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

18 April 2025

Revision 2

# 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, relies on an expert to interpret the result, and as such not always accessible, especially in research contexts. We aim 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 project we propose will use the Breast Invasive Carcinoma Collection dataset from The Cancer Genome Atlas (TCGA-BRCA). This dataset includes RNA sequencing data as well as clinical data for over 1000 patients. We decided to focus on a binary classification (TNBC versus non-TNBC), favoring simplicity over complexity, and prioritizing explainability, fairness and reproducibility of the model. We intend to evaluate multiple machine learning algorithms and select the most effective one, and also to apply interpretability techniques such as feature importance analysis.

The anticipated outcome is a transparent and accurate classification tool that provides an addition to current methods to show a TNBC diagnosis in research and potentially even clinical settings, 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 women. 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. About 15% of breast cancers are TNBC and these have a significantly worse prognosis than other types of breast cancer.

TNBC is defined by the lack of expression of three hormone receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). 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.

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. This is considered the clinical gold standard, although it 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 focuses on subtype classification of TNBC, or involves commercial tools with limited transparency and unclear ethics, 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.

Existing research on this topic often has little or no mention of ethics principles such as explainability, fairness and reproducibility. Especially in a healthcare setting, such principles are not just desirable but downright essential. 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 (yes or no) using RNA sequencing data alone. It is this gap that we want to fill with this project.

We intend 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). Our 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. We hope this model can have potential as an additional diagnostic tool in both research and clinical settings.

# 2. Literature Review

We have reviewed the existing literature on TNBC, with a focus on studies that combine RNA sequencing and/or machine learning for diagnosis or classification.

Dass et al. [1] extensively reviewed current and potential future methods of diagnosing TNBC. They showed that the current clinical gold standard for diagnosis is using IHC 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* [2], 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. [3], 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 that this was not useful for predictions.

Looking more into TNBC vs. non-TNBC classification, Davis et al. [4] 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. [5] 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, we have learned 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. We want to address this gap by training a machine learning model using publicly available RNA sequencing data. In particular, we would like to build on the work of Kothari et al. [3] 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 our research, we have set up the following research questions:

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 the best 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 our model and ensure that we focus 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 accurately and ethically classify triple negative breast cancer (TNBC) status using only RNA sequencing data, which can provide an additional diagnostic tool besides existing methods such as IHC.

Connected to the research questions above, we have defined the following SMART objectives. We assume a project runtime of 6 weeks (12 May to 22 June):

1. *Model setup and first evaluation:* Prepare the data and set up a simple 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 reproducibility, and have a first draft of the final report by the end of week 5 (15 June).

This leaves the final week for finalizing the report as well providing some room for delays.

# 5. Significance of the Project

As we’ve established so far, triple negative breast cancer (TNBC) is a challenge in both diagnosis and treatment as it does not respond well to common therapies. 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, we avoid the 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 but effective machine learning model that can be used for research and possible even clinical purposes, especially where RNA sequencing data is already available.

# 6. Methodology

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

## 6.1 Project Workflow

Our 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

***(Kan iemand hier misschien een diagram van maken?)***

## 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:* To explore model explainability, depending on how feasible it is to integrate this *(Is dit haalbaar?)***
* *Git (via Github):* Version control and collaboration

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

## 6.3 Data acquisition and preprocessing

As mentioned before, we will use the TCGA-BRCA dataset, obtained through the Genomic Data Commons (GDC) portal. To make the data usable, we will follow the following preprocessing steps:

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

We have already performed the first two steps 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 the gene expression data. To achieve a usable selection of features, we will both use existing literature (such as Kothari et al. [3]) and do our own selection using statistical or model-based methods, such as looking at the variance of features. **We also intend to look into dimensionality reduction techniques such as Principal Component Analysis (PCA)** ***(Is dit te doen?)***, although we are aware of the potential implications for explainability.

## 6.5 Model selection and development

We intend to compare the performance of at least three common machine learning algorithms. The primary candidates are:

* *Logistic Regression:* due to its interpretability and hopefully providing a good 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

We will consider testing adding additional models, for example XGBoost, if time and performance permits.

