Deep learning for residual disease stratification in early Breast Cancer

Louis-Oscar Morel[1]
Supervised by Pr. Sylvain Ladoire[2], Pr. Jens Rittscher[3] and Pr. Simon Lord[4]

Thesis submitted for the degree of
Doctor of Philosophy in Cancer Science

Linacre College
University of Oxford
October 2022 - October 2025

1. Cancer Research UK Oxford Centre, Medical Sciences Division, University of Oxford, Oxford OX3 7DQ, UK
2. Department of Medical Oncology, Georges François Leclerc Cancer Centre, Dijon, France
3. Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, OX3 7DQ, UK
4. Department of Oncology, University of Oxford, UK (S.L.), Oxford, OX3 7DQ, UK

Abstract

Breast cancer remains the leading cause of cancer-related deaths among women globally. Despite advances in treatment, the decision-making process for administering neoadjuvant systemic chemotherapy (NAC) still poses significant challenges. This research aims to improve the characterisation of residual disease (RD) in surgical specimens of breast cancer patients following NAC. Utilising a deep learning approach to analyse histopathological information from Whole Slide Images (WSIs), this study seeks to predict overall survival (OS) and disease-free survival (DFS) more accurately. While RD is a known prognostic factor, its presence does not unequivocally predict patient outcomes, as evidenced by variations in patient response to NAC. Our study will focus on the PRIMUNEO dataset, encompassing 500 patients from various cancer centres, and will later extend to the CGFL dataset for external validation. The goal is to enhance the stratification of breast cancer treatment, particularly in the post-neoadjuvant setting, considering the direct assessment of chemosensitivity and the need for potential treatment adjustments based on RD.

Overall number of words (exclusive of bibliography, appendices, diagrams and tables): 43,939

Acknowledgements

This work would not have been possible without the guidance, generosity, and trust of many people to whom I am deeply indebted.

I am profoundly grateful to Professor Jens Rittscher, my supervisor, for his scientific mentorship and steady encouragement. His clarity of thought and engineering perspective helped me simplify my analyses in the most meaningful way, and his numerous ideas on artificial intelligence shaped the direction of this work far beyond its original scope.

I also wish to thank Professor Simon Lord, who supported this project at its inception and provided valuable clinical insight and encouragement at key moments of its development.

My deepest gratitude goes to Professor Sylvain Ladoire, my French supervisor, mentor, and friend. For more than five years, since my master's studies, he has shared his time, his data, his confidence, and his unwavering belief in this project. His scientific integrity and friendship have guided every stage of this journey, and I owe him more than words can convey.

I warmly thank the pathologists who made this work possible: Dr Laurent Arnould, Dr Carlo Pescia, and my father, Dr Henri Philippe Morel, for their passion, precision, and the countless hours they devoted to reviewing and discussing the histological patches.

To my colleagues and friends from the Dijon team, thank you for your teamwork, your humour, and your resilience. Nothing would have been achieved without your daily commitment and collective spirit.

I am especially grateful to my godfather, Pierre-Olivier, for his friendship, advice, and the many moments of laughter and support we shared throughout this long journey. His constant encouragement and perspective offered both grounding and joy when they were most needed.

Finally and above all, I wish to thank my wife, whose love and patience have accompanied me through this long and demanding journey that began with my seven-year academic hiatus.

We met in the midst of it, built our life together, and were blessed with our wonderful daughter, Hélène, who fills my heart with joy and meaning. To both of them, I dedicate this thesis.

Declaration of the Use of AI-Assisted Tools

Portions of this thesis (including text editing, grammar improvement, and non-scientific language refinement) were assisted by artificial intelligence tools. All scientific content, research design, data analysis, interpretation, and conceptual work presented in this thesis are entirely my own. AI tools were not used to generate novel scientific ideas, analysis pipelines, or results.

I take full responsibility for the accuracy, originality, and integrity of all scientific content presented.

Glossary of Abbreviations

Abbreviation	
ADC	Antibody-Drug Conjugate
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
AUC	Area Under the Curve
BCE	Binary Cross-Entropy
CDK4/6	Cyclin-Dependent Kinases 4 and 6
CGFL	Centre Georges-François Leclerc
CI	Confidence Interval
CNIL	Commission Nationale de l'Informatique et des Libertés
CPS+EG	Clinical-Pathologic Stage + Estrogen receptor status and Grade
ctDNA	Circulating tumour DNA
CTC	Circulating Tumour Cells
DCNN	Deep Convolutional Neural Network
DFS	Disease-Free Survival
DL	Deep Learning
DNA	Deoxyribonucleic Acid
eBC	Early Breast Cancer
EBCTCG	Early Breast Cancer Trialists' Collaborative Group
ECIBC	European Commission Initiative on Breast Cancer
ER	Estrogen Receptor
ESMO	European Society for Medical Oncology
H&E	Haematoxylin and Eosin
HER2	Human Epidermal Growth Factor Receptor 2
HR	Hormone Receptor
HRD	Homologous Recombination Deficiency

IHC	Immunohistochemistry
MIL	Multiple Instance Learning
ML	Machine Learning
NAC	Neoadjuvant Chemotherapy
NGS	Next-Generation Sequencing
NN	Neural Network
OS	Overall Survival
PAM50	Prediction Analysis of Microarray 50
pCR	Pathological Complete Response
PCR	Polymerase Chain Reaction
PgR	Progesterone Receptor
RCB	Residual Cancer Burden
RF	Random Forest
SBR	Scarff-Bloom-Richardson (grading system)
SSL	Self-Supervised Learning
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TILs	Tumour-Infiltrating Lymphocytes
TNM	Tumour-Node-Metastasis (staging system)
TNBC	Triple-Negative Breast Cancer
WSI	Whole Slide Image

Contents

Chapter 1: Introduction and literature review.

1.1 Overview of the thesis

Despite major advances in systemic therapy, residual disease after neoadjuvant chemotherapy (NAC) remains one of the most powerful predictors of poor outcome in early breast cancer. Patients achieving pathological complete response (pCR) exhibit excellent long-term survival, whereas those with residual invasive disease face a substantially increased risk of recurrence and death, even under optimal systemic therapy.

However, the mechanisms that determine why some tumours respond completely while others persist remain poorly understood and identified. Current post-NAC stratification frameworks, based on residual cancer burden (RCB) or tumour cellularity, capture only a part of the prognosis. At the same time, access to molecular testing (e.g. genomic or transcriptomic assays) remains limited or delayed in many clinical settings. Together, these limitations underscore a major unmet need: the ability to extract reliable, biologically meaningful prognostic information directly from routine histology, both before and after treatment.

This thesis develops and evaluates a set of complementary deep-learning frameworks that connect morphology to molecular status, treatment response, and post-therapy risk in early breast cancer.

Across the different parts of this work, we move from molecular inference (mutation and signature prediction) to therapeutic prediction (chemosensitivity) and finally to outcome modelling (residual disease and survival), building an integrative computational framework for post-NAC risk stratification.

1.1.1 Thesis structure

The thesis is organised in five chapters:

- **Chapter I - Introduction and Literature review**: introduces the clinical problem and some background, and then provides some details on the contributions from this work.
- **Chapter II - Clinical Cohorts and Data Acquisition**: provides the technical framework and describes the datasets (PRIMUNEO, CGFL Neoadj, TCGA), preprocessing pipelines, and computational infrastructure used throughout the thesis.

It details the acquisition of whole-slide images (WSIs), the annotation and quality-control procedures , and the design of reproducible train–validation–test splits.

- **Chapter III – Predicting survival from post-NAC surgical histology**: This chapter establishes the methodological foundation for the entire thesis by systematically evaluating different survival-modelling strategies on post-NAC surgical specimens.

  We compared classical survival models (Cox proportional hazards, Random Survival Forests, Survival Support Vector Machine) with deep-learning approaches such as DeepSurv, DeepHit, and end-to-end Vision Transformer (ViT) pipelines, quantifying their predictive performance and robustness on internal and external cohorts.

  The results revealed a consistent pattern: while complex multi-task architectures achieved high apparent performance on controlled datasets like TCGA (Tumor Cancer Genome Atlas), their generalisability collapsed when applied to real-world data.

  In contrast, simpler pipelines like foundation-model embeddings coupled with Cox regression proved to be more stable, interpretable, and reproducible. In computational pathology, parsimony and methodological transparency often outperform architectural sophistication.

- **Chapter IV – Predicting gene and variant mutations from histology:** Building on previous work published in Scientific Reports (*Morel et al.2023*) before this thesis, we extended the gene mutation prediction pipeline using histology to a more clinically relevant solution. This chapter introduces MultiVarNet, a novel mutation variant-aware deep-learning architecture designed to predict not only gene-level mutation status but also specific protein-level variants (e.g., KRAS p.G12C, BRAF p.V600E, IDH1 p.R132H, PIK3CA p.E545K). Cross-cancer transfer experiments and domain-adversarial (multi-DA) training revealed, however, that most morpho-molecular signatures are tissue-specific and context-dependent, failing to generalise beyond their originating cancer type except in shared lineages such as LGG/GBM for IDH1 gene mutations.

