


Predicting Pathologic Complete Response to Neoadjuvant Treatment in HER2-positive Breast Cancer using Interpretable Classification


Sergio Peñafiel

(Department of Computer Science, Faculty of Physical and Mathematical Sciences, University of Chile, Santiago, Chile

 <https://orcid.org/0000-0002-0025-7805>, spenafie@dcc.uchile.cl)


Esteban Ramírez

(Department of Computer Science, Faculty of Physical and Mathematical Sciences, University of Chile, Santiago, Chile,

 <https://orcid.org/0009-0004-4370-6844>, esteban.ramirez@ug.uchile.cl)

Nelson Baloian

(Department of Computer Science, Faculty of Physical and Mathematical Sciences, University of Chile, Santiago, Chile

 <https://orcid.org/0000-0003-1608-6454>, nbaloian@dcc.uchile.cl)

Isabel Saffie

(Breast Oncologic and Reconstructive Surgery Unit, Instituto Oncológico Fundación Arturo López Pérez, Santiago, Chile, isabel.saffie@falp.org)

Paulo Luz

(Medical oncology department, Centro hospitalar universitário do Algarve, Faro, Portugal, p_luz@msn.com)

Inti Paredes

(Medical Informatics and Data Science unit, Department of Cancer Research, Instituto Oncológico Fundación Arturo López Pérez, Santiago, Chile
inti.paredes@falp.org)

Abstract: Breast cancer is a significant global health problem, and HER2-positive breast cancer accounts for a substantial proportion of cases. The combination of Trastuzumab and Pertuzumab monoclonal antibodies with chemotherapy has demonstrated effectiveness in achieving pathologic complete response (pCR) among HER2-positive breast cancer patients. This study aims to develop an interpretable machine learning model to predict pCR in patients undergoing this neoadjuvant treatment. Previous studies have explored predictors of pCR and utilized statistical techniques, but no prior research has applied machine learning to this specific treatment. This work proposes a rule-based interpretable method based on Dempster-Shafer theory. The model is trained using a dataset of 390 patients, with 57% achieving pCR. The performance of the model is compared with other classification algorithms, demonstrating its moderate but promising results. This work highlights the importance of combining accuracy and interpretability in healthcare applications, providing insights into the factors influencing treatment response in HER2-positive breast cancer patients.

Keywords: Breast Cancer, Pathologic Complete Response, Interpretability, Supervised Learning, Expert Systems, Dempster-Shafer Theory

Categories: I.2.1, I.2, J.3

DOI: 10.3897/jucs.164692

1 Introduction

Breast cancer is the most frequently diagnosed cancer worldwide and the leading cause of cancer-related mortality among women, making it the fifth deadliest cancer overall [Sung et al., 2021]. Given its prevalence and severity, breast cancer constitutes a major global health issue.

There are three principal subtypes of breast cancer, classified according to the presence or absence of molecular markers for estrogen receptors, progesterone receptors, and human epidermal growth factor receptor 2 (HER2): hormone receptor-positive/HER2-negative (approximately 70% of cases), HER2-positive (15%–20%), and triple-negative breast cancer (15%, characterized by the absence of all three markers) [Waks and Winer, 2019]. This study specifically focuses on HER2-positive breast cancer, which, despite its aggressive progression, is often responsive to targeted therapy.

Neoadjuvant therapy for early HER2-positive breast cancer, encompassing the combination of anti-HER2 agents—such as the monoclonal antibodies Trastuzumab and Pertuzumab—with conventional chemotherapy, has demonstrated high effectiveness. This treatment strategy results in a considerable proportion of patients achieving pathologic complete response (pCR), which is associated with improved long-term clinical outcomes [Loibl and Gianni, 2017, Schneeweiss et al., 2013].

In line with this therapeutic approach, Saffie et al. [Saffie Vega et al., 2022] studied a cohort of 94 women with non-metastatic HER2-positive breast cancer treated at Fundación Arturo López Pérez (FALP), with the objective of identifying clinical and histopathological variables influencing the likelihood of achieving pCR.

Although machine learning (ML) techniques have been widely applied in various healthcare contexts, their use in predicting response to this specific treatment in HER2-positive breast cancer remains limited. Furthermore, while highly accurate models have been developed, they are often difficult to interpret and are commonly referred to as black-box models. In many cases, there is an inherent trade-off between accuracy and interpretability [Abdullah et al., 2021].

Interpretability is defined as the extent to which a machine learning model can provide understandable and transparent explanations for its predictions. In medical applications, interpretability is crucial, as it enables clinicians and researchers to evaluate and trust model outputs. Interpretability can be categorized as either local or global. Local interpretability, also known as explainability, refers to the capacity of the model to explain individual predictions, which may help in understanding specific patient outcomes. Global interpretability, by contrast, aims to describe the overall structure and decision-making logic of the model [Molnar, 2025]. The latter is particularly valuable in clinical research, as it can enhance understanding of disease mechanisms and contribute to the refinement of treatment strategies by offering insights aligned with domain knowledge [Rudin, 2019].

The objective of this study is to develop a globally interpretable model capable of predicting whether a patient with HER2-positive breast cancer will achieve pCR following neoadjuvant therapy that includes chemotherapy and dual HER2 blockade.

Such a model can help identify key predictive variables and support clinical decision-making through transparent and explainable results.

The proposed approach utilizes a rule-based, interpretable machine learning method previously applied in healthcare research [Peñafiel et al., 2020b]. The model was trained on a dataset comprising 390 patients, 57% of whom achieved pCR after completing the treatment.

