AI in Healthcare

Capstone Project Proposal

Detection of Triple Negative Breast Cancer from RNA sequencing data using Machine Learning

Ruben Holthuijsen (1064459)

Kevin Hartman (1044032)

Sander van Swieten (1063788)

Vince van Doorn (1061669)

Victor de Sousa Gama (0929470)

Group 1

12 May 2025

Revision 14

# Abstract

This proposal addresses the challenge of diagnosing triple negative breast cancer (TNBC). 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 focus lies on 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, and which will be useful in cases where RNA sequencing data is available. All code, visualizations, and documentation will be made openly available to support reproducibility and future use.

# 1. Introduction

Breast cancer is one of the most common cancers in the world, and is a major cause of death for many people, in particular among (young) women [1]. In 2022 female breast cancer was the second most common cancer with 2.3 million cases [2, 3]. As with all cancers, early and accurate diagnosis and classification are important for providing effective treatment. This project focuses on triple negative breast cancer (TNBC), one of the most aggressive and difficult to treat types of breast cancer due to lack of molecular targets [4]. About 15% of breast cancers are TNBC and these have a significantly worse prognosis than other types of breast cancer [5, 6].

A lack of expression of three hormone receptors defines TNBC: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) [7]. Because of this lack of expression, TNBC does not respond to targeted hormone therapies, which limits treatment options and leads to a poorer prognosis for patients.

The typing of TNBC done by Lehmann in 2011 [8], and cited recently [9-11], creating subtypes of TNBC, specifically basal-like 1 (BL1), basal-like 2 (BL2), mesenchymal-like (MES), mesenchymal/stem-like (MSL), immunomodulatory (IM), and luminal androgen receptor (LAR) [Fig. 1]. The subtypes have their own treatment modality. As ‘Lehmann typing’ is considered homogenous, recent TNBC typing is done using ‘Fudan typing’ [12, 13] to indicate the aforementioned subtypes. Another option for profiling for subtypes is ‘PAM50 intrinsic molecular sub­types’ [14]. For a recent extensive review on subtyping see the work of Asleh et al. [15].

Currently, diagnosis of TNBC is most commonly performed using immunohistochemistry (IHC), a technique that identifies the presence or absence of ER, PR and HER2 from a tissue sample. Using IHC is considered the current clinical standard, although some contend there is no reliable biomarker [16]. IHC is time consuming, depends on a skilled operator, and not always available. Therefore, there is a need for additional diagnostic methods, especially in terms of speed, objectivity and reproducibility.

In the past years, machine learning in particular has emerged as a powerful tool in research and diagnostics, especially in very large datasets such as with genomic data obtained from RNA sequencing. Recent improvements in this area have shown a lot of promise for new possibilities of cancer classification. However, most research on this topic often focuses on TNBC subtypes or uses commercial tools that lack transparency raising ethical concerns, or uses large amounts of additional data that leads to complex models that lack explainability and might not be practical in all settings.

|  |
| --- |
| A diagram of a cell cycle  AI-generated content may be incorrect. |
| **Fig. 1** Classification and therapeutic options for TNBC. ADC: antibody‒drug conjugates; AR: androgen receptor; LAR: luminal androgen receptor; M: mesenchymal; MSL: mesenchymal-stem-like; PARP: poly-adenosine diphosphate ribose polymerase; PI3K: phosphoinositol-3 kinase; TKI: tyrosine kinase inhibitor; and TNBC: triple-negative breast cancer **Source:** Fig. 1 in [17] |

Although RNA sequencing provides a rich source of data for classification tasks, it also has its challenges, such as the high dimensionality of gene expression profiles, the biological heterogeneity among patients, and class imbalance in datasets, which leads to both practical and ethical concerns in its use.

Following ethical principles is not just desirable but essential, especially in a healthcare setting. 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.

As such, there seems to be a gap in the creation of a simple, explainable and reproducible machine learning model that can classify TNBC status binary using RNA sequencing data alone. It is this gap that the project needs to fill.