  These findings delineate the boundaries of morpho-molecular generalisation, suggesting that histology-based AI may ultimately evolve toward companion diagnostic specificity where each model is validated for a particular biomarker, tissue, and acquisition setting.

- **Chapter V – From morphology to therapy: predicting chemosensitivity and post-treatment prognosis.** Chapter V translates the molecular and methodological insights into an analytical validation for NAC response prediction and post-treatment stratification. Two complementary frameworks are presented:

  Chemo-prAIdict Breast, trained on diagnostic biopsies, predicts intrinsic chemosensitivity to anthracycline–taxane regimens. In TNBC, the model identifies patients likely to respond to standard chemotherapy alone, supporting de-escalation from carboplatin or pembrolizumab; conversely, low predicted chemosensitivity may guide escalation or trial enrolment. In luminal ER+/HER2− tumours, poor predicted response could inform endocrine-based strategies, while in HER2+ disease, it may influence early intensification or sequencing of dual HER2 blockade.

  A post-NAC residual-disease network, applied to surgical slides, predicts disease-free survival (DFS) and overall prognosis. High-risk predictions correspond to residual invasive carcinoma, necrosis, and high-grade cytology; low-risk profiles reflect fibrotic regression and minimal residual tumour.

- **Chapter VI - General discussion**. The final chapter synthesises the conceptual and methodological results of this work. It emphasises that while tumour morphology does encode actionable biological information, its predictive power is inherently context-specific, limited by data provenance, cohort heterogeneity, and evolving therapeutic standards.

  The discussion argues that digital pathology is transitioning toward a companion-diagnostic era, where models are validated for specific clinical tasks, scanners, and contexts rather than as universal predictors.

  It also outlines future research directions: large-scale multicentric validation, multimodal integration with spatial transcriptomics and proteomics, uncertainty quantification, and the design of transparent regulatory pathways for clinical AI.

  Finally, it positions biologically informed supervision, parsimony, and interpretability as the key pillars for building trustworthy, clinically relevant AI in pathology.

1.1.2 What this work adds

This thesis develops a framework for predicting molecular alterations, treatment response, and post-therapy outcomes directly from routine histology. Its contributions lie in showing how morphology-based deep learning can complement existing diagnostic and therapeutic pathways in precision oncology.

Mutation variant-level modelling: connecting morphology to therapeutic actionability

Most prior image-based mutation-prediction studies treat gene-level mutation status as a single binary endpoint, overlooking the clinical reality that therapeutic decisions depend on specific variants rather than genes as a whole.

In this thesis, the *MultiVarNet* framework introduces variant-aware supervision, decomposing each gene into its protein-specific alterations (e.g., *KRAS p.G12C*, *BRAF p.V600E*, *IDH1 p.R132H*, *PIK3CA p.E545K*). This fine-grained approach reveals that morphological correlates are stronger when training aligns with functionally distinct variants, improving discrimination and interpretability without architectural complexity. Clinically, this shift mirrors precision-therapy logic:

- KRAS p.G12C mutations are now druggable with covalent inhibitors such as sotorasib or adagrasib[135,136] but other KRAS variants (p.G12D, p.G12V) remain resistant.
- BRAF p.V600E is actionable across multiple organs with BRAF/MEK inhibition, while non-V600 mutations are not.
- PIK3CA activating variants (E545K, H1047R) guide the use of alpelisib in HR-positive breast cancer.
- IDH1 p.R132H defines glioma subtypes eligible for IDH1 inhibitors (ivosidenib, vorasidenib).

By learning variant-specific morphological cues, *MultiVarNet* could therefore support *front-line triage*: flagging likely drug-eligible cases when sequencing is unavailable, delayed, or economically constrained. The method does not replace genotyping, but provides a low-cost, scalable prescreen that prioritises confirmatory testing, particularly valuable in low-resource or time-sensitive settings.

Defining the boundaries of morpho-molecular generalisation

Throughout Chapter IV, we explored the unique morphological nature of each variant-cancer pair. Pairwise cross-cancer tests and Leave-One-Cancer-Out (LOCO) domain-adversarial

training largely failed to generalise for PIK3CA and TP53, with the notable exception of IDH1 across LGG/GBM, tumours with shared lineage and histology[1]. Even with gradient-reversal, encouraging tissue-invariance did not rescue performance, suggesting that mutation-linked morphology is tissue-program bound and potentially microenvironment-dependent[2]. This limited transferability may also reflect mutation clonality and subclonality, which could influence the strength and visibility of morphological correlates: clonal driver mutations are more likely to shape consistent histopathological patterns than heterogeneous subclonal variants. These findings caution against universal "pan-cancer" claims and strongly support the development of problem-specific, context-aware models[168,169].

From morphology to therapy: guiding neoadjuvant and post-NAC decisions

The clinical translation of this work materialises through two independent yet complementary frameworks:

1. Chemo-prAIdict Breast: predicting intrinsic chemosensitivity from diagnostic biopsies before neoadjuvant chemotherapy (NAC).
    - In triple-negative breast cancer (TNBC), the tool identifies patients likely to respond to standard anthracycline–taxane backbones, potentially avoiding unnecessary carboplatin or pembrolizumab escalation when predicted pCR probability is already high.

      Conversely, low predicted chemosensitivity may justify treatment intensification, such as the addition of platinum salts, immune checkpoint blockade [3], or inclusion in novel-agent trials.
    - In luminal ER+/HER2- tumours, poor predicted chemosensitivity could support chemotherapy de-escalation in favour of endocrine-based neoadjuvant strategies (e.g., CDK4/6 inhibitors + AI), consistent with the European Society of Medical Oncology (ESMO) 2023 guidance.
    - In HER2+ disease, the model may inform early use of dual HER2 blockade (trastuzumab + pertuzumab) or alternative sequencing of taxane phases.
2. Hence, Chemo-prAIdict Breast could act as a baseline precision-triage tool, optimising treatment intensity before systemic therapy begins.
3. Post-NAC residual-disease analysis applied to surgical specimens after therapy to predict disease-free survival (DFS) and overall prognosis.

- High-risk predictions correspond morphologically to residual invasive tumour, necrosis, and high nuclear grade, features aligned with poor outcomes despite current regimens.
- Such patients could benefit from adjuvant escalation: capecitabine, olaparib for BRCA1/2 carriers, or T-DM1 in HER2+ disease [4–6].
- Conversely, patients with low predicted risk, often with fibrotic regression and minimal residual invasion, might be candidates for therapy de-escalation or shorter adjuvant durations.

Together, these two frameworks delineate a continuum of AI-driven decision support:

- Chemo-prAIdict Breast informs what treatment to start;
- The residual-disease model informs what treatment to continue or stop.

Both rely solely on standard Hematoxylin and Eosine (H&E) slides, removing the logistic and cost barriers of molecular assays, and thus could realistically integrate into multidisciplinary tumour-board workflows.

Human-centred interpretability and biological consistency

A pathologist-guided interpretability pipeline confirmed that both networks base their predictions on morphologically credible features (cellularity, nuclear pleomorphism, fibrosis, necrosis) rather than artefactual cues.

The recurrence of these motifs across pre- and post-treatment contexts implies that tumour micro-architecture reflects stable biological traits influencing both chemosensitivity and recurrence risk. For instance, a dense fibrotic stroma is adverse before therapy (reflecting poor drug diffusion) but favourable after therapy (marking regression), a duality consistent with histopathological regression frameworks [7].

This observation opens new research avenues: whether the same stromal or immune signatures observed morphologically correspond to transcriptomic programmes (e.g., TGF-β-rich fibrosis, interferon-γ-driven immune response) measurable by spatial-omics or proteomics.

Methodological transparency and analytical reproducibility

Each network was trained under a fully transparent analytical-validation framework: fixed data partitions, nested cross-validation, independent external validation, class-imbalance correction, and pre-defined evaluation metrics (AUC with 95% CI, p-values, and calibration). Such rigour ensures that results are analytically valid and not artefacts of split leakage or overfitting. This procedural clarity positions these models not as prototypes but as analytical tools ready for further testing and on the path for regulatory-grade validation.

Publications throughout this DPhil

- Morel, LO. et al.,. (2024). MultiVarNet - Predicting Tumour Mutational Status at the Protein Level. In: Linguraru, M.G., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. Lecture Notes in Computer Science, vol 15003. Springer, Cham. https://doi.org/10.1007/978-3-031-72384-1_30
- Sylvain Ladoire et al., (ASCO Abstract, 2025) Chemo-praidict eBC. J Clin Oncol 43, e15156-e15156(2025). DOI:10.1200/JCO.2025.43.16_suppl.e15156
- Morel LO et al., (2025). Chemo-prAIdict Breast: a deep learning solution for predicting residual disease on biopsies of breast cancer patients treated with neoadjuvant chemotherapy. *European Journal of Cancer*, vol. 234. 5 February 2026, 116222. https://doi.org/10.1016/j.ejca.2026.116222

Other publications

- Nicolas Dumas, Valentin Derangère, Laurent Arnould, Sylvain Ladoire, Louis-Oscar Morel, Nathan Vinçon (2022). Inter-Semantic Domain Adversarial in Histopathological Images. https://doi.org/10.48550/arXiv.2201.09041
- Morel, LO., Derangère, V., Arnould, L. et al. Preliminary evaluation of deep learning for first-line diagnostic prediction of tumor mutational status. Sci Rep 13, 6927 (2023). https://doi.org/10.1038/s41598-023-34016-y

1.2 Introduction and literature review

Breast cancer is the leading cause of cancer-related deaths for women in Europe and worldwide[8]. In 2018, 2.1 million women were newly diagnosed with breast cancer worldwide according to the World Health Organization, causing the death of 630,000 women[9].