2 Related Work

2.1 Predictors of pCR to Neoadjuvant Therapy

Pathological complete response (pCR) to neoadjuvant therapy is a strong prognostic marker in breast cancer, particularly in HER2-positive and triple-negative subtypes. Numerous studies have explored clinical, histopathological, and molecular predictors of pCR in this context. This section summarizes key findings from the literature, grouped by the type of predictor.

Clinical and Molecular Predictors. Hormone receptor status has consistently emerged as a relevant clinical and molecular predictor. Saffie *et al.* [Saffie Vega et al., 2022] analyzed a cohort of 94 women with HER2-positive breast cancer treated with neoadjuvant chemotherapy plus dual anti-HER2 blockade. Using multiple logistic regression, they identified hormone receptor positivity and HER2/Luminal subtype as factors associated with a lower probability of achieving pCR. Other studies have confirmed that patients with hormone receptor-negative tumors are more likely to achieve pCR compared to those with hormone receptor-positive tumors [Schneeweiss et al., 2013]. In a subset analysis of HER2 IHC 2+/amplified tumors, estrogen receptor (ER) negativity also emerged as a significant predictor of pCR [Katayama, 2021].

Histopathological Predictors. Tumor grade is another factor associated with pCR. In a large-scale study involving 500 HER2-positive breast cancer patients, histological grade 3 was found to be an independent predictor of pCR, highlighting the role of tumor aggressiveness in treatment response [Katayama, 2021].

HER2 Expression Levels. Several studies have examined the intensity of HER2 protein expression as a predictor of response. Katayama *et al.* [Katayama, 2021] showed that tumors with HER2 immunohistochemistry (IHC) 3+ expression exhibited significantly higher pCR rates (52%) than those classified as HER2 IHC 2+ with confirmed HER2 amplification by FISH (20%). This indicates that higher HER2 overexpression correlates with a better treatment response. Similarly, a retrospective study of 119 patients with high-risk HER2-positive breast cancer found that moderate HER2 IHC expression was associated with a significantly lower likelihood of achieving pCR ($P = 0.0078$) and worse breast cancer-specific survival ($P = 0.0015$) compared to strong HER2 expression [Hännikäinen, 2023]. These findings emphasize the prognostic and predictive importance of quantifying HER2 expression intensity.

2.2 Supervised and Unsupervised Machine Learning

Machine learning algorithms learn from training data to make predictions or classifications when processing new inputs. There are two main approaches: supervised learning and unsupervised learning [Russell and Norvig, 2016].

In supervised learning, algorithms learn from labeled datasets where the desired output is known. By working with labeled data, the model can accurately classify new

data or predict outcomes. This type of learning is broadly divided into two problem types: classification, where the algorithm assigns data to specific categories or classes, and regression, where the output is a numerical value focusing on understanding the relationship between dependent and independent variables.

Conversely, unsupervised learning algorithms analyze and cluster unlabeled datasets, aiming to find inherent patterns within the data. This type of machine learning primarily encompasses two categories: clustering, where the algorithm groups unlabeled data based on their similarities, and association, where the model uses various rules to discover relationships between variables in a dataset [Alpaydin, 2020].

2.3 Interpretable Classification

Classification is a type of problem addressed in supervised learning. Many algorithms can be used for classification tasks, such as Classification and Regression Trees, Support Vector Machines, and Artificial Neural Networks. These algorithms can be broadly divided into two categories: those that are interpretable and those that are not.

Non-interpretable methods, often called black boxes, do not provide insights into how a class is obtained from the input. Support Vector Machines and Artificial Neural Networks are examples of non-interpretable classifiers. In contrast, interpretable classifiers offer information on how a decision was made and which features were important in the task.

The lack of interpretability poses a significant challenge in real-world environments like healthcare. Allowing diagnosis or treatment decisions to be made by black-box models violates the principles of evidence-based medicine [London, 2019]. Experts utilizing a machine learning model for clinical decisions must be able to explain and justify their choices.

However, the most accurate algorithms are often the least interpretable. For example, Artificial Neural Network models [Abdullah et al., 2021] generally exhibit superior performance across many problems, yet they do not explain the decision path taken to generate an outcome. Consequently, developing a model with both high performance and strong interpretability remains a challenge in computer science but is highly valuable in fields such as healthcare.

In this work, we present an interpretable classifier based on Dempster-Shafer theory using the Gradient Descent algorithm (DSGD). This method can handle missing values and allows for the incorporation of expert knowledge to improve predictions.

2.4 Decision Trees

A decision tree [Breiman et al., 1984] is a method used for both classification and regression tasks, based on simple rules. This algorithm recursively splits data into subsets based on the rule that provides the most information gain. In classification tasks, impurity is typically measured by the Gini index or entropy, while in regression, it is usually measured by the mean squared error. The recursive splitting process builds a binary tree structure, where an inner node represents a rule and its branches represent the outcomes of that rule. If a sample does not meet a condition, it follows one branch; otherwise, it follows the other, depending on the specific implementation. Finally, the leaves in a decision tree represent the predicted values.

The tree structure is inherently interpretable because one can trace the path a sample takes to generate a result, thereby understanding which features contribute to the final prediction.

2.5 Dempster-Shafer Theory

The Dempster-Shafer Theory (DST) [Shafer, 1976] is a mathematical framework for reasoning with uncertainty and incomplete data. It is often considered a generalization of Bayesian theory because it explicitly accounts for uncertainty beyond classical Bayesian models.

