

## ORIGINAL RESEARCH ARTICLE

# Predicting breast cancer prognosis using $\gamma\delta$ T cell gene signatures

**Jia Weng<sup>1†</sup>, Jiacheng Weng<sup>2†</sup>, Antony Kam<sup>1</sup>, Shining Loo<sup>3</sup>, Lina Zhou<sup>4</sup> , Rencai Fan<sup>5</sup> , Runwei Guan<sup>6\*</sup>, Shicheng Li<sup>7,8\*</sup> , and Kai Chen<sup>9\*</sup>**

<sup>1</sup>Department of Biosciences and Bioinformatics, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

<sup>2</sup>Department of Oncology, Suzhou Xiangcheng People's Hospital, Suzhou, Jiangsu, China

<sup>3</sup>Wisdom Lake Academy of Pharmacy, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

<sup>4</sup>Health Management Center, The Second Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China

<sup>5</sup>Department of Medical Oncology, Fujian Medical University Union Hospital, Fuzhou, Fujian, China

<sup>6</sup>Thrust of Artificial Intelligence, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China

<sup>7</sup>Computing Science and Artificial Intelligence College, Suzhou City University, Suzhou, Jiangsu, China

<sup>8</sup>Department of Oncology, The Second Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China

<sup>9</sup>Department of Oncology, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China

<sup>†</sup>These authors contributed equally to this work.

### \*Corresponding authors:

Runwei Guan  
(runwayrwan@hkust-gz.edu.cn)  
Shicheng Li  
(lishcheng@126.com)  
Kai Chen  
(cky9920@163.com)

**Citation:** Weng J, Weng J, Kam A, Loo S, Zhou L, Fan R, *et al.* Predicting breast cancer prognosis using  $\gamma\delta$  T cell gene signatures. *Artif Intell Health*. 2026;3(2):025470105. doi: 10.36922/AIH025470105

**Received:** November 19, 2025

**Revised:** December 29, 2025

**Accepted:** January 12, 2026

**Published online:** February 3, 2026

**Copyright:** © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

**Publisher's Note:** AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Abstract

Gamma delta ( $\gamma\delta$ ) T cells exert a pivotal anti-tumor role in the breast cancer (BC) tumor microenvironment, highlighting the importance of investigating their prognostic value for improved patient stratification. We analyzed single-cell ribonucleic acid sequencing data to cluster immune cells and identify marker genes. Prognostic features were selected using Least Absolute Shrinkage and Selection Operator regression in the Cancer Genome Atlas–Breast Invasive Carcinoma cohort and validated across five external Gene Expression Omnibus cohorts. The expression of these prognostic genes was further validated by immunohistochemistry (IHC) in an in-house cohort of BC patients. These selected features were used to construct machine learning models, with the best-performing model undergoing hyperparameter tuning to optimize its performance.  $\gamma\delta$  T cells were identified as one of the major immune cell populations in BC. Twelve  $\gamma\delta$  T cell-associated genes were selected based on their prognostic significance in external validation cohorts. The random forest (RF) model achieved the highest accuracy (0.835) after hyperparameter tuning. In external validation, the final model showed the highest performance with area under the curve/accuracy values of 0.81/0.849. IHC analysis confirmed dysregulated expression of key signature proteins in BC tissues. In conclusion, we developed an efficient prognostic RF model based on a 12-gene signature from  $\gamma\delta$  T cells, which may serve as a clinically valuable tool for risk stratification in BC patients.

**Keywords:** Breast cancer;  $\gamma\delta$  T cell; Single-cell sequencing; Machine learning; Prognosis prediction

## 1. Introduction

Breast cancer (BC) has remained the leading cause of cancer-related death in women for decades.<sup>1</sup> Clinically, BC patients are grouped into different subtypes based on the expression of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor type 2 (HER2). This molecular classification has enabled molecular-targeted therapy, such as the HER2-targeted monoclonal antibody trastuzumab, which has brought substantial benefits to BC patients.<sup>2</sup> However, conventional subtype-based classification largely neglects the influence of the tumor microenvironment (TME). In recent years, tumor immunotherapy has offered new hope for BC treatment and prevention through immune checkpoint blockers.<sup>3</sup> Current strategies targeting inhibitory receptors, such as programmed cell death protein 1, benefit only a subset of BC patients.<sup>4</sup> Therefore, identifying novel biomarkers is urgently needed to extend the clinical efficacy of immunotherapy to a broader population of BC patients.

Growing evidence suggests that tumor-infiltrating lymphocytes (TILs) are key components of the TME and can profoundly influence cell growth in some tumor types.<sup>5</sup> In the TME, individual gamma delta ( $\gamma\delta$ ) T cells play a prominent role in mediating immune responses against various pathogens and cancer cells. Meanwhile, recent studies have demonstrated that  $\gamma\delta$  T cells are involved in cancer immune surveillance, although their effects vary across different cancer types.<sup>6</sup> For example, higher circulating levels of  $\gamma\delta$  T cells have been significantly correlated with improved overall survival (OS) in patients with acute leukemia.<sup>7</sup> Moreover, high  $\gamma\delta$  T cell levels have been associated with favorable clinical outcomes in patients with gastric cancer and colorectal cancer.<sup>8</sup> Unlike conventional alpha-beta ( $\alpha\beta$ ) T cells,  $\gamma\delta$  T cells recognize antigens in a major histocompatibility complex (MHC)-independent manner, allowing rapid activation in response to cellular stress. Their role in cancer, however, is dualistic: while they can exert potent cytotoxicity via interferon-gamma production, specific subsets—particularly interleukin-17-producing  $\gamma\delta$  T cells—have been implicated in promoting an immunosuppressive TME and facilitating metastasis. Despite this potential importance, the prognostic value of  $\gamma\delta$  T cell-associated gene signatures in BC remains largely unexplored.

Single-cell ribonucleic acid (RNA) sequencing (scRNA-seq) is an efficient approach for dissecting molecular mechanisms underlying tumor heterogeneity and evolution, and it provides insights that can inform precision medicine.<sup>9</sup> scRNA-seq analysis of immune cells contributes to the identification of molecular characteristics that

provide a novel perspective on cancer immunotherapy.<sup>10</sup> Previous studies have derived gene prognostic signatures from immune cell markers using single-cell RNA data, predominantly relying on Cox proportional hazards (Cox) models. Artificial intelligence (AI) is a branch of computer science that aims to perform cognitive tasks traditionally associated with human intelligence. Compared with linear models such as the Cox model, AI algorithms often demonstrate superior performance in predicting clinical outcomes from large datasets.<sup>11</sup> For example, our previous study developed an XGBoost model to predict thyroid-stimulating hormone levels in a large population undergoing routine physical examination.<sup>12</sup> Given the complexity and size of scRNA-seq datasets, AI models may outperform traditional linear models by integrating multiple features.

Therefore, this study aimed to identify a novel prognostic signature for BC based on  $\gamma\delta$  T cell markers by leveraging both single-cell and bulk transcriptomic data. Specifically, we first characterized the gene expression profile of  $\gamma\delta$  T cells within the BC microenvironment. Subsequently, we developed and validated a robust prognostic model using advanced machine learning algorithms, with the ultimate goal of providing a new tool for clinical risk stratification in BC patients.

## 2. Methods

### 2.1. Data collection

Bulk RNA sequencing data and corresponding clinical information for BC were obtained from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases. In total, 4,671 samples were included in this study. Single-cell RNA transcript sequencing data from six BC samples in GSE195861 were downloaded from the GEO database and were used to identify  $\gamma\delta$  T cell marker genes. Bulk cancer transcriptomic data (spliced transcripts alignment to a reference counts) and associated clinical profiles of BC patients were obtained from the University of California, Santa Cruz Xena platform (<https://xenabrowser.net/>). For external validation, five independent microarray datasets—GSE3143 ( $n = 158$ ), GSE202203 ( $n = 2,915$ ), GSE20685 ( $n = 327$ ), GSE19536 ( $n = 110$ ), and GSE35629 ( $n = 55$ )—were procured from the GEO database. Each dataset was normalized separately using Z-score scaling to rigorously test the model's robustness and generalizability. This approach was chosen to prioritize the assessment of the model's generalizability across diverse cohorts rather than direct data integration, thus offering a more stringent assessment of its robustness and ensuring a clearer methodological context for the reader.

## 2.2. Analysis of single-cell data

The R package Seurat (version 4.0) was used for standardizing downstream processing for scRNA-seq data. Specifically, genes detected in more than three cells and cells with more than 200 detected genes were selected. Cells with >20% mitochondrial reads were excluded. The LogNormalize package was used for data normalization. The T-distributed stochastic neighbor embedding (t-SNE) method was used for nonlinear dimensionality reduction. Cell populations were visualized on a two-dimensional map. The “FindVariableFeatures” function was used to identify the top 2,000 highly variable genes. The “FindAllMarkers” function was used to identify significant genes for each cluster by setting an absolute  $\log_2$  fold change of 1 and a minimum percentage of 0.25. In addition, the “SingleR” package was employed for cluster annotation of different immune cell types, using the HumanPrimaryCellAtlasData() reference dataset.

## 2.3. Feature selection

The most critical step in optimizing AI model performance is the selection of informative features. The least absolute shrinkage and selection operator (LASSO) regression model was used for variable selection. Lambda ( $\lambda$ ) parameters were utilized to calculate the coefficients of selected variables.<sup>13</sup> Five-fold internal cross-validation and random seed of 12345 were applied, and the 1-standard-error  $\lambda$  value was chosen. To validate the prognostic abilities of the chosen variables in external datasets, the stepwise multivariate Cox regression analysis was conducted. A risk formula was constructed as a linear combination of the selected gene expression levels weighted by the regression coefficients derived from the stepwise Cox regression model. The risk score was then calculated for each sample using this formula. Patients were stratified into high-risk and low-risk groups according to the median cutoff. For comparison with machine learning models, the predictive accuracy of this linear risk model was calculated. Kaplan–Meier survival curves were generated using the R package survminer.

