IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Unveiling The Precision Of Single-Cell Rna Sequencing And Machine Learning In Breast Cancer Gene Detection

[1]Ms. Ruheena Hashmi Syed Mubeen Hashmi, [2]Dr. Dnyaneshwari D. Patil, [3]Mrs. Anjali Rokde[3]

[1]Assistant Professor, Dept. of Management Science & Computer Studies, Maulana Azad College of Arts, Science and Commerce, Chh. Sanbhaji Nagar (Aurangabad), Maharashtra, India

[2]Assistant Professor, Dept. of Computer Science & IT, G. Y. Pathrikar College of CS & IT, MGM University, Chh. Sambbhaji Nagar, Maharashtra, India

[3]Assistant Professor, Dept. of Computer Science & IT, G. Y. Pathrikar College of CS & IT, MGM University, Chh. Sambbhaji Nagar, Maharashtra, India

Abstract: This study evaluates the performance of six machine learning classifiers on publicly available scRNA-seq datasets. The dataset was preprocessed using normalization, scaling, and feature selection. Gradient Boosting achieved the highest accuracy (82% on the test set before balancing), followed by Random Forest (75%) post-balancing. Hyperparameter tuning further optimized model performance, with the best model achieving 90.77% training accuracy and 82% test accuracy. Cross-validation confirmed model robustness, ensuring generalizability and reduced overfitting. The study emphasizes the importance of feature selection, ensemble learning, and data augmentation techniques in improving classification performance. Future research should explore deep learning models and multi-omics data integration for enhanced breast cancer subclassification.

Keywords: Breast Cancer Subclassification, Machine Learning, sc-RNA-seq, Random Forest, Gradient Boosting, SMOTE, Hyperparameter Tuning.

I. INTRODUCTION

Breast cancer is a prevalent and heterogeneous disease that affects millions of women worldwide. It is characterized by the abnormal growth of cells in breast tissue, forming tumors. Due to its high incidence and potential mortality, significant research has been conducted to understand its underlying mechanisms and genetic factors.

Advancements in breast cancer research have led to new detection methods utilizing single-cell RNA sequencing (scRNA-seq) and machine learning techniques. Machine learning has significantly contributed to identifying key clinical features and genetic markers associated with breast cancer, enabling more personalized and targeted therapies. Additionally, scRNA-seq has provided deeper insights into gene

expression patterns within the tumor microenvironment, leading to the discovery of potential therapeutic targets.

Understanding the expression of a gene panel in the tumor microenvironment, as revealed by scRNA-seq, offers valuable insights. Leveraging machine learning techniques facilitates the detection of mutations in specific genes associated with breast cancer. These developments highlight the importance of continued research efforts to improve breast cancer detection and treatment.

The advent of single-cell sequencing has revolutionized breast cancer detection and understanding. By attributing unique DNA and RNA signatures to tumor cells based on their presence in specific microenvironments, scRNA-seq has enabled the identification of gene expression signatures related to metastatic burden and spatial orientation within primary breast cancer tissue.

As breast cancer research continues to advance, integrating these technologies holds great promise for improving breast cancer diagnosis, prognosis, and treatment outcomes. In this study, we integrate scRNA-seq datasets, including GSE235168 and GSE161529, to assess their performance in breast cancer classification. Various machine learning techniques such as Support Vector Machine (SVM), Decision Tree, Random Forest, and K-Nearest Neighbor (KNN) play pivotal roles in analyzing scRNA-seq data.

2. Literature Review

This research uncovered a novel approach for detecting breast cancer by developing an advanced deep-learning model that analyzes gene expression of RNA-Seq data. The gene selection problem was solved by a combination of Harris Hawk Optimization (HHO) and Whale Optimization (WO) which was the foremost model out of 6 other optimization algorithms, earning an astounding score of 99.0%. Furthermore, the study provides paired breast cancer tissue samples for examination thus advancing the method to early detection and customized treatment of the disease [2].

This study investigates cell type annotation algorithms that utilize reference datasets. They include methods that compare query and reference cells like scmap and SingleR or employ other machine learning models for pattern recognition such as SVM, RF, scPred, CaSTLe, and scANVI. General models get the work done quicker, but they often underperform when it comes to handling multi-dimensional and highly complex data. [3].