Numerous risk factors for breast cancer are divided into two main categories: reproductive and nonreproductive. It's crucial to consider these factors as they can influence how patients are managed. Among reproductive risks, factors such as early onset of menstruation, delayed menopause, usage of oral contraceptives, and external hormone therapies are known to heighten breast cancer risk[10]. Conversely, breastfeeding for an additional 12 months reduces the risk by 4%, and each full-term pregnancy lowers it by 7%. Women with children have a 30% decreased risk compared to those without children[11]. On the non-reproductive side, age and genetic predisposition are significant. About 80% of breast cancer cases occur in women over 50, and 5% are linked to genetic factors. Obesity tends to decrease the risk by 20% in premenopausal women but doubles the risk in postmenopausal women[12].

Breast cancer management follows a three-step approach. The first step is screening, which enables the early identification of breast cancer. If an anomaly is detected, a sample is collected through a core needle biopsy. This sample is used to determine whether it is cancerous and to identify its type based on the microscopic analysis of the tissue (histology) and molecular characterisation. Following this diagnosis, the third step involves treatment and follow-up.

The European Commission Initiative on Breast Cancer (ECIBC) recommends a mammography screening for women aged 50-69 at average risk every 2 years. A review of randomized controlled mammography trials in the UK estimated a 20% relative reduction in breast cancer mortality among women aged 50-70. Furthermore, recent research has shown that initiating screening at 40 years old could reduce breast cancer mortality by as much as 50%[13–15].

Diagnosis is determined through histological analysis. Utilizing this characterization, pathologists differentiate between benign lesions and carcinoma. In the morphological examination of breast carcinoma, two crucial assessments are made. The first is to ascertain whether the tumour is confined to the ductal-lobular system (carcinoma in situ) or has

penetrated the surrounding stroma (invasive carcinoma). The second determination is whether the carcinoma is of the ductal or lobular variety. The ducts and lobules are key components of breast anatomy, particularly involved in milk production and delivery. As shown on **Figure 1**, ducts are tube-like structures in the breast that transport milk from the lobules, where it is produced, to the nipple. During breastfeeding, milk flows through a network of these ducts to reach the nipple and feed the infant. Lobules are also known as mammary glands. They are the milk-producing glands in the breast. Each breast contains hundreds of lobules. They produce milk in response to hormonal signals, particularly during pregnancy and breastfeeding. Both ducts and lobules are supported by connective tissue and fat within the breast, and they play crucial roles in the breast's primary function of lactation[16].

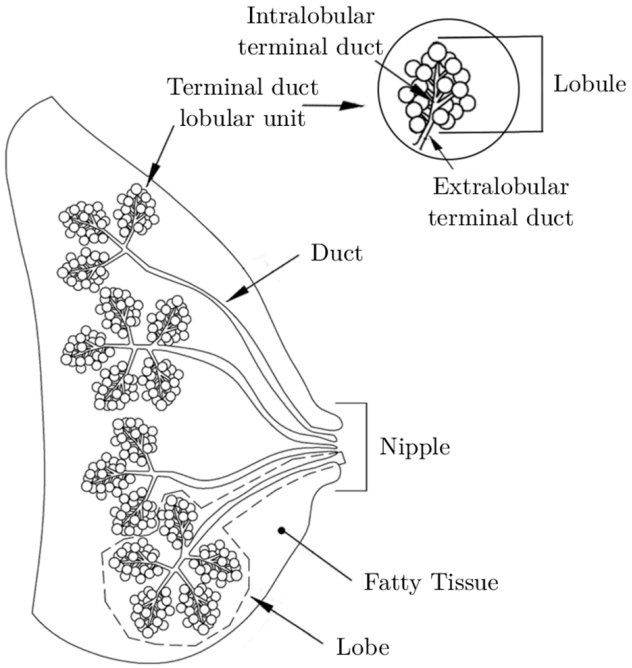


Figure 1: Anatomy of the human breast and organisation of the ductal-lobular system.
Schematic representation of the macroscopic and microscopic structures of the breast. The breast is composed of multiple lobes separated by fatty tissue, each lobe containing several lobules that converge toward the nipple through a ductal system. Each lobule includes terminal duct–lobular units (TDLUs), which are the functional and pathological core of breast tissue. TDLUs consist of intralobular terminal ducts surrounded by clusters of acini (lobules), which drain into extralobular ducts. Most breast carcinomas originate from epithelial cells lining the terminal ducts and lobules, making these regions the key sites of malignant transformation. Understanding this structural organisation is essential for interpreting the histopathological basis of breast cancer development and its representation in whole-slide

images. *Image adapted from Oliveira, Barbara (2018). "Towards Improved Breast Cancer Diagnosis Using Microwave Technology and Machine Learning."*

The prognostic significance of the invasive criterion is far more important than the ductal vs lobular. According to the American Cancer Society[17], invasive breast cancer represents around 80% of the cases. Ductal carcinoma in situ (DCIS), accounting for approximately 20% of all breast cancer cases, typically offers a favorable prognosis, with most patients achieving effective treatment outcomes. DCIS makes up about 80% of all in situ breast cancers, and lobular carcinoma in situ LCIS about 12%. Invasive ductal carcinoma (IDC) is the most prevalent type of invasive breast cancer, comprising about 70-80% of all invasive breast cancer cases. In contrast, invasive lobular carcinoma (ILC) accounts for approximately 10-15% of invasive breast cancers. More precisely, the differentiation between in situ and invasive is defined through the AJCC (American Joint Committee on Cancer) staging system[18]. It uses the TNM classification to describe the extent of cancer spread. "T" denotes the size of the tumour and whether it has invaded nearby tissue, "N" describes the extent of spread to nearby lymph nodes, and "M" indicates the presence of metastasis to distant body parts. This staging ranges from Stage 0, which represents non-invasive cancers like DCIS (ductal carcinoma in situ), to Stage IV, which indicates cancer has spread to other parts of the body.

		Stage	Primary tumour (T)*	Regional lymph node status (L)	Distant metastasis (M)
T- Tumour		**0**	Tis	N0	M0
T1	Tumour ≤ 2 cm	**I**	T1	N0	M0
T2	Tumour ≥ 2 cm but ≥ 5 cm		T0	N1	M0
T3	Tumour ≥ 5 cm	**IIA**	T1	N1	M0
T4	Tumour of any size with direct extension to chest wall or skin		T2	N0	M0
N- Lymph node		**IIB**	T2	N1	M0
N0	No cancer in regional node		T3	N0	M0
N1	Regional movable metastasis	**III A**	T0	N2	M0
N2	Non-movable regional metastases		T1	N2	M0
N3	Cancer in the internal mammary lymph nodes		T2	N2	M0
M- Metastasis			T3	N1/N2	M0
M0	No distant metastases	**III B**	T4	Any N	M0
M1	Distant metastases	**III C**	Any T	N3	M0
		IV	Any T	Any N	M1

Criteria for staging breast tumours according to the UICC ICD-10 TNM classification.

*Size measurements are for the tumour's greatest dimension.

Figure 2 : AJCC TNM staging system for breast cancer (8th edition). The TNM classification (Tumour–Node–Metastasis) summarises the anatomical extent of breast cancer and remains the cornerstone of clinical and pathological staging. The table outlines the 8th edition criteria established by the American Joint Committee on Cancer (AJCC), which combine tumour size (T), regional lymph node involvement (N), and presence of distant metastasis (M) to define overall stage groups (0–IV). *Adapted from the AJCC Cancer Staging Manual, 8th Edition (UICC ICD-10 TNM classification).*

The second crucial histological classification is the SBR (Scarff-Bloom-Richardson) grading system[19], which evaluates the microscopic appearance of the cancer cells, specifically looking at three factors: the degree of tubule formation, the rate of mitotic activity (how quickly the cancer cells are dividing), and the nuclear grade (abnormalities in the size and shape of the cell nucleus). These factors are scored, and the scores are then combined to assign a grade from 1 to 3. Grade 1 (well-differentiated) tumours are generally slower growing and have a better prognosis compared to Grade 3 (poorly differentiated) tumours, which are more aggressive and have a worse prognosis. Both systems are essential as they provide detailed

information about the cancer's aggressiveness and spread, helping clinicians to tailor treatment strategies effectively.