The idea is to build a binary classifier for TNBC status, trained on RNA sequencing data from The Cancer Genome Atlas (TCGA), more specifically the Breast Invasive Carcinoma Collection (TCGA-BRCA). The goal is to explore whether gene expression data is by itself sufficient to accurately identify TNBC status, and to do so in a way that emphasizes ethical principles such as transparency, reproducibility and algorithmic fairness. This model has the potential to be an additional diagnostic tool in both research and clinical settings.

# 2. Literature Review

Having reviewed existing literature on TNBC, whilst being aware of recent developments using multi-omics [18-20], this review has a focus on studies that use RNA sequencing and/or machine learning for diagnosis or classification.

Dass et al. [21] extensively reviewed current and potential future methods of diagnosing TNBC. They showed that using IHC is the current clinical standard for diagnosis, despite its limitations [22], for identifying ER, PR and HER2 status, and that these are what determines TNBC status. However, they noted that this method of diagnosis is time consuming and depends on a skilled operator. They emphasized that there is a need for faster and more objective technologies for diagnosis of TNBC.

One of the methods reviewed in Dass et al. was the *nCounter® Breast Cancer 360™ Panel* [23], a commercial tool that uses RNA expression levels from 770 genes to classify breast cancer. Although this tool looks promising in using RNA for classification, further research shows no mention of employing machine learning, and it being a commercial tool without any focus on explainability limits its usability for research purposes.

Further searching for TNBC and machine learning brought up Kothari et al. [24], who have also used the TCGA-BRCA dataset for TNBC classification. Their primary goal was identifying which genes are highly correlated with TNBC status and classification of subtypes, with a special focus on prognosis and survival. They have found 20 genes with strong potential but emphasized the need for further research. In addition to RNA sequencing data, they used methylation and miRNA data as well (leading to them using a smaller cohort from the TCGA-BRCA dataset due to limited availability of that data), but found this was not useful for predictions.

Looking more into TNBC vs. non-TNBC classification, Davis et al. [25] published a review of the genomic characteristics of TNBC cases, and noted that many subtypes of TNBC are very close to subtypes of non-TNBC, showing that classification is complex and that there is a lot of overlap on the gene level between TNBC and non-TNBC. Their review focuses on the implications of this for the development of targeted therapies for different subtypes of TNBC, but does not mention machine learning.

Finally, Bissanum et al. [26] have researched the molecular classification of TNBC, also noting that it is heterogeneous and challenging to treat. They mention that there is currently no established way of classifying subtypes of TNBC and have used gene expression analysis to develop a classification method using machine learning. However, their focus was on classifying cases that were already known to be TNBC into subtypes, rather than the classification of TNBC versus non-TNBC.

From this literature review, it becomes apparent that RNA sequencing and machine learning have been used in the classification of TNBC and breast cancer in general, but there remains a gap in the development of a simple, transparent and reproducible classification model that focuses only on identifying TNBC status (yes or no) using RNA sequencing data alone. Previous studies have often focused on subtype classification, relied on additional data that might not have been predictive, or referred to commercial tools that have unclear ethics and lack transparency. To address this gap, the solution would be to train a machine learning model using publicly available RNA sequencing data. In particular the work of Kothari et al. [24] who have identified promising genes to use as features.

# 3. Problem Statement and Research Questions

Triple negative breast cancer (TNBC) is an aggressive type of breast cancer that does not respond to common treatment options such as hormone therapy, which results in poor prognosis for patients. The diagnosis of TNBC is most commonly done by immunohistochemistry (IHC), a method that is time consuming, requires interpretation by an expert, and is not always accessible. RNA sequencing, combined with machine learning, seems to be a promising alternative for more scalable and objective diagnosis.

To guide the research, the following research questions have been made:

1. How accurately can TNBC status be predicted from RNA sequencing data using a machine learning model?
2. Which gene features are most useful for training a machine learning model to distinguish TNBC from non-TNBC cases?
3. What type of machine learning algorithm gives highly accurate results for building a TNBC detection model?
4. To what extent can the machine learning model meet important ethical principles such as explainability, fairness and reproducibility?

These questions will guide the technical development of the model and ensure that the focus lies on both model performance and ethical considerations throughout the project.