## 6.6 Performance evaluation

We will use 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 (TNBC is under-represented)
* F1 score: mean of precision and recall
* ROC / AUC

We will add a visualization of a confusion matrix for each model to obtain additional insight, which will help us to select the most suitable model in the end.

## 6.7 Explainability and ethical evaluation

In addition to model performance, we intend to evaluate the model 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 values) *(Haalbaar? Andere opties?)*** to understand how features/genes influence predictions. Additionally we will assess fairness 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. We will document all decisions we make based on our findings for transparency and reproducibility.

# 7. Dataset

The dataset we intend to use is from The Cancer Genome Atlas (TCGA), specifically the Breast Invasive Carcinoma Collection (TCGA-BRCA) [6]. 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, which are used to determine the patient’s status for estrogen receptor (ER), progesterone receptor (PR) and HER2 protein expression. From this, 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, we found that 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 we cannot determine TNBC status.

Additionally, we found that 1094 patients 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 end, we have 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, we will use the RNA sequencing and clinical data from the TCGA-BRCA dataset. As this is highly personal healthcare data, it raises various ethical questions. We will address the most important ethical issues (privacy, fairness, bias and explainability) 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. We will follow best practices by not storing data publicly or redistributing the dataset. Instead, we will clearly document the data source and how it was accessed, making 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, we aim 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 we will use has an imbalance between TNBC cases (115) and non-TNBC cases (862), which could potentially lead to a biased model. To mitigate this, we will use strategies like oversampling the TNBC cases or adjusting model weights to ensure the model does not favor 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, we will aim to build 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

We plan on splitting the project into three phases. Since we have six weeks (12 May - 22 June), there are two weeks available for every phase. We will coordinate our work on Github, Microsoft Teams (for communication), and of course our weekly meetings at school.

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

In the preparation phase, we intend to prepare our development environment, obtain the data and train an initial model (including basic validation). The division of tasks between team members is as follows:

* ***Ruben (deels al gedaan):***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 that others 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, committing it to Git and being responsible to ensure that others can also run the notebook locally.
* ***(Vince?):*** Perform initial exploratory data analysis to understand the RNA sequencing files, and creating scripts for filtering features, then using these for selecting the genes that were identified as promising in the literature. The scripts for this will be committed to Git as well.
* ***(Victor?):*** Implement a basic machine learning model using the selected features, building on the work of the others, again committing the exact code used (including random state) to Git for reproducibility.
* ***(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. As with everything else, this too will be committed to Git.

We are aware that many of these tasks are necessarily sequential. To prevent one team member from holding up the work, it will be the shared responsibility of all other team members to step in and help others when necessary. To ensure delays are prevented as much as possible, we will put an additional (tentative) deadline on 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, we will focus on identifying the most useful features and model algorithms. After the first phase, we will have an initial model (and all necessary script to build it) that we will build upon in different ways. As a result, the following tasks can mostly be done concurrently. However, there is an earlier 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 (such as PCA) to identify features (again, genes) 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.

***(NB: Ik zelf [Ruben] ben rond Pinksteren een weekje weg, dat is rond de deadline van deze fase. Daarom wil ik in deze fase graag één van de eerste twee taken oppakken)***

Once this phase is completed, we will have trained multiple models using different sets of features, and should have a general idea of the most promising features and model. Finally, we should have a first draft of the report.

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

In the final phase, we intend to fine-tune the most promising model (or models), evaluate the fairness and explainability, and present all findings to conclude the project.

* ***(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?):*** Using SHAP and/or ELI5 methods to analyze the model’s interpretability and to determine 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), and 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 Powerpoint presentation for reporting our findings.

Across all phases, all members will participate in discussions (weekly in school) and help make major desisions such as feature and model selection, and the contents of the final report and presentation. Everyone is responsible for their own tasks, the sharing of created content (such as Jupyter notebooks), and ensuring others can reproduce the results locally. By collaborating this way, we hope to reach equal contribution from all team members while ensuring reproducibility of all work.

# 10. Deliverable

At the end of the project, we will deliver an extensive written report that includes our 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 we 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.

Finally, we will prepare a Powerpoint presentation for our final presentation, however this might need to be finalized after the main project deadline of June 22. At the latest, it will be added to the repository on the date we give our final presentation.

# References

[1] 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>.

[2] "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).

[3] 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.

[4] 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.

[5] 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.

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