Let X be the set of all possible states of a system, known as the frame of discernment. A mass assignment function, or simply mass, m , is a function that satisfies:

$$m : 2^X \rightarrow [0, 1], \quad m(\emptyset) = 0, \quad \sum_{A \subseteq X} m(A) = 1 \quad (1)$$

where A is a subset of X and \emptyset is the empty set. The term $m(A)$ can be interpreted as the exact likelihood of outcomes belonging to set A , and not any of its proper subsets. The knowledge about the system is considered to be encoded in these masses.

The plausibility metric is defined as the total amount of evidence that can support a particular outcome. Its formulation is as follows:

$$Pl_m(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad (2)$$

Dempster's Rule (DR) [Shafer, 2016] allows for combining multiple sources of evidence, each expressed by its mass assignment function, within the same frame of discernment. According to the rule, given two mass assignment functions m_1 and m_2 , a new combined mass assignment function m_c can be constructed using the following formula:

$$\begin{aligned} m_c(A) &= m_1(A) \oplus m_2(A) \\ &= \frac{1}{1 - K} \sum_{B \cap C = A \neq \emptyset} m_1(B)m_2(C) \end{aligned} \quad (3)$$

where K is a constant representing the degree of conflict between m_1 and m_2 , given by the expression:

$$K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \quad (4)$$

2.6 Classifier based on Dempster-Shafer Theory

The Dempster-Shafer Gradient Descent (DSGD) model is an interpretable machine learning method designed to balance classification accuracy with global interpretability [Peñafiel et al., 2020a]. Unlike traditional black-box models, DSGD leverages rule-based learning to provide transparency in decision-making while maintaining competitive classification performance. A key advantage of DSGD is its ability to effectively handle missing data, as it assigns mass functions to subsets of possible outcomes rather than requiring complete certainty about individual data points. This makes it particularly useful in real-world scenarios such as medical diagnosis and financial decision-making, where interpretability is crucial and data may be incomplete. For example, an application of the DSGD model in the medical field is presented in [Peñafiel et al., 2020b], where it is used to predict the occurrence of a stroke within one year, extensively utilizing the model's features for handling missing data and incorporating expert knowledge.

DSGD is built upon the Dempster-Shafer Theory and gradient descent optimization. Within this framework, rule-based mass assignment functions can be either defined by experts or derived from training data by partitioning attributes into possible categories. For continuous attributes, partitioning can be based on statistical measures such as the mean (μ) and standard deviation (σ). For instance, a 3-break division for an input feature space might categorize values into ranges as follows:

$$X \leq \mu - \sigma, \quad \mu - \sigma < X \leq \mu + \sigma, \quad X > \mu + \sigma,$$

Each partition is associated with an initial belief mass function reflecting evidence from the data.

During training, each rule, defined by a specific condition on the features, is refined by optimizing its corresponding mass values. This is achieved by comparing the predicted class—computed from the combined mass values of activated rules—to the actual class label using a loss function J . The mass values are then iteratively updated via gradient descent:

$$m_{t+1}^{(i)} = m_t^{(i)} - \alpha \frac{\partial J}{\partial m^{(i)}},$$

where α is the learning rate. A projection function π_C is subsequently applied to ensure that the updated mass assignments remain valid within the Dempster-Shafer framework:

$$m_{t+1}^{(i)} = \pi_C \left(m_t^{(i)} - \alpha \frac{\partial J}{\partial m^{(i)}} \right).$$

Once trained, the classifier applies these interpretable rules for prediction. A feature vector x is evaluated against the activated rules, which contribute their mass functions. These are then combined using Dempster's Rule:

$$m_f = \bigoplus_{m \in M_x} m.$$

The belief function for each class is then computed, and the final predicted class is determined by selecting the one with the highest belief:

$$\hat{y} = \arg \max Bel(m_f).$$

This process improves classification accuracy while maintaining transparency, allowing domain experts to validate and refine decisions.

Regarding interpretability, DSGD offers global interpretability. This is achieved because, once trained, the DSGD model can identify the rules that contribute most significantly to the prediction of a given class. The model generates a ranking by applying a strategy to assess the importance of each rule in the class prediction. As a result, it becomes possible to clearly determine which rules explain the model's predictions at a global level.

3 Method

This section describes the data utilized and the methodology employed for model development and rule generation.

3.1 Data Description

The dataset included HER2-positive early breast cancer patients diagnosed and treated in 8 hospitals (from Portugal, Spain, and Chile) between January 2018 and December 2021. The study enrolled patients who received neoadjuvant therapy with dual HER2 blockade (trastuzumab and pertuzumab), followed by surgery. Patients with disease progression during neoadjuvant therapy were excluded.

The clinical and histopathological characteristics considered were: age, menopausal status, T and N values from the clinical TNM (cTNM) staging system, initial disease stage, tumor grade, presence of estrogen receptor (ER) and progesterone receptor (PR) overexpression, and Ki67 proliferative index.

The cTNM classification system is a framework for cancer staging, describing the anatomical extent of cancer based on three key components: T (tumor size), N (lymph node involvement), and M (presence of metastasis) [Cserni et al., 2018].

Finally, the primary outcome variable to predict was whether or not a pathological complete response (pCR) was achieved after the combined neoadjuvant therapy.

The study included a total of 390 patients from the 8 hospitals, with a mean age of approximately 53 years. Of these, 65.1% had hormone receptor-positive tumors and 54.9% presented regional lymph node involvement. In terms of chemotherapy backbone, 59.7% of the patients had received an anthracycline-based regimen. Overall, pCR was achieved in 56.7% of the patients.