This paper analyzes the need to combine the different methods of dealing with challenges found in scRNA-seq data analysis. The use of multiple techniques improves cell type characterization, diversity of analyses, and biological interpretation. Machine learning, AI, and statistical techniques are going to guide scRNA-seq research moving forward. Moving forward, there should be more focus on the refinement of algorithms, coping with data complexity, and broadening applicability for biological purposes. The utilization of these ways will help with understanding cellular mechanisms and ultimately, biomedical innovations [4].

This analysis compares 13 supervised algorithms for classifying scRNA seq data in terms of size, effectiveness, computation time, accuracy, and overall performance. Elastic Net with Interactions outperformed for small and medium datasets. Naive Bayes was also successful for medium datasets. XGBoost was successful for large datasets, but required more computation time. Ensemble techniques did not always outperform single methods. With regard to time sensitive approaches, Linear Discriminant Analysis (LDA)

was the fastest. The importance of feature selection and algorithm choice depending on dataset size for accuracy versus efficiency of the model was highlighted [5].

This literature review focuses on the impact of machine learning and AI on breast cancer subclassification. The model BreCML remarkably managed to breach the existing paradigm by single cell transcriptome analysis for identifying breast cancer subpopulations and marker genes. Other classifiers were comparably outperformed by an RNA-Seq gene expression deep learning model with a fusion gene selection strategy. Cell type classification proved to be more productive with cell-based annotation algorithms than clustering. The study pointed out the need for cross-methods for scRNA-seq data processing and modification of the algorithms to increase their complexity for the incoming data. Out of the 13 supervised learning algorithms analyzed, ElasticNet, Naïve Bayes, and XGBoost were highlighted as the best performing algorithms depending on the dataset size. Novelty of the study was the use of the Decision Tree for classification, for which it was shown to discriminate breast cancer patients from healthy women.

The proposed research study considered six separate algorithms available in genetic expression namely: 1) Logistic Regression 2) Random Forest 3) Support Vector Machine 4) K-Nearest Neighbor 5) Gradient Boosting 6) Decision Tree the combined machine learning algorithms generally tend to provide better results than a single one. Thus, to be truthful, we have created six algorithms that voted on the basis of the six individual algorithms identified and evaluated their performance against the standalone algorithms.

3. Methodology

The research study considered six separate algorithms available in genetic expression namely: 1) Logistic Regression 2) Random Forest 3) Support Vector Machine 4) K-Nearest Neighbor 5) Gradient Boosting 6) Decision Tree the combined machine learning algorithms generally tend to provide better results than a single one. Thus, to be truthful, we have created six algorithms that voted on the basis of the six individual algorithms identified and evaluated their performance against the standalone algorithms.