## 2.4. Validation with independent cohorts

Five independent GEO cohorts were used for external validation. Risk scores for each patient in these cohorts were calculated using the previously described formula. The optimal cutoff was obtained with the survminer R package, and patients in the independent cohort were also stratified into high-risk and low-risk groups based on this optimal cutoff. Univariate Cox regression analysis was conducted to investigate the correlations between the gene signature and the OS of patients with BC from the GEO dataset. A forest graph was plotted based on the Cox

results using GraphPad Prism (version 9.0). Moreover, all AI models developed in this study were validated in the external cohorts following parameter tuning.

## 2.5. Biological function analysis

To better understand the molecular mechanism underlying the chosen variables, biological function analysis was conducted using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. First, genes most closely associated with the selected variables were identified using the STRING tool and visualized in a protein-protein-interaction (PPI) network with Cytoscape<sup>14</sup> (version 3.0). Subsequently, GO and KEGG analyses were conducted on the network genes using the R package ClusterProfiler.<sup>15</sup> To explore TME scores of the input genes, sample stromal, immune, and tumor purity scores were assessed using the Estimation of STromal and Immune cells in Malignant Tumor tissues using Expression data (ESTIMATE) algorithm.

## 2.6. Artificial intelligence model development

After identifying survival-related variables, derivation and inter-validation cohorts were constructed. To determine the optimal strategy, we tested and compared the accuracy of 3-fold, 5-fold, and 10-fold cross-validation. For example, in 10-fold cross-validation, BC patients were randomly divided into 10 subgroups. To develop the derivation cohort, nine subsets (90%) were combined, and the other subgroup (10%) was reserved as the validation set. The process was repeated 10 times, ensuring that all samples in the cohort could serve as both training and validation sets. To improve model interpretability and identify the most influential prognostic genes, SHapley Additive exPlanations (SHAP) analysis was applied. This game-theoretic approach quantifies the contribution of each of the 12 genes to the model's risk predictions, ensuring transparency in the decision-making process of the final random forest (RF) model.

In this study, five commonly used machine learning models were trained to assess the clinical outcomes in BC patients, including logistic regression (LR), support vector machine (SVM), RF, eXtreme gradient boosting (XGBoost), and decision tree (DT). Models were trained using the scikit-learn package in Python. Parameters of the five models were first set to default values and subsequently fine-tuned for the best-performing model. For LR, the default  $\lambda$  parameter in the glmnet package was selected through cross-validation. The SVM model was trained with the default kernel. XGBoost and DT models similarly picked variables that contributed to accurate survival prediction, while the RF model partitioned patients into several subgroups based on the top-ranked

variables. Among RF parameters, the number of trees and the splitting criteria were considered critical and were optimized in this study.

## 2.7. Immunohistochemistry (IHC) validation of key prognostic genes

A total of 15 paraffin-embedded BC tissue samples were obtained from the Second Affiliated Hospital of Soochow University. This study was approved by the Institutional Ethics Committee of The Second Affiliated Hospital of Soochow University. IHC was performed on 4- $\mu$ m-thick sections. Briefly, sections were deparaffinized in xylene and rehydrated through a graded ethanol series. Antigen retrieval was performed by heating the slides in 10 mM citrate buffer (pH 6.0) at 95°C for 20 minutes. Endogenous peroxidase activity was blocked by incubating the sections with 3% hydrogen peroxide for 10 minutes. Non-specific binding was blocked using 10% normal goat serum for 30 minutes at room temperature. The sections were then incubated overnight at 4°C with primary antibodies against: thymidylate synthase (TYMS) (MedChemExpress, USA), tubulin alpha 1B (TUBA1B) (MedChemExpress, USA), deoxyuridine triphosphatase (DUT) (MedChemExpress, USA), and high mobility group box 2 (HMGB2) (MedChemExpress, USA). The following day, sections were incubated with a horseradish peroxidase-conjugated goat anti-rabbit/mouse secondary antibody

(MedChemExpress, USA) for 1 h at room temperature. The signal was visualized using a diaminobenzidine substrate kit (MedChemExpress, USA), and the sections were subsequently counterstained with hematoxylin.

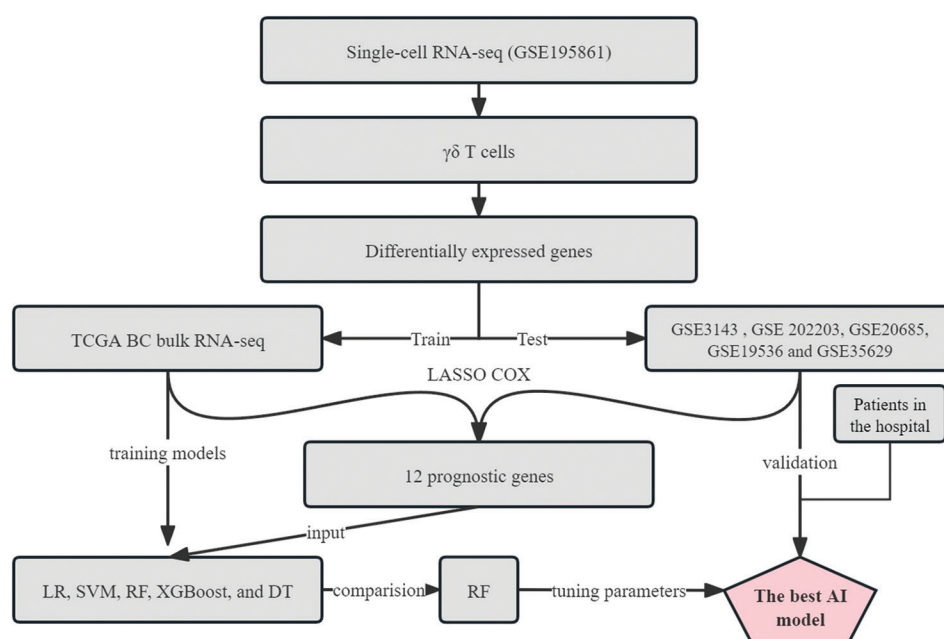
## 3. Results

The workflow of this study is illustrated in Figure 1.

### 3.1. Single-cell ribonucleic acid profiling and clustering

In this study, the scRNA-seq cohort ( $n = 6$ ) served as a discovery platform to characterize the molecular landscape of T cells, and the generalizability of the identified markers was subsequently validated in large-scale cohorts comprising over 4,000 patients. Following data preprocessing and screening, a total of 19,717 high-quality cells were selected from the six samples after quality control (Figure 2A and B). The correlation between the number of detected genes per cell (nFeature) and sequencing depth (nCount) is illustrated in Figure 2C and D.

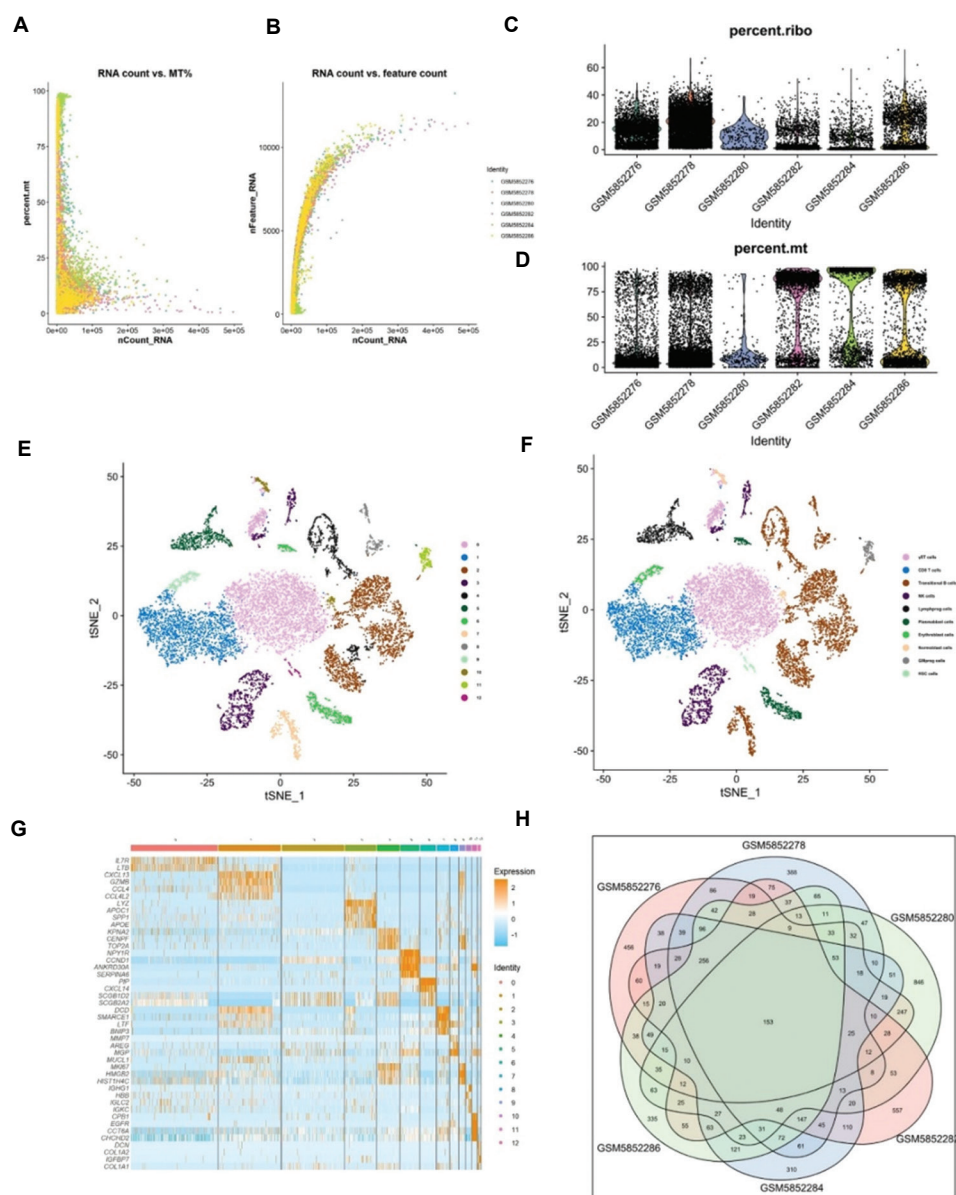
The t-SNE clustering method was used to reduce the dimensionality and to visualize the reduced-dimensional single-cell RNA data with the 2,000 highly variable genes, and all cells were successively classified into 12 subclasses using Seurat (Figure 2E). Batch effects were corrected using Harmony. These clustered 12 classes were then annotated