# 4. Aim and Objectives

The aim is to develop a machine learning model that can improve the accuracy of triple negative breast cancer (TNBC) status classification using only RNA sequencing data, while adhering to ethical principles.

Connected to the research questions above, assuming a project runtime of 6 weeks (12 May to 22 June), the following SMART objectives are defined:

1. *Model setup and first evaluation:* Prepare the data and set up a model for binary TNBC classification using a subset of genes that were identified as promising in existing literature and perform a basic evaluation of its performance by the end of week 2 (25 May).
2. *Feature testing and selection:* Test with different genes and different feature selection strategies, such as literature based, statistic filtering and model selection, then evaluate and compare the results and decide what features to use by the end of week 3 (1 June).
3. *Algorithm comparison and final selection:* Compare at least three machine learning algorithms using standard performance metrics and validation methods and determine the most suitable one for this project by the end of week 4 (8 June).
4. *Ethical evaluation and report writing:* Research and document the model’s explainability, fairness and bias, and reproducibility, and have a first draft of the final report by the end of week 5 (15 June).

The final week for finalizing the report will be used as a buffer week to aid in potential unforeseen challenges such as delays of (additional) data access, team members’ availability, or unexpected technical difficulties. This period allows for additional testing, revisions, or troubleshooting, making sure that the final deliverables can be completed and submitted on time without sacrificing quality.

# 5. Significance of the Project

As mentioned earlier in the document, triple negative breast cancer (TNBC) is a challenge in both diagnosis and treatment as it does not respond well to common therapies due lack of related receptor markers [27]. Currently, diagnosis relies heavily on immunohistochemistry (IHC), which is time-consuming, requires expert interpretation, and is not always available. This shows the need for simpler, more accessible diagnostic tools.

The goal of this project is to develop a machine learning model that can classify TNBC status using only RNA sequencing data. As RNA sequencing is widely used in research, this project is particularly useful for cases where this data is already available. By focusing on a binary (yes/no) classification, the solution avoids model complexity that comes with subtype classification, making it as simple and straightforward as possible to implement and interpret.

The creation of a straightforward, accessible diagnostic tool, which can be easily integrated into existing research workflows, is what makes this project significant. The simplicity of a ‘yes’ or ‘no’ classification makes it feasible to use for both research and clinical settings, where a fast and accurate decision is very important. Our focus on a transparent and reproducible machine learning model, with interpretable and reliable results, plays well with the increasing interest in ethical AI tools in healthcare research.

In summary, this project has the potential to make a meaningful contribution to TNBC diagnosis by providing a simple yet effective machine learning model that can support research and, potentially, even clinical diagnostics, especially where RNA sequencing data is already available.

# 6. Methodology

The following is an outline of the technical workflow of the project, with details on the tools, techniques and strategies that will be used to develop a transparent and reproducible machine learning model for classifying TNBC using RNA sequencing data.

## 6.1 Project Workflow

The project will use the following basic workflow, based on the needs for using the TCGA-BRCA dataset and the goal of building a binary classifier using machine learning:

1. *Data acquisition:* Data will be obtained from the Genetic Data Commons (GDC)
2. *Data preprocessing:* This includes cleaning, filtering, and label generation (TNBC/non-TNBC)
3. *Feature selection:* Based on literature and statistical methods
4. *Model development:* Training and testing multiple machine learning algorithms
5. *Model evaluation:* Using performance metrics and validation techniques such as cross-validation
6. *Interpretability and fairness analysis:* Analyzing the generated models for transparency and validity with different demographic groups
7. *Documentation and reporting*: All methods and findings will be described in the final report

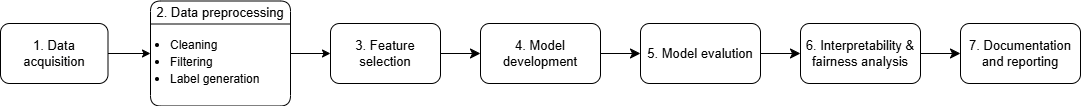
****

Fig. 2 Workflow

## 6.2 Tools

The project will be built in Python, using the following tools and libraries:

* *Jupyter Notebooks:* Interactive development and technical documentation
* *Numpy and Pandas:* Data manipulation, preprocessing, and analysis
* *Scikit-learn:* Model building, feature selection and model evaluation
* *Matplotlib and Seaborn:* Visualization of the findings
* *SHAP and/or ELI5 and/or LIME* [28]*:* To explore model explainability, depending on how feasible integration is
* *Git (via GitHub):* Version control and collaboration

All experiments will be documented in a Jupyter Notebook to ensure reproducibility, and use version control in Git to track changes during development. All project files and history will be available on GitHub.

