RNA sequencing-based Machine Learning for Detection of Triple Negative Breast Cancer

Ruben Holthuijsen (1064459) • Kevin Hartman (1044032) • Sander van Swieten (1063788) • Vince van Doorn (1061669) • Victor de Sousa Gama (0929470)

**Abstract**

TNBC is an aggressive type of breast cancer that is difficult to treat and has a poor prognosis. It is most commonly diagnosed using immunohistochemistry (IHC), but this method is time-consuming and relies on an expert to interpret the result. The goal is to offer a simpler and more scalable diagnostic tool by building a machine learning model that can classify TNBC status using only RNA sequencing data.   
The proposed project will use the Breast Invasive Carcinoma Collection dataset from The Cancer Genome Atlas (TCGA-BRCA). The result is a binary classification (TNBC versus non-TNBC), favouring simplicity over complexity, and prioritizing explainability, fairness and reproducibility of the model. The intention is to evaluate multiple machine learning algorithms and select the most effective one. The anticipated outcome is a transparent and accurate classification tool that makes use of explainable algorithms like Logistic Regression, Random Forest, and Support Vector Machines (SVM), which in turn provides an addition to current methods to show a TNBC diagnosis in research, to be applied when RNA sequencing data is available. All code, visualizations, and documentation will be made openly available to support reproducibility and future use.

**Keywords** Machine Learning · Triple Negative Breast Cancer · RNA sequence · Detection

**1 Introduction**

Breast cancer is one of the most common cancers in the world [1] in particular among women age 39 or younger [2] and among women it is the primary cause of cancer-related mortality [3]. In 2022 female breast cancer was the second most common cancer with 2.3 million cases [4, 5]. A subtype of breast cancer, triple negative breast cancer (TNBC), occurs in 10-20% of cases [6, 7]. Notable characteristics include low socio-economic background and ethnicity [8, 9].

What defines TNBC is a lack of expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) in tumor cells [3, 9, 10]. Gene mutations and highly rearranged genomes characterize TNBC [11]. In addition, TNBC lacks molecular targets for therapy, is heterogeneous , high aggressivity, and proliferates more [12], leading to poor prognosis [13].

The influence of solid tumors on the body's systemic immune environment throughout cancer progression remains incompletely understood [14]. The tumor microenvironment (TME) plays a pivotal role in tumor genesis and tumor progression [15-17].

Diagnostic methods for TNBC include MRI, mammography, and, primarily, IHC [8]. Recently, liquid biopsy is emerging as a tool for diagnosing cancer [7].

Initial therapeutic decisions are guided by PD-L1 expression and BRCA mutations [8]. Although (neoadjuvant) chemotherapy is the primary treatment modality for breast cancer in the early stage [16, 18-20]. Immunotherapy is recognized as additional treatment, yet the approach is in early stages and TNBC lacks enough immune cell infiltration making the treatment moderately effective [14, 17, 21].

This study builds a binary classifier for TNBC status, based ethical principles such as transparency, reproducibility and algorithmic fairness. As pointed out by Pinto et al., there is evidence for ethnicity specific determinants, and there exists a gap in genomic data between ethnicities [5].

Diagnostic tools based on machine learning must be transparent and understandable to be able to safely and ethically adopt them into practice, as well as to earn the trust of doctors [22, 23]. Using RNA sequencing data given the accuracy and high sensitivity of RNA sequencing data [24]. The data used is The Cancer Genome Atlas (TCGA), more specifically the Breast Invasive Carcinoma Collection (TCGA-BRCA), given the detailed view the dataset offers [25, 26].

**2 Literature Review**

In assessing the current research on TNBC this study frames the developments for treatment of TNBC into four main areas. A point of note in the approach for literature review is that only open access literature has been reviewed. Below, each area is summarized, after which this study is positioned within the framework developed.

**Tumor microenvironment (TME) / tumor immune microenvironment (TIME)**

Comprehensive understanding of the TME is the basis for immunotherapy [27]. A key role of TME is the genesis, development, and metastasis of cancer. Within the TME different types of (non-cancerous) cells, including immune, endothelial, and adipocytes cells, continuously interact with cancer cells, thereby having a pronounced effect on growth and progression of a tumor [28, 29]. Regarding progression, and treatment, the TIME plays an increasing role although the understanding of the processes is incomplete [11]. Within the context of subtypes, the TIME has also been included in recent research [6].