**Figure 1.** Overview of the study workflow

Abbreviations: AI: Artificial intelligence; BC: Breast cancer; Cox: Cox proportional hazards; DT: Decision tree;  $\gamma\delta$ : Gamma delta, LASSO: Least absolute shrinkage and selection operator; LR: Logistic regression; RF: Random forest; RNA-seq: Ribonucleic acid sequencing; SVM: Support vector machine; TCGA: The Cancer Genome Atlas; XGBoost: eXtreme gradient boosting.





**Figure 2.** Single-cell ribonucleic acid sequencing analysis identifies  $\gamma\delta$  T cells and their marker genes. (A) Correlation analysis between percentage of mitochondrial genes (percent.mt) and total UMI counts (nCount). (B) Correlation analysis between the number of detected genes per cell (nFeature\_RNA) and nCount. (C) and (D) Quality control of sequenced cells illustrated by violin plots showing RNA features (nFeature\_RNA) and absolute UMI counts (nCount\_RNA). (E) t-SNE plot of 19,717 cells from six BC samples, colored by 12 cell clusters. (F) Annotation of cell types based on known marker genes. (G) Heatmap showing the top marker genes in cell clusters. (H) Venn analysis of dysregulated genes across clusters. Abbreviations: BC: Breast cancer; t-SNE: T-distributed stochastic neighbor embedding;  $\gamma\delta$ : Gamma delta.

to known cell types using the SingleR package in R software (Figure 2F). The major cell types, ranked from highest to lowest proportion, were  $\gamma\delta$  T cells, myeloid dendritic cells, classical monocytes, plasmablasts, and other immune cells. Among all cells, clusters 0, 1, and 9 were characterized as  $\gamma\delta$  T cells. Across all groups, 310 significantly differentially expressed genes were identified, and the top significant differential genes for each cluster are displayed in the heatmap (Figure 2G). A Venn diagram of highly expressed

genes across samples is shown in Figure 2H. Marker genes specific to  $\gamma\delta$  T cells were subsequently used for variable selection in AI models to predict clinical outcomes in BC patients.

### 3.2. Characterization of $\gamma\delta$ T cells

To further characterize the  $\gamma\delta$  T cells that predominate in the BC microenvironment, we examined their expression

profiles and functional states. Based on the clustering in Figure 2, we further characterized the identified  $\gamma\delta$  T cell populations (clusters 0, 1, and 9). Figure 3A displays a heatmap illustrating the expression levels of canonical markers across all identified cell clusters, facilitating the annotation and characterization of distinct cell populations within the scRNA-seq data. Among these clusters, a cluster corresponding to  $\gamma\delta$  T cells was identified. To further validate this identification, we examined the expression of specific markers associated with  $\gamma\delta$  T cells. As shown in Figure 3B, the  $\gamma\delta$  T cell cluster exhibited high expression of key markers, including CD3 delta chain, T cell receptor gamma variable 9, and T cell receptor delta variable 2.

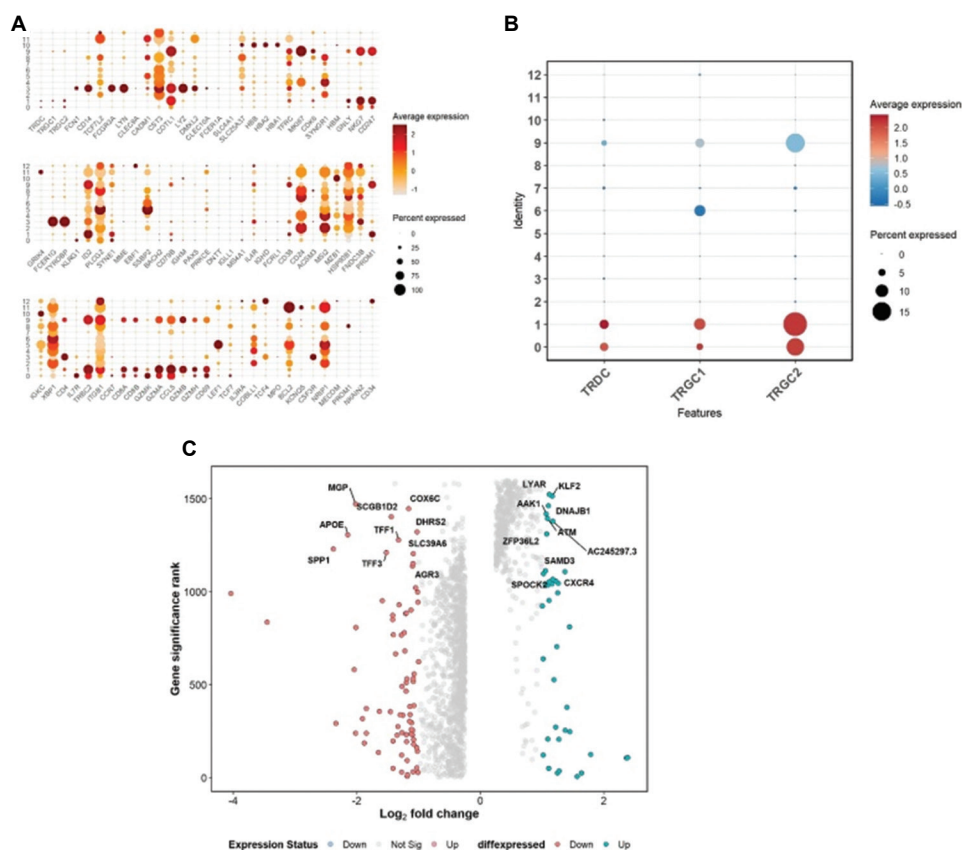
To explore the functional state of the identified  $\gamma\delta$  T cells, we performed a differential gene expression analysis. The volcano plot in Figure 3C highlights the differentially expressed genes within the  $\gamma\delta$  T cell population, with significantly upregulated genes shown in red and significantly downregulated genes shown in blue. This

analysis reveals a distinct transcriptional signature of the  $\gamma\delta$  T cells, suggesting their potential functional roles in the BC TME.

### 3.3. Cell communications

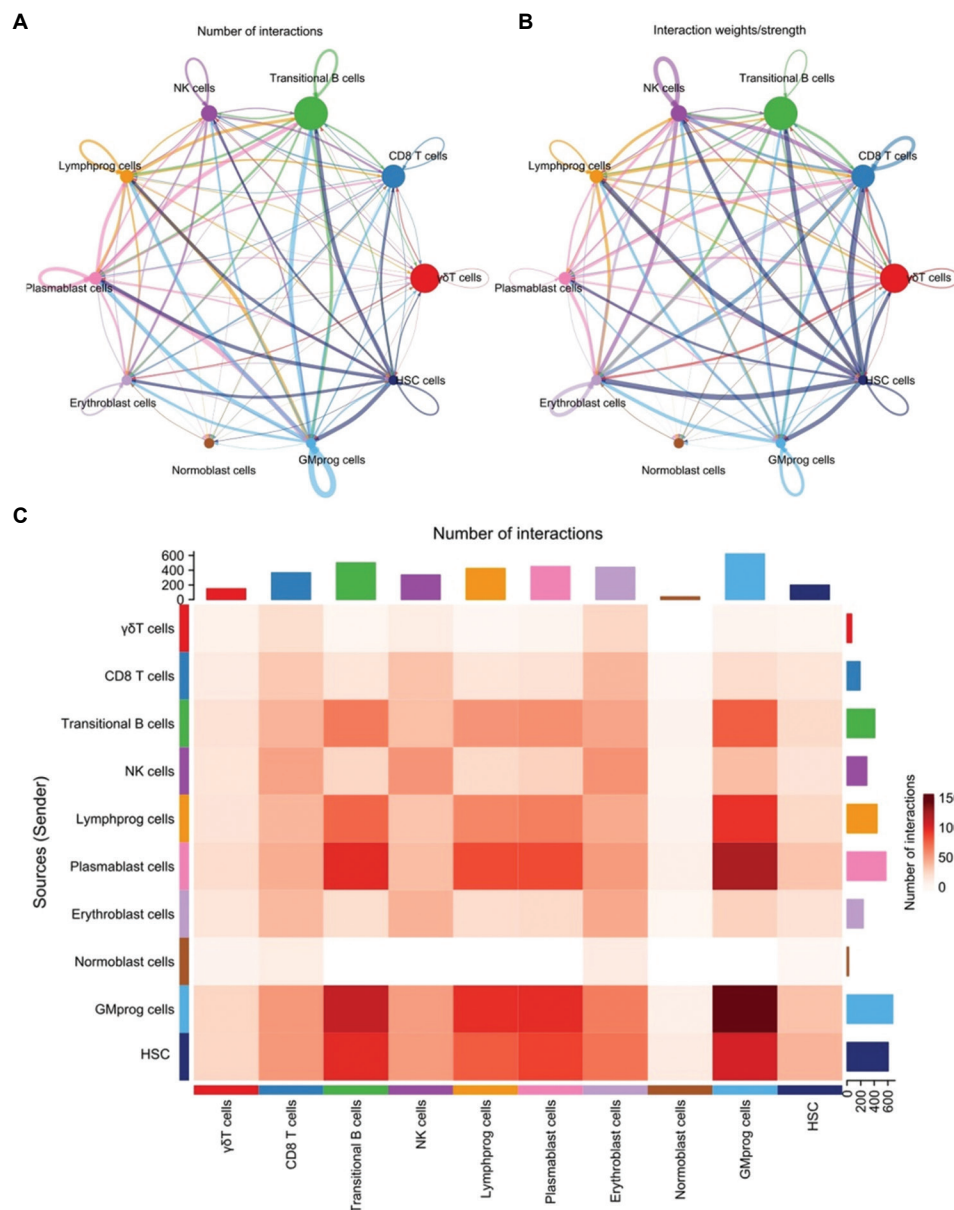
Understanding how  $\gamma\delta$  T cells communicate with other cells in the TME may provide insights into their prognostic significance. The overall interaction network is summarized in Figure 4A, illustrating the number and strength of connections among all identified cell types. Significant crosstalk was observed among various populations, with myeloid cells emerging as a central communication hub, exhibiting strong interactions with T cells, B cells, and mast cells.