The dataset consists of two types of attributes: categorical and numerical (continuous). Table 1 presents the summary statistics for numerical variables, while Table 2 summarizes the distribution of categorical variables. These tables provide an overview of the dataset, detailing the frequency distribution of categorical variables and highlighting the central tendency, dispersion, and missing values for numerical variables.

Attribute	Description	Mean	Std Dev	Median	Min - Max	Missing values (%)
Age	Age of the patient at treatment	52.85	11.18	52.5	22.7 - 84.4	0.00
ER	Estrogen receptor expression (%)	46.41	45.00	45.0	0.0 - 100.0	28.21
PR	Progesterone receptor expression (%)	26.81	36.43	1.0	0.0 - 100.0	27.95
Ki67	Ki67 protein expression (%)	43.00	22.39	40.0	0.2 - 90.0	28.97

Table 1: Description of numerical continuous variables in the dataset

As previously explained, rule generation depends on the type of variable. For categorical variables, a new rule is created for each category, excluding null values. For example, some rules derived from categorical attributes are “*Stage = III*” and “*cT = T2*”.

In the case of continuous variables, two approaches were used: For biomarker expressions such as ER, PR, and Ki67, experts generally agree that overexpression is

Attribute	Description	Category	Frequency	(%)
Menopausal status	Indicates the menopausal condition at the treatment	Peri-postmenopausal	156	40.0
		Premenopausal	114	29.2
		<i>null</i>	120	30.8
cT	"T" value of the clinical TNM staging system	T2	206	52.8
		T3	86	22.1
		T4	45	11.5
		T1	41	10.5
		<i>null</i>	12	3.1
cN	"N" value of the clinical TNM staging system	N0	176	45.1
		N1	169	43.3
		N2	21	5.4
		N3	12	3.1
		<i>null</i>	12	3.1
Stage	Cancer stage grouping	II	232	59.5
		III	125	32.1
		I	24	6.2
		<i>null</i>	9	2.3
Grade	Grade of the tumoral cells	2	132	33.8
		3	121	31.0
		1	5	1.3
		<i>null</i>	132	33.8
pCR	Pathological complete response. Target value.	Yes	221	56.7
		No	169	43.3

Table 2: Description of categorical variables in the dataset

defined as values exceeding 20%. Consequently, a rule is defined for overexpression (e.g., $ER \geq 20\%$) and another for normal expression (e.g., $ER < 20\%$). Regarding the age attribute, the model is allowed to determine four breakpoints through statistical analysis, using the mean and standard deviation of the distribution.

All the aforementioned rules apply to creating single-attribute rules. For this experiment, two-attribute rules for biomarker expression are also included. These rules are generated by combining all possible pairs of the previously defined single-attribute rules. For example, one rule could specify overexpressed PR while maintaining normal expression in ER, and it will be written as $PR \geq 20\%$ and $ER < 20\%$.

3.2 Problem Statement

The challenge addressed in this work is to determine whether or not a patient with HER2-positive cancer, treated with the described neoadjuvant therapy, will achieve a pathological complete response after treatment, based on their clinical and histopathological variables. This problem can be understood as a binary classification task where the input is the patient's clinical information and the output is one of two options: the patient will achieve pCR or the patient will not achieve pCR.

3.3 Model

The DSGD model by Peñafiel *et al.* [Peñafiel et al., 2020a] demonstrates performance comparable to other methods such as Support Vector Machine (SVM) or the K-nearest neighbors algorithm (KNN). Furthermore, the DSGD algorithm allows for both manual definition and automatic generation of rules. In this work, some basic rules were added manually. The model was trained using Mean Squared Error (MSE) as the loss function and the Adam algorithm [Kingma, 2017] as the optimizer, with a learning rate of 0.0015.

4 Results

4.1 Model Performance

To evaluate the model's performance and prevent overfitting, we employed 3-fold cross-validation, as described in [Domingos, 2012]. This technique involves training and testing the model on different subsets of the available data to assess its generalization ability. In our study, each fold used approximately 66.7% of the data for training and 33.3% for testing. We repeated this process three times, each time using a different portion for testing, thereby obtaining a more reliable estimate of the model's performance.

Table 3 presents the mean performance of the DSGD model across the three experiments, while Figure 1 displays the aggregate confusion matrix obtained from all experiments.

Indicator	Value
Accuracy	0.61
AUC ROC	0.61
F1-score	0.69
Precision	0.63
Recall	0.76

Table 3: Performance of the DSGD model

To compare performance, we conducted a 3-fold cross-validation experiment with other classification algorithms. Their hyperparameters were tuned using the GridSearchCV technique [Géron, 2017], which systematically searches for the optimal set of hyperparameters yielding the best performance.

For the DSGD model, we fine-tuned several hyperparameters through empirical testing and leveraging prior knowledge of the dataset. The final configuration was:

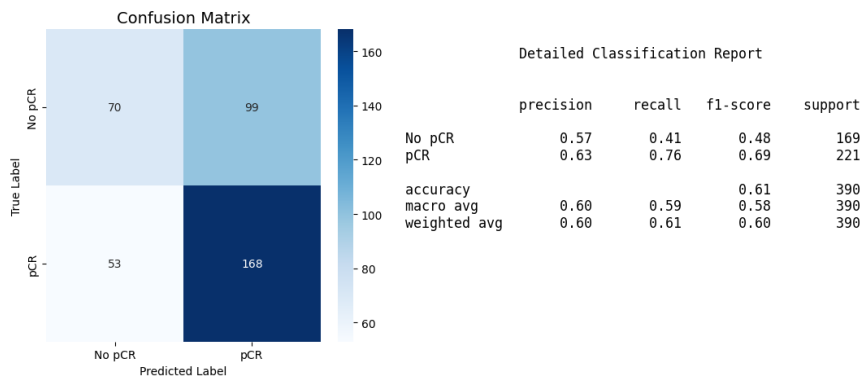


Figure 1: Confusion Matrix and detailed classification report of all 3-fold experiments