**Subtyping**

Within different subtypes of TNBC the heterogeneity of TME is notably pronounced [6, 29]. Initial therapeutic decisions are based on a limited set of biomarkers (PD-L1, gBRCA, ki-67).

As TNBC is heterogeneous, to improve diagnosis and treatment there are multiple classifications of subtypes. Classifications like PAM50, Lehmann, Burstein, Jézéquel, and Fudan University Shanghai Cancer Center (FUSCC). Where Lehmann and Burstein have similarities, Jézéquel puts emphasis on immune-relative factors, and FUSCC is recognized as a system tailored specifically for Chinese patients [8]. As for subtyping and ethnicity, there is a gap in genomic information between ethnicities, especially regarding the African continent where a higher TNBC incidence is observed, and the continent has the highest genomic diversity [5]. In addition to the aforementioned classifications, other subtyping has been suggested based on metabolics [30]. Recent developments in different approaches in determining subtypes are differential sparse canonical correlation analysis network (DSCCN) [31] and Multi-Omics Adaptive Integration Method with Graph Learning and Self Attention (MoAGL-SA) [26]. The heterogeneity of TNBC as revealed by transcriptomic and proteomics underlines the importance of a multi-omics approach [32].

**Multi-omics**

By applying the technology of multi-omics, enabling simultaneous analysis of different layers, both the broadening of the scope, and understanding of, the heterogeneity is facilitated. Use of multi-omics include mapping intra-tissue interactions, DNA mutations, protein expression levels, epigenetic modifications, and intercellular communications [33]. The variability of the omics data also plays a role in benchmarking studies [34], although the use of multi-omics comes with challenges like integration of data [35]. Spatial omics can dissect further differences of cells in tissue [36]. Different types of spatial omics have been developed and can be defined into two categories. On the one hand deciphering characterizations of regions (spacecraft-like) and outlining the structure of the tumor (telescope-like). An example of spacecraft-like is Laser Capture Microdissection (LCM) which discovered a novel biomarker [37].

**Biomarkers and prognosis**

While the main treatment modality remains chemotherapy, the in-depth research into the molecular heterogeneity of TNBC provides new perspectives for other forms of therapy like immune modulators [16]. Trying to find targets for treatment in, for example, differential expressed genes (DEGs) [38], extracellular vesicles (EVs) derived from plasma [7], or Oxidative phosphorylation (OXPHOS) [13].

**Current study**

Given the complexity of the TME environment and the multi-omics approach, this study tries to reduce complexity and aims to create a classification model that identifies TNBC status based on (predictive) biomarkers using RNA sequencing data. Given the background portrayed through this literature review, this study follows by setting out the methods, results, discussion and conclusion.

**3 Methods and Material**

**3.1 Overall Description of the Proposed Model**

Table 1

|  |  |
| --- | --- |
| Software used |  |
| anaconda-navigator | 2.6.6 |
| boruta\_py | 0.3 |
| matplotlib | 3.9.2 |
| numpy | 1.26.4 |
| pandas | 2.2.2 |
| python | 3.12.3 |
| scikit-learn | 1.5.1 |
| seaborn | 0.13.2 |

Making use of available literature, the approach used in this study follows the Fair, Accountable and Transparent [FAT] Flow framework [39] and inspiration from Hassan et al. [40]. A visual representation is shown in figure 1.

The elementary event that will be used as a classifier is TNBC status. The two possible values are True (TNBC) and False (nTNBC), which allow two complementary and disjoint equivalence classes [41].

Technology used is Python 3.12.3, Jupyter Notebook 7.2.2, both are packaged by Anaconda version 2.6.6. For a complete list of software and versions used see table 1.

Clinical data is filtered for TNBC status and used to get the RNA sequence data. Missing data and imbalance is addressed. Data is normalized using <xyz>. The featureset is split into a testset and validationset with a ratio of 8:2. Correlation is applied, and LASSO is used to do something extra.