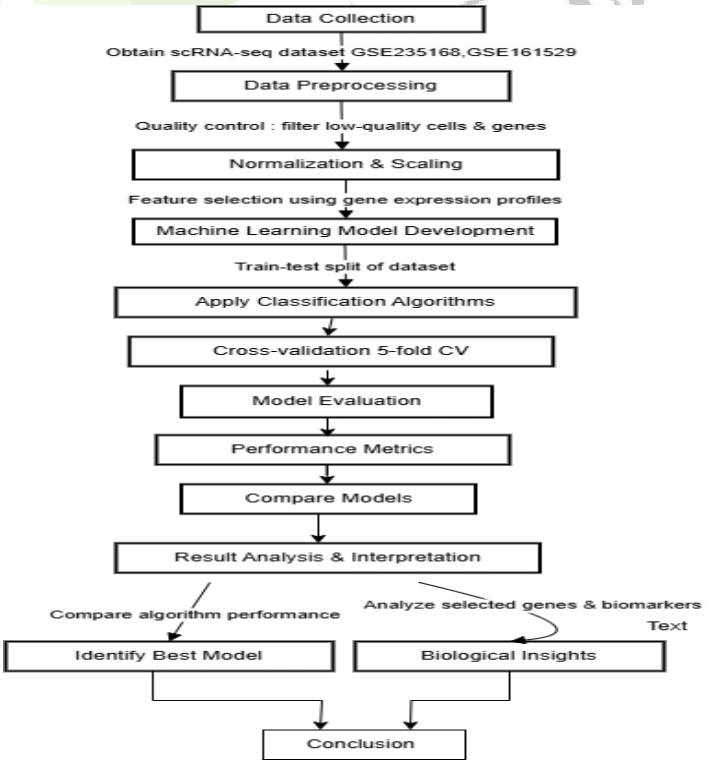

Figure 1: Workflow Steps of research methodology

3.1 Data Collection

The dataset utilized in the research is a publicly accessible dataset from the Gene Expression Omnibus GEO repository accessed through the NCBI portal. It is a dataset that is accessible to the public and contains gene expression data, which is a very important aspect of both biomedical research and machine learning applications. The description of the dataset, its characteristics, and the study's relevance to the dataset are discuss in the following sections.

The datasets GSE235168, GSE161529 were used to evaluate all classification methods for sc-RNA-seq data. For the details of these simulated datasets.

3.1.1 Dataset GSE235168

The gene expression dataset GSE235168, hosted on the Gene Expression Omnibus platform, provides valuable insights into the transcriptional landscape of a specific biological system. This dataset includes 25 patients with various molecular subtypes such as ER+ and Triple Negative mammary breast cancer.

3.1.2 Dataset GSE169246

The GSE169246 dataset, available through the Gene Expression Omnibus (GEO) database. This dataset includes immune cells from primary or metastatic tumor tissues and peripheral blood of 22 advanced TNBC patients. The dataset focuses on triple-negative breast cancer (TNBC), an aggressive form of breast cancer that lacks expression of the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2).

3.2 Data Preprocessing

Preprocessing is a crucial step in single-cell RNA sequencing (scRNA-seq) analysis, ensuring data quality and preparing it for classification models. The following steps outline the methodology for handling raw scRNA-seq data, filtering relevant features, and preparing a consolidated dataset for classification. scRNA-seq datasets typically consist of three key files stored in the following format.

- **TSV (Tab-Separated Values) format such as Feature file (.tsv) which** Contains gene names and their corresponding identifiers.
- **Matrix file:** Stores expression levels of genes across different cells.
- **Barcode file:** Lists unique cell barcodes for identifying individual cells in the dataset.

These files are first extracted and stored in a designated folder for further processing.

Steps for processing:

1. Extract the .tsv, barcode and matrix files in to a folder
2. Read the feature of .tsv file, matrix, relevant barcode file.
3. Add the feature information to a variable object.
4. Filter the object with the marker genes that exist in it.
5. Convert the filter object in to csv file.

6. Merge all files and create a single csv file.
7. Apply Classifiers on the file.

3.3 Normalization and Scaling

Normalization is required because raw gene expression values can be affected by factors such as sequencing depth, technical noise, and cell size. The goal is to scale expression levels within and across cells to make them comparable. remove the missing values from the dataset.

After normalization scaling is performed to ensure that the features (gene expression values) have a consistent range and distribution, scaling is done after normalization. This is important since machine learning algorithms work better when features are on a similar scale.

3.4 Model development and apply classification.

After the normalization and scaling we structured the dataset on tumor type and their respective genes. To developed the model the data was split into training and testing sets, followed by applying six classifiers: 1) Logistic Regression 2) Random Forest 3) Support Vector Machine 4) K-Nearest Neighbor 5) Gradient Boosting 6) Decision Tree to evaluate their performance in tumor classification.