- **Initial learning rate (lr):** 0.005
- **Minimum number of epochs (min_iter):** 1
- **Maximum number of epochs (max_iter):** 200
- **Minimum variation of loss to consider convergence (min_dloss):** 0.00001
- **Number of workers:** 4
- **Loss function (lossfn):** Mean Squared Error (MSE)
- **Optimization method (optim):** Adam
- **Precompute rules:** Enabled

These parameters were set after multiple experiments to ensure the model's convergence and stability. This setup provided the most stable and generalizable results across the three cross-validation folds.

For the Decision Tree, we explored criterion = ['gini', 'entropy'] and max_depth = [1, 3, 5, 10], selecting entropy and depth 3 as optimal. For KNN, we tested K = [3, 5, 7, 11, 15, 21, 25], with K=21 yielding the best performance. For SVM, we evaluated polynomial kernels with degree = [1, 2, 3], choosing degree 1. For XGBoost, we tuned booster = ['gbtree', 'dart'], max_depth = [3, 5, 7], and learning_rate = [0.01, 0.1, 0.2], selecting gbtree with default learning rate and depth 3. The performance metrics shown correspond to the best configuration found for each model.

Table 4 presents the performance of the DSGD model alongside other models, with metrics derived from the mean of the 3-fold experiments. Additionally, Figure 2 displays the ROC curves of the models, illustrating the trade-off between the true positive rate (sensitivity or recall) and the false positive rate (1-specificity) across different classification thresholds.

From the results, we can observe that the DSGD model exhibits moderate performance, likely due to the limited and imbalanced nature of the dataset. However, for this specific problem, it has demonstrated one of the best performances in terms of both accuracy and ROC curve area when compared to other models. Its metrics are even

Model	Accuracy	AUC ROC	F1-score
DSGD	0.61	0.61	0.69
Decision Tree	0.52	0.53	0.54
KNN	0.52	0.49	0.64
SVM	0.56	0.56	0.72
XGBoost	0.56	0.59	0.60

Table 4: Performance of all methods applied for predicting the pCR of patients

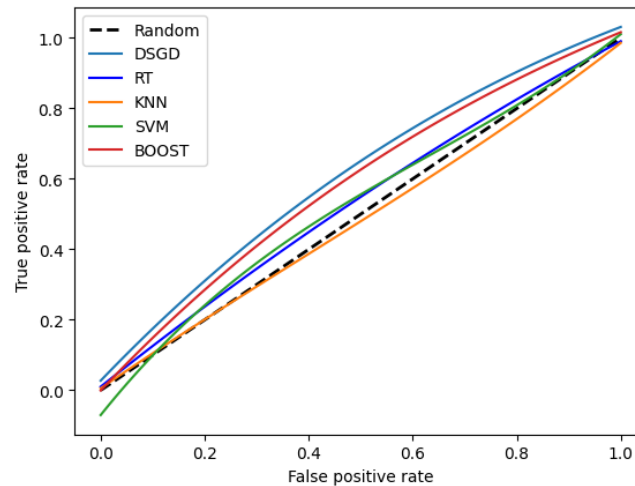


Figure 2: Average ROC curves for the methods used in the problem

comparable to a top-performing binary classifier, such as SVM, but with the added advantage of interpretability.

Figure 1 also provides various metrics for assessing the model's performance in classifying each class. Given the class imbalance, using the F1-score is preferable for comparing performance. The F1-score for pCR is approximately 0.7, while for no pCR, it is approximately 0.5. This indicates that the model demonstrates better classification performance for patients achieving pCR.

4.2 Generated Rules

In addition to its performance, the model is capable of identifying the most important rules for predicting whether a patient will achieve a pathological complete response after neoadjuvant treatment.

During the training phase, the masses of rules were adjusted to minimize prediction error. Mass values are assigned to each individual rule for each possible outcome: no pCR, pCR, and uncertainty. The mass measures the contribution of a single rule to a specific outcome.

Table 5 shows the most important rules for predicting the class of patients achieving pCR (Mass pCR) and those who do not (Mass No pCR).

Rule	Mass No pCR	Mass pCR	Uncertainty
$ER \geq 20\%$ and $PR \geq 20\%$	0.279	0.000	0.721
$Ki67 < 20\%$ and $ER \geq 20\%$	0.259	0.000	0.741
Grade 1	0.231	0.000	0.769
$ER \geq 20\%$	0.206	0.000	0.794
Pre-menopausal	0.000	0.244	0.756
Stage III	0.000	0.206	0.794
cN = cN2	0.000	0.206	0.794
cT = cT1	0.000	0.204	0.796

Table 5: Most important rules for the 'No pCR' and 'pCR' classes, and their uncertainty

The model suggests (Table 5) that patients with higher proportions of estrogen and progesterone receptors, lower levels of the Ki67 index, and a lower grade of breast cancer are less likely to achieve a pathological complete response to treatment.

Conversely, the rules presented in Table 5 indicate that premenopausal patients and those in an advanced, non-metastatic stage (Stage III) are more likely to achieve pCR. Furthermore, the results suggest a tendency towards achieving pCR after treatment in patients with cN2 and cT1 in the cN and cT categories, respectively.