Models used are SVM, Random Forest and Logistic Regression. <further substantiate the why>.

Lastly, analysis is applied <uitbreiden>.

**3.2 Data collection**

This study uses the TCGA-BRCA on the GDC portal (https://portal.gdc.cancer.gov/projects/TCGA-BRCA) as train- and testset. All data in this dataset are cancer samples. The number of cases [n] is 1097. Determining which cases have TNBC is based on the clinical data. Dimensions used in the clinical data to determine TNBC status is ‘er\_status\_by\_ihc', 'pr\_status\_by\_ihc' and 'her2\_status\_by\_ihc’. Cases missing data in any of the three dimensions are dropped. There remained 863 cases with TNBC and 116 cases with nTNBC.

Data from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) is also used, specifically GSE65216, GSE76250, GSE53752, GSE10893, as validationset. These datasets are also used in a study by Liu et. al. [20], to uniform the data the same techniques are applied. The genes in these datasets sometimes use synonyms for the names of genes used in TCGA-BRCA. To asses for the names the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/kegg2.html) and Human Protein Atlas (HPA) (https://www.proteinatlas.org) was consulted.

**3.3 Data PreProcessing**

Follow data cleaning practices [42], applying imputation using <method>. Deduplication is applied to the TCGA-BCRA dataset for genes (e.g. CD99), removing the duplicate that has zero sum as result. Outliers are managed by applying logarithm transformation [43]. To make features more comparable standardization is applied [43, 44] and remove the mean to center features on zero [44, 45].

**3.4 Feature Selection and Feature Extraction**

Given the heterogenous nature of TNBC and the lack of distinct biomarkers multiple approaches are used for feature selection, keeping in mind the heuristic “n / p > 5” [35]. To address imbalance in the dataset augmentation is applied using Synthetic Minority Oversampling Technique (SMOTE) [46, 47]. The SMOTE algorithm creates a better balance by synthesizing new samples from the minority class.

Two selections are based on literature. One selection is based on the list of twenty genes suggested by Kohari et al [48]. Another selection is based on assessing 65 open access research papers for insights into possible, relevant, biomarkers and oncogenes that are related to TNBC (results table / supplement). For example, very recent research suggests the LRPPRC gene as a distinct marker for TNBC [13].

