in the research cohort showed that PIK3CA had the highest mutation frequency. PIK3CA is a catalytic subunit of the key proto-oncogene PI3K in the PI3K-Akt signaling pathway. Mutation of PIK3CA can lead to enhanced kinase activity, which in turn continuously stimulates downstream AKT (50), increases cell invasion and metastasis, and promotes tumor development. PIK3CA is located on chromosome 3, with a total of 20 exons, and 80% of PIK3CA mutations occur in the two hotspot regions of the helical region and the kinase region. The three most common mutations are H1047R on exon 20, and E542k and E545K on exon 9. Many studies have confirmed the existence of PIK3CA mutations in various human solid tumors, and its positive rate in breast cancer can reach 30-40% (51–53). This result has also been confirmed in our study. There were a total of 980 samples in our research cohort, of which 322 samples had PIK3CA mutation, with a positivity rate of 32.86%. This indicates that

TABLE 1 Comparison of different obfuscations in terms of their transformation capabilities.

Gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
MMP9	-2.16300	6.53699	-22.74219	2.30436E-95	4.0453E-91	206.0615
SPP1	-1.57532	7.40448	-16.51827	2.41593E-55	2.1206E-51	114.9107
GZMK	1.37200	3.62741	16.30778	4.08304E-54	1.7919E-50	112.115
CD8A	1.13855	3.83799	15.98659	2.9213E-52	8.7254E-49	107.8928
ITM2A	1.16847	5.33790	15.98503	2.9822E-52	8.7254E-49	107.8724
CD27	1.11183	3.84288	15.47268	2.41841E-49	6.065E-46	101.2505
GZMA	1.23115	4.50821	15.34996	1.17782E-48	2.5846E-45	99.68555
CD3D	1.20354	4.59038	14.98294	1.27527E-46	2.4875E-43	95.05518
CD2	1.18523	4.97137	14.40682	1.70525E-43	2.1383E-40	87.94169
CD3E	1.04785	3.94658	14.37973	2.38055E-43	2.786E-40	87.61205
CCL19	1.79801	5.54616	13.89503	8.65444E-41	7.2347E-38	81.78719
SELP	1.03400	3.15338	13.45229	1.66462E-38	8.3492E-36	76.59265
ACKR1	1.54096	4.60494	13.44081	1.90484E-38	9.2887E-36	76.45952
SELL	1.09100	4.24668	13.42028	2.42319E-38	1.1497E-35	76.22184
CD79A	1.30591	3.78360	12.93794	6.41371E-36	2.3457E-33	70.71375
TNFRSF17	1.04404	2.42067	12.93594	6.56205E-36	2.351E-33	70.69118
NKG7	1.04241	4.27307	12.85253	1.69539E-35	5.9525E-33	69.75416
IL33	1.12487	3.51383	12.67761	1.22287E-34	3.4084E-32	67.80383
IGHM	1.72575	6.90470	12.48471	1.0554E-33	2.6852E-31	65.67671
C7	1.29024	3.28276	12.26408	1.20419E-32	2.7454E-30	63.27449
CCL5	1.00413	6.27016	12.24827	1.43191E-32	3.2227E-30	63.1036
IGKC	1.59129	8.71681	12.10846	6.57321E-32	1.3418E-29	61.60006
MS4A1	1.08781	2.23466	11.89404	6.62999E-31	1.1999E-28	59.32029
IGLL5	1.39436	4.96581	11.54786	2.58583E-29	3.9473E-27	55.7075
IGLV1-44	1.46962	5.91042	10.94526	1.24059E-26	1.5446E-24	49.62315
MFAP4	1.03166	5.87060	10.67499	1.81439E-25	1.9784E-23	46.98055
CXCL9	1.28566	6.23401	10.33092	5.10202E-24	4.9484E-22	43.69549
CHRDL1	1.11266	3.84926	10.09488	4.77912E-23	4.216E-21	41.49378
CCL21	1.21797	4.35651	9.90108	2.90555E-22	2.4761E-20	39.71812
IGHD	1.06524	2.92884	9.80581	6.98186E-22	5.6744E-20	38.85589
MMP13	-1.08928	4.17293	-9.37678	3.31451E-20	2.3557E-18	35.06133
MMP12	-1.03738	2.61418	-9.27906	7.82465E-20	5.2831E-18	34.21744
IGLV6-57	1.17385	4.73478	9.11611	3.22241E-19	2.0203E-17	32.82725
COL11A1	-1.15667	5.69851	-9.08603	4.17467E-19	2.5625E-17	32.57302
ADH1B	1.16813	3.34122	9.08188	4.32622E-19	2.6462E-17	32.53801
MMP1	-1.25655	4.14037	-8.85569	2.96053E-18	1.6292E-16	30.6501
IGHG1	1.06179	7.57753	7.55540	8.29792E-14	2.9428E-12	20.61969
CXCL13	1.04036	4.59093	7.26950	6.51845E-13	2.0958E-11	18.60688