4.3 Decision Tree Comparison

The Decision Tree method was one of the worst-performing models in the group; however, it is, like the DSGD model, another directly interpretable method. Figure 3 shows the best resulting Decision Tree (0.55 accuracy) from the 3-fold experiments.

Each node in the tree represents a decision rule applied to a feature threshold. The node displays the entropy value (H), the proportion of the samples used to generate the split (N), the predicted class (pCR or No pCR), and its corresponding confidence percentage.

The nodes are color-coded according to their predicted class: orange for 'No pCR' and blue for 'pCR', where darker colors indicate a higher confidence level in the prediction. If the decision rule at a node is *true*, the algorithm follows the left branch; otherwise, it follows the right branch. For example, in the root node, the decision rule $PR \leq 20\%$ determines the first split. If $PR \leq 20\%$, the left branch is taken; otherwise, the right branch is followed.

The entropy (H) and the proportion of samples (N) at each node reveal how uncertainty evolves across the levels of the decision tree. At the root node, high entropy indicates significant class uncertainty. The first split ($PR \leq 20\%$) does not strongly differentiate the classes, as both child nodes retain high entropy.

At different levels, certain splits are more effective. For example, $Age \leq 46.8$ significantly reduces entropy ($H = 0.544$), while $Ki67 \leq 20\%$ also improves class separation ($H = 0.966$). In contrast, some branches, such as $Age \leq 60.95$ ($H = 0.966$), maintain high uncertainty.

In the leaf nodes, the entropy reaches $H = 0$ in two cases, indicating perfect separation, but these nodes contain very few samples ($N < 3\%$), suggesting possible overfitting.

This analysis highlights that age and Ki67 contribute to better class separation, while some splits remain uncertain, requiring further refinement. The results show similar

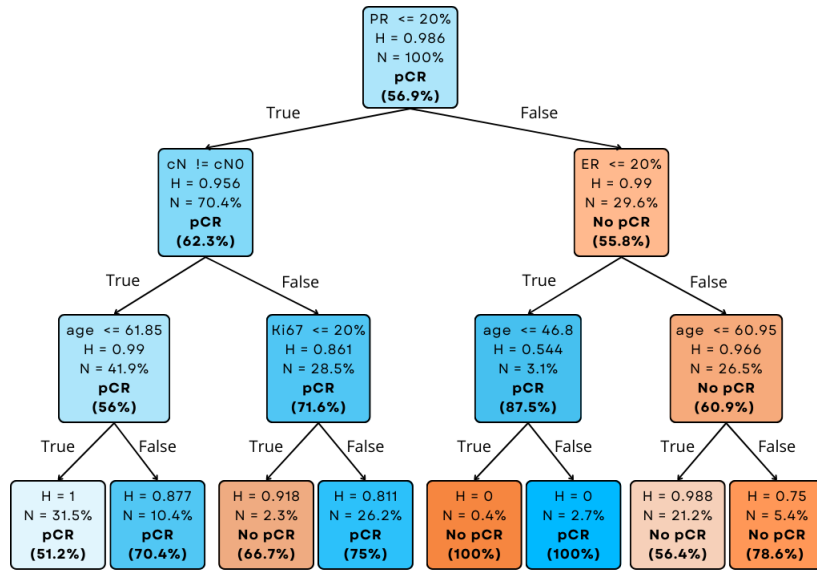


Figure 3: Decision Tree results for the problem

rules that can be read in Table 5. For example, patients with low levels of progesterone, estrogen receptor, and a high presence of Ki67 protein tend to show pCR. However, the DSGD model does not directly generate rules about the age of patients.

4.4 Comparison with Other Predictive Approaches for pCR

It is important to note that while various studies explore predictors of pathological complete response (pCR), direct comparison of results is often challenging due to differences in patient cohorts, cancer subtypes, and analytical methodologies. Nevertheless, examining related research provides valuable context for our findings.

The endeavor to predict pCR in breast cancer patients undergoing neoadjuvant therapy has been approached through various methods. For instance, as mentioned previously, [Saffie Vega et al., 2022] conducted a retrospective statistical analysis on 94 women with HER2-overexpressing breast cancer in Chile who received neoadjuvant combination chemotherapy with Trastuzumab and Pertuzumab. Their study focused on identifying factors affecting therapeutic response rather than proposing a predictive classifier, partly because the limited dataset size constrained the inclusion of more variables in their regression models. Their findings indicated that the absence of hormone receptor expression is a strong predictor of achieving pCR; specifically, patients with HER2/Luminal (hormone receptor-positive) breast cancer had an 8.15 times higher odds of not achieving a complete pathological response compared to those with a pure HER2 profile. Additionally, they reported that an increased HER2 copy number and a higher HER2/centromere ratio, determined by FISH, were both significantly associated with a greater likelihood of achieving pCR.

In a different context, [Mastrantoni et al., 2023] centered their research on a distinct patient cohort—those with hormone-receptor (HoR)-positive/HER2-negative early breast cancer—and employed machine learning models to predict pCR following neoadjuvant chemotherapy. Although their dataset pertains to a cancer subtype different from our HER2-positive focus, making direct comparisons of predictive factors or pCR rates inappropriate, their work is relevant because it demonstrates the application of similar machine learning strategies (including Random Forest, k-nearest neighbor, and support vector machines) in the broader field of pCR prediction. This use of analogous computational techniques helps to validate the general approach of using machine learning to identify potential responders to neoadjuvant treatment. Mastrantoni et al. identified Ki67, ER/PgR status, age, and nodal status as the variables carrying the highest importance in their models for HoR-positive/HER2-negative breast cancer.