If time permits, research of relevance for additional tools like XGBoost to explore alternatives to the machine learning algorithms available in Scikit Learn might be considered.

## 6.3 Data acquisition and preprocessing

As mentioned before, the TCGA-BRCA dataset will be used, obtained through the Genomic Data Commons (GDC) portal. To make the data usable, the following preprocessing steps have been decided:

1. Joining of the RNA and clinical datasets using patient UIDs
2. Labeling and filtering patients with known TNBC status
3. Normalizing gene expression values
4. Applying data balancing techniques, such as oversampling of TNBC cases

The first two step have already been performed in preparation of this proposal. See the section on *Dataset* below for our findings.

## 6.4 Feature selection

Feature selection is very important due to the high dimensionality of gene expression data. To mitigate issues that might arise due to a small number of samples *n* and high number of features *p*, the rule of thumb suggested by Morabito et al. that *n/p* should be equal or bigger than five is considered during feature selection in order to minimize chances on overfitting and maximize statistical power [29].

To achieve a usable selection of features, usage of existing literature (such as Kothari et al. [24]) and our own selection using statistical or model-based methods, such as looking at the variance of features. Additionally considering dimensionality reduction techniques such as Principal Component Analysis (PCA), although there are potential implications for explainability.

## 6.5 Model selection and development

Of at least three common machine learning algorithms the performance is compared. The primary candidates are:

* *Logistic Regression:* due to its interpretability and providing a baseline
* *Random forest:* known for robustness and built-in feature importance
* *Support Vector Machine (SVM):* for its support for high-dimensional binary classification, potentially allowing a wider selection of features

Testing additional models, for example XGBoost, if time and performance permits, are considered.

## 6.6 Performance evaluation

The primary performance evaluation will be done by applying cross validation with stratified sampling to preserve class imbalance within folds. The most important metrics in evaluation are:

* Accuracy
* Precision and Recall: especially relevant due to the dataset’s class imbalance as TNBC is under-represented
* F1 score: mean of precision and recall
* ROC / AUC

Visualization of a confusion matrix will be added for each model to obtain additional insight, which will assist in selecting the most suitable model in the end.

## 6.7 Explainability and ethical evaluation

In addition to model performance, the model is evaluated on fairness and explainability, to make sure it functions in an ethical way. This includes looking into tools for feature importance analysis (such as SHAP or LIME values)to understand how the model gets its results. Additionally, by evaluating model performance across different subgroups (e.g. by sex or race, as are available from clinical data) to check for any potential bias, the fairness will be assessed. All decisions made based on our findings for transparency and reproducibility will be documented.

# 7. Dataset

The dataset used is from The Cancer Genome Atlas (TCGA), specifically the Breast Invasive Carcinoma Collection (TCGA-BRCA) [30-32]. This project provides RNA sequencing data for 1095 patients and clinical data for 1097 patients. The clinical data includes the results from immunohistochemistry (IHC) tests from which a patient’s TNBC status can be derived. It also includes metadata such as age, sex and ethnicity.

Limitations are the relatively small size of the dataset, as well as missing values that make it impossible to determine some patients’ TNBC status. With further exploration of the data, a negative TNBC status (non-TNBC) can be derived for 863 patients (having a positive IHC test result for at least one of ER, PR and HER2), while positive TNBC status can be derived for 116 patients (having a negative IHC test result for all of ER, PR and HER2). As the clinical data file contains data for 1097 patients in total, this means that there are 118 patients for whom TNBC status cannot be determined.