Features		Score
Tubule and gland formation	Majority of tumour (>75%)	1
	Moderate degree (10–75%)	2
	Little or none (< 10%)	3
Nuclear pleomorphism	Small, regular uniform cells	1
	Moderate increase in size and variability	2
	Marked variation	3
Mitotic count (dependent on microscopic field area, e.g. for field area 0.264 mm^2 with field diameter 0.58 mm)	0–9	1
	10–19	2
	>20	3

Figure 3: Scarff–Bloom–Richardson (SBR) grading system for invasive breast carcinoma. The SBR system evaluates tumour differentiation based on three histological criteria: tubule formation, nuclear pleomorphism, and mitotic count. Each feature is scored from 1 (well differentiated) to 3 (poorly differentiated), and their sum defines the histological grade (I–III). Higher scores indicate poorer differentiation and more aggressive tumour behaviour. The SBR grade remains a key prognostic factor in breast cancer and a reference for evaluating AI-based morphological predictions.

Standard pathology reports must also include the identification of several molecular biomarkers. The most important are estrogen receptor (ER), progesterone receptor (PgR), human epidermal growth factor receptor 2 (HER2) and a proliferation marker such as Ki-67, and are assessed by immunohistochemistry (IHC) according to the American Society of Clinical Oncology/College of American Pathologists guidelines[20] and HER2 status[21]. These markers define the three molecular categories of breast cancer, (1) HER2-amplified, (2) HR+/HER2- or luminal tumours, (3) triple negative breast cancer (TNBC).

HER2-amplified tumours have an overexpression of the HER2 protein, which is involved in cell growth and survival. HER2-positive cancers, accounting for about 20% of cases, are typically more aggressive but respond well to targeted therapies that inhibit the HER2 protein, such as trastuzumab (Herceptin) and pertuzumab (Perjeta)[22]. HER2-low cancers express

HER2 but at lower levels and make up approximately 55% of all breast cancers. HER2-low is defined as IHC 1+ or 2+ and in situ hybridisation (ISH) negative in metastatic disease[23]. Although HER2-low status is currently only relevant in the metastatic setting, current ESMO guidelines recommend reporting it in early breast cancer (EBC) as a basis for possible future therapeutic decisions, including novel agents such as antibody-drug conjugates (ADC)[24].

Luminal tumours express hormonal receptors but not the HER2 protein (HR+/HER2-). They account for 70-75% of all breast cancer cases[25]. We distinguish luminal A tumours from luminal B. Luminal A tumours usually have higher levels of hormone receptor expression compared to luminal B tumours, are typically low grade, HER2 negative and have low proliferation (low ki-67, usually under 10%). Luminal B tumours display hormone receptor positivity; however, they can show varying levels of estrogen receptor (ER) and progesterone receptor (PgR) expression. They are often of higher grade and exhibit increased proliferation compared to tumours classified as Luminal A[8]. Luminal tumours do not respond well to chemotherapy, as only a small subset of tumours can achieve a pathological complete response (pCR)[1] after treatment[26,27].

Triple negative breast cancer represents a heterogeneous group of breast cancers that lack expression of ER, PgR, and HER2. It constitutes about 15-20% of all breast cancer cases. TNBC can be further classified into molecular subtypes based on gene expression profiles, including basal-like, claudin-low, androgen receptor-positive, and mesenchymal stem-like subtypes[28]. It is typically associated with a poorer prognosis compared to other breast cancer subtypes due to its aggressive nature, higher rates of metastasis, and limited targeted treatment options. However, recent advances in treatment protocols have significantly improved its prognosis. For instance, the baseline pathologic complete response rate has increased from 35% with traditional cytotoxic chemotherapy utilizing anthracyclines and taxanes, to 50% with carboplatin regimens (paclitaxel)[29], and even up to 63% with pembrolizumab + paclitaxel escalation in the KEYNOTE-522 clinical trial[30].

[1] A pathological complete response is defined as the microscopic disappearance of the tumour on the tissue after chemotherapy (pT0N0).

1.2.1 Early Breast Cancer (eBC) management in a neoadjuvant setting and residual disease stratification

Treatment recommendations for early breast cancer are tailored to each patient's histological and molecular profile. Locoregional treatment often involves breast conserving surgery, followed by postoperative radiotherapy to target any remaining cancer cells. The decision regarding (neo)adjuvant systemic treatment is informed by factors such as the individual's risk of relapse and the predicted sensitivity to different treatment modalities. This may include chemotherapy, hormone therapy, targeted therapy, or a combination thereof, aiming to reduce the risk of recurrence and improve long-term outcomes. By integrating both locoregional and systemic approaches, clinicians strive to optimise treatment strategies while minimising the risk of disease progression.

Adjuvant or neoadjuvant systemic chemotherapy is a standard treatment for high-risk early breast cancer (eBC) and consists of performing chemotherapy after or before the surgery, respectively. Patients with HER2-amplified or triple negative (TNBC) breast cancer are recommended to receive chemotherapy in the vast majority of cases, and when there is a significant risk of metastatic relapse for ER+/HER2- tumours. The decision to administer systemic chemotherapy, either in an adjuvant or neoadjuvant setting, is made by oncologists based on individual patient risk factors for relapse. A neoadjuvant chemotherapy[31] can be proposed to the patients in order to reduce the size of the tumour before the surgery, with the aim to facilitate a conservative surgery, but also to administer systemic treatment very early on, with the aim of treating any micro metastatic disease as quickly as possible. In the Early Breast Cancer Trialists' Collaborative Group (EBCTCG)[26] meta-analysis, there was no significant difference between patients treated in a neoadjuvant scheme in terms of distant recurrence, breast cancer mortality or death from any other cause compared to same standard adjuvant chemotherapies. Another advantage of NAC is the ability to evaluate the chemosensitivity of each patient's tumour to the standard chemotherapy on the surgical specimen: Thus, patients with a pathological complete response (pCR, defined by no residual invasive cancer cells both in breast and axillary lymph nodes: pT0 and pN0[27]) after NAC have significantly better relapse-free, and overall survival than patients with residual disease (RD) on the surgical specimen, especially for HER2-amplified and TN breast cancer[5].

Patients who achieve a pathologic complete response have a reduced risk of disease relapse, potentially making them suitable candidates for treatment de-escalation strategies.

Conversely, individuals with residual disease following surgery are at heightened risk of disease recurrence and might derive benefit from supplementary post-neoadjuvant chemotherapy treatments (**Figure 4** adapted from Agostinetto et al. (2022)[32]). Some systemic post-NAC therapies have already been approved in clinical practice. For these two HER2-amplified and TNBC cancer subtypes, patients with a RD can be selected after surgery for post operative adjuvant treatment intensification with T-DM1 and capecitabine respectively[33]. Concerning HR+/HER2- eBC subtype, pCR achievement is much rarer, with less prognostic impact. Nevertheless, the identification of tumours genuinely resistant to standard systemic chemotherapy would help direct clinicians toward therapeutic de-escalation. Recent studies propose combining endocrine therapy with CDK4/6 inhibitors as a potential approach in this scenario[34,35].

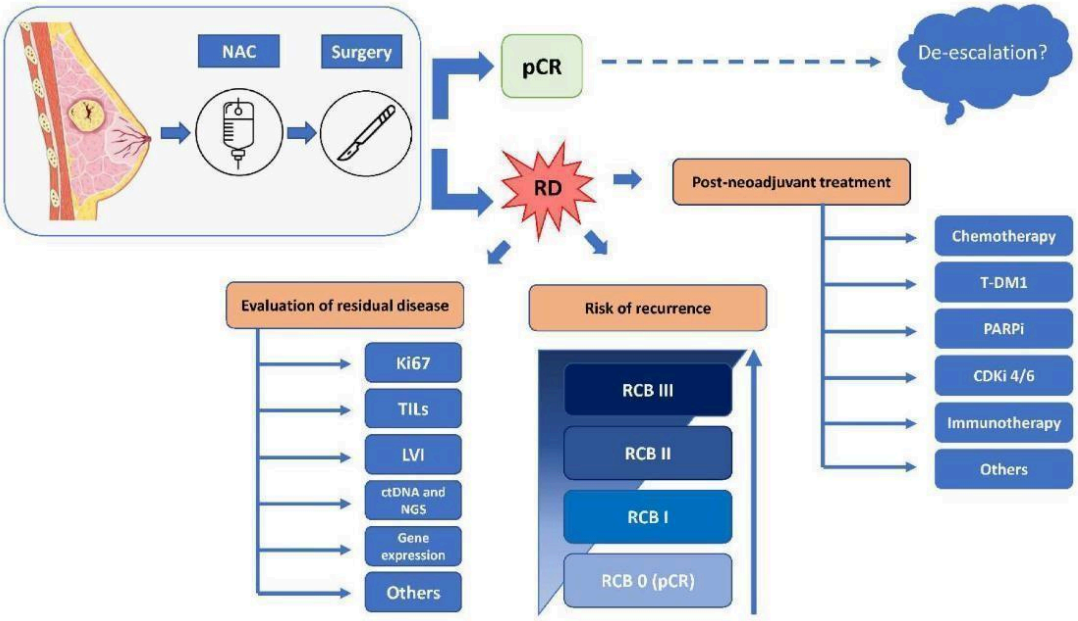


Figure 4: Simplified overview of the post-neoadjuvant treatment workflow in breast cancer. After neoadjuvant chemotherapy (NAC), the surgical specimen is examined to determine whether a pathological complete response (pCR) or residual disease (RD) remains. Pathologists evaluate multiple parameters, including Residual Cancer Burden (RCB) score, Ki-67 proliferation index, Tumour-Infiltrating Lymphocytes (TILs), lymphovascular invasion (LVI), and, when available, genomic or transcriptomic data such as Next-Generation Sequencing (NGS) or gene-expression signatures. The level of residual disease stratifies patients into four RCB classes (0–III), which correlate with the risk of recurrence. This pathological risk assessment directly guides post-neoadjuvant therapy: Patients achieving pCR (RCB 0) may be considered for treatment de-escalation. Those with residual invasive disease

(RCB I–III) are candidates for therapy escalation, including adjuvant capecitabine, PARP inhibitors, T-DM1, CDK4/6 inhibitors, or immunotherapy, depending on subtype and molecular profile. This figure summarises the decision-making framework underpinning the residual-disease prediction models developed in Chapter V. *Adapted from Agostinetto et al. (2022).*

While the personalised approach based on the status of the post-NAC surgical specimen is becoming the current gold standard, it also introduces significant risks. For patients who achieve a pCR, there's a risk of not treating them despite a genuine risk of relapse. This risk stands at 10% for HER2-positive patients, 15% for triple-negatives and over 20% for ER+/PR+ patients. For those with residual disease, there's a potential of an unnecessary therapeutic escalation[32,36] and thus potential unnecessary adverse events.