Condition employed: pvalue <0.05, | logFC | >1, and 95% confidence interval.
The results were obtained by using the R "limma" package Bayesian test.

PIK3CA can be used as a prognostic molecular biomarker and therapeutic target for breast cancer. The development and use of drugs targeting PIK3CA to block the PI3Ks pathway will play an effective role in the treatment of breast cancer. In recent years, the deep learning framework in the field of artificial intelligence is gradually being applied in various disciplines and industries. Although the interpretability of deep learning frameworks is still debated, their ability to solve bioinformatics problems requires further investigation. It has shown powerful functions in bioinformatics such as protein structure prediction (54, 55), protein–protein interaction prediction (56, 57), RNA structure prediction (58, 59), drug small molecule interaction prediction (60, 61), and drug design (62–64). Based on the identification of breast cancer immune subtypes, we designed and developed a subtype ICS–A classifier based on a deep learning framework.

In order to improve the prediction accuracy of the model, we used 70% of the data for model training and 30% of the data for model testing. The number of iterations epoch is 1000

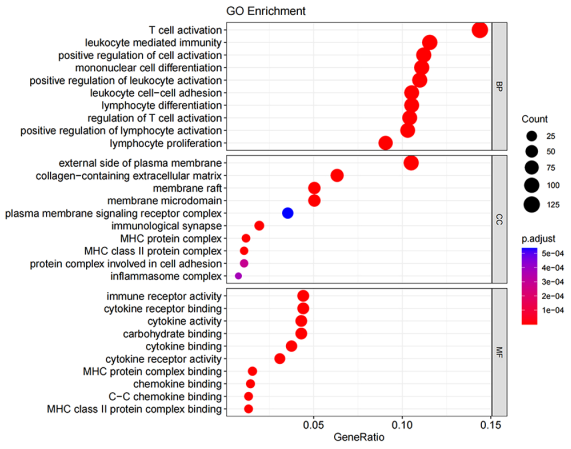

FIGURE 10
Gene ontology term enrichment for DEGs of ICS–A and ICS–B subtypes. The result shows the top 10 significantly enriched terms on the BP, CC, and MF.

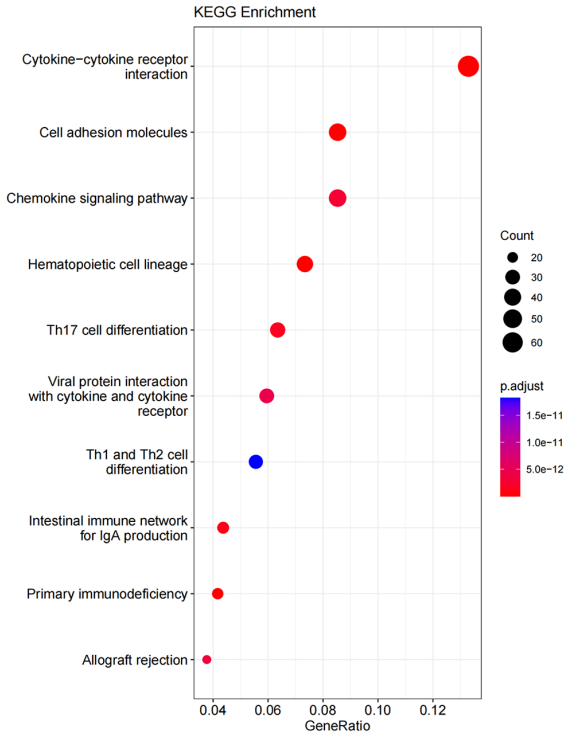

FIGURE 11
KEGG pathway enrichment for DEGs of ICS-A and ICS-B subtypes. The result shows the top 10 significantly enriched pathways.