Additionally, the dataset has 1094 patients that can be matched to the available RNA sequencing data files, leaving three patients without RNA data: one each from the TNBC, non-TNBC and unknown groups.

In the dataset there are 977 usable sets of patient data, of whom 115 are known to be TNBC and 862 are known to be non-TNBC. This imbalance may in part be caused by the fact that for the TNBC group, measurements for all of ER, PR and HER2 need to be available and negative. By contrast, for the non-TNBC group, only one of the measurements needs to be available and positive. This is a known limitation of the dataset.

# 8. Ethics

For this project, both the RNA sequencing and clinical data from the TCGA-BRCA dataset will be used. As this is highly personal healthcare data, it raises various ethical questions. Ethical issues regarding privacy, fairness and bias, and explainability are addressed to ensure the safe and ethical usability of the model especially in the context of its potential deployment in research and clinical settings.

## 8.1 Privacy

The dataset used in this project contains sensitive patient information, making privacy a primary ethical concern. Although all TCGA data is de-identified and anonymized, it is still essential to handle data responsibly throughout the project. Adhering to best practices the study does not store source data publicly or redistributing the dataset. Instead, the data source and its use (ie. how it was accessed) will be documented. This practice also makes sure that the project is fully reproducible without duplicating the data itself. This approach allows for potential correction or deletion of data by TCGA, for example in the case of errors being found or if a patient revokes their consent. By treating even anonymized data as sensitive the study aims to uphold the highest standards of privacy.

## 8.2 Fairness and Bias

In machine learning models, bias can be introduced through unbalanced data, systemic inequalities present in the dataset, and improper feature selection. The dataset used has an imbalance between TNBC cases (115) and non-TNBC cases (862), which could potentially lead to a biased model. To mitigate this, strategies like oversampling the TNBC cases or adjusting model weights are used to ensure the model does not favour one class over the other. Furthermore, feature selection will be fully limited to RNA sequencing data. Clinical metadata such as age, sex and race will not be selected as features, but will be used to select different slices of testing data when validating the model, to ensure the model’s performance does not vary unfairly within different groups.

## 8.3 Explainability

In healthcare, one of the most important ethical considerations is the need for explainability. The black-box nature of some machine learning models can undermine the trust of doctors and patients. To address this, this study builds a model that is as interpretable as possible, using explainable machine learning techniques such as feature importance analysis and existing explainability tools. These methods should help make the model’s predictions more transparent and easier to understand by both doctors and patients, to support informed decision making.

# 9. Plan, Team Collaboration, and Contribution of Each Member

The project is split into three phases; preparation, feature and model selection, and ethical evaluation and final report. Each phase has two weeks lead time as the duration of the project is six weeks (12 May - 22 June). Work is managed using GitHub for the deliverables, Microsoft Teams for communication, and our weekly meetings at school for discussing project status, dependencies, and impediments. In each phase tasks are divided between team members as follows, where the name signifies the ‘leader’ for that task

## 9.1 Phase 1: Preparation (deadline: 25 May)

In the preparation phase, development environment is prepared, the data obtained, and an initial model is trained (including basic validation). The result of this phase is an initial model (and all necessary script to build it) that is built upon in different ways:

* ***Ruben:***Obtain and pre-process RNA sequencing and clinical data (including generation of TNBC label) and setting up the GitHub repository with the initial structure and documentation of this process. Also, ensuring all participants can reproduce and access the data in the same way.
* ***Sander:*** Setting up the development environment using Jupyter, creating an initial notebook that can load the data that was obtained, ensure the notebook can run locally.
* ***Vince:*** Perform initial exploratory data analysis to understand the RNA sequencing files and creating scripts for filtering features. Select genes that were identified as promising in the literature.
* ***Victor:*** Implement a basic machine learning model using the selected features.
* ***Kevin:*** Evaluate the generated baseline model using common metrics (such as accuracy, precision, recall and F1 score), perform cross validation and visualize and document the results.

Many of these tasks are (necessarily) sequential. The study is a shared responsibility; if impediments occur team members are expected to step in and help. To ensure delays are prevented as much as possible, an additional (tentative) deadline is set for the first three tasks at the end of week one (18 May), depending on when this proposal is accepted.