5 Discussion

This section discusses the main findings of the DSGD model, highlighting its performance, interpretability, and comparison with existing literature.

5.1 Model Performance and Potential Improvements

The DSGD model generally outperforms other models in terms of predictive performance within this context. However, there remains room for improvement. One approach to enhance its performance would be to incorporate a larger dataset, as increased data volume and a greater number of study cases have been shown to improve machine learning algorithm efficacy [Domingos, 2012]. Additionally, integrating more detailed features, such as those describing HER2 gene amplification in breast cancer, including the HER2/centromere ratio and HER2 copy number, which have demonstrated a predictive role for pCR in prior studies [Arnould et al., 2007], could further improve the model's accuracy.

5.2 Interpretability of the Model

A notable feature of the presented model is its inherent interpretability, a crucial aspect for problems of this nature in healthcare. The model's transparency allows for easier comprehension of its underlying decision rules, thereby facilitating the practical application of this knowledge by clinicians.

5.3 Rules and Comparison with Literature

The rules generated by the trained model offer several interesting insights for discussion. One of the most significant findings is that patients with both estrogen and progesterone receptor-positive tumors are less likely to achieve pCR after treatment, as illustrated in Table 5. Consistent with the discussion by Saffie et al. [Saffie Vega et al., 2022], the presence of hormone receptor positivity indicates more luminal tumor characteristics. Moreover, the heterogeneity of HER2/Luminal breast cancer is known to be a predictor of lower response to neoadjuvant therapies. Further supporting these findings, Schneeweiss et al. [Schneeweiss et al., 2013] demonstrated a higher pCR rate in patients with hormone

receptor-negative tumors compared to those with hormone receptor-positive tumors. Thus, the results obtained align well with prior research.

Another intriguing rule indicates that patients with $Ki67 < 20\%$ and $ER \geq 20\%$ are less likely to achieve a pCR. This feature is also discussed in [Saffie Vega et al., 2022], where it is noted that patients with $Ki67 > 20\%$ are more likely to achieve a pCR compared to those with $Ki67 < 20\%$. This aligns with the findings of [Mastrantoni et al., 2023], which identified $Ki67$ as one of the most important attributes for prediction, thereby validating this observation. Furthermore, it is well-established that tumors exhibiting high $Ki67$ expression are generally more aggressive and show a higher responsiveness to both cytotoxic and targeted treatments [Luporsi et al., 2012].

In addition to these, Table 5 suggests that low-grade or well-differentiated tumors are less likely to achieve a pathological complete response to the treatment. While supporting evidence for this specific observation is limited, it can be hypothesized that the lack of pCR in low-grade tumors may be attributed to their less aggressive nature, resulting in a diminished response to the specific neoadjuvant therapy.

Table 5 also indicates that cancer in more advanced stages tends to exhibit pCR, which is consistent with the findings of Saffie et al. [Saffie Vega et al., 2022]. This result can be explained by considering that more advanced cancers are often less luminal and more proliferative, leading to a higher likelihood of achieving pCR.

5.4 Controversial Findings and Limitations

One rule presented suggests that smaller tumors (cT1) tend to show a pathological complete response after neoadjuvant treatment. It is important to clarify that while tumor size is a factor in cancer staging, it is not the sole determinant. Other variables and criteria are also used to assess tumor stage. Despite the model indicating tumor size as an important factor in response to this type of treatment, existing literature generally contradicts this notion, stating that tumor size is not a reliable predictor of pCR [Livingston-Rosanoff et al., 2019, Houvenaeghel et al., 2022].

Similarly, another important rule suggests that women in a premenopausal state tend to have a pathological complete response to neoadjuvant treatment. This rule raises controversy, as there is currently insufficient evidence in the broader literature to support it unequivocally. Unfortunately, the available dataset does not allow for a detailed analysis of the molecular profile of the cancer to establish a direct connection between premenopausal states and less luminal tumor types. Consequently, further studies are required to investigate any potential relationship between premenopausal states and pCR in neoadjuvant treatments.

5.5 Implications for Further Research

Finally, some of the less directly explainable rules highlight one of the most valuable aspects of interpretability within the utilized model. The presence of statements that lack an immediate clear interpretation can serve as guidance and motivation for further research in the field. This characteristic renders the model a valuable tool for knowledge discovery, prompting deeper investigation into underlying biological mechanisms.

6 Conclusions

In this study, we presented a novel model for predicting the effectiveness of neoadjuvant therapy based on clinical and histopathological variables. Our model outperformed all

other tested models, demonstrating its superior ability to accurately predict pathological complete response (pCR). Nevertheless, there is still potential for improvement, particularly by incorporating more patient data and integrating new features.

The rules derived from the training process reveal the impact of certain variables on pCR in breast cancer patients. Some of these variables, such as hormone receptor expression and cancer stages, are directly supported by previous research as pCR predictors. Other rules, like the association with premenopausal status or certain cancer grades, lack clear explanations in existing literature. Therefore, further data and analysis are required to establish a definitive correlation for these. Nevertheless, all the rules consistently indicate that the heterogeneity of breast cancer serves as an indicator of a lower likelihood of achieving pCR.

Ethical Statement

Informed consent was obtained from all participants prior to their participation in the study.