It is crucial to identify low-risk residual disease. Firstly, pinpointing patients with a high residual risk of relapse could lead to a reduction in the sample size required for (neo)adjuvant trials, thereby reinvigorating clinical research in breast cancer[37]. Secondly, better identification of such patients could narrow the target population for which new drugs should be approved in the adjuvant setting, resulting in significant cost savings. Approving drugs based on large randomised trials in broad populations may lead to reimbursement requests for large populations where most patients do not require the drug, resulting in substantial additional costs. For instance, if we consider a drug costing 50,000 euros per year for a disease with an incidence of around 10,000 women/year in France (such as HER2-positive breast cancer), the annual cost of the drug could reach 500 million euros without patient selection[38]. Thirdly, such identification could significantly reduce toxicity in women with breast cancer. Finally, it could facilitate the restructuring of healthcare delivery for breast cancer patients by directing higher-risk patients to comprehensive cancer centres.

To improve this binary stratification, assessment of the surgical specimen after chemotherapy is now systematic and can include a number of characteristics, such as the Residual Cancer Burden score (RCB), the Neo-Bioscore, ki67 scoring, TILs, lymphovascular invasion (LVI), sequencing signature and ctDNA[39–44]. However, these biomarkers are imperfect and lack prognostic value[45,46].

The Neo-Bioscore is derived from the Clinical-Pathologic Stage plus Oestrogen/Grade (CPS+EG) staging system. They are designed specifically for HR+/HER2- breast cancer patients treated with neoadjuvant chemotherapy[47]. Neo-Bioscore incorporates HER2 status

into the previously developed CPS+EG staging. These scores give valuable information by giving a stratified survival prediction but does not quantify residual disease. It only focuses on the diagnosis information by integrating the clinical stage, the biopsy pathological stage and the following tumour markers: ER status, Grade 3 and HER2 status. Although interesting, the results might lack refinement. As described in Valderrama and Morel et al.[170], using clinical information within each molecular subtypes of early breast cancer does not give any significant information on the therapeutic response. These results validated on 12 cancer centres are robust and show that the Neo-bioscore accuracy comes from the molecular stratification rather than the clinical staging.

The RCB scoring system is currently regarded as the gold standard for analysing surgical specimens after neoadjuvant chemotherapy (NAC). It quantifies residual disease post-treatment and assigns a continuous score, categorised into four classes (RCB-0 to RCB-3), to predict long-term survival outcomes. RCB is proven to provide significant prognostic value across various breast cancer subtypes, thereby enhancing its clinical utility. However, there is potential for further enhancement. For instance, the univariable hazard ratio (HR) associated with a one-unit increase in RCB ranges from 1.55 (95% CI 1.41–1.71) for hormone receptor-positive, HER2-negative patients, to 2.16 (1.79–2.61) for hormone receptor-negative, HER2-positive groups, indicating variability across subtypes and moderate performances. Additionally, the evaluation process of RCB scoring can be quite burdensome for pathologists due to the lengthy and detailed analysis required for each node and surgical slide.

Gene expression signatures are currently widely used for predicting response to primary systemic therapy. Genomic tests analyse tumours to see how active or modified certain genes are. These modifications from normal state affect the behaviour of the tumour, predicting how likely it is to grow and spread. The Oncotype DX Breast Recurrence Score Test for people diagnosed with early-stage, oestrogen receptor-positive, HER2-negative invasive breast cancer[48]. Other signatures exist, such as EndoPredict, MammaPrint, Prosigna and others[49–51], but OncotypeDx is the most widely used. The tests vary in the genes they target and the technologies they use, yet each is capable of identifying a group with low to intermediate risk who may not need adjuvant chemotherapy. Gene expression assays should be reserved for cases where clinical risk is intermediate, as determined by ImmunoHistoChemestry (IHC) biomarkers and clinical factors. It is advisable to use just one type of assay for each patient[52]. However, little work has been done with these tests on a neoadjuvant setting. In recent

studies, a high OncotypeDx risk score derived from a pre-treatment tumour sample appears to correlate with a pathological complete response following neoadjuvant chemotherapy. This suggests that the score could also be useful for identifying patients who are most likely to benefit from NAC. However, these molecular signatures seem to be much less efficient in a neoadjuvant setting compared to solutions predicting pCR using histology[170]. Moreover, the relationship between this score and the risk of relapse, as well as its analysis in post-NAC surgical specimens, has not yet been established.

Circulating tumour DNA (ctDNA) and circulating tumour cells (CTCs) are two types of biomarkers detectable in blood samples, often referred to collectively in the context of "liquid biopsies". Both ctDNA and CTCs offer valuable insights into the nature and progression of cancer, including breast cancer, though they differ significantly in their origins and applications[53].

ctDNA refers to small fragments of DNA from cancer cells that are found freely circulating in the bloodstream[54]. These DNA fragments can be analysed to detect mutations and other genetic alterations in the tumour, providing information about the tumour's characteristics without needing a traditional tissue biopsy. The advantages of ctDNA testing include its non-invasive nature and the ability to provide real-time monitoring of tumour evolution and response to treatment. In a large meta-analysis[55], ctDNA detection was associated with worse DFS at baseline [HR 2.98, 95% confidence interval (CI) 1.92-4.63], after neoadjuvant therapy (HR 7.69, 95% CI 4.83-12.24), and during follow-up (HR 14.04, 95% CI 7.55-26.11). However, the detection of ctDNA may have low sensitivity since it occurs only in a subset of residual tumours. Nonetheless, due to its inherent characteristics, the specificity of ctDNA identification will be high. As described by the ESMO meta-analysis, ctDNA detection sensitivity and specificity for BC recurrence ranged from 0.31 to 1.0 and 0.7 to 1.0, respectively. The mean lead time from ctDNA detection to overt recurrence was 10.81 months (range 0-58.9 months).

On the other hand, CTCs are intact cancer cells that have shed from the primary tumour or metastatic sites and entered the bloodstream. While ctDNA analysis focuses on genetic material, CTC assays can examine the cells themselves, potentially offering information on the tumour cells' phenotype and other characteristics. CTCs can also be used to assess disease burden and are important for understanding the metastatic process, as they are thought to be directly involved in the spread of cancer to other parts of the body. The Food and Drug Administration (FDA) recently approved the CellSearch® and mesenchymal CTCs Parsortix® PC1 systems[56]. As for cfDNA, CTCs sensitivity is low because of its scarcity in

peripheral blood. In Trapp et al. (2019)[57], 1087 patients were followed for detection of CTCs 2 years after completion of adjuvant chemotherapy. Overall, 43 of 1087 patients (4.0%) died during the follow-up period and 21 (48.8%) of these patients presented with CTCs at the time of the 2-year follow-up. As mentioned by the authors "As a standalone tool for intensified follow-up, CTC counts may prove to be neither sufficiently sensitive nor sufficiently specific".

These results are very promising, yet not sufficient. The first drawback is the aforementioned low sensitivity, but the cost is also becoming a rising issue. As an example, standard molecular biology tests such as Polymerase Chain Reaction (PCR) or Next Generation Sequencing (NGS) in non-small cell lung cancer (NSCLC) are often prohibitively expensive[58]. This high cost deters physicians from prescribing these tests, adversely affecting patient care[59]. Recent research in NSCLC suggests that nearly half of the patients miss out on effective targeted therapies due to these financial barriers[60]. PCR price for EGFR in France is 237.55€ (CCAM quotation). Large NGS panels are €1,503.90 in France (RIHN N453) and $2,987.85 in the USA (CPT81445/450). FoundationOne CDx is $3,500, Veracyte test between $3,000 and $4,000, new liquid biopsy test will be around $10,000 to be profitable for high risk venture capitalists. As described in JAMA, current liquid biopsy screening strategies are not yet cost effective (here in colorectal cancer)[61].