## 9.2 Phase 2: Feature and Model selection (deadline: 8 June)

In this phase, the most useful features and model algorithms are identified. As a result, most tasks in this phase can be done concurrently. However, there is an additional (tentative) deadline of 1 June for the first two tasks, involving feature selection (see also the SMART goals as defined in section 4, ‘Aim and Objectives’):

* ***Vince:*** Applying statistical techniques such as correlation analysis and filtering by variance to identify candidate features (genes in the RNA sequencing data) for TNBC classification. In the second week, the results (if they are promising) will be combined with the third task.
* ***Ruben:*** Exploring automated methods of dimensionality reduction to identify features in an alternative way. The third task can also build on this in the second week.
* ***Victor:*** Training different types of algorithms, initially using the features selected based on literature, but in such a way that features found through other methods can be easily swapped out. Features will be re-selected in the second week, based on the findings of the first week.
* ***Sander:*** Comparing the results from different feature selection methods (literature, statistical, automated), and different machine learning algorithms, as provided by the work of the first phase and the other tasks. Creating visualizations of the resulting metrics.
* ***Kevin:*** Creating an initial draft of the final report, writing out the methods and findings so far, including preprocessing steps, feature and model selection, and using the visualizations that have been created.

Once this phase is completed, the result is multiple models trained using different sets of features, having a general idea of the most promising features and model, and having a first draft of the final report.

## 9.3 Phase 3: Ethical evaluation and Reporting of findings (deadline: 22 June)

In the final phase, the most promising model is (or models are) fine-tuned, evaluate the fairness and bias, and explainability, and present all findings to conclude the project. This phase has a first, intermediate, deadline of 15 June in order to have a week to reflect on all work done in the study:

* ***Vince:*** Applying hyperparameter tuning to fine-tune the most promising model (or models), based on the findings from phase two, to further improve its performance.
* ***Sander:*** Analyzing the model’s interpretability and determining which features are most influential in predictions.
* ***Ruben:*** Evaluating fairness and generalization of the fine-tuned model by re-training it based on different subgroups (such as sex and race), documenting any disparities or bias.
* ***Kevin:*** Writing of the final report, combining input from all members and ensuring a coherent structure.
* ***Victor:*** Design additional visual materials for the final report and design a presentation to report on the findings of this study.

Across all phases, all members will participate in discussions (weekly in school) and help make major decisions such as feature and model selection, and the contents of the final report and presentation. Team members are responsible for their tasks, the sharing of created content (such as Jupyter notebooks), and ensuring others can reproduce the results locally. By collaborating this way, it is expected the result has an equal contribution from all team members while ensuring reproducibility of all work.

# 10. Deliverable

At the end of the project, a written report is delivered that includes methodology, results, and conclusions, fitting with the objectives outlined in this proposal. All related materials, including source code, Jupyter notebooks, visualizations, and intermediate results, will be made available on Github via the following link: <https://www.github.com/rhlt/tnbc>

In keeping with the ethical principles outlined here, this repository will not include the raw data or any trained models, but will include all materials needed for full reproducibility, including the final report and final presentation. The presentation is expected to be completed after the main project deadline of June 22. At the latest, it will be added to the repository on the date the final presentation is given.

# References

[1] S. Zhu *et al.*, "Recent advances in targeted strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 16, no. 1, p. 1d, 2023/08/28 2023, doi: 10.1186/s13045-023-01497-3.

[2] W. H. Organization. "Global cancer burden growing, amidst mounting need for services." <https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services> (accessed 05/10/2025, 2025).

[3] 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.

[4] R. S. Bouzid *et al.*, "Molecular subtyping and target identification in triple negative breast cancer through immunohistochemistry biomarkers," *BMC Cancer,* vol. 25, no. 1, p. 2, 2025/03/13 2025, doi: 10.1186/s12885-025-13832-7.

[5] A. C. Association. "Triple-negative Breast Cancer." <https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/triple-negative.html> (accessed 05/10/2025, 2025).