Next, we dissected the key signaling pathways governing this network. As shown in Figure 4B, pathways such as annexin family of proteins, secreted phosphoprotein 1 (SPP1), macrophage migration inhibitory factor, and midkine were highly active, suggesting their crucial roles



**Figure 3.** Identification and characterization of  $\gamma\delta$  T cells. (A) Heatmap displaying the expression of canonical markers for major cell lineages across all identified clusters. The color scale represents the scaled expression levels for each marker. (B) Violin plots showing the expression of key  $\gamma\delta$  T cell marker genes (*CD3D*, *TRGV9*, and *TRDV2*) in the identified  $\gamma\delta$  T cell cluster. (C) Volcano plot of differentially expressed genes in  $\gamma\delta$  T cells. Red dots represent significantly upregulated genes, and blue dots represent significantly downregulated genes. The x-axis represents the  $\log_2$  fold change, and the y-axis represents the  $-\log_{10} p$ -value.

Abbreviation:  $\gamma\delta$ : Gamma delta.



**Figure 4.** Intercellular communication network analysis in the breast cancer tumor microenvironment. (A) Bubble plot summarizing the number and strength of interactions between different cell types. Bubble size is proportional to the number of interactions, and color intensity indicates communication strength. (B) Bubble plot showing the role of major signaling pathways across different cell types. Color intensity corresponds to the communication probability of each pathway. (C) Bubble plot displaying the key ligand-receptor pairs between sender and receiver cells. Color represents the interaction score.

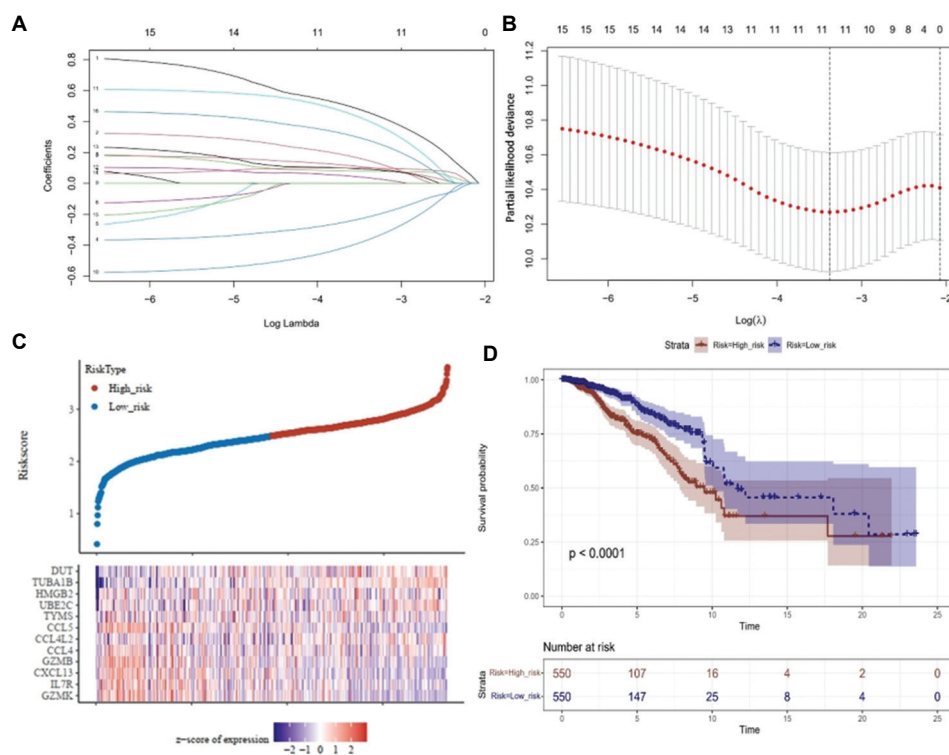
Abbreviations: CD8: Cluster of differentiation 8;  $\gamma\delta$ : Gamma delta; GMprog: Granulocyte-monocyte progenitor; HSC: Hematopoietic stem cell; Lymphprog: Lymphoid progenitor; NK: Natural killer.

in mediating intercellular communication. To pinpoint the specific molecular drivers of these signals, ligand-receptor (L-R) pairs were analyzed. Figure 4C highlights the most significant L-R interactions, such as the annexin A1-formyl peptide receptor 1 pair between myeloid cells and other immune populations, and the SPP1-(integrin subunit alpha V+integrin subunit beta 1) interaction,

revealing precise molecular mechanisms underlying the observed cellular crosstalk.

### 3.4. Variable selection

In Figure 5A, the coefficients of independent variables decreased to zero as the LASSO penalty parameter ( $\lambda$ ) gradually increased. The corresponding confidence



**Figure 5.** Variable selection and prognostic risk score construction using the least absolute shrinkage and selection operator regression. (A) The trajectory of independent variables: the horizontal axis represents the log value of lambda ( $\lambda$ ) while the vertical axis represents the coefficients of independent variables. (B) Confidence interval value of each lambda. (C) Distribution of risk score across patients and heatmap generation based on the significant genes. (D) Kaplan–Meier curves comparing high-risk and low-risk groups based on the prognostic gene signature.

intervals for each  $\lambda$  are shown in Figure 5B. Based on these results, the prognostic gene signature was constructed as follows: risk score =  $(-0.0424) \times \text{granzyme K} (GZMK) + (-0.01) \times \text{interleukin-7 receptor} (IL7R) + (-0.0415) \times \text{C-X-C motif chemokine ligand 13} (CXCL13) + (0.0335) \times \text{granzyme B} (GZMB) + (0.1718) \times \text{C-C motif chemokine ligand 4} (CCL4) + (0.057) \times \text{C-C motif chemokine ligand 4-like 2} (CCL4L2) + (-0.1203) \times \text{C-C motif chemokine ligand 5} (CCL5) + (-0.2746) \times \text{TYMS} + (0.1171) \times \text{ubiquitin conjugating enzyme E2 C} (UBE2C) + (-0.1496) \times \text{HMGB2} + (0.4131) \times \text{TUBA1B} + (0.2303) \times \text{DUT}$ . Risk scores for BC patients in the TCGA dataset were calculated to determine their distribution and association with mortality rates. Patients were stratified into high and low-risk groups (Figure 5C). Kaplan–Meier survival analysis demonstrated that patients in the high-risk group exhibited poorer clinical outcomes compared with those in the low-risk group (Figure 5D).

### 3.5. Variable validation

The prognostic performance of the 12-gene signature was validated in five independent BC cohorts. In GSE20685, hazard ratio (HR) = 1.634 (95% confidence

interval [CI] = 1.048–2.548,  $p=0.031$ ); in GSE3143, HR = 2.887, (95% CI = 1.555–5.361,  $p=0.001$ ); in GSE19536 HR = 2.713 (95% CI = 1.492–4.933,  $p=0.001$ ); in GSE202203, HR = 2.088 (95% CI = 1.700–2.566,  $p < 0.001$ ), and in GSE35629, HR = 6.476 (95% CI = 2.393–17.527,  $p=0.001$ ) (Figure 6A). In all five cohorts, patients stratified into the high-risk group had shorter OS than those in the low-risk group (Figure 6B–F).