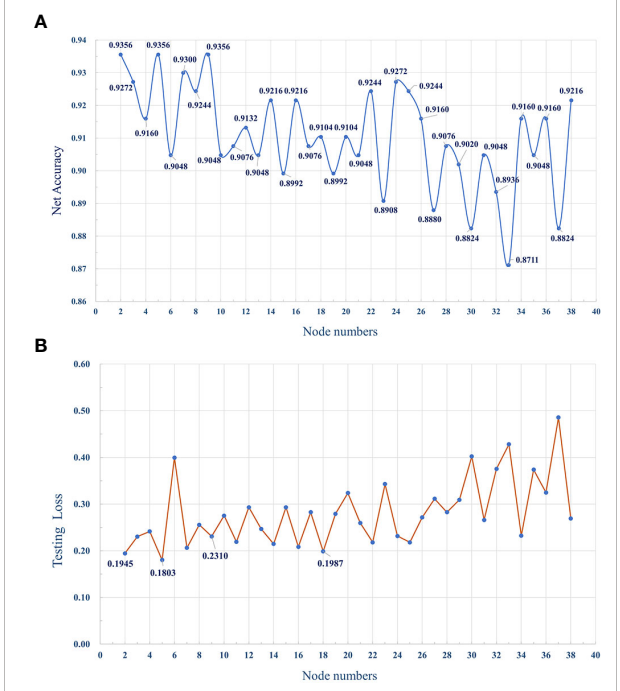

FIGURE 12
(A) The dotted line graph shows the accuracy of the MLP classifier, with different number of nodes in the hidden layer. **(B)** The average loss of the model with different numbers of hidden layer nodes during testing.

times. The loss and accuracy of the model during the training process are shown in Figures 14, 15 shown. As the number of training increases, the loss of the model decreases and tends to stabilize. The accuracy improves continuously with the increase of training times, and tends to stabilize after more than 200 times. Further, comparative experiments were conducted using other machine learning models, and Table 2 shows the highest accuracy of each model on the test dataset. The results show that Naive Bayes has the lowest accuracy (89.36%), and the accuracy rates of SVM, RF, MLP are 92.99%, 91.59% and 93.56%, respectively. The prediction accuracy of the MLP model on the test dataset is slightly higher than that of the SVM by 0.57%. However, with tuning of the MLP hyperparameters (eg, number of model layers, number of iterations for training), the prediction performance could be improved. Figure 12A shows the accuracy of models trained with different numbers of nodes in the hidden layer. When the number of nodes were 2, 5 and 9 respectively, the accuracy was the highest (93.56%), while with the number of nodes at 33, the accuracy was the worst (87.11%); However, when the number of hidden layer nodes is 5, the model obtains the smallest loss during testing(Figure 12B). Hence, the number of nodes in the hidden layer of the model was

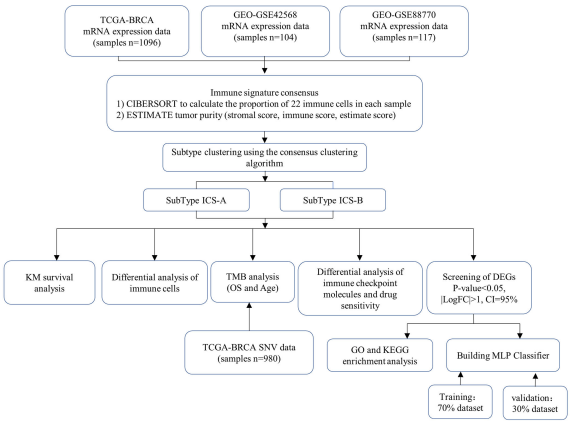

FIGURE 13
Overview of the study design.

finally determined to be 5. This classifier can effectively assist clinicians in the diagnosis and subtype identification of breast cancer.