[6] 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, p. 2, 2025/03/25 2025, doi: 10.1186/s12967-024-05946-6.

[7] 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, p. 1, 2023/09/01 2023, doi: 10.1007/s12609-023-00492-4.

[8] B. D. Lehmann *et al.*, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *The Journal of clinical investigation,* vol. 121, no. 7, pp. 2750-2767, 2011.

[9] S. Zhu *et al.*, "Recent advances in targeted strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 16, no. 1, p. 2, 2023/08/28 2023, doi: 10.1186/s13045-023-01497-3.

[10] Y. Li *et al.*, "Recent advances in therapeutic strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 15, no. 1, p. 2, 2022/08/29 2022, doi: 10.1186/s13045-022-01341-0.

[11] 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.

[12] Y. Z. Jiang *et al.*, "Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies," (in eng), *Cancer Cell,* vol. 35, no. 3, pp. 428-440.e5, Mar 18 2019, doi: 10.1016/j.ccell.2019.02.001.

[13] Y. R. Liu *et al.*, "Comprehensive transcriptome analysis identifies novel molecular subtypes and subtype-specific RNAs of triple-negative breast cancer," (in eng), *Breast Cancer Res,* vol. 18, no. 1, p. 33, Mar 15 2016, doi: 10.1186/s13058-016-0690-8.

[14] 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, p. 3, 2023/06/03 2023, doi: 10.1007/s10911-023-09540-2.

[15] 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.

[16] 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, p. 11, 2025/03/25 2025, doi: 10.1186/s12967-024-05946-6.

[17] Y. Li *et al.*, "Recent advances in therapeutic strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 15, no. 1, p. 3, 2022/08/29 2022, doi: 10.1186/s13045-022-01341-0.

[18] 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.

[19] A. Jose *et al.*, "Integration of pan-omics technologies and three-dimensional in vitro tumor models: an approach toward drug discovery and precision medicine," *Molecular Cancer,* vol. 23, no. 1, pp. 1-24, 2024/03/09 2024, doi: 10.1186/s12943-023-01916-6.

[20] 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.

[21] S. A. Dass *et al.*, "Triple Negative Breast Cancer: A Review of Present and Future Diagnostic Modalities," *Medicina,* vol. 57, no. 1, p. 62, 2021. [Online]. Available: <https://www.mdpi.com/1648-9144/57/1/62>.

[22] 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, p. 19, 2022/09/01 2022, doi: 10.1186/s13046-022-02476-1.

[23] "nCounter® Breast Cancer 360™ Panel." NanoString Technologies. <https://www.nanostring.com/products/gene-expression-panels/gene-expression-panels-overview/ncounter-breast-cancer-360-panel> (accessed 16 April, 2025).

[24] 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.

[25] S. L. Davis, S. G. Eckhardt, J. J. Tentler, and J. R. Diamond, "Triple-negative breast cancer: bridging the gap from cancer genomics to predictive biomarkers," *Therapeutic Advances in Medical Oncology,* vol. 6, no. 3, pp. 88-100, 2014, doi: 10.1177/1758834013519843.

[26] R. Bissanum, S. Chaichulee, R. Kamolphiwong, R. Navakanitworakul, and K. Kanokwiroon, "Molecular Classification Models for Triple Negative Breast Cancer Subtype Using Machine Learning," *J Pers Med,* vol. 11, no. 9, Sep 1 2021, doi: 10.3390/jpm11090881.

[27] Y. Li *et al.*, "Recent advances in therapeutic strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 15, no. 1, p. 1, 2022/08/29 2022, doi: 10.1186/s13045-022-01341-0.

[28] 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, p. 21, 2025/04/10 2025, doi: 10.1186/s12967-025-06446-x.

[29] 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, p. 20, 2025/04/10 2025, doi: 10.1186/s12967-025-06446-x.

[30] W. Lingle *et al.* *The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA)*, doi: 10.7937/K9/TCIA.2016.AB2NAZRP.

[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, p. 5, 2024/03/27 2024, doi: 10.1186/s12859-024-05749-y.

[32] 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, p. 2, 2023/06/03 2023, doi: 10.1007/s10911-023-09540-2.