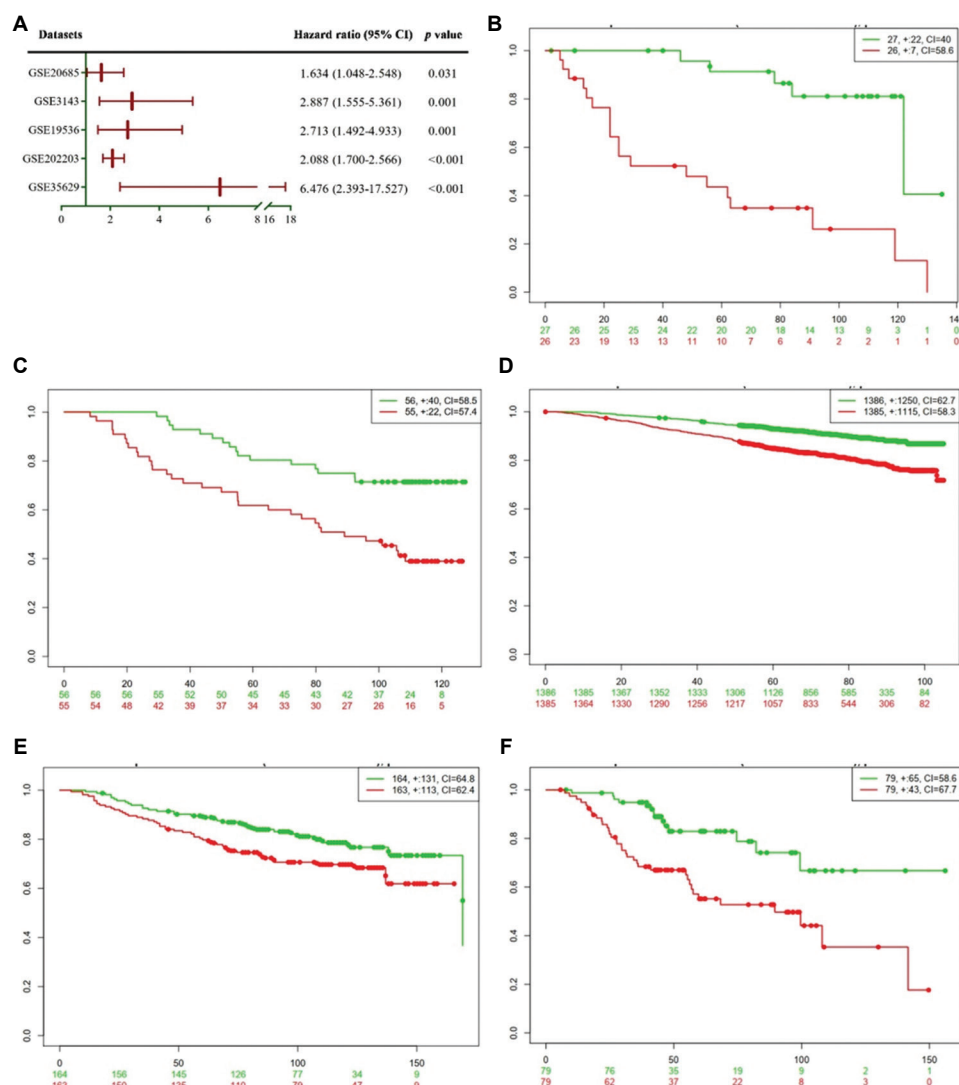
### 3.6. TME scores of the variables

TME score analysis showed that higher expression levels of *CCL4*, *CCL4L2*, *CCL5*, *CXCL13*, *GZMB*, and *IL7R* were associated with increased immune scores, stromal scores, and overall ESTIMATE scores (Figure 7).

### 3.7. Function enrichment of variables

To explore the potential biological functions of the selected variables, PPI network analysis was first conducted using the STRING database and visualized with Cytoscape. As shown in Figure 8A, a PPI network with 251 nodes and 395 edges was built. The highest co-expression score was observed between cyclin B2 and cell division cycle 20 (CDC20) (0.988), followed by





**Figure 6.** Validation of the 12-gene prognostic signature in five independent GEO datasets. (A) Hazard ratios and 95% confidence intervals for overall survival in the five validation datasets. Kaplan–Meier survival curves comparing high-risk and low-risk groups in (B) GSE35629 ( $n = 55$ ), (C) GSE19536 ( $n = 110$ ), (D) GSE202203 ( $n = 2,915$ ), (E) GSE20685 ( $n = 327$ ), and (F) GSE3143 ( $n = 158$ ).

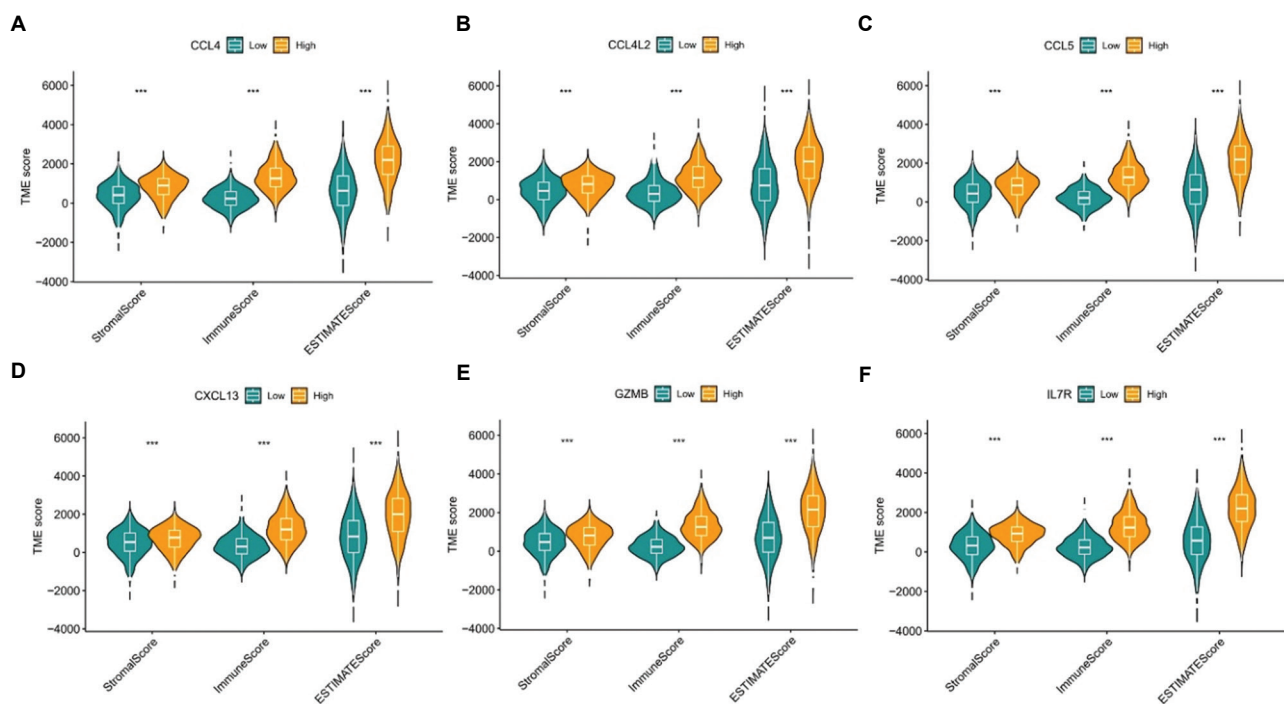
Abbreviation: GEO: Gene Expression Omnibus.

cyclin A2–mitotic arrest deficient 2 like 1 (0.984) and CDC20–UBE2C (0.983).

Subsequently, GO and KEGG enrichment analyses were conducted to investigate the biological functions and pathways associated with genes in the network. As shown in Figure 8B, most genes were significantly enriched in immune-related GO terms, such as immune response-activating cell surface receptor signaling pathway and immune response-activating signal transduction. Figure 8C presents the top 30 enriched KEGG pathways, which were mainly involved in the cytokine-cytokine receptor interactions and T cell-related signaling pathways.

### 3.8. Model training

Based on the 12-gene signature, we next explored optimal machine learning configurations to improve model stability. Different cross-validation strategies were tested and compared to identify the most effective approach. As shown in Figure 9A, the RF model with 10-fold cross-validation achieved the highest accuracy (0.824) and was therefore chosen as the candidate model. Subsequently, the RF model was further optimized through hyperparameter tuning. As shown in Figure 9B, a configuration with 90 trees and 15 splits yielded the best performance, achieving an accuracy of 0.835.



**Figure 7.** Association between  $\gamma\delta$  T cell-related gene expression and tumor microenvironment scores. Immune scores, stromal scores, and Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data scores stratified by the expression levels of (A) *CCL4*, (B) *CCL4L2*, (C) *CCL5*, (D) *CXCL13*, (E) *GZMB*, and (F) *IL7R*. \*\*\* $p < 0.005$ .

Abbreviation:  $\gamma\delta$ : Gamma delta.

### 3.9. Model validation

In the previous step, an optimized RF model was trained using the training cohort and achieved a best accuracy of 0.835 following hyperparameter tuning. The model was subsequently validated in five independent GEO datasets of BC patients. Owing to the relatively small sample sizes in some datasets, five-fold cross-validation was performed within each external cohort. The results showed that the fine-tuned RF model exhibited robust and consistent performance across all validation cohorts. Specifically, the area under the curve/accuracy values of the model is 0.81/0.849 in GSE20685, 0.75/0.812 in GSE3143, 0.75/0.807 in GSE19536, 0.80/0.841 in GSE202203, and 0.78/0.821 in GSE35629 (Figure 10).