In conclusion, the identification of breast cancer subtypes based on the immune signature in the tumor microenvironment can assist clinicians to effectively and accurately assess the progression of breast cancer and formulate different treatment strategies for different subtypes. In the present study, we detailed the immune infiltration landscape of the study cohort and demonstrated the clinical utility of immune–based subtyping. Further, this study explored the differences in immune checkpoint molecules, DEGs, and pathway enrichment between the two subtypes, and revealed that TMB in breast cancer patients was associated with OS and age. These findings have the potential to provide a new approach for the targeted therapy of breast cancer and lay

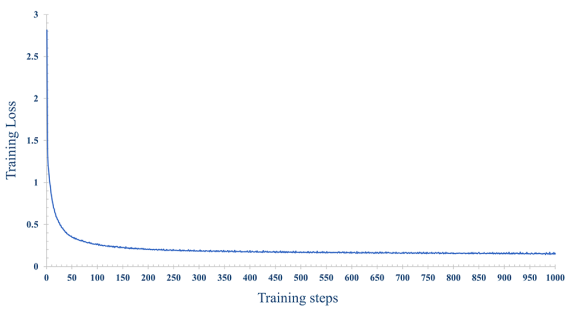

FIGURE 14
The change process of the loss, with the increase of training epochs, when the number of nodes is 5 in the hidden layer.

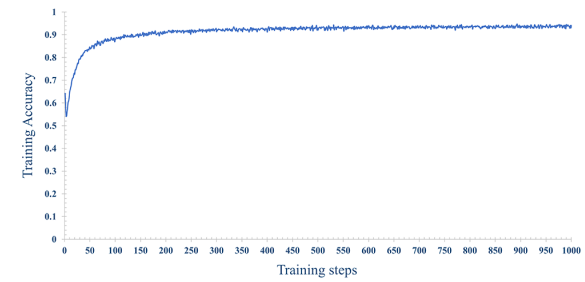

FIGURE 15
The change process of the accuracy rate, with the increase of training epochs, when the number of nodes is 5 in the hidden layer.

TABLE 2 Accuracy comparison of machine learning models.

Model Types	Testing Accuracy
SVM Model	92.99%
Naive Bayes Model	89.36%
Random Forest Model	91.59%
MLP Model	93.56%

a theoretical basis for the use of chemotherapy drugs for patients. Finally, we developed a subtype classifier with high robustness and accuracy, which can effectively assist clinicians in medical diagnosis.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

XY and YZ contributed to conception and design of the study. XY and XS contributed to the collection and collation of data. XY and WJ performed the statistical analysis. XY and XX wrote the first draft of the manuscript. XS, WJ, and HP wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work is supported in part by the National Natural Science Foundation of China (No. 81871508) and the Major Program of Shandong Province Natural Science Foundation (ZR2019ZD04).

Acknowledgments

We thank all participants for their contributions in this study. Thanks to The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) for their data support services.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.943874/full#supplementary-material

References

1. Yousefi H, Maheronnaghsh M, Molaei F, Mashouri L, Reza Aref A, Momeny M, et al. Long noncoding rnas and exosomal lncrnas: classification, and mechanisms in breast cancer metastasis and drug resistance. *Oncogene* (2020) 39:953–74. doi: 10.1038/s41388-019-1040-y

2. Adelaida A, Elena G-MD, Laura R, Andrei NS, Radu JC, Ioan L. Patient-reported quality of life 3 months after breast reconstruction. *Chirurgia (Bucharest Romania: 1990)* (2021) 116:232–7. doi: 10.21614/chirurgia.116.2.232

3. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistic. *CA: Cancer J Clin* (2019) 69:438–51. doi: 10.3322/caac.21583

4. Joseph C, Papadaki A, Althobiti M, Alsaleem M, Aleskandarany MA, Rakha EA. Breast cancer intratumour heterogeneity: current status and clinical implications. *Histopathology* (2018) 73:717–31. doi: 10.1111/his.13642

5. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *cell* (2011) 144:646–74. doi: 10.1016/j.cell.2011.02.013

6. Kalluri R, Zeisberg M. Fibroblasts in cancer. *Nat Rev Cancer* (2006) 6:392–401. doi: 10.1038/nrc1877

7. Straussman R, Morikawa T, Shee K, Barzily-Rokni M, Qian ZR, Du J, et al. Tumour micro-environment elicits innate resistance to raf inhibitors through hgf secretion. *Nature* (2012) 487:500–4. doi: 10.1038/nature11183