References

- [Abdullah et al., 2021] Abdullah, T. A. A., Zahid, M. S. M., and Ali, W. (2021). A review of interpretable ml in healthcare: Taxonomy, applications, challenges, and future directions. *Symmetry*, 13(12):2439.
- [Alpaydin, 2020] Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press.
- [Arnould et al., 2007] Arnould, L., Arveux, P., Couturier, J., Gelly-Marty, M., Loustalot, C., Ettore, F., Sagan, C., Antoine, M., Penault-Llorca, F., Vasseur, B., Fumoleau, P., and Coudert, B. P. (2007). Pathologic Complete Response to Trastuzumab-Based Neoadjuvant Therapy Is Related to the Level of HER-2 Amplification. *Clinical Cancer Research*, 13(21):6404–6409.
- [Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.
- [Cserni et al., 2018] Cserni, G., Chmielik, E., Cserni, B., and Tot, T. (2018). The new tm-based staging of breast cancer. *Virchows Arch*, 472:697–703.
- [Domingos, 2012] Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87.
- [Géron, 2017] Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media.
- [Houvenaeghel et al., 2022] Houvenaeghel, G., de Nonneville, A., Cohen, M., Viret, F., Rua, S., Sabiani, L., Buttarelli, M., Charaffe, E., Monneur, A., Jalaguier-Coudray, A., Bannier, M., Sabatier, R., and Gonçalves, A. (2022). Neoadjuvant chemotherapy for breast cancer: Evolution of clinical practice in a french cancer center over 16 years and pathologic response rates according to tumor subtypes and clinical tumor size: Retrospective cohort study. *J Surg Res (Houst)*, 5(3):511–525.
- [Hännikäinen, 2023] Hännikäinen, E. N., M. J. . K. P. (2023). Predictors of successful neoadjuvant treatment in her2-positive breast cancer. *Oncology letters*, 26:434.
- [Katayama, 2021] Katayama, A., M. I. S. S. (2021). Predictors of pathological complete response to neoadjuvant treatment and changes to post-neoadjuvant her2 status in her2-positive invasive breast cancer. *Modern Pathology*, 34:1271–1281.
- [Kingma, 2017] Kingma, Diederik P., J. B. (2017). Adam: A method for stochastic optimization.

- [Livingston-Rosanoff et al., 2019] Livingston-Rosanoff, D., Schumacher, J., Vande Walle, K., Stankowski-Drengler, T., Greenberg, C. C., Neuman, H., and Wilke, L. G. (2019). Does tumor size predict response to neoadjuvant chemotherapy in the modern era of biologically driven treatment? a nationwide study of us breast cancer patients. *Clin Breast Cancer*, 19(6):e741–e747.
- [Loibl and Gianni, 2017] Loibl, S. and Gianni, L. (2017). Her2-positive breast cancer. *The Lancet*, 389(10087):2415–2429.
- [London, 2019] London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1):15–21.
- [Luporsi et al., 2012] Luporsi, E., André, F., Spyrtos, F., and et al. (2012). Ki-67: level of evidence and methodological considerations for its role in the clinical management of breast cancer: analytical and critical review. *Breast Cancer Research and Treatment*, 132(3):895–915.
- [Mastrantoni et al., 2023] Mastrantoni, L., Garufi, G., Di Monte, E., Arcuri, G., Frescura, V., Maliziola, N., Rotondi, A., Giordano, G., Carbognin, L., Fabi, A., et al. (2023). 276p development and external validation of an artificial intelligence (ai)-based machine learning model (ml) for predicting pathological complete response (pCR) in hormone-receptor (hor)-positive/her2-negative early breast cancer (ebc) undergoing neoadjuvant chemotherapy (nct). *Annals of Oncology*, 34:S294.
- [Molnar, 2025] Molnar, C. (2025). *Interpretable Machine Learning*. 3 edition.
- [Peñafiel et al., 2020a] Peñafiel, S., Baloian, N., Sanson, H., and Pino, J. A. (2020a). Applying dempster–shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Systems with Applications*, 148:113262.
- [Peñafiel et al., 2020b] Peñafiel, S., Baloian, N., Sanson, H., and Pino, J. A. (2020b). Predicting stroke risk with an interpretable classifier. *IEEE Access*, 9:1154–1166.
- [Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.
- [Russell and Norvig, 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson Education.
- [Saffie Vega et al., 2022] Saffie Vega, I., Sapunar Zenteno, J., Buscaglia Fernandez, F., Reyes Cosmelli, F., Lagos Chavez, R., and Chahuán Manzur, B. (2022). Predictors of pathologic complete response to neoadjuvant treatment in her2-overexpressing breast cancer: a retrospective analysis using real-world data. *Ecancermedicalscience*, 16:1338.
- [Schneeweiss et al., 2013] Schneeweiss, A., Chia, S., Hickish, T., Harvey, V., Eniu, A., Hegg, R., Tausch, C., Seo, J., Tsai, Y.-F., Ratnayake, J., McNally, V., Ross, G., and Cortés, J. (2013). Pertuzumab plus trastuzumab in combination with standard neoadjuvant anthracycline-containing and anthracycline-free chemotherapy regimens in patients with her2-positive early breast cancer: a randomized phase ii cardiac safety study (tryphaena). *Annals of Oncology*, 24(9):2278–2284.
- [Shafer, 1976] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.
- [Shafer, 2016] Shafer, G. (2016). Dempster’s rule of combination. *International Journal of Approximate Reasoning*, 79:26–40. 40 years of Research on Dempster-Shafer Theory.
- [Sung et al., 2021] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.
- [Waks and Winer, 2019] Waks, A. G. and Winer, E. P. (2019). Breast Cancer Treatment: A Review. *JAMA*, 321(3):288–300.