In conclusion, there are limited options available for enhancing the stratification of residual disease. Existing methods are predominantly traditional, employing standard pathological approaches that offer only a basic level of granularity. Alternatively, some solutions have been developed for different applications and later adapted to address this specific issue. However, inherently, these adapted solutions are not ideally suited for this purpose and tend to deliver suboptimal performance.

In recent years, deep learning techniques have been increasingly applied to various image analysis tasks in digital pathology, including detection and subtyping of tumours, counting cells, classifying cell types, and RNA sequencing[62–65]. These methods have also demonstrated potential in predicting mutational status from digitised, hematoxylin and eosin-stained (H&E) tissue in whole slide images[66] (WSI). Such technologies could be pivotal for refining diagnostic accuracy and treatment precision. By employing deep learning for WSI analysis, clinicians could gain deeper insights into the histological and molecular profiles of tumours, which is crucial for tailoring patient-specific therapeutic strategies. This approach not only promises to enhance the evaluation of chemosensitivity and tumour behaviour but

would also support the ongoing shift towards more personalised and effective cancer care. Ultimately, integrating deep learning with current diagnostic practices could significantly improve outcomes by enabling more accurate predictions of treatment response and relapse, thus optimising both locoregional and systemic treatment plans for patients with early breast cancer. During the course of this thesis, we will extensively utilise these methodologies to improve the stratification of the breast cancer residual disease after neoadjuvant systemic therapy.

1.2.2 Predictive deep learning for histopathology

The advent of digital pathology, propelled by advancements in slide scanning technology since the early 2010s, has revolutionised medical diagnostics. This transformation has been further accelerated by the deep learning boom since 2012[67], fostering a surge in literature focused on automating medical diagnosis processes. These technological advancements have paved the way for more precise and efficient analysis methods, which are essential in the fight against diseases.

The initial models in digital pathology relied on manually extracted features, such as colour, intensity, and texture, attributes well-known to pathologists[68]. Following this, a second generation of models emerged, which were more refined and utilised characteristics traditionally evaluated by medical experts. These models analysed aspects such as the size and appearance of nuclei and the presence of perinuclear vacuoles, incorporating prior human knowledge into their frameworks[69]. Deep Convolutional Neural Networks (DCNNs), introduced in pathology image analysis by Ciresan et al. (2013)[70], excel in automatically extracting morphological features directly related to diagnostic tasks, without requiring detailed human input. This capability not only streamlines the analytical process but also unveils characteristics previously unrecognised by human experts, offering potentially better predictions of tumour types.

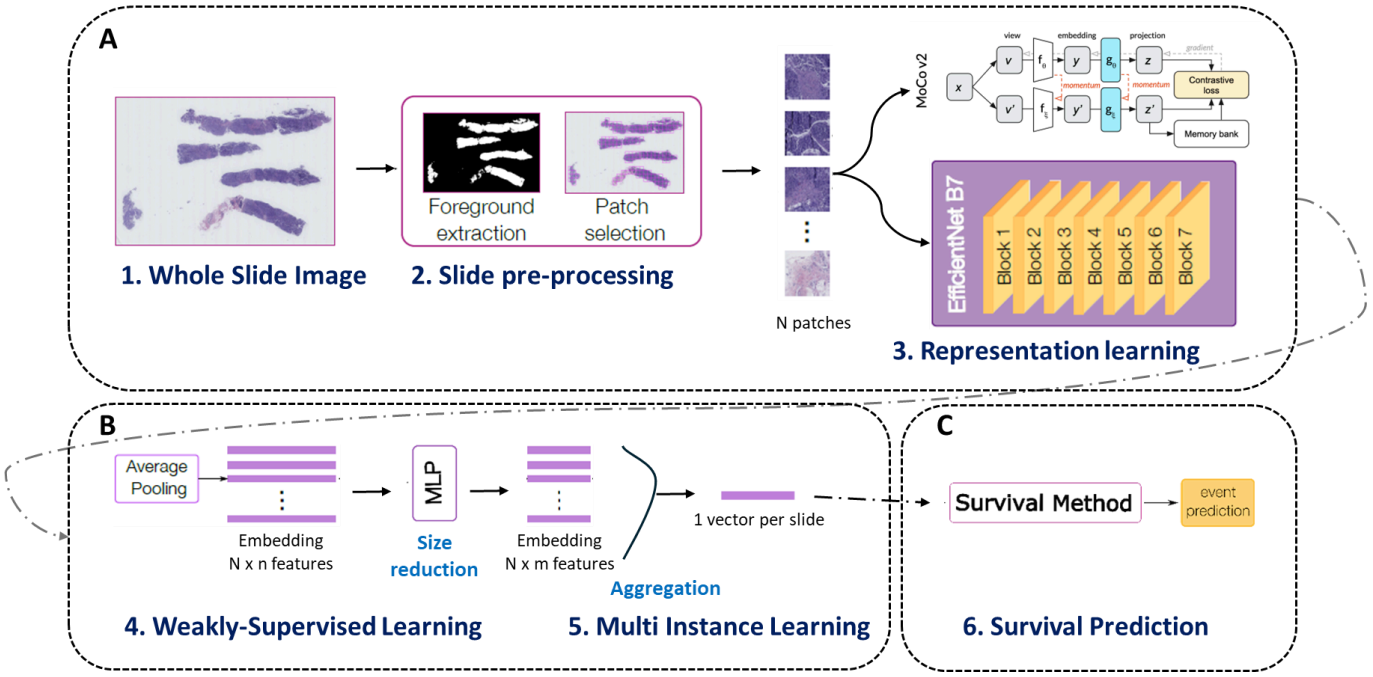


Figure 5: Deep Learning Pipeline for Survival Prediction. A. Preprocessing and Embedding: The core component of any deep learning algorithm in computational pathology (used for segmentation, cancer classification, or various virtual molecular biology tasks). This stage involves preprocessing the WSI, highlighted in parts 1 and 2, where slides are preprocessed to extract relevant patches. In part 3, each patch is transformed into an embedding, which is then input into a multi-layer perceptron (MLP) to predict the desired outcomes. **B. Weakly-Supervised Learning Framework:** This framework involves the prediction for individual patches as outlined in part 4, followed by the aggregation of scores from all patches into a single descriptor for the slide, as shown in part 5. This step is crucial for synthesising localised data into a global prediction metric. **C. Event Prediction for Clinical Outcomes:** Specific to survival and relapse prediction, this final part integrates the slide descriptor into a survival prediction model to assess the risk of clinical events, such as patient survival or relapse, as depicted in part 6. Each of these sections represents a distinct research domain within computational pathology, offering multiple methodologies to enhance the performance of the overall framework.

DCNNs mimic the human visual cortex's hierarchical structure, allowing for nuanced analysis of histology images[71]. Their ability to discern subtle differences in tissue structure is vital for identifying pathological changes. The first significant applications were a mitosis detection model and a breast cancer diagnosis model[70] trained on H&E stained whole slide images (WSI), demonstrating the potential of DCNNs in practical diagnostic settings. Despite

their capabilities, DCNNs face challenges due to the immense size and complexity of digital biopsy images, often requiring the division of these images into smaller segments called patches. Histopathology images are notably large, often reaching gigabytes in size with dimensions up to 100,000 x 100,000 pixels, captured at various microscopic zoom levels. DCNNs like EfficientNet are typically designed to process much smaller images, such as 224x224 pixels for EfficientNetB0 or 600x600 for EfficientNetB7[72]. Consequently, each histopathology slide is divided into numerous smaller segments, known as patches (**part 2, Figure 5.A**). These patches facilitate the automation of tasks traditionally performed by pathologists, enabling localised annotation and performance assessment of the automated systems (**part 3, Figure 5.A**). Unfortunately, annotating each individual patch is overly labour-intensive and not feasible for large-scale application, highlighting the need for more efficient approaches in this field. Furthermore, these localised annotations are not necessarily obtainable, particularly when one wishes to predict characteristics not known by doctors, such as for the classification of new disease subtypes or the prediction of a prognosis or a mutational status (obtained by molecular biology).

Recent developments have seen an increase in the use of weakly supervised learning (WSL) and multi-instance learning (MIL) to manage the vast amount of unlabeled data in digital pathology[73]. These methods, which aggregate patch-level predictions to infer slide-level diagnoses, are essential for analysing WSI where exhaustive labelling is impractical. Weakly-Supervised Learning corresponds to the identification of suspicious regions of a WSI image when the training data only contains labels at the scale of the overall image. In this context, WSIs are treated as groups of patches, or tiles, of which the package which constitutes their aggregation has a unique label, for example “healthy”, “cancer”, “PIK3CA mutation” (**part 4, Figure 5.B**). The aggregation of these patches to predict a global label (**part 5, Figure 5.B**) is called “Multi Instance Learning” (MIL)[74,75]. Traditionally, in the MIL framework, a collection of instances, or a 'bag', is classified as positive if at least one instance within the bag, such as a patch in histopathological analysis, is labelled positive. However, this classification criterion may vary based on the specific histological features or output being predicted. The objective in this context is to utilise bag-level labels to deduce a set of instance-level rules that can accurately classify individual patches. In the realm of cancer biopsy image analysis, the ability to establish such rules is crucial for pinpointing localised regions containing abnormal cells within large-scale Whole Slide Images. Consequently, even when a slide is globally labelled, this approach enables the identification of heterogeneity within the slide, distinguishing areas of tumour tissue from healthy regions.