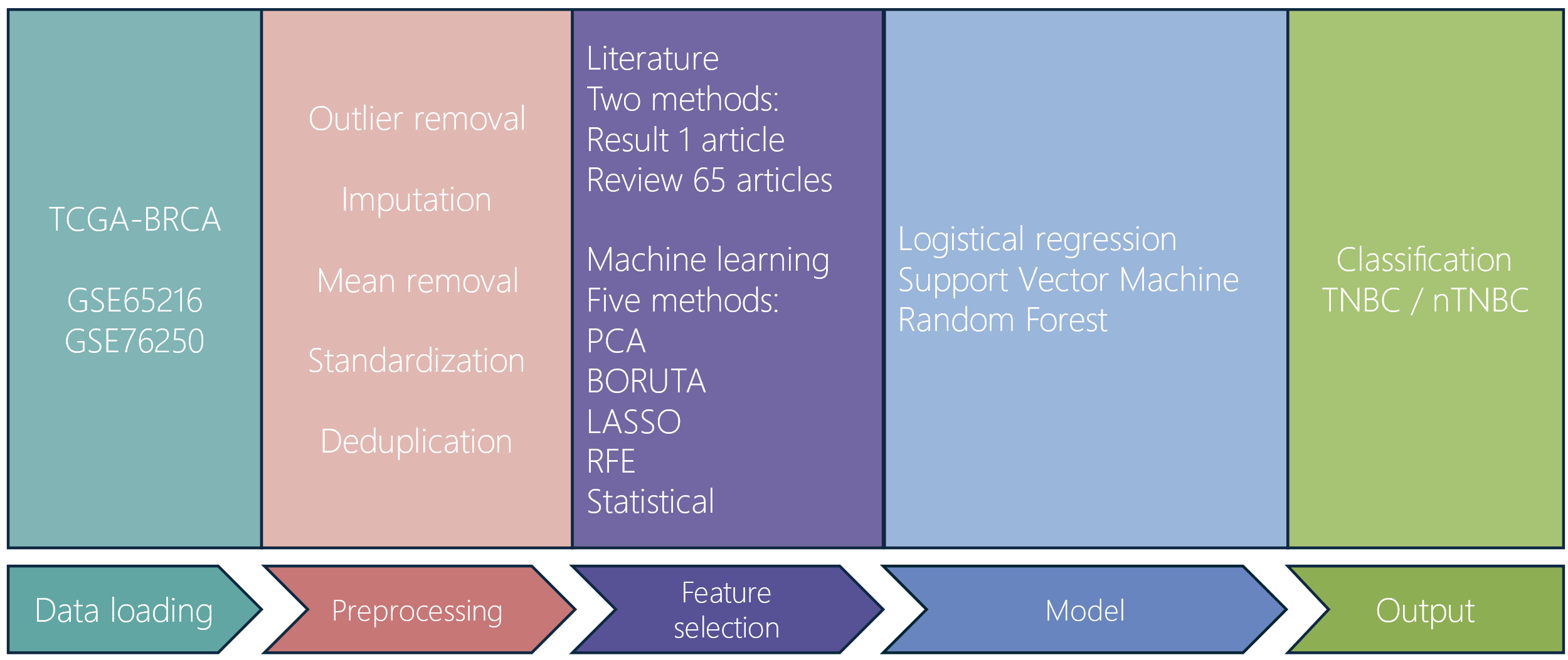
Using selection-based machine learning feature selection, this study employs <x> methods. Following practices in machine learning studies, this study uses the Boruta package as a way to select features [40, 49, 50]. The Boruta algorithm uses random forest classification algorithm and, to cite from the original paper, “classification is performed by voting of multiple unbiased weak classifiers […] [t]he importance measure of an attribute is obtained as the loss of accuracy of classification caused by the random permutation of attribute values between objects” [51]. To contrast the Random Forest used by Boruta, a Recursive Feature Elimination (RFE) with SVM is used. To cite from Artasanchez and Joshi, RFE is “a greedy optimization algorithm with a stated goal of finding the best performing feature subset […] features are then ranked based on the order of their elimination” [43]. Having used a Random Forest and a Support Vector Machine, an attempt at feature selection is made using Least Absolute Shrinkage Selection Operator (LASSO) as a linear model. Applying statistical methods for feature selection is done using VINCE APPROACH.

Figure 1 – Overview of framework used

A transformation based method is applied by using principal component analysis (PCA) to reduce dimensionality and extract features [52].

See table 2 for the resulting selected features.

Aanvullen deduplicaton + aanpassen FS

**3.5 Classification and Prediction Modelling**

Three models are used for classification. Logistic regression, random forest, and SVM [53].

The models, as with feature selection, are trained and tested with and without SMOTE.

To account for the class imbalance, datasets are stratified upon splitting.

Models validation uses 5-fold cross validation. To guarantee reproducibility, the random\_state parameter is consistently set to 42 across all models.

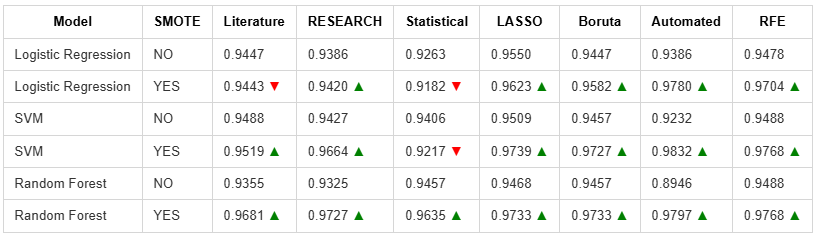


Figure n – Models Mean CV accuracy. Red and green arrows points the decrease/increase in accuracy between model and respective feature set comparing SMOTE to its non-SMOTE counterpart.

Logistic regression shows variability depending on the feature sets. By applying SMOTE, accuracy improves for most feature sets. Especially the Automated feature set shows a notably improved accuracy resulting in an accuracy of 0.9780 (with SMOTE) compared to 0.9386 (without SMOTE). However, the Literature and Statistical feature sets shows a (slight) decrease in accuracy from 0.9447 (without SMOTE) to 0.9443 (with SMOTE) for Literature and 0.9263 (without SMOTE) to 0.9182 (with SMOTE) for Statistical.