### 3.10. Protein expression of the signature genes in BC tissues

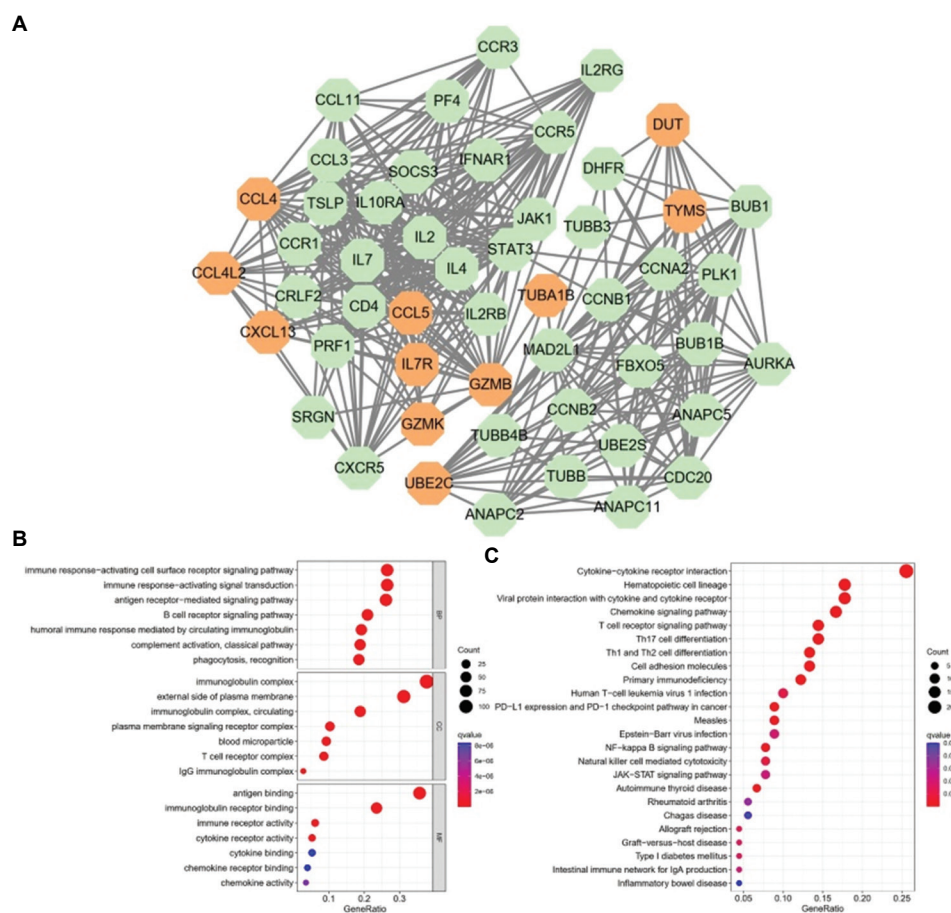
To confirm whether the prognostic value of our 12-gene signature at the RNA level is recapitulated at the protein level, we validated four most influential genes in the RF model using SHAP analysis (Figure 11A), including *TYMS*, *TUBA1B*, *DUT*, and *HMGB2* (Figure 11B-E), for IHC validation in our BC cohort. As shown in Figure 11, IHC staining revealed that the protein expression of *TYMS*, *TUBA1B*, and *DUT* was significantly elevated in BC

tumor cells. In contrast, *HMGB2* showed low expression in tumor tissues. These findings confirm that the RNA-level prognostic signature is reflected at the protein level, supporting its potential clinical relevance.

## 4. Discussion

With the rapid development of cancer immunotherapy, researchers have identified some predictive biomarkers for immunotherapy response. In some immunotherapy-related studies, the TME has been shown to play an important role in therapeutic efficacy, highlighting the urgency of exploring TME-related biomarkers.<sup>16</sup> In recent years, an increased number of studies have suggested that intertumoral heterogeneity exists across various cancer types,<sup>17</sup> which poses a significant challenge for identifying effective immunotherapeutic molecular targets. Especially for BC, few reliable TME-based biomarkers are available for prognosis prediction. Fortunately, advances in scRNA-seq and AI technologies provide powerful tools to dissect the molecular characteristics of TIL within the TME, thereby enhancing the predictive performance of prognostic models.

In clinical practice, predicting survival is important because it can influence treatment decisions.<sup>18</sup> Several studies have applied AI technologies to genomic or

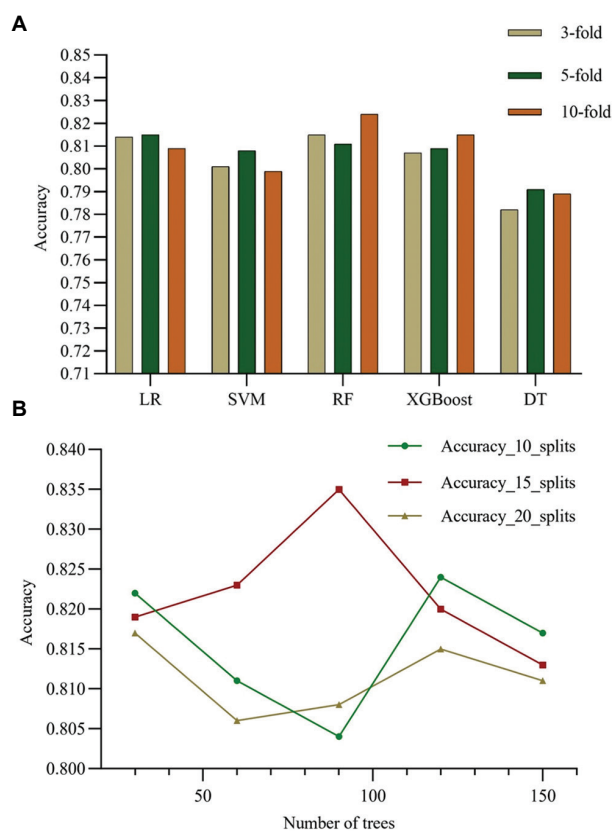


**Figure 8.** Functional enrichment analysis of the 12-gene signature. (A) Protein-protein interaction network generated from the 12 prognostic genes. Bubble maps showing the top (B) Gene Ontology terms and (C) Kyoto Encyclopedia of Genes and Genomes pathways.

transcriptomic data to predict the prognosis of cancer patients.<sup>19</sup> Traditionally, the Cox regression model, a linear model, has been used to explore correlations between clinical variables and outcomes. However, as a linear model, the Cox model may fail to capture complex, nonlinear relationships among variables.<sup>20</sup> By contrast, machine learning models are inherently nonlinear, allowing them to account for such complexities more effectively. The advantages of AI and machine learning models are increasingly recognized across various oncological domains. For instance, in non-small cell lung cancer, an artificial neural network integrating clinical and genomic data outperformed logistic regression in predicting the efficacy of epidermal growth factor receptor-tyrosine kinase inhibitor therapy, demonstrating the strength of nonlinear models in complex clinical decision-making.<sup>21</sup> Similarly, deep learning models have shown remarkable accuracy in diagnosing acute leukemia from cell images, and ensemble methods like XGBoost have achieved high precision in forecasting short-term survival for colorectal

cancer patients, underscoring their utility in prognostic tasks.<sup>22,23</sup> Furthermore, studies comparing machine learning and deep learning for COVID-19 mortality prediction confirmed that these models effectively leverage combined clinical and radiomic features. In breast cancer, techniques such as the Box-Cox transformation have been shown to optimize machine learning models by handling data skewness and improving predictive accuracy.<sup>24,25</sup> Collectively, these findings from diverse clinical scenarios reinforce the rationale for employing advanced, nonlinear AI approaches in our study. In the present work, we demonstrated that our nonlinear AI model significantly outperformed the traditional linear Cox regression model in prognosis prediction, highlighting the advantage of machine learning in handling complex transcriptomic data.

In this study, scRNA-seq data in GSE195861 were used to characterize the BC heterogeneity. Dimensionality reduction using the t-SNE clustering method and clustering via Seurat revealed 12 distinct cell clusters.



**Figure 9.** Parameter tuning of the RF model. (A) Comparison of model performance using different cross-validation methods. (B) Hyperparameter tuning of the RF model with different numbers of splits.

Abbreviations: LR: Logistic regression; SVM: Support vector machine; RF: Random forest; XGBoost: eXtremeGradient Boosting; DT: Decision tree.