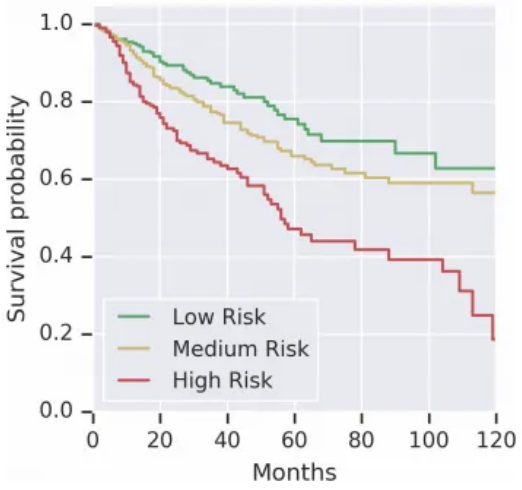
This method not only enhances the precision of diagnostic processes but also provides a nuanced understanding of the tissue's pathology. Various aggregation methods are used, including averaging scores, median scores, calculating the 90th percentile, or employing additional neural networks to optimise patch aggregation for predicting the slide's overall label[76,77]. Recent studies have demonstrated that the application of WSL and MIL yields impressive results for identifying specific areas in biopsy images. Nonetheless, strongly supervised models typically outperform weakly supervised counterparts in tasks where localised annotations provided by medical experts are available. For instance, on the Camelyon16 challenge, which focuses on the detection of breast cancer metastases, the state-of-the-art[78,79] strongly supervised model achieves an Area Under the Receiver Operating Characteristic curve (AUC) of 99.3%. In contrast, weakly supervised learning approaches on the same dataset reach only 91.4% AUC without detailed annotations[80]. However, there are numerous research questions in histology, such as predicting gene expression from transcriptomic data and other previously cited examples where the relevant markers may not be visually discernible on the slide. In these scenarios, while annotations might highlight useful regions of interest, they alone are insufficient for resolving the task. Thus, weakly supervised learning becomes indispensable, serving as a crucial methodology for advancing machine learning applications in histology[81].

To reduce the difference in performance between weakly supervised learning (WSL) and strongly supervised learning, Dehaene et al used "Self-Supervised Learning" (SSL) techniques (**part 3, Figure 5.A**)[82]. One significant limitation of WSL is that its neural networks are pre-trained on varied semantic domains. For example, DCNNs pre-trained on ImageNet are primarily skilled in recognizing animal images. The filters they learn are specialised for this task. Although it is possible to fine-tune the weights of these DCNNs to better suit specific tasks, SSL offers the advantage of enabling unsupervised pre-training on a large dataset of cancer images. This significantly enhances the semantic relevance of the filters learned, leading to a marked improvement in prediction performance at the core of the "foundation models" revolution. For instance, using this approach on the Camelyon16 dataset, the AUC improved from 91.4% to 98.7%, as reported by Dehaene et al. in 2021. Many techniques in the SSL field have been developed, such as the MoCo, SimCLR, SwAV and the new foundation models which are pre-trained on millions of histology images[83–88].

The last part of our pipeline (**part 6, Figure 5.C**) is dedicated to the main goal of this thesis, which concentrates on predicting overall survival (OS) and disease-free survival (DFS) in patients with early-stage breast cancer. By doing so, we introduce an additional layer to

deep learning-based histology analysis by integrating survival analysis with computer vision techniques. As it complexifies the analysis workflow and data acquisition, literature is rarer and more exploratory than the standard WSL framework.

Survival analysis is a statistical method[89] traditionally used in medicine to determine the probability of an event of interest (often time-to-event data such as death or relapse) over time. It allows us to analyse a proportion of population surviving up at a given time, the rate at which a population is dying and helps understand the impact of covariates on survival or find differences between populations. In the context of oncology, survival models are crucial for predicting patient outcomes based on clinical and pathological data, with the overall goal to distinguish between subsets of patients based on their characteristics (either clinical data or image-based features). A good example of this is provided by Wulczyn et al. (2020)[90], where they used a deep learning method to stratify patients in two categories depending on their prognosis (**Figure 6**).



Kaplan Meier curves for three risk groups [Wulczyn2020]

Figure 6. Kaplan–Meier survival curves illustrating risk-group stratification. Example of Kaplan–Meier analysis showing survival probability over time across three risk categories (low, medium, high). Each curve represents the cumulative proportion of patients surviving beyond a given time point, with steeper declines indicating poorer outcomes. The separation between curves demonstrates effective prognostic discrimination, a principle used throughout this thesis to evaluate survival-prediction models. *Adapted from Wulczyn et al., 2020.*

Survival analysis or time-to-event analysis is complex, as it has its own methodologies and that complete data is not always available. Indeed, participants may be followed for shorter periods than the study, they can drop out, they can die (considering the outcome is not death). Missing data in survival analysis is normal and introduces the concept of censoring.

Censoring refers to a situation where the exact time of the event of interest (such as death, failure, relapse, etc.) is not known for some subjects in the study. Censoring can occur in several forms:

- **Right Censoring**: This is the most common type of censoring in survival analysis. Right censoring happens when the study ends, or a participant drops out before experiencing the event of interest. The exact time of the event is unknown but is known to occur after the last observed time.
- **Left Censoring:** Left censoring occurs when the event of interest has already happened before the participant enters the study. In this case, the exact time of the event is unknown but is known to have occurred before a certain time.
- **Interval Censoring**: Interval censoring occurs when the event is known to have happened between two time points, but the exact time is unknown. This is often the case in medical follow-ups where patients visit at scheduled intervals.
- **Random Censoring**: Random censoring assumes that the reason for the censoring is independent of the survival prospects or any other variables. This type does not systematically bias the survival times.

Censoring can complicate the analysis because the exact event times are not known for all subjects. However, it is a critical concept because it reflects a realistic aspect of longitudinal and survival studies. Ignoring censored data or incorrectly handling it can lead to biassed estimates and conclusions. Survival analysis methods are specifically designed to handle censored data effectively. Indeed, they make use of censored data in the estimate of the probability of event. They assume that censoring is independent or unrelated to the likelihood of developing the event of interest, hence participants whose data are censored would have the same distribution of failure times if they were actually observed[91]. Survival methods estimate survival functions for a given sample. These methods are used for risk stratification which will help pathologists design classifications as for cancer staging and cancer grading.

The integration of deep learning into survival analysis, especially in the analysis of histological slides, has created new opportunities for more accurate and personalised prognostic evaluations. However, this approach extends beyond simple binary classification. As previously mentioned, some patients may not experience the event of interest by the study's conclusion, or they may be lost to follow-up during the study period. These instances are referred to as right-censored data, which must be carefully considered in our analysis. Traditional regression models are inappropriate as they would require omitting right-censored

cases. Similarly, binary classification models, which might categorise patients as alive or dead after a certain number of years, fail to utilise the actual survival times effectively.

The most common method of modelling survival is the Cox Proportional Hazards (CPH) model[92]. It is a statistical technique used to explore the relationship between the survival time of patients and one or more predictor variables, working as a regression model. In this model, the hazard function is assumed to be a baseline hazard function, multiplied by a function of the covariates (predictor variables):

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Where:

- *h(t, X)* is the hazard at time t for a given vector of covariates X.
- *h0(t)* is the baseline hazard, representing the hazard for an individual with all covariates equal to zero.
- β1 ,β2,...,βp are the coefficients of the covariates *X*1, *X*2,..,*Xp*. Can be represented as β^{Tx}.

The model is called "proportional hazards" because the effect of an increase in a predictor variable is to multiply the hazard by a factor that is constant over time, meaning the hazards for different individuals are proportional. The Cox Proportional Hazards model remains a cornerstone in survival analysis due to its robustness and the detailed insights it provides into the effects of various covariates on survival times. It is widely used across many fields of research, particularly in clinical studies to determine the influence of treatment procedures or other risk factors on patient outcomes. However, these models often require assumptions about the data's hazard functions and may not capture complex interactions in high-dimensional datasets, such as those derived from histological slides. For instance, traditional Cox models can only handle the linear condition in the risk function[93]. DeepSurv[94] method has been proposed to replace the exponential part β^{Tx} by a multilayer perceptron (MLP). However, this model was designed for structured data such as clinical tables. Zhu et al.[95] added a convolutional part before the fully connected layer predicting the risk to generate better image-based features for the survival prediction task, as shown by **Figure 7**.

