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 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, 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, and is a major cause of death for many people, in particular among (young) women [1] and people from low socio-economic backgrounds [2]. In 2022 female breast cancer was the second most common cancer with 2.3 million cases [3, 4]. 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, as TNBC proliferates more [5] and difficult to treat types of breast cancer due to lack of molecular targets [6]. About 15% of breast cancers are TNBC and these have a significantly worse prognosis than other types of breast cancer [7, 8].

A lack of expression of three hormone receptors defines TNBC: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) [9].

Although neoadjuvant chemotherapy is the primary treatment modality for breast cancer in the early stage and used in conjunction with immunotherapy [10-12], TNBC does not respond to targeted hormone therapies due to the lack of related receptor markers [13], which limits treatment options and leads to a poorer prognosis for patients. Gene mutations and highly rearranged genomes characterize TNBC [14]. Initial therapeutic decisions are guided by PD-L1 expression and BRCA mutations [2].

The typing of TNBC done by Lehmann in 2011 [15], and cited recently [16-18], 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’ [18-20] to indicate the subtypes. Another option for profiling for subtypes is ‘PAM50 intrinsic molecular sub­types’ [21]. For a recent extensive review on subtyping see the work of Asleh et al. [22].

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. Recently liquid biopsy is emerging as a tool for diagnosing cancer [23]. Using IHC is considered the current clinical standard, although some contend there is no reliable biomarker [24].

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 given the accuracy and high sensitivity of RNA sequencing data [25]. 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, or uses large amounts of additional data that leads to complex models that lack explainability and might not be practical in all settings.

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 essential, especially in a healthcare setting. As pointed out by Pinto et al., there is evidence for ethnicity specific determinants, and there exists a gap in genomic data between ethnicities [26]. The study of Asleh et al. showed the prevalence of the Chinese population in the LAR subtype [22]. 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 [27, 28].

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.

This study proposes 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), given the detailed view the dataset offers [21, 29]. The goal is to explore if gene expression data by itself is sufficient to accurately identify TNBC status, and to do so in a way that emphasizes ethical principles such as transparency, reproducibility and algorithmic fairness.

**2 Literature Review**

Having reviewed existing literature on TNBC, whilst being aware of recent developments using multi-omics [25, 30, 31], this review has a focus on studies that use RNA sequencing and/or machine learning for diagnosis or classification.

Dass et al. [32] 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 [33], 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* [34], 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. [35], 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. [36] 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, the research of molecular classification of TNBC is heterogeneous and challenging to treat [21, 37]. 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 lack transparency. To address this gap, this study proposes as a solution the training of a machine learning model using publicly available RNA sequencing data.

**3 Methods and Material**

**3.1 Overall Description of the Proposed Model**

AsjfasklfjskfskfsjfksjgksalfjaslfjsalfjslkfdsjklgdAslkfjsklfjsdkldsjklgdsjklgdsjkllasjfkasjfsjkfjsklfjdsklfjsklfjsklfjsklfjsdklfsajklfsajklfsajklfsajklfsajkfjasklfsajkfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfj

sklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdkl

Klasjfasklfjsklfjsfklsjfasklfjaslfksajflaskjfladskfjaskkfjaslkfjlskajfaskljflaskfjkaslfjklasjfklasjfklsjfklsfkls

Gfdsaskldfdasflasjfksfasjkfjsjfashjkfashjkfashjkfsahjfsajkfhsakfsk

**3.2 Datasets**

Here is some more

**3.3 Data PreProcessing**

Asfsgdgf

**3.4 Feature Selection**

In this study multiple approaches are used for feature selection. Using the list of twenty proteins suggested by Kohari et al is one [35]. Another approach is using correlation by machine learning. Third approach is done by reviewing 35 research papers of the past 3 years in order to assess the latest insights for relevant features. For example, very recent research suggests LRPPRC as a distinct marker for TNBC [38]. See table 1 for the resulting selected features.

Some general info on the FeatureSets on 18th May:

FeatureSet1 = the three markers that determine if BC is TNBC

FeatureSet2 = the result of reviewing 65 papers, based on human selection, selecting the markers that the papers said is a biomarker, close to a biomarker, or considered an oncogene.

FeatureSet3 = all genes with more than 1 reference, thus excluding a recently found biomarker LRPPRCC [38].

FeatureSet4 = all genes. This does not adhere to the rule “n/p > 5”, as n = 997 and p = 290.

**3.5 Data Integration**

Asfdsfsfsfs

**3.6 Classification and Prediction Modelling**

Asfasfsd

**3.7 Implementation**

Asfdsfsfsdfgs

**3.8 Evaluation**

Asdfsdg

NogwattekstNogwattekstNogwattekstNogwattekstNogwattekstkstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwat

tekstNogwattekskstNogwattekstNogwattekskstNo

gwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwatteks

**4 Results and Discussion**

kstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwatteks

**Hier kunnen subparagrafen komen**

**5 Conclusions**

‘Wat we denken / vinden / zien / gevonden hebben’

**5.1 Innovations**

Oftewel, hier komt ‘what can be researched next’

Hierboven is 2 kolommen. Dit is weer normaal Hierboven is 2 kolommen. Dit is weer normaal Hierboven is 2 kolommen. Dit is weer normaal Hierboven is 2 kolommen. Dit is weer normaal

Weer ‘terug

Page break after this

# 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] 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.

[3] 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).

[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] 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.

[6] 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.

[7] 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).

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[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, pp. 1-26, 2022/09/01 2022, doi: 10.1186/s13046-022-02476-1.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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>.

[33] 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.

[34] "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).

[35] 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.

[36] 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.

[37] 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.

[38] 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.