8. Shihab I, Khalil BA, Elemam NM, Hachim IY, Hachim MY, Hamoudi RA, et al. Understanding the role of innate immune cells and identifying genes in breast cancer microenvironment. *Cancers* (2020) 12:2226. doi: 10.3390/cancers12082226

9. Gajewski TF, Schreiber H, Fu Y-X. Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* (2013) 14:1014–22. doi: 10.1038/ni.2703

10. Pagés F, Galon J, Dieu-Nosjean M-C, Tartour E, Sautès-Fridman C, Fridman W. Immune infiltration in human tumors: A prognostic factor that should not be ignored. *Oncogene* (2010) 29:1093–102. doi: 10.1038/onc.2009.416

11. Fridman WH, Pagés F, Sautès-Fridman C, Galon J, et al. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* (2012) 12:298–306. doi: 10.1038/nrc3245

12. Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, et al. Intratumoral t cells, recurrence, and survival in epithelial ovarian cancer. *New Engl J Med* (2003) 348:203–13. doi: 10.1056/NEJMoa020177

13. Sato E, Olson SH, Ahn J, Bundy B, Nishikawa H, Qian F, et al. Intraepithelial cd8+ tumor-infiltrating lymphocytes and a high cd8+/regulatory t cell ratio are associated with favorable prognosis in ovarian cancer. *Proc Natl Acad Sci* (2005) 102:18538–43. doi: 10.1073/pnas.0509182102

14. Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, et al. Effector memory t cells, early metastasis, and survival in colorectal cancer. *New Engl J Med* (2005) 353:2654–66. doi: 10.1056/NEJMoa051424

15. Mlecnik B, Tosolini M, Kirilovsky A, Berger A, Bindea G, Meatchi T, et al. Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction. *J Clin Oncol* (2011) 29:610–8. doi: 10.1200/JCO.2010.30.5425

16. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* (2013) 4:1–11. doi: 10.1038/ncomms3612

17. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* (2015) 12:453–7. doi: 10.1038/nmeth.3337

18. Pesce S, Trabanelli S, Di Vito C, Greppi M, Obino V, Guolo F, et al. Cancer immunotherapy by blocking immune checkpoints on innate lymphocytes. *Cancers* (2020) 12:3504. doi: 10.3390/cancers12123504

19. Yang W, Lei C, Song S, Jing W, Jin C, Gong S, et al. Immune checkpoint blockade in the treatment of malignant tumor: current statue and future strategies. *Cancer Cell Int* (2021) 21:1–14. doi: 10.1186/s12935-021-02299-8

20. Foster JB, Barrett DM, Karikó K. The emerging role of *in vitro*-transcribed mrna in adoptive t cell immunotherapy. *Mol Ther* (2019) 27:747–56. doi: 10.1016/j.ymthe.2019.01.018

21. Zhao H, Xu J, Li Y, Guan X, Han X, Xu Y, et al. Nanoscale coordination polymer based nanovaccine for tumor immunotherapy. *ACS Nano* (2019) 13:13127–35. doi: 10.1021/acsnano.9b05974

22. Qin L, Zhang H, Zhou Y, Umeshappa CS, Gao H. Nanovaccine-based strategies to overcome challenges in the whole vaccination cascade for tumor immunotherapy. *Small* (2021) 17:2006000. doi: 10.1002/smll.202006000

23. Seliger B, Massa C. Immune therapy resistance and immune escape of tumors. *Cancers* (2021) 13:551. doi: 10.3390/cancers13030551

24. Jia X, Yan B, Tian X, Liu Q, Jin J, Shi J, et al. Cd47/sirpαPathway mediates cancer immune escape and immunotherapy. *Int J Biol Sci* (2021) 17:3281. doi: 10.7150/ijbs.60782

25. Cerezo M, Robert C, Liu L, Shen S. The role of mrna translational control in tumor immune escape and immunotherapy resistance. *Cancer Res* (2021) 81:5596–604. doi: 10.1158/0008-5472.CAN-21-1466