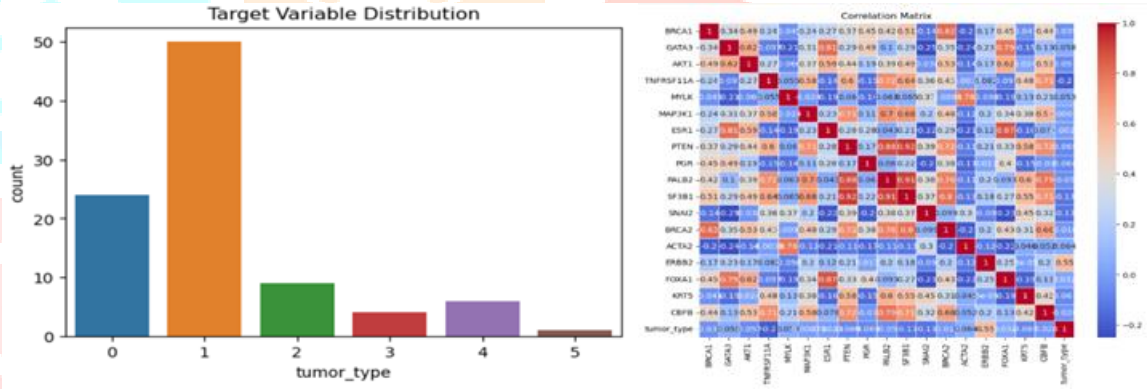

Figure2: a) Distribution of tumor type in the dataset. b) Heatmap of feature correlations in the dataset

The figure 2: a) visualizing the 'tumor type' column in the dataset provides insights into the frequency distribution of different tumor types. Each bar in the plot represents the number of samples corresponding to a specific tumor type.

Figure 2: b) describe a correlation matrix, a table of the correlations between all pairs of numerical variables in a dataset. The values range from -1 to 1:

- +1 indicates a perfect positive correlation (when a variable increases, the other also increases).
- -1 indicates a perfect negative correlation (when one variable increases, the other decreases).
- 0 indicates no correlation (no relationship between the variables).

This heatmap helps detect features with a strong correlation which can enhance feature selection in machine learning models. Highly correlated features are usually redundant and might cause overfitting, and uncorrelated ones can be more relevant for classification or prediction.

Highly correlated features such as ‘PALB2‘, ‘SF3B1‘, and ‘FOXA1‘were eliminated to reduce issues with multicollinearity, which can negatively impact the performance of the model because they are highly

correlated with each other. The removal of such features increases model generalization and minimizes repetition within the dataset. Apply the classifications on the combined dataset.

4. Results

This segment offers an in-depth analysis of classification techniques with reference to a defined benchmark. Alongside these, we also performed a train and test model of the classification accuracy for the datasets with differing sample amounts. Further, we performed evaluation of the gene-selection technique on some simulated datasets. The results question the performance of different classifiers against scRNA-seq data and how feature selection influences the model accuracy.

4.1 Classification Performance

The performance of different classification algorithms on two types of datasets with different sizes were examined. This section, Compares the performance of the classification. Display the results of each classification criterion are shown to compare the performance of the methods.

Table 2: Classification performance of different dataset

Sr. No	Classifiers	Accuracy
1	Logistic Regression	68%
2	Random Forest	78%
3	SVM	71%
4	KNN	71%
5	Gradient Boosting	82%
6	Decision Tree	64%

From the above table Best Model is Gradient Boosting with Accuracy: 82%.

To improve the performance of the machine learning model we are used optimization technique and synthesization technique. For optimized technique we are used a hyperparameter tuning process used, and for synthesized technique we are used SMOTE technique.

- The hyperparameter tuning process used, such as Gradient Boosting, XGBoost, or Random Forest. The process involves cross-validation with 5 folds for each combination of different hyperparameter settings, resulting in total model fits. The best performing hyperparameter set is selected, including learning rate, and sample fraction. The model achieved 90.77% accuracy on the training dataset, indicating a well-fitted model. Further evaluation on a test dataset is necessary to confirm generalization and avoid overfitting.

The classification applies on the test dataset with hyperparameter following report generated.

Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.71	0.71	0.71	7
1	0.94	1.00	0.97	15
2	0.50	0.33	0.40	3
3	0.00	0.00	0.00	1
4	1.00	1.00	1.00	2
accuracy			0.82	28
macro avg	0.63	0.61	0.62	28
weighted avg	0.81	0.82	0.81	28

Figure 3: Classification report on Test set accuracy with 5-fold cross validation

Best Model Accuracy of Train dataset: 90.77%

5- fold Cross-Validation Test Set Accuracy: 82%

Classification report shows Test Set Accuracy is 82%.The top performer is appraised by best model through cross-validation, which is employed to avoid overfitting as well as to make sure its generalizability and robustness. For cross validation we have to specifies 5-fold cross validation i.e. dataset is spilt into 5 subsets, and the model is trained and validated 5 times on different portions of the data. Calculate the average accuracy over all the folds and then computes the average accuracy to assess model performance. This means the model achieved an average accuracy of **91.23%** across all 5 folds.

- SMOTE (Synthetic Minority Over-sampling Technique) is used to generate synthetic samples for the minority class. SMOTE (Synthetic Minority Over-sampling Technique) is a popular oversampling method used in machine learning to address class imbalance in datasets. It synthetically generates new data points for the minority class by interpolating between existing samples, rather than simply duplicating them. This technique helps to balance the dataset, reducing bias and improving classification performance. By increasing the number of samples in the minority class, SMOTE helps models learn better decision boundaries instead of treating minority class instances as outliers.

Table 4: Classification performance of different dataset after SMOTE technique

Sr. No	Classifiers	Accuracy
1	Logistic Regression	57%
2	Random Forest	75%
3	SVM	60%
4	KNN	57%
5	Gradient Boosting	61%
6	Decision Tree	75%

Best Model: Random Forest with Accuracy: 75%

After the SMOTE technique applied on the test dataset following report generated.

Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.71	0.71	0.71	7
1	0.94	1.00	0.97	15
2	0.50	0.33	0.40	3
3	0.00	0.00	0.00	1
4	1.00	1.00	1.00	2
accuracy			0.82	28
macro avg	0.63	0.61	0.62	28
weighted avg	0.81	0.82	0.81	28

Figure 4: Classification report on Test set accuracy with 5-fold cross validation after SMOTE

Best Model Accuracy on Train Set: 99.43%

5- fold Cross-Validation Test Set Accuracy: 75%

SMOTE is applied to balance the dataset by generating synthetic examples for the minority class. This improves classification model performance by preventing bias toward the majority class and ensuring better generalization. The output shows the class distribution before and after applying SMOTE, confirming that the dataset is now balanced.

4.2 Discussion on results

The study improved a machine learning model's classification performance by utilizing hyperparameter tuning and data synthesization techniques like SMOTE. Hyperparameter tuning improved models like Gradient Boosting, XGBoost, and Random Forest, achieving a training accuracy of 90.77%. Cross-validation ensured the model did not overfit to the training data, achieving 91.23% accuracy. SMOTE helped balance the dataset by generating synthetic samples for the minority class, leading to improved classification accuracy. However, test accuracy did not increase significantly, possibly due to overfitting on synthetic data.

Comparison of Classifiers:

The classification results before and after SMOTE are summarized as follows:

Table 5: Comparision of all classifiers before and after SMOTE

Sr. No	Classifiers	Accuracy (Before SMOTE)	Accuracy (After SMOTE)
1	Logistic Regression	68%	57%
2	Random Forest	78%	75%
3	SVM	71%	60%
4	KNN	71%	57%
5	Gradient Boosting	82%	61%
6	Decision Tree	64%	75%

The **Random Forest classifier remained the best-performing model**, indicating its robustness in handling both imbalanced and balanced datasets.

5. Conclusion

The study examined the impact of machine learning algorithms on breast cancer subclassification using scRNA-seq data. Gradient Boosting achieved the highest accuracy (82% before data balancing), while Random Forest performed best (75%) after applying SMOTE to handle class imbalance. Hyperparameter tuning improved model accuracy, with the best model achieving 90.77% accuracy on the training set and 82% on the test set. Applying SMOTE balanced the dataset, reducing bias towards the majority class. However, training accuracy increased to 99.43%, while test accuracy remained at 75%. The Random Forest classifier remained the top-performing model in both imbalanced and balanced datasets, demonstrating its robustness despite sensitivity to imbalanced data and non-linearity in feature space. The study suggests optimized feature selection, ensemble learning approaches, and advanced augmentation techniques for improved classification performance.