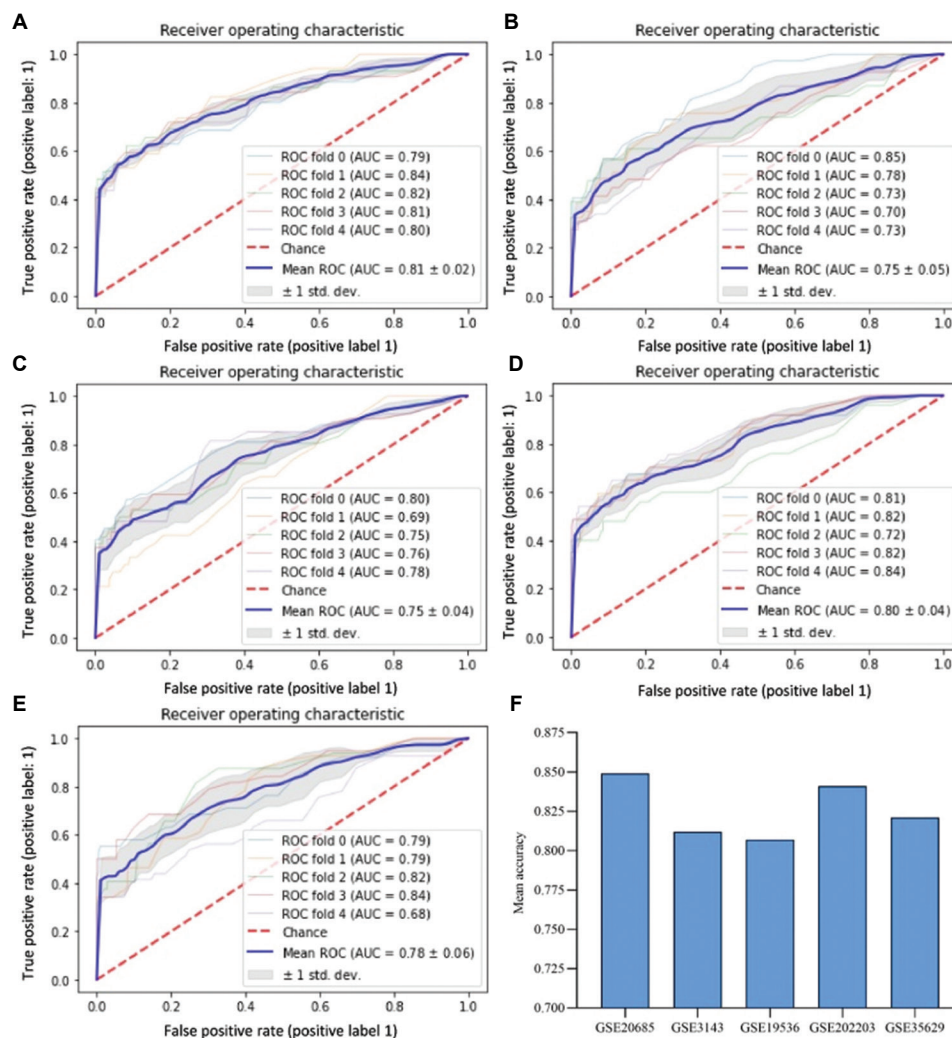
Our single-cell analysis revealed that  $\gamma\delta$  T cells constitute the predominant immune cell population in the BC microenvironment, highlighting their potential immunological significance in BC. This high prevalence is notable, as previous studies have primarily focused on  $\alpha\beta$  T cells and tumor-associated macrophages in BC.<sup>26,27</sup>  $\gamma\delta$  T cells have attracted attention in oncology for their potential use in immunotherapy, as well as their unique ability to recognize antigens independently of classical antigen presentation.<sup>28</sup> They enable the rapid activation of anti-tumor responses through direct cytotoxicity and cytokine production. The unexpected abundance of these cells in BC suggests they may play a more central role in the anti-tumor immunity than previously appreciated. This provides a strong rationale for focusing on  $\gamma\delta$  T cell-derived markers in prognostic signature development, as their prevalence likely indicates functional importance in tumor progression and patient outcomes. To identify  $\gamma\delta$  T cell-based markers, LASSO regression analysis was conducted to select genes significantly associated with

survival in BC patients. The prognostic value of these genes was then validated with external validation from the GEO database. Despite the wide disparity in sample sizes (ranging from  $n=55$  to  $n=2,915$ ), our model consistently demonstrated predictive performance. This stability, observed even in the smallest dataset (GSE35629), confirms the robustness of the  $\gamma\delta$  T cell-related signature. The  $\gamma\delta$  T cell-related genes could help predict the outcome of BC patients and were chosen as the features for downstream machine learning models.

The 12-gene signature encapsulates diverse biological functions that collectively determine patient prognosis. First, genes such as *GZMB* and *GZMK* reflect the direct cytotoxic potential of tumor-infiltrating  $\gamma\delta$  T cells, mediating anti-tumor immunity through granule secretion. Second, the chemokine cluster, including *CXCL13*, *CCL4*, and *CCL5*, is pivotal for orchestrating the immune microenvironment; notably, *CXCL13* serves as a critical organizer of tertiary lymphoid structures, which are often associated with favorable immunotherapy responses in BC. In parallel, the signature also integrates tumor proliferation and metabolism markers such as *TYMS*, *DUT*, and *UBE2C*, whose elevated expression, as confirmed by our IHC validation, correlates with aggressive tumor growth. The predictive accuracy of the model likely stems from its ability to capture the dynamic interplay between host immune defense (cytotoxicity and immune cell recruitment) and intrinsic tumor aggressiveness. This dualistic nature of the 12-gene signature resonates with emerging therapeutic paradigms in supramolecular cancer therapy, where host-guest interactions enable precise drug delivery to modulate the TME while minimizing systemic toxicity.<sup>29-31</sup> To facilitate clinical translation, the 12-gene signature could be developed into a standardized targeted assay, utilizing platforms such as a targeted quantitative polymerase chain reaction panel or NanoString nCounter, analogous to established prognostic tools. By inputting the expression data into the pre-trained model, clinicians could obtain a personalized risk score to identify patients requiring intensified treatment strategies, thus optimizing the benefit–risk ratio of adjuvant therapies.

Recently, there has been mounting evidence to suggest that our prognostic genes may also serve as potential immunotherapy targets for cancers. For example, *TUBA1B* has been reported as a prognostic biomarker for hepatocellular carcinoma,<sup>32</sup> while *HMGB2* was identified as an independent prognostic factor for BC following radical resection.<sup>33</sup> Additionally, elevated *UBE2C* expression has been correlated with poorer cancer prognosis. However, these studies were limited by small sample sizes, reliance





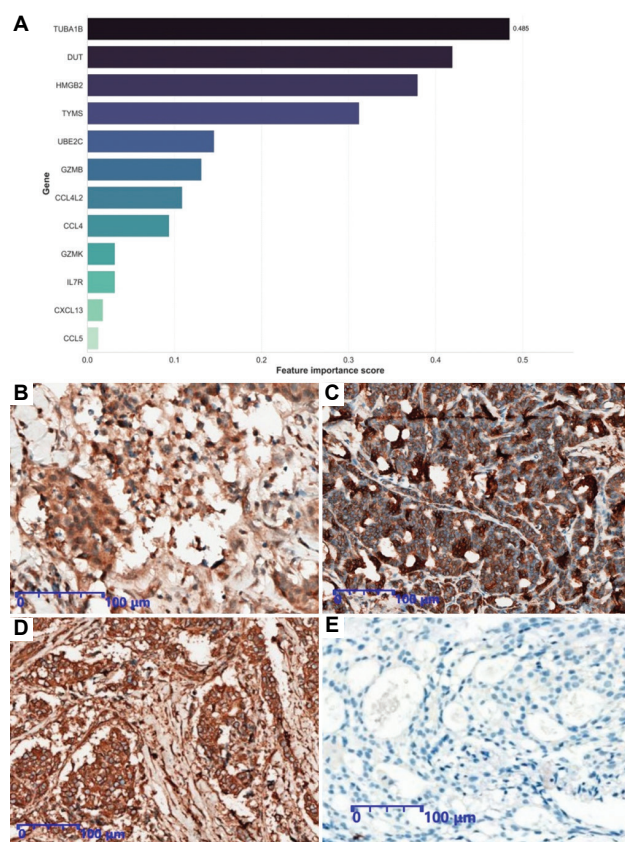
**Figure 10.** Validation of the fine-tuned random forest model. Receiver operating characteristic curves of the model in (A) GSE20685, (B) GSE3143, (C) GSE19536, (D) GSE202203, and (E) GSE35629. (F) Classification accuracies of the model in the five datasets. Abbreviations: AUC: Area under the curve; ROC: Receiver operating characteristic; std dev: Standard deviation.

on linear models, and the absence of rigorous external verification. Our findings intersect with established immunosuppressive circuits within the BC TME.

While the enrichment of cytokine-related genes suggests active immune surveillance, these pathways often function as double-edged swords. For instance, while *CCL5* and *CXCL13* are classical chemoattractants for effector T cells, their chronic overexpression may recruit regulatory T cells and myeloid-derived suppressor cells, contributing to an exhausted or suppressive immune niche. This interpretation aligns with our intercellular communication analysis, which pinpointed myeloid cells as a central hub of TME interactions. Consequently, the high-risk profile captured by our signature likely reflects a TME in which

cytotoxic  $\gamma\delta$  T cell activity is functionally compromised by concurrent myeloid-driven suppression and regulatory feedback loops.

Our study has several limitations that should be acknowledged. First,  $\gamma\delta$  T cells exhibit considerable heterogeneity, and distinguishing among  $\gamma\delta$  subsets was constrained by the absence of full-length variable-diversity-joining sequencing in the standard 3' scRNA-seq data. Second, our prognostic model focused primarily on OS due to the available clinical annotations in TCGA and GEO; future prospective studies are needed to validate its predictive value for progression-free survival and specific immunotherapy responses. Finally, although the identified markers were robustly validated in large bulk cohorts, the



**Figure 11.** Protein-level validation of the 12-gene signature using IHC. (A) Four key genes identified as most influential in the random forest model by SHapley Additive exPlanations analysis. IHC staining of the selected genes in breast cancer tissues showed that the protein levels of (B) thymidylate synthase, (C) tubulin alpha 1B, and (D) deoxyuridine triphosphatase were significantly elevated, while (E) high mobility group box 2 expression was low in tumor tissues. Scale bar: 100  $\mu$ m, magnification: 200 $\times$  for B-C.

Abbreviation: IHC: Immunohistochemistry.

relatively small size of the single-cell discovery cohort ( $n = 6$ ) may limit the detection of rare or transient  $\gamma\delta$  T cell subpopulations.

## 5. Conclusion

In summary, we developed and rigorously validated a novel 12-gene prognostic signature derived from  $\gamma\delta$  T cell markers using an RF model. This AI-driven signature demonstrates superior predictive performance compared with traditional linear models and holds significant potential as a tool for improving risk stratification and personalizing treatment for patients with BC. The IHC analysis further corroborated our findings, confirming at the protein level the differential expression of key genes identified by our AI model. This concordance between model predictions and IHC data strongly supports the clinical relevance of our gene signature. Future work integrating spatial transcriptomics

and  $\gamma\delta$  T-cell single-cell T cell receptor sequencing could further refine the immunological mechanisms underlying our prognostic model, enabling more precise and biologically informed patient management.

