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

Ruben • Kevin • Sander • Vince • Victor

**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, 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.

**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 and difficult to treat types of breast cancer due to lack of molecular targets [5]. About 15% of breast cancers are TNBC and these have a significantly worse prognosis than other types of breast cancer [6, 7].

A lack of expression of three hormone receptors defines TNBC: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) [8]. Although neoadjuvant chemotherapy is the primary treatment modality for breast cancer in the early stage [9], TNBC does not respond to targeted hormone therapies due to the lack of related receptor markers, which limits treatment options and leads to a poorer prognosis for patients.

The typing of TNBC done by Lehmann in 2011 [10], and cited recently [11-13], 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’ [13-15] to indicate the subtypes. Another option for profiling for subtypes is ‘PAM50 intrinsic molecular sub­types’ [16]. For a recent extensive review on subtyping see the work of Asleh et al. [17].

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 [18]. 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.

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 [19, 20].

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 [16, 21]. 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 [22-24], this review has a focus on studies that use RNA sequencing and/or machine learning for diagnosis or classification.

Dass et al. [25] 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 [26], 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* [27], 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. [28], 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. [29] 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 [16, 30]. 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, 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**

AsjfasklfjskfskfsjfksjgksalfjaslfjsalfjslkfdsjklgdAslkfjsklfjsdkldsjklgdsjklgdsjkllasjfkasjfsjkfjsklfjdsklfjsklfjsklfjsklfjsdklfsajklfsajklfsajklfsajklfsajkfjasklfsajkfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdklgjdslfjsklfgjsdkl

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 [28]. 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 [31]. See table 1 for the resulting selected features.

**3.5 Data Integration**

Asfdsfsfsfs

**3.6 Classification and Prediction Modelling**

Asfasfsd

**3.7 Implementation**

Asfdsfsfsdfgs

**3.8 Evaluation**

Asdfsdg

NogwattekstNogwattekstNogwattekstNogwattekstNogwattekstkstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwattekstNogwattekskstNogwat

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| Source |  |  |  |  |  |  |
| Protein |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Source |  |  |  |  |  |  |
| Protein |  |  |  |  |  |  |