REFERENCES

[1] Ke S, Huang Y, Wang D, Jiang Q, Luo Z, Li B,Yan D and Zhou J (2024) BreCML:identifying breast cancer cell state in scRNA-seq via machine learning. Front. Med. 11:1482726. doi: 10.3389/fmed.2024.1482726.

[2] Abrar Yaqoob1 · Navneet Kumar Verma1 · Rabia Musheer Aziz2 · Mohd Asif Shah3,4,5 RNA-Seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. Journal of Cancer Research and Clinical Oncology (2024) 150:455 https://doi.org/10.1007/s00432-024-05968-z.

[3] Erik Christensen†,Ping Luo†, Andrei Turinsky, MiaHusi´c, Alaina Mahalanabis, Alaine Naidas, Juan Javier Diaz-Mejia, Michael Brudno Evaluation of single-cell RNA seq labelling algorithms using cancer datasets Briefings in Bioinformatics, 2023, 24(1), 1–16 https://doi.org/10.1093/bib/bbac561.

[4] Megha Patel, Nimish Magre, Himanshi Motwani, and Nik Bear Brown, Advances in Machine Learning, Statistical Methods, and AI for Single-Cell RNA Annotation Using Raw Count Matrices in scRNA-seq Data arXiv:2406.05258v1 [q-bio.OT] 7 Jun 2024.

[5] Cao, X., Xing, L., Majd, E., He, H., Gu, J., & Zhang, X. (2022). A Systematic Evaluation of Supervised Machine Learning Algorithms for Cell Phenotype Classification Using Single-Cell RNA Sequencing Data. Frontiers in Genetics, 13, 836798. https://doi.org/10.3389/fgene.2022.836798

[6] Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T.,& Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biology. (2019) 20:194 https://doi.org/10.1186/s13059-019-1795-z.

[7] Mamta Jadhav1, Zeel Thakkar2, Prof. Pramila M. Chawan3 Breast Cancer Prediction using Supervised Machine Learning Algorithms, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 10 | Oct 2019 www.irjet.net p-ISSN: 2395-0072.

[8] T.Gowri1, Dr.S.Geetha2 International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072.

[9] Chen, L., Wang, W., Zhai, Y., & Deng, M. (2020). Deep soft K-means clustering with self-training for single-cell RNA sequence data. NAR Genomics and Bioinformatics, 2(2). Oxford University Press (OUP).

[10] Chen, L., Zhai, Y., He, Q., Wang, W., & Deng, M. (2020). Integrating Deep Super-vised, Self-Supervised and Unsupervised Learning for Single-Cell RNA-seq Clustering and Annotation. Genes **2020**, *11*(7), 792; **https://doi.org/10.3390/genes11070792**

[11] Chen, R., Yang, L., Goodison, S. & Sun, Y. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics 36, 1476–1483 (2020).

[12] ZhaoX, WuS,Fang N, etal. Evaluation of single-cell classifiers for single-cell RNA sequencing datasets. BriefBioinform2020;21:1581–95.

[13] Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., et al. (2019). A Comparison of Automatic Cell Identification Methods for Single-Cell Rna-Sequencing Data. Genome Biol. 20, 194. doi:10.1186/s13059-019-1795-z.

[14] Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., and Powell, J. E. (2019). ScPred: Accurate Supervised Method for Cell-Type Classification from Single-Cell RNA-Seq Data. Genome Biol. 20, 264. doi:10.1186/s13059-019-1862-5.

[15] Vieira, A. F. & Schmitt, F. An update on breast cancer multigene prognostic tests - emergent Clinical biomarkers. Front. Med. 5, 248 (2018).

[16] Brueffer, C. et al. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome analysis network—breast initiative. JCO Precis. Oncol. 2, 1–18 (2018).

[17] SavasP, VirassamyB,YeC,etal. Single cell profiling of breast cancer T cell Sreevalsan tissue-resident memory subset associated with improved prognosis. NatMed 2018; 24:986–93.

[18] Ohnstad, H. O. et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast Cancer Res. 19, 120 (2017).

[19] Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).