Random Forest shows great potential as it consistently yields high accuracy results across all feature sets. By applying SMOTE , accuracy improves for all feature sets. With the most noticeable being the Automated feature set. Improving from 0.8946 to a staggering 0.9797. The random forest has no additional parameters as input

As with Random Forest, Support Vector Machines also shows quite an improvement in accuracy when trained with SMOTE across almost all feature sets and again notably the Automated feature set benefits greatly from it. Improving from 0.9232 (without SMOTE) to 0.9832 (with SMOTE). Implementation for this study is with the parameter ‘probability’ set to True.

**3.6 Implementation**

Asfdsfsfsdfgs

**3.7 Evaluation**

To generate metrics, Python and Matplotlib were used. Specifically, Precision, Recall, and F-Measure [54], the ROC curve, AUC, accuracy, sensitivity, and F1 served as indicators for evaluation of the model.

In order to understand the models LIME and SHAP are used [35]. Both these algorithms can make the output of models more explainable.

**4 Results**

Hier komen vooral de grafiekjes met de resultaten van de verschillende aanpakken

**5 Discussion**

Onze reflectie op het resultaat.

**6 Conclusions**

In conclusion, the model developed using RNA sequencing data demonstrates <xyz>.

**6.1 Innovations**

Oftewel, hier komt ‘what can be researched next’

Page break after this

# References

[1] M. Lupi *et al.*, "TBC1 domain-containing proteins are frequently involved in triple-negative breast cancers in connection with the induction of a glycolytic phenotype," *Cell Death & Disease,* vol. 15, no. 9, p. 647, 2024/09/04 2024, doi: 10.1038/s41419-024-07037-2.

[2] X. Hu *et al.*, "Naphthoquinone-derived ZSW-4B induces apoptosis in triple-negative breast cancer via AMPK signalling activation," *Scientific Reports,* vol. 14, no. 1, p. 28559, 2024/11/19 2024, doi: 10.1038/s41598-024-79592-9.

[3] G. Wang *et al.*, "Integrating multi-omics data reveals the antitumor role and clinical benefits of gamma-delta T cells in triple-negative breast cancer," *BMC Cancer,* vol. 25, no. 1, p. 623, 2025/04/07 2025, doi: 10.1186/s12885-025-14029-8.

[4] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," (in eng), *CA Cancer J Clin,* vol. 71, no. 3, pp. 209-249, May 2021, doi: 10.3322/caac.21660.

[5] R. J. Pinto *et al.*, "Coding and regulatory somatic profiling of triple-negative breast cancer in Sub-Saharan African patients," *Scientific Reports,* vol. 15, no. 1, p. 10325, 2025/03/25 2025, doi: 10.1038/s41598-025-94707-6.

[6] M. Aine *et al.*, "The DNA methylation landscape of primary triple-negative breast cancer," *Nature Communications,* vol. 16, no. 1, p. 3041, 2025/03/28 2025, doi: 10.1038/s41467-025-58158-x.

[7] G. H. Tamarindo *et al.*, "Distinct proteomic profiles of plasma-derived extracellular vesicles in healthy, benign, and triple-negative breast cancer: candidate biomarkers for liquid biopsy," *Scientific Reports,* vol. 15, no. 1, p. 12122, 2025/04/09 2025, doi: 10.1038/s41598-025-95232-2.

[8] Z. Chen *et al.*, "Classifications of triple-negative breast cancer: insights and current therapeutic approaches," *Cell & Bioscience,* vol. 15, no. 1, p. 13, 2025/02/01 2025, doi: 10.1186/s13578-025-01359-0.

[9] R. S. Bouzid *et al.*, "Molecular subtyping and target identification in triple negative breast cancer through immunohistochemistry biomarkers," *BMC Cancer,* vol. 25, no. 1, pp. 1-16, 2025/03/13 2025, doi: 10.1186/s12885-025-13832-7.