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

## Feature selection (See revised below, this to doublecheck)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Source |  | Gene | Source |  | Gene | Source |  | Gene | Source |
| DCLK1 | [17, 32, 33] |  | BRCA1 | [2, 17, 23, 34-36] |  | ERBB3 | [2, 23] |  | COL1A2 | [16] |
| FOXC1 | [17, 32] |  | BRCA2 | [17, 23, 34-37] |  | IGFBP5 | [23] |  | TOP2A | [16, 37] |
| CD8A | [17, 23, 32, 34-36] |  | BRD4 | [34] |  | MYCN | [23] |  | VAV3 | [16] |
| CD8B | [17, 23, 32, 34-36] |  | ATR | [17, 34] |  | MDM2 | [2, 23] |  | TFF1 | [16] |
| PARP1-16 | [9, 17, 23, 32, 34, 35] |  | WEE1 | [17, 34] |  | LAMA5 | [23] |  | ACTG2 | [16] |
| RAD51 | [32] |  | MTOR | [16, 34, 35] |  | SDCBP | [23] |  | CKB | [16] |
| CDK1 | [17, 24, 34, 38] |  | PALB2 | [17, 34] |  | SMAD4 | [23] |  | NME1 | [16] |
| CDK2 | [34] |  | CXCR4 | [34] |  | DKK3 | [23] |  | ASHA1 | [16] |
| CDK4 | [17, 31, 32, 34, 39] |  | TP53 | [2, 17, 23, 34] |  | DKK4 | [23] |  | NOP16 | [16] |
| CDK6 | [2, 17, 31, 32, 34, 39] |  | CTLA4 | [17, 34, 35] |  | CD4 | [23] |  | S100A16 | [16] |
| AKT1 | [17, 32] |  | CCR5 | [34] |  | EZH2 | [17, 23] |  | PHLDA2 | [16] |
| AKT2 | [17] |  | ROR1 | [34] |  | ARHGEF2 | [23] |  | FGFR2 | [17] |
| PTEN | [2, 16, 17, 32, 34, 35, 37, 39] |  | CD276 | [34] |  | YTHDF1 | [23] |  | CHEK2 | [17] |
| PIK3CA | [2, 17, 32, 34, 40] |  | RRM2 | [24] |  | POSTN | [41] |  | ATM | [17] |
| EGFR | [2, 16, 17, 32, 34, 35, 39] |  | MCM7 | [24] |  | MMP2 | [37, 41] |  | RAD51C | [17] |
| KRAS | [32] |  | BUB1B | [24] |  | MMP9 | [41] |  | VTCN1 | [17] |
| FGFR1 | [17, 32, 34] |  | CCNA2 | [24] |  | MMP13 | [41] |  | PGR | [2, 17] |
| FGFR2 | [32] |  | FXR | [24] |  | MYC | [2, 17, 41] |  | PIP | [17] |
| NOTCH1 | [17, 32, 35] |  | LRPPRC | [31] |  | SOX2 | [17, 41] |  | ERBB2 | [2, 17] |
| NOTCH2 | [32, 35] |  | RB1 | [2, 17, 31] |  | NANOG | [41] |  | CDKN2A | [17] |
| NOTCH3 | [17, 32, 35] |  | MAOA | [42] |  | HSP90AA1 | [34] |  | PDGFRA | [17, 37] |
| NOTCH4 | [32, 35] |  | ADH1B | [42] |  | HSP90AB1 | [34] |  | B2M | [17] |
| HES1-7 | [32] |  | ADH1C | [42] |  | HSP90B1 | [34] |  | ASXL1 | [17] |
| STAT3 | [17, 31, 32, 34] |  | AOC3 | [42] |  | JAK1 | [17, 34] |  | ASXL2 | [17] |
| HDAC1-11 | [32, 34, 35] |  | TAT | [42] |  | JAK2 | [34] |  | ASXL3 | [17] |
| YTHDC1 | [32] |  | PCK1 | [42] |  | ESR1 | [2, 16] |  | VEGF | [17] |
| SMAD3 | [32] |  | CD44 | [17, 35] |  | FOXA1 | [2, 16, 17] |  | VRK1 | [17] |
| LAG3 | [17, 32, 34] |  | CXCR2 | [35] |  | XBP1 | [2, 16, 17] |  | PKN1 | [17] |
| ICOS | [32] |  | FUT8 | [35] |  | MYBL1 | [16] |  | PRKD2 | [17] |
| TACSTD2 | [32] |  | MUC1 | [2, 17, 35] |  | MYBL2 | [16] |  | TIGIT | [17] |
| SORT1 | [32] |  | HER3 | [35] |  | FGFR4 | [16] |  | CD274 | [17, 36] |
| GPX4 | [32] |  | MAPK14 | [24] |  | INPP4B | [2, 16, 17] |  | CCND1 | [2, 17] |
| MTHFD2 | [32] |  | GATA3 | [2, 16, 17, 37, 40] |  | TRIM29 | [16] |  | RAD50 | [17] |
| ACSL3 | [32] |  | IGF1 | [17] |  | MAP3K1 | [17] |  | TERF1 | [17] |
| RPS6 | [17] |  | EIF4A2 | [17] |  | SPDEF | [17] |  | POLR1D | [17] |
| POLD4 | [17] |  | CDC25A | [17] |  | CDCA7 | [17] |  | BIRC3 | [17] |
| HMCES | [17] |  | LIG3 | [17] |  | WNT3 | [17] |  | SMAD9 | [17] |
| MAP2K4 | [2, 17] |  | RAC1 | [17] |  | ITGAV | [17] |  | RSPO1 | [17] |
| ITGB3 | [17] |  | FRS2 | [17] |  | NGFR | [17] |  | RARA | [17] |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Source |  | Gene | Source |  | Gene | Source |  | Gene | Source |
| YOD1 | [38] |  | PECAM1 | [37] |  | CDH5 | [37] |  | PDGFRB | [37] |
| KRT18 | [37] |  | CLDN5 | [37] |  | MKI67 | [37] |  | AMIGO2 | [37] |
| FLRT2 | [37] |  | IGFBP7 | [37] |  | AEBP1 | [37] |  | EDNRA | [37] |
| COL1A1 | [37] |  | CCL2 | [37] |  | COL11A1 | [37] |  | FAP | [37] |
| PCOLCE | [37] |  | CXCL8 | [37] |  | COL12A1 | [37] |  | KRT14 | [37] |
| SPARC | [37] |  | EPCAM | [37] |  | CXCL1 | [37] |  | KRT15 | [37] |
| SERPINE2 | [37] |  | COL5A1/3 | [37] |  | FGF2 | [2, 37] |  | KRT16 | [37] |
| COL1A1 | [37] |  | COL1A2 | [37] |  | COL3A1 | [37] |  | COL4A1 | [37] |
| KRT17 | [37] |  | NCOR1 | [2] |  | CCNT1 | [2] |  | RAD17 | [2] |
| MDGA2 | [2] |  | ITGAE |  |  | SCUBE2 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