26. Mazzolini G, Murillo O, Atorrasagasti C, Dubrot J, Tirapu I, Rizzo M, et al. Immunotherapy and immunoescape in colorectal cancer. *World J gastroenterol: WJG* (2007) 13:5822. doi: 10.3748/wjg.v13.i44.5822

27. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *nature* (2000) 406:747–52. doi: 10.1038/35021093

28. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci* (2001) 98:10869–74. doi: 10.1073/pnas.191367098

29. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* (2011) 121:2750–67. doi: 10.1172/JCI45014

30. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SA, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* (2015) 21:1688–98. doi: 10.1158/1078-0432.CCR-14-0432

31. Jézéquel P, Loussouarn D, Guérin-Charbonnel C, Campion L, Vanier A, Gouraud W, et al. Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast Cancer Res* (2015) 17:1–16. doi: 10.1186/s13058-015-0550-y

32. Horr C, Buechler SA. Breast cancer consensus subtypes: A system for subtyping breast cancer tumors based on gene expression. *NPJ Breast Cancer* (2021) 7:1–13. doi: 10.1038/s41523-021-00345-2

33. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* (2017) 318:517–8. doi: 10.1001/jama.2017.7797

34. Wang H, Pujos-Guillot E, Comte B, de Miranda JL, Spiwok V, Chorbev I, et al. Deep learning in systems medicine. *Briefings Bioinf* (2021) 22:1543–59. doi: 10.1093/bib/bbaa237

35. Zhou F, Yin M-M, Jiao C-N, Zhao J-X, Zheng C-H, Liu J-X. Predicting mirna-disease associations through deep autoencoder with multiple kernel learning. *IEEE Trans Neural Networks Learn Syst* (2021). doi: 10.1109/TNNLS.2021.3129772

36. Zhang S, Zhang E, Long J, Hu Z, Peng J, Liu L, et al. Immune infiltration in renal cell carcinoma. *Cancer Sci* (2019) 110:1564–72. doi: 10.1111/cas.13996

37. Liu X, Shang X, Li J, Zhang S. The prognosis and immune checkpoint blockade efficacy prediction of tumor-infiltrating immune cells in lung cancer. *Front Cell Dev Biol* (2021) 9:1983. doi: 10.3389/fcell.2021.707143

38. Hu B, Shi X, Du X, Xu M, Wang Q, Zhao H. Pattern of immune infiltration in lung cancer and its clinical implication. *Clinica Chimica Acta* (2020) 508:47–53. doi: 10.1016/j.cca.2020.04.036

39. Mami-Chouaib F, Blanc C, Corgnac S, Hans S, Malenica I, Granier C, et al. Resident memory t cells, critical components in tumor immunology. *J immunother Cancer* (2018) 6:1–10. doi: 10.1186/s40425-018-0399-6

40. Rosenberg J, Huang J. Cd8+ t cells and nk cells: parallel and complementary soldiers of immunotherapy. *Curr Opin Chem Eng* (2018) 19:9–20. doi: 10.1016/j.coche.2017.11.006

41. Liu Q, Yang C, Wang S, Shi D, Wei C, Song J, et al. Wnt5a-induced m2 polarization of tumor-associated macrophages *via* il-10 promotes colorectal cancer progression. *Cell Communication Signaling* (2020) 18:1–19. doi: 10.1186/s12964-020-00557-2

42. Jiang Y, Li Y, Zhu B. T-Cell exhaustion in the tumor microenvironment. *Cell Death Dis* (2015) 6:e1792–2. doi: 10.1038/cddis.2015.162

43. Noy R, Pollard JW. Tumor-associated macrophages: from mechanisms to therapy. *Immunity* (2014) 41:49–61. doi: 10.1016/j.immuni.2014.06.010

44. Condeelis J, Pollard JW. Macrophages: obligate partners for tumor cell migration, invasion, and metastasis. *Cell* (2006) 124:263–6. doi: 10.1016/j.cell.2006.01.007

45. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* (2019) 51:202–6. doi: 10.1038/s41588-018-0312-8

46. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol Cancer Ther* (2017) 16:2598–608. doi: 10.1158/1535-7163.MCT-17-0386