[10] M. Matossian, N. Chen, and R. Nanda, "Exploiting Therapeutic Vulnerabilities in Triple-Negative Breast Cancer: Successes, Challenges, and Opportunities," *Current Breast Cancer Reports,* vol. 15, no. 3, pp. 266-278, 2023/09/01 2023, doi: 10.1007/s12609-023-00492-4.

[11] S. Roostee *et al.*, "Tumour immune characterisation of primary triple-negative breast cancer using automated image quantification of immunohistochemistry-stained immune cells," *Scientific Reports,* vol. 14, no. 1, p. 21417, 2024/09/13 2024, doi: 10.1038/s41598-024-72306-1.

[12] J. C. Martin *et al.*, "Aryl hydrocarbon receptor suppresses STING-mediated type I IFN expression in triple-negative breast cancer," *Scientific Reports,* vol. 14, no. 1, p. 5731, 2024/03/08 2024, doi: 10.1038/s41598-024-54732-3.

[13] Q. Xue *et al.*, "LRPPRC confers enhanced oxidative phosphorylation metabolism in triple-negative breast cancer and represents a therapeutic target," *Journal of Translational Medicine,* vol. 23, no. 1, pp. 1-15, 2025/03/25 2025, doi: 10.1186/s12967-024-05946-6.

[14] N. A. M. Bakker *et al.*, "Triple-negative breast cancer modifies the systemic immune landscape and alters neutrophil functionality," *npj Breast Cancer,* vol. 11, no. 1, p. 5, 2025/01/23 2025, doi: 10.1038/s41523-025-00721-2.

[15] Y.-C. Chien, J.-Y. Wu, L.-C. Liu, and Y.-L. Yu, "Capsanthin inhibits migration and reduces N-linked glycosylation of PD-L1 via the EZH2-PD-L1 axis in triple-negative breast cancer brain metastasis," *Cell Death Discovery,* vol. 11, no. 1, p. 85, 2025/03/04 2025, doi: 10.1038/s41420-025-02368-1.

[16] W. Ren *et al.*, "Comprehensive analysis of metabolism-related gene biomarkers reveals their impact on the diagnosis and prognosis of triple-negative breast cancer," *BMC Cancer,* vol. 25, no. 1, p. 668, 2025/04/11 2025, doi: 10.1186/s12885-025-14053-8.

[17] Y. Chen *et al.*, "Intratumoral microbiota-aided fusion radiomics model for predicting tumor response to neoadjuvant chemoimmunotherapy in triple-negative breast cancer," *Journal of Translational Medicine,* vol. 23, no. 1, p. 352, 2025/03/20 2025, doi: 10.1186/s12967-025-06369-7.

[18] X. Xiong, X. Wang, C.-C. Liu, Z.-M. Shao, and K.-D. Yu, "Deciphering breast cancer dynamics: insights from single-cell and spatial profiling in the multi-omics era," *Biomarker Research,* vol. 12, no. 1, p. 107, 2024/09/18 2024, doi: 10.1186/s40364-024-00654-1.

[19] H. Bischoff, M. Espié, and T. Petit, "Neoadjuvant Therapy: Current Landscape and Future Horizons for ER-Positive/HER2-Negative and Triple-Negative Early Breast Cancer," *Current Treatment Options in Oncology,* vol. 25, no. 9, pp. 1210-1224, 2024/09/01 2024, doi: 10.1007/s11864-024-01251-y.

[20] M. Liu, L. Zheng, Y. Zhang, and J. Tian, "Mechanistic insights into pachymic acid’s action on triple-negative breast Cancer through TOP2A targeting," *Scientific Reports,* vol. 15, no. 1, p. 2856, 2025/01/22 2025, doi: 10.1038/s41598-025-87286-z.

[21] B. Zhu *et al.*, "Injectable supramolecular hydrogel co-loading abemaciclib/NLG919 for neoadjuvant immunotherapy of triple-negative breast cancer," *Nature Communications,* vol. 16, no. 1, p. 687, 2025/01/15 2025, doi: 10.1038/s41467-025-55904-z.

[22] T. Vidalis, "Artificial Intelligence in Biomedicine: A Legal Insight," *BioTech*, vol. 10, no. 3*,* doi: 10.3390/biotech10030015.