# Feature selection REVISED

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Gene | Source |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Some notes on the genes / proteins

* High POSTN expression correlates with poor survival in patients with TNBC [41]
* Targeting periostin [POSTN] can significantly suppress TNBC progression [41]
* Notably, AOC3 and PCK1 were identified as genes significantly associated with poor overall survival (OS) [42]
* Based on these results, it is hypothesized that AOC3 and PCK1 may serve as potential biomarkers for predicting the prognosis of TNBC patients.[42]
* LRPPRC served as a distinct biomarker for TNBC [31]
* Recent studies have shown that the LRPPRC is overexpressed in TNBC [31]
* *ESR1, GATA3, FOXA1, XBP1 and cMYB*, in ER+/luminal-like subtypes [16]
* *FGFR4, HER1/EGFR* and loss of *PTEN* and *INPP4B* in the HER2 subtype [16]
* This higher mutational burden is nevertheless characterized by very few consistently mutated genes other than TP53, highlighting the complex repertoire of somatic mutations that underpin the heterogeneity of TNBC [17]
* EGFR, FGFR2 and MYC amplifications as well as PTEN losses are more frequent in TNBC but appear to be more clinically-relevant when analyzed in the context of TNBC transcriptomic subtypes [17]
* TNBC is known to display several BRCAness features through somatic *BRCA1/2* mutations or germline/somatic alterations in homologous repair-related genes (e.g., *PALB2*, *ATM*, *CHEK2*)
* Compared to the TCGA dataset, this study highlighted distinct patterns in the Chinese TNBC population including higher frequencies of *PIK3CA* mutations and LAR subtype [17]
* BLIA tumors were characterized with high expression of the kinases PKN1 and PRKD2 and an increased activity of STAT family proteins (STAT1, STAT2, STAT5A) as inferred from transcriptional factor analysis of phospho-proteomic data [17]
* BLIS tumors displayed high phosphorylation of CDK1 and VRK1, suggesting that they may benefit from CDK1/2 inhibitors.
* The LAR subtype was characterized by increased activity of PI3K/AKT pathway as demonstrated by high phosphorylation of AKT1 and AKT2, high phosphorylation of the RB protein while retaining low levels of E2F, and an elevated phosphorylation of the MAP2K4 kinase which was notably observed as downregulated in LAR at the RNA level [17]
* In particular, the mesenchymal TNBC subtype was characterized with DNA repair signalling, as demonstrated by elevated phosphorylation of ATR and ATM which was not observed at the RNA level [17]
* Since defective DNA damage repair mechanisms are a hallmark of TNBC, the activity of DNA damaging agents, particularly platinum-based chemotherapy, has been investigated [17]
* High expression of YOD1 was a critical feature of TNBC and was associated with poor prognosis of TNBC. [38]
* In breast cancer, higher expression of IL-6 and DCLK1 predicts shorter recurrence-free survival in TNBC but not in luminal A, luminal B and HER2 subtypes, which suggests the critical roles of DCLK1 and IL-6 in tumor recurrence of TNBC. [33]
* The TNBC population of MDA-MB-231 exhibited upregulation of AMIGO2, which is implicated in cancer metastasis, and enhanced translational initiation [37]
* Tumor: KRT18 and EPCAM [37]
* Amplification (MYC, PIK3CA, CDK6, MDM2); Deletion (MAP2K4, TP53, NCOR1, BRCA, PTEN, INPP4B) [2]
* Nearly 75% of TNBC is classified as BL [2]
* The occurrence of the LAR subtype is less likely in patients under the age of 50 [2]
* The MES subtype is characterized by a distinctive overexpression of genes typically associated with osteocytes and adipocytes, along with the essential insulin-like growth factor 1 (IGF-1). This subtype is marked by significant cellular signaling related to cell cycle regulation, mismatch repair, and DNA damage response mechanisms. [2]