47. Osipov A, Lim SJ, Popovic A, Azad NS, Laheru DA, Zheng L, et al. Tumor mutational burden, toxicity, and response of immune checkpoint inhibitors targeting pd (l) 1, ctla-4, and combination: a meta-regression analysis. *Clin Cancer Res* (2020) 26:4842–51. doi: 10.1158/1078-0432.CCR-20-0458

48. Sha D, Jin Z, Budczies J, Kluck K, Stenzinger A, Sinicrope FA. Tumor mutational burden as a predictive biomarker in solid tumors. *Cancer Discov* (2020) 10:1808–25. doi: 10.1158/2159-8290.CD-20-0522

49. Erbe R, Wang Z, Wu S, Xiu J, Zaidi N, La J, et al. Evaluating the impact of age on immune checkpoint therapy biomarkers. *Cell Rep* (2021) 36:109599. doi: 10.1016/j.celrep.2021.109599

50. Berenjeno IM, Piñeiro R, Castillo SD, Pearce W, McGranahan N, Dewhurst SM, et al. Oncogenic pik3ca induces centrosome amplification and tolerance to genome doubling. *Nat Commun* (2017) 8:1–15. doi: 10.1038/s41467-017-02002-4

51. Martínez-Sáez O, Chic N, Pascual T, Adamo B, Vidal M, González-Farré B, et al. Frequency and spectrum of pik3ca somatic mutations in breast cancer. *Breast Cancer Res* (2020) 22:1–9. doi: 10.1186/s13058-020-01284-9

52. Goncalves MD, Hopkins BD, Cantley LC. Phosphatidylinositol 3-kinase, growth disorders, and cancer. *New Engl J Med* (2018) 379:2052–62. doi: 10.1056/NEJMra1704560

53. André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, et al. Alpelisib for pik3ca-mutated, hormone receptor–positive advanced breast cancer. *New Engl J Med* (2019) 380:1929–40. doi: 10.1056/NEJMoa1813904

54. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. *Nature* (2021) 596:590–6. doi: 10.1038/s41586-021-03828-1

55. Baek M, Baker D. Deep learning and protein structure modeling. *Nat Methods* (2022) 19:13–4. doi: 10.1038/s41592-021-01360-8

56. Sledzieski S, Singh R, Cowen L, Berger B. D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst* (2021) 12:969–82. doi: 10.1016/j.cels.2021.08.010

57. Kong R, Wang F, Zhang J, Wang F, Chang S. Codockpp: a multistage approach for global and site-specific protein–protein docking. *J Chem Inf modeling* (2019) 59:3556–64. doi: 10.1021/acs.jcim.9b00445

58. Townshend RJ, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, et al. Geometric deep learning of rna structure. *Science* (2021) 373:1047–51. doi: 10.1126/science.abe5650

59. Eismann S, Townshend RJ, Thomas N, Jagota M, Jing B, Dror RO. Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure Function Bioinf* (2021) 89:493–501. doi: 10.1002/prot.26033

60. Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. K Deep: protein–ligand absolute binding affinity prediction *via* 3d-convolutional neural networks. *J Chem Inf modeling* (2018) 58:287–96. doi: 10.1021/acs.jcim.7b00650

61. Ragoza M, Masuda T, Koes DR. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chem Sci* (2022) 13:2701–13. doi: 10.1039/D1SC05976A

62. Bai Q, Liu S, Tian Y, Xu T, Banegas-Luna AJ, Pérez-Sánchez H, et al. Application advances of deep learning methods for *de novo* drug design and molecular dynamics simulation. *Wiley Interdiscip Reviews: Comput Mol Sci* (2021) 12:e1581. doi: 10.1002/wcms.1581

63. Krishnan SR, Bung N, Vangala SR, Srinivasan R, Bulusu G, Roy A. *De novo* structure-based drug design using deep learning. *J Chem Inf Modeling* (2021), 62 21:5100–5109. doi: 10.1021/acs.jcim.1c01319

64. Jing Y, Bian Y, Hu Z, Wang L, Xie X-QS. Deep learning for drug design: An artificial intelligence paradigm for drug discovery in the big data era. *AAPS J* (2018) 20:1–10. doi: 10.1208/s12248-018-0210-0