[23] B. C. Stahl, D. Schroeder, and R. Rodrigues, "Right to Life, Liberty and Security of Persons," in *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*, B. C. Stahl, D. Schroeder, and R. Rodrigues Eds. Cham: Springer International Publishing, 2023, pp. 63-78.

[24] P. Tan *et al.*, "Application of omics technologies in studies on antitumor effects of Traditional Chinese Medicine," *Chinese Medicine,* vol. 19, no. 1, pp. 1-20, 2024/09/09 2024, doi: 10.1186/s13020-024-00995-x.

[25] M. M. O. Ortiz and E. R. Andrechek, "Molecular Characterization and Landscape of Breast cancer Models from a multi-omics Perspective," *Journal of Mammary Gland Biology and Neoplasia,* vol. 28, no. 1, pp. 1-13, 2023/06/03 2023, doi: 10.1007/s10911-023-09540-2.

[26] L. Cheng *et al.*, "MoAGL-SA: a multi-omics adaptive integration method with graph learning and self attention for cancer subtype classification," *BMC Bioinformatics,* vol. 25, no. 1, p. 364, 2024/11/23 2024, doi: 10.1186/s12859-024-05989-y.

[27] Z. Guo *et al.*, "Tumor microenvironment and immunotherapy for triple-negative breast cancer," *Biomarker Research,* vol. 12, no. 1, p. 166, 2024/12/31 2024, doi: 10.1186/s40364-024-00714-6.

[28] J. Wu, J. Li, H. Xu, N. Qiu, X. Huang, and H. Li, "Periostin drives extracellular matrix degradation, stemness, and chemoresistance by activating the MAPK/ERK signaling pathway in triple–negative breast cancer cells," *Lipids in Health and Disease,* vol. 22, no. 1, pp. 1-14, 2023/09/16 2023, doi: 10.1186/s12944-023-01912-1.

[29] Y. Liu *et al.*, "Advances in immunotherapy for triple-negative breast cancer," (in eng), *Mol Cancer,* vol. 22, no. 1, p. 145, Sep 2 2023, doi: 10.1186/s12943-023-01850-7.

[30] L. Weng, J. Zhou, S. Guo, N. Xu, and R. Ma, "The molecular subtyping and precision medicine in triple-negative breast cancer---based on Fudan TNBC classification," *Cancer Cell International,* vol. 24, no. 1, p. 120, 2024/03/30 2024, doi: 10.1186/s12935-024-03261-0.

[31] Y. Huang, P. Zeng, and C. Zhong, "Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning," *BMC Bioinformatics,* vol. 25, no. 1, pp. 1-19, 2024/03/27 2024, doi: 10.1186/s12859-024-05749-y.

[32] K. Asleh, N. Riaz, and T. O. Nielsen, "Heterogeneity of triple negative breast cancer: Current advances in subtyping and treatment implications," *Journal of Experimental & Clinical Cancer Research,* vol. 41, no. 1, pp. 1-26, 2022/09/01 2022, doi: 10.1186/s13046-022-02476-1.

[33] X. Liu *et al.*, "Spatial multi-omics: deciphering technological landscape of integration of multi-omics and its applications," *Journal of Hematology & Oncology,* vol. 17, no. 1, pp. 1-24, 2024/08/24 2024, doi: 10.1186/s13045-024-01596-9.

[34] E. Brombacher, O. Schilling, and C. Kreutz, "Characterizing the omics landscape based on 10,000+ datasets," *Scientific Reports,* vol. 15, no. 1, p. 3189, 2025/01/25 2025, doi: 10.1038/s41598-025-87256-5.

[35] A. Morabito, G. De Simone, R. Pastorelli, L. Brunelli, and M. Ferrario, "Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review," *Journal of Translational Medicine,* vol. 23, no. 1, pp. 1-26, 2025/04/10 2025, doi: 10.1186/s12967-025-06446-x.

[36] J. Qian *et al.*, "Identification and characterization of cell niches in tissue from spatial omics data at single-cell resolution," *Nature Communications,* vol. 16, no. 1, p. 1693, 2025/02/16 2025, doi: 10.1038/s41467-025-57029-9.