Hier verder met lijstje

* Nguyen QH et al. used scRNA-seq to obtain the transcriptome data of 25,790 primary human mammary epithelial cells from 7 individuals and found that normal mammary epithelial cells could be divided into three cell types: basal (KRT14+), luminal-1 (KRT18+SLPI+), luminal-2 (KRT18+ANKRD30A+). [36]
* The classical typing of breast cancer includes immunohistochemistry typing and PAM50 typing. [36]
* Wu Qet al. first found that SCUBE2 was highly expressed in luminal breast cancer and was related to bone metastasis and verified through experiments that this gene could promote bone metastasis of official breast cancer [36]
* HR + /HER2‑ breast cancer HR + /HER2- breast cancer accounts for about 70% of all breast cancer cases. [36]
* TNBC typically exhibits a high density of TILs, predominantly composed of CD3 + T cells and CD20 + B cells. [43]
  + KEVIN: CD3+ == CD3G && CD3D && CD3E
  + KEVIN: CD20+ == MS4A1
* Infiltration of CD3 + T cells is most common in TME and plays a major role. CD3 + T cells are mainly composed of CD8 + T cells and CD4 + T cells, with regulatory T lymphocytes (Tregs) and natural killer (NK) cells constituting less than 1%. The proportion of CD8 + T cells in TNBC is higher than in other breast cancer subtypes. [43]
  + KEVIN: CD3+ == CD3G && CD3D && CD3E
  + KEVIN: CD8+ == CD8A && CD8B
  + KEVIN: CD4+ == CD4s
* As expected, a majority of tumors were noted to have p53 alterations with a substantial frac­tion also harboring PIK3CA mutations. [16]
* For example, the Luminal B subtype is overrepresented with ATM loss and Cyclin D1 and MDM2 amplification[16]
* Proteomes from Luminal, HER2 + and basal-like contain an enrichment of E2F and MYC targets as well as G2M checkpoint proteins, however, basal-like tumors are distinguished by immune markers in special MHC class proteins [16]
* Inter­estingly, over 80% of MMTV-PyMT tumors have a V483M mutation in *Ptprh*, a phosphatase targeting EGFR and other kinases.[16]
* There are clear similarities between the cancer cell lines and breast cancer, with shared mutations in key genes such as p53, RB and PI3K[16]
* Furthermore, a total of 58 genes were iden­tified to be epigenetically regulated across a panel of 45 cell lines . Some of them, such as COL1A2, TOP2A, VAV3, CDKN2A, and TFF1, have validated roles in tumor development.[16]
* The prevalence of TNBC in women with a *BRCA1* mutation, the multitude of molecular alterations in TN tumors, and the histopathological similarities shared between TNBC and *BRCA1*-mutated BC (such as deficiencies in DNA repair systems and disruptions of homologous recombination) have sparked interest in platinum salts for this subtype [9].
* A subset of TNBC exhibits PDL-1 expression on both the tumor and tumor-infiltrating lymphocytes (TILs), indicating a higher mutational burden compared to other BC subtypes [9]
* Trop-2 and topoisomerase-1 (TOPO1) expression is present in 56–80% of primary and metastatic TNBC tumors [39]
* ROR-1 is an attractive novel target that is highly expressed on TNBC and is minimally present or absent on healthy tissues.[39]
* TNBC has the highest mutation rate compared to other BC subtypes, with EGFR, FGFR2 and MYC amplifications as well as PTEN loss more frequent, although other mutations including TP53 exist [39]
* A quarter of TNBC patients will have activating mutations in the PI3K/AKT/PTEN/mTOR pathway which drives tumor progression and promotes survival [39]
* Growth factors drive diverse signaling pathways in cancer development and many TNBC have EGFR over-expression [39]
* Low expression of superoxide dismutatase 3 (SOD3) is common in TNBC; epigenetic silencing of SOD3 via methylation represents a novel therapeutic target in TNBC [39]
* Statistically, TNBC accounts for 10 to 25% of incident BC depending on patient ethnic background [13]
* In-depth gene expression profiling was carried out by Lehmann and his colleagues who identified six different subgroups of TNBC: Luminal Androgen Receptor (LAR), Mesenchymal (MES), mesenchymal stem-like (MSL), Basal-like1 (BL1), Basal-like1 (BL2) and immunomodulatory (IM) .Further research revealed that the stromaassociated tumor cells and infiltrating lymphocytes were the source of the IM and MSL subtype transcripts. Consequently they revised their classification and grouped the subtypes into four main categories: BL1, BL2, MES and LAR subtypes. In another work, used mRNA profiling to identify four subgroup of TNBC; LAR, MES, BL immune-suppressed (BLIS), and BL immune-activated (BLIA) subtypes. Actually, the MES described by Burstein almost completely corresponded to the IM and MSL subtypes of Lehmann et al. [13]
* Similarly, showed that the disease-free survival of every molecular TNBC subtypes is decreasing in the following order BLIA > MES > LAR > BLIS. More recently, using surrogate IHC, Leeha et al. showed different overall survival rates between TNBC subgroups, with the BLIS subtype having the worst outcome. + see ethnicity [13]
* BL1 subtype: EGFR expression is negative and CK5 and/or CK14 expression is positive. BL2 subtype: EGFR expression that is positive regardless of CK5 or CK14 positivity. Mesenchymal subtype (MES): Vimentin expression is positive and E-cadherin, Claudins 3 and 7 have a diminished expression. LAR subtype is associated with the positive expression of AR.[13]
* Cytokeratins are the main IHC markers of BL BC, especially the CK14 which has proven to be the most accurate one. [13]
* We found that 79.68% of the TNBCs have a p53 mutated-type pattern.[13]
* Our results confirmed that high γδT cell infiltration is associated with a favor­able prognosis in patients with TNBC but not in patients with hormone receptor-positive or human epidermal growth factor receptor 2 (HER2). The γδT cell infiltration may induced by somatic gene mutations including Mucin 16 (*MUC16)*, Dedicator Of Cytokinesis 11 *(DOCK11)*, Laminin Subunit Alpha 2 *(LAMA2)*, and Ryanodine Receptor 1 (*RYR1)*. [44]
* Concretely, γδ T cells indicated high expres­sion of *CD3D*, *TRDC*, *TRGC1*, and *TRGC2*, with almost no expression of CD4 (Fig. 2C).[44]
* Our results indicate that a high infiltration of γδT cells in TNBC is closely associ­ated with somatic mutations in *MUC16*, *RYR1*, *LAMA2*, *and DOCK11*. Interestingly, *MUC16* (CA125) is a well-known biomarker in multiple cancers, including ovarian and pancreatic cancer[44]
* *LAMA2* has been identified as a tumor suppressor in breast cancer and is among the top 10 most frequently mutated genes in inflam­matory breast cancer[44]
* Additionally, TNBC is more common in younger women and certain racial groups, and immune cell infiltration in the tumor microenvironment also exhibits characteristics different from those of other subtypes [45]