## Acknowledgments

None.

## Funding

This research was supported by the Postgraduate Research Scholarship (PGRS FOS2211JM02), Research Development Funding (RDF-22-02-002) from Xi'an Jiaotong-Liverpool University (Suzhou, China), the Gusu Health Talent Research Fund (GSWS2023097), the Second Affiliated Hospital of Soochow University Pre-research Project for Doctoral and Returned Overseas Students (SDFEYBS2210), and the Su-zhou Applied Basic Research Technology Innovation Project (SYW2024096).

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* Jia Weng, Jiacheng Weng

*Data curation:* Jia Weng, Rencai Fan

*Formal analysis:* Jia Weng, Shining Loo

*Funding acquisition:* Shicheng Li, Kai Chen

*Investigation:* Jia Weng, Jiacheng Weng, Antony Kam, Shining Loo, Lina Zhou

*Methodology:* Jia Weng, Antony Kam, Shining Loo

*Project administration:* Shicheng Li, Kai Chen

*Resources:* Shicheng Li, Jia Weng

*Software:* Jia Weng, Rencai Fan

*Supervision:* Shicheng Li, Kai Chen

*Validation:* Jia Weng, Jiacheng Weng, Antony Kam, Lina Zhou

*Visualization:* Jia Weng, Rencai Fan

*Writing—original draft:* Jia Weng, Jiacheng Weng

*Writing—review & editing:* All authors

## Ethics approval and consent to participate

This study was approved by the Institutional Ethics Committee of The Second Affiliated Hospital of Soochow University (JD-HG-2023-63). The requirement for informed consent from patients was obtained by the Ethics Committee of The Second Affiliated Hospital of Soochow University due to the nature of the study.

## Consent for publication

Written informed consent for publication was obtained from participants.

## Availability of data

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

## References

- Britt KL, Cuzick J, Phillips KA. Key steps for effective breast cancer prevention. *Nat Rev Cancer*. 2020;20:417-436.  
doi: 10.1038/s41568-020-0266-x
- Cameron D, Piccart-Gebhart MJ, Gelber RD, *et al*. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: Final analysis of the HERceptin adjuvant (HERA) trial. *Lancet*. 2017;389(10075):1195-1205.  
doi: 10.1016/S0140-6736(16)32616-2
- Emens LA. Breast cancer immunotherapy: Facts and hopes. *Clin Cancer Res*. 2018;24(3):511-520.  
doi: 10.1158/1078-0432.ccr-16-3001
- Li CW, Lim SO, Chung EM, *et al*. Eradication of triple-negative breast cancer cells by targeting glycosylated PD-L1. *Cancer Cell*. 2018;33(2):187-201.e10.  
doi: 10.1016/j.ccell.2018.01.009
- Roychowdhury A, Jondhale M, Saldanha E, *et al*. Landscape of toll-like receptors expression in tumor microenvironment of triple negative breast cancer (TNBC): Distinct roles of TLR4 and TLR8. *Gene*. 2021;792:145728.  
doi: 10.1016/j.gene.2021.145728
- Sebestyen Z, Prinz I, Déchanet-Merville J, Silva-Santos B, Kuball J. Translating gammadelta (γδ) T cells and their receptors into cancer cell therapies. *Nat Rev Drug Discov*. 2020;19(3):169-184.  
doi: 10.1038/s41573-019-0038-z
- Spiegel JY, Patel S, Muffly L, *et al*. CAR T cells with dual targeting of CD19 and CD22 in adult patients with recurrent or refractory B cell malignancies: a phase 1 trial. *Nat Med*. 2021;27(8):1419-1431.  
doi: 10.1038/s41591-021-01436-0
- Kabelitz D, Serrano R, Kouakanou L, Peters C, Kalyan S. Cancer immunotherapy with γδ T cells: Many paths ahead of us. *Cell Mol Immunol*. 2020;17(9):925-939.  
doi: 10.1038/s41423-020-0504-x
- Qian J, Olbrecht S, Boeckx B, *et al*. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res*. 2020;30:745-762.  
doi: 10.1038/s41422-020-0355-0
- Van der Leun AM, Thommen DS, Schumacher TN. CD8(+) T cell states in human cancer: Insights from single-cell analysis. *Nat Rev Cancer*. 2020;20:218-232.  
doi: 10.1038/s41568-019-0235-4
- Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell NAO. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*. 2021;13(1):152.  
doi: 10.1186/s13073-021-00968-x
- Cheng X, Li S, Deng L, *et al*. Predicting elevated TSH levels in the physical examination population with a machine learning model. *Front Endocrinol (Lausanne)*. 2022;13:839829.  
doi: 10.3389/fendo.2022.839829
- Lu Y, Zhou Y, Qu W, Deng M, Zhang C. A lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*. 2011;27:2406-2413.  
doi: 10.1093/bioinformatics/btr410
- Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498-2504.  
doi: 10.1101/gr.1239303
- Wu T, Hu E, Xu S, *et al*. Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*. 2021;2(3):100141.  
doi: 10.1016/j.xinn.2021.100141
- Fukumura D, Kloepper J, Amoozgar Z, Duda DG, Jain RK. Enhancing cancer immunotherapy using antiangiogenics: Opportunities and challenges. *Nat Rev Clin Oncol*. 2018;15(5):325-340.  
doi: 10.1038/nrclinonc.2018.29
- Risom T, Wang X, Liang J, *et al*. Deregulating MYC in a model of HER2+ breast cancer mimics human intertumoral heterogeneity. *J Clin Invest*. 2020;130:231-246.  
doi: 10.1172/jci126390
- Costa-Pinheiro P, Montezuma D, Henrique R, Jerónimo C. Diagnostic and prognostic epigenetic biomarkers in cancer. *Epigenomics*. 2015;7(6):1003-1015.  
doi: 10.2217/epi.15.56
- Sammut SJ, Crispin-Ortuzar M, Chin SF, *et al*. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601(7894):623-629.  
doi: 10.1038/s41586-021-04278-5
- Matsuo K, Purushotham S, Jiang B, *et al*. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am J Obstet Gynecol*. 2019;220(4):381.e1-381.e14.  
doi: 10.1016/j.ajog.2018.12.030
- Liang X, Guan R, Zhu J, *et al*. A clinical decision support system to predict the efficacy for EGFR-TKIs based on artificial neural network. *J Cancer Res Clin Oncol*.

- 2023;149(13):12265-12274.  
doi: 10.1007/s00432-023-05104-3
22. Ansari S, Navin AH, Sangar AB, Gharamaleki JV, Danishvar S. A customized efficient deep learning model for the diagnosis of acute leukemia cells based on lymphocyte and monocyte images. *Electronics*. 2023;12:322.  
doi: 10.3390/electronics12020322
23. Lee J, Cho Y, Kyung Y, Kim E. Precision forecasting in colorectal oncology: Predicting Six-month survival to optimize clinical decisions. *Electronics*. 2025;14:880.  
doi: 10.3390/electronics14050880
24. Verzellesi L, Botti A, Bertolini M, *et al.* Machine and deep learning algorithms for COVID-19 mortality prediction using clinical and radiomic features. *Electronics*. 2023;12:3878.  
doi: 10.3390/electronics12183878
25. Alshamrani SS. Machine learning techniques improving the box-cox transformation in breast cancer prediction. *Electronics*. 2025;14:3173.  
doi: 10.3390/electronics14163173
26. Zhang Y, Zhong F, Liu L. Single-cell transcriptional atlas of tumor-associated macrophages in breast cancer. *Breast Cancer Res*. 2024;26(1):129.  
doi: 10.1186/s13058-024-01887-6
27. Lv J, Liu Z, Ren X, *et al.*  $\gamma\delta$  T cells a key subset of T cell for cancer immunotherapy. *Front Immunol*. 2025;16:1562188.  
doi: 10.3389/fimmu.2025.1562188
28. Jin C, Lagoudas GK, Zhao C, *et al.* Commensal microbiota promote lung cancer development via  $\gamma\delta$  gammadelta T cells. *Cell*. 2019;176:998-1013.e16.  
doi: 10.1016/j.cell.2018.12.040
29. Yan M, Wu S, Wang Y, *et al.* Recent progress of supramolecular chemotherapy based on host-guest interactions. *Adv Mater*. 2024;36(21):e2304249.  
doi: 10.1002/adma.202304249
30. Geng H, Lin W, Liu J, Pei Q, Xie Z. Choline phosphate lipid-hitchhiked near-infrared BODIPY nanoparticles for enhanced phototheranostics. *J Mater Chem B*. 2023;11(24):5586-5593.  
doi: 10.1039/d3tb00175j
31. Dai Y, Sun J, Zhang X, *et al.* Supramolecular assembly boosting the phototherapy performances of BODIPYs. *Coord Chem Rev*. 2024;517:216054.  
doi: 10.1016/j.ccr.2024.216054
32. Hu X, Zhu H, Chen B, *et al.* Tubulin Alpha 1b Is associated with the immune cell infiltration and the response of HCC patients to immunotherapy. *Diagnostics*. 2022;12:858.  
doi: 10.3390/diagnostics12040858
33. Fu DA, Li J, Wei J, *et al.* HMGB2 is associated with malignancy and regulates Warburg effect by targeting LDHB and FBP1 in breast cancer. *Cell Commun Signal*. 2018;16(1):8.  
doi: 10.1186/s12964-018-0219-0