[37] S. Lee, G. Kim, J. Lee, A. C. Lee, and S. Kwon, "Mapping cancer biology in space: applications and perspectives on spatial omics for oncology," *Molecular Cancer,* vol. 23, no. 1, pp. 1-27, 2024/01/30 2024, doi: 10.1186/s12943-024-01941-z.

[38] W. Chen *et al.*, "Construction of the bromodomain-containing protein-associated prognostic model in triple-negative breast cancer," *Cancer Cell International,* vol. 25, no. 1, p. 18, 2025/01/18 2025, doi: 10.1186/s12935-025-03648-7.

[39] D. Martens, *Data Science Ethics: Concepts, Techniques, and Cautionary Tales*: Oxford University Press, 2022. [Online]. Available: <https://doi.org/10.1093/oso/9780192847263.001.0001>.

[40] A. M. Hassan, S. M. Naeem, M. A. A. Eldosoky, and M. S. Mabrouk, "Multi-omics-based Machine Learning for the Subtype Classification of Breast Cancer," *Arabian Journal for Science and Engineering,* vol. 50, no. 2, pp. 1339-1352, 2025/01/01 2025, doi: 10.1007/s13369-024-09341-7.

[41] J. J. Cuadrado-Gallego and Y. Demchenko, "Supervised Classification," in *Data Analytics: A Theoretical and Practical View from the EDISON Project*. Cham: Springer International Publishing, 2023, pp. 335-404.

[42] M. Plaue, "Elements of data organization," in *Data Science: An Introduction to Statistics and Machine Learning*, M. Plaue Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, pp. 11-34.

[43] A. A. P. Joshi, "Feature Selection and Feature Engineering," in *Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3.x , Second Edition*. Birmingham: Packt Publishing Ltd., 2020.

[44] M. Plaue, "Multivariate statistics," in *Data Science: An Introduction to Statistics and Machine Learning*, M. Plaue Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, pp. 161-181.

[45] A. A. P. Joshi, "Classification and Regression Using Supervised Learning," in *Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3.x , Second Edition*. Birmingham: Packt Publishing Ltd., 2020.

[46] X. Song *et al.*, "Use of ultrasound imaging Omics in predicting molecular typing and assessing the risk of postoperative recurrence in breast cancer," *BMC Women's Health,* vol. 24, no. 1, p. 380, 2024/07/02 2024, doi: 10.1186/s12905-024-03231-8.

[47] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.,* vol. 18, no. 1, pp. 559–563, 2017.

[48] C. Kothari *et al.*, "Machine learning analysis identifies genes differentiating triple negative breast cancers," *Scientific Reports,* vol. 10, no. 1, p. 10464, 2020, doi: 10.1038/s41598-020-67525-1.

[49] Y. Li, X. Wu, D. Fang, and Y. Luo, "Informing immunotherapy with multi-omics driven machine learning," *npj Digital Medicine,* vol. 7, no. 1, p. 67, 2024/03/14 2024, doi: 10.1038/s41746-024-01043-6.

[50] S. Li *et al.*, "Artificial intelligence learning landscape of triple-negative breast cancer uncovers new opportunities for enhancing outcomes and immunotherapy responses," *Journal of Big Data,* vol. 10, no. 1, p. 132, 2023/08/26 2023, doi: 10.1186/s40537-023-00809-1.

[51] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software,* vol. 36, no. 11, pp. 1 - 13, 09/16 2010, doi: 10.18637/jss.v036.i11.

[52] M. Plaue, "Unsupervised machine learning," in *Data Science: An Introduction to Statistics and Machine Learning*, M. Plaue Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, pp. 249-276.

[53] U. Qamar and M. S. Raza, "Classification," in *Data Science Concepts and Techniques with Applications*, U. Qamar and M. S. Raza Eds. Cham: Springer International Publishing, 2023, pp. 127-166.

[54] Z. Qi, H. Wang, and Z. Dong, "Impacts of Dirty Data on Classification and Clustering Models," in *Dirty Data Processing for Machine Learning*, Z. Qi, H. Wang, and Z. Dong Eds. Singapore: Springer Nature Singapore, 2024, pp. 7-37.