|  |
| --- |
|  |
| **Fig. 1 [**Visuals slightly modified from] **Source** Fig. 1 in [17], p. 5 |

## Feature imputation

Kevin: Applicable? Volgt?

# 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] 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.

[6] 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).

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[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] 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>.

[26] 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.

[27] "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).

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] S. Zhu *et al.*, "Recent advances in targeted strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 16, no. 1, pp. 1-36, 2023/08/28 2023, doi: 10.1186/s13045-023-01497-3.

[33] H. Liu *et al.*, "Targeting DCLK1 attenuates tumor stemness and evokes antitumor immunity in triple-negative breast cancer by inhibiting IL-6/STAT3 signaling," *Breast Cancer Research,* vol. 25, no. 1, p. 43, 2023/04/17 2023, doi: 10.1186/s13058-023-01642-3.

[34] Y. Li *et al.*, "Recent advances in therapeutic strategies for triple-negative breast cancer," *Journal of Hematology & Oncology,* vol. 15, no. 1, pp. 1-30, 2022/08/29 2022, doi: 10.1186/s13045-022-01341-0.

[35] 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.

[36] 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.

[37] S. J. Hachey *et al.*, "Targeting tumor–stromal interactions in triple-negative breast cancer using a human vascularized micro-tumor model," *Breast Cancer Research,* vol. 26, no. 1, p. 5, 2024/01/05 2024, doi: 10.1186/s13058-023-01760-y.

[38] Z. Han *et al.*, "Deubiquitylase YOD1 regulates CDK1 stability and drives triple-negative breast cancer tumorigenesis," *Journal of Experimental & Clinical Cancer Research,* vol. 42, no. 1, pp. 1-17, 2023/09/04 2023, doi: 10.1186/s13046-023-02781-3.

[39] 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.

[40] 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.

[41] 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.

[42] S. Gong *et al.*, "Comprehensive analysis of the metabolomics and transcriptomics uncovers the dysregulated network and potential biomarkers of Triple Negative Breast Cancer," *Journal of Translational Medicine,* vol. 22, no. 1, p. 1016, 2024/11/11 2024, doi: 10.1186/s12967-024-05843-y.

[43] 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.

[44] 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.

[45] 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.