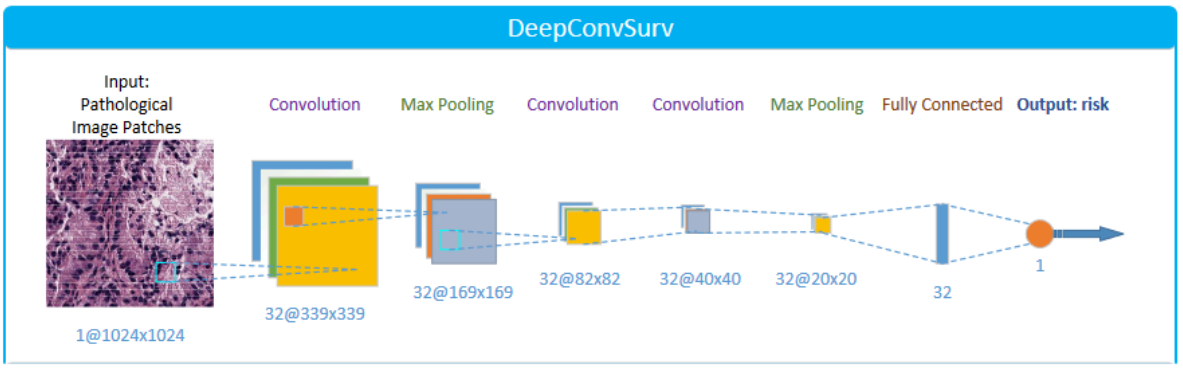


Figure 7. DeepConvSurv architecture for image-based survival prediction. Schematic representation of the DeepConvSurv model, a convolutional neural network (CNN) designed to predict patient survival directly from histopathological image patches. The network sequentially applies convolution and max-pooling operations to extract multi-scale morphological features, followed by fully connected layers producing a single continuous risk score. Unlike standard CNN classifiers, DeepConvSurv optimises a negative partial log-likelihood loss, adapting the Cox proportional hazards framework for survival analysis. This approach exemplifies direct end-to-end prediction of prognostic risk from pathology images, as referenced in this literature review.

Adapted from Zhu et al., DeepConvSurv: Deep Convolutional Neural Networks for Survival Analysis, Bioinformatics, 2016.

As the main approach in survival prediction uses coxPH derived models as a loss function, some authors suggested developing other losses. The concordance index (C-index) is the predominant metric used in time-to-event prediction models, calculated as the ratio of concordant pairs to the total number of possible pairs for assessment. A pair is deemed concordant when the model accurately predicts that one patient is at higher risk than another. The C-index is 1 in the case of perfect ordering and 0.5 for random sorting. For deep learning models, the loss function must be differentiable to facilitate efficient training using gradient-based optimizers. Meier et al. (2020)[96] designed two specific losses to train their NNs : i) the Uno loss to maximise the concordance index, which specifically considers the ability of the model to rank patients by their survival times accurately, and ii) the Logrank loss, based on the Logrank test statistic, which is used to compare the survival distributions between groups divided based on their predicted risks. Its optimization focuses on maximising the difference in survival times across these groups, aiming to create a model that is effective at distinguishing between different risk levels in a way that is statistically significant in terms of survival curves. If the new approaches are interesting, comparisons

between DeepSurv derived methods and C-Index or Logrank test optimization lack sufficient literature. They are very similar as the first one predicts the risk score directly (without the linearity assumption of the traditional CoxPH method) and the latter emphasises the model's ability to rank patients accurately according to their risk of an event, which is based on the risk score.

DeepHit[97] is a more recent approach that takes a different path by focusing on directly predicting the survival time distribution for each individual, rather than modelling the hazard function or relative risks. DeepHit uses a multi-task learning setup where one part of the network predicts the survival time distribution across discrete time intervals, and another part may predict auxiliary tasks like the cause of death in competing risk scenarios. It aims to model the entire survival distribution rather than just the hazard function. This allows it to provide detailed predictions about the likelihood of survival at each time point, which is particularly useful in medical settings where detailed time-to-event predictions are needed.

While deep learning models offer significant advancements in survival prediction, they also present challenges such as the need for large annotated datasets, computational complexity, and the interpretability of the model outputs. Furthermore, the integration of such models into clinical practice requires rigorous validation and explanation to ensure they provide reliable and actionable insights. Future research in this area could focus on improving the interpretability of deep survival models, integrating multimodal data (e.g., combining histological images with genetic information), and developing federated learning approaches to train models on decentralised datasets while maintaining patient privacy. Over the past decade, digital pathology has rapidly evolved, and deep learning now enables us to extract prognostic and therapeutic signals directly from routine histological slides, a shift we explored here through the lens of post-NAC breast cancer specimens.

1.2.3 Research aims

The aim of this research is to better characterise the residual disease (RD) on surgical specimens (SS) of breast cancer patients after neoadjuvant chemotherapy (NAC) using a deep learning approach that utilises histopathological information. Our goal is to better stratify patients according to their risk of recurrence and survival. The characterisation of the residual disease must be fine-tuned and take into account tumour heterogeneity (molecular subtypes

and if possible genomic alterations). The ultimate goal would be to apply these insights clinically to improve patient treatment plans.

1.3 Proposed plan

A vital consideration is that neural networks, which derive predictions from Whole Slide Images (WSI), are essentially 'black boxes'. They have several limitations, such as sensitivity to data shifts and adversarial examples and an inability to explain their predictions. These issues are key roadblocks for the incorporation of predictive AI into clinical practice, as they undermine the trust clinicians place in such systems. In this thesis, we plan to develop a deep learning pipeline dedicated to analysing Whole Slide Images (WSIs) of tumours. Our methodology aims to encompass a comprehensive scope to align closely with clinical practice. To achieve this, our strategy will follow three interlinked pathways:

i) **End-to-end post-NAC surgical specimen stratification**: We will design a neural network that performs end-to-end analysis of post-NAC surgical specimens WSIs, thereby facilitating stratification using deep learning methodologies.

ii) **Integration of Molecular Data and Inference of Missing Molecular Information**: Leveraging available molecular data, we will utilise neural networks to infer absent molecular biology information based on morpho-molecular correlations. Besides advancing tissue stratification, our method aims to elucidate potential molecular underpinnings, bridging the divide between histology and molecular biology.

iii) **Identification of Prognostic Correlated Pathological Markers**: Our last objective is to develop a tool that pinpoints markers traditionally used by pathologists that have the highest correlation with prognosis. Such an endeavour ensures that our solution maintains a degree of interpretability, essential for both clinical utility and trustworthiness.

Chapter 2: Material and Methods.

In this thesis, we will use two main data types as inputs for our prediction tasks: biopsy and surgical specimens whole slide images. They come from 3 different datasets: the PRIMUNEO dataset, the Centre Georges François Leclerc (CGFL) breast cancer neoadjuvant dataset and the Tumor Cancer Genome Atlas (TCGA) project. The cohorts and experimental setups are described in this **Material and Methods** section. Specific adaptations are covered in the dedicated sections throughout this thesis.

2.1 Neoadjuvant early Breast Cancer Datasets

Our main task is to stratify the residual disease of post-NAC early breast cancer patients. To do so, our main dataset are sourced from two distinct eBC patients cohorts: i) the PRIMUNEO dataset, a French multicentric prospective dataset from the PRIMUNEO study (ClinicalTrials.gov Identifier: NCT01513408), conducted between May 2012 and February 2015 at different cancer centers in France, and dedicated to the identification of predictive/prognostic histopathological factors in eBC patients treated with standard NAC (this study was funded by a grant from the French Ministry of Health PHRC-K2011); and ii) the CGFL breast cancer neoadjuvant dataset, a retrospective single-center database generated from the neoadjuvant treated population in a single French cancer center (Centre Georges François Leclerc, Dijon) between the early 2000s and 2022.

This study was conducted in compliance with the Declaration of Helsinki and received approval from the Institutional Review Board and the CNIL (French national commission for data privacy). Informed consent was obtained from all participants and/or their legal guardians.

Variables	PRIMUNEO (N=326)	CGFL breast cancer neoadjuvant database (N=490)	P value
Age, median, range	51 [25 – 79]	52 [23 - 88]	0.4591
Menopausal status			0.9046

Premenopausal	163 (52.9%)	246 (53.4%)	
Postmenopausal	145 (47.1%)	215 (46.6%)	
Missing	18	29	
Breast surgery			0.6483
Radical	158 (48.5%)	239 (50.1%)	
Conservative	168 (51.5%)	238 (49.9%)	
Missing	0	13	
Type of NAC			< 0.001
Others	3 (0.9%)	20 (4.1%)	
Taxanes	44 (13.5%)	97 (19.8%)	
Anthracyclines	2 (0.6%)	81 (16.5%)	
Anthracyclines and taxanes	277 (85.0%)	292 (59.6%)	
Ki-67 percentage			< 0.001
≤ 14%	20 (16.3%)	8 (15.7%)	
> 14% and < 30%	33 (26.8%)	15 (29.4%)	
≥30%	70 (56.9%)	28 (59.9%)	
Missing	203	439	
Mitotic index			< 0.001
Low (0 to 6 mitoses)	49 (36.0%)	217 (67.6%)	
Medium (7 to 12 mitoses)	37 (27.2%)	59 (18.3%)	
High (> 12 mitoses)	50 (36.8%)	45 (14%)	
Missing	190	169	

tumour Molecular Subtype			< 0.001
HER2+	80 (24.5%)	220 (44.9%)	
ER+/HER2-	138 (42.3%)	156 (31.8%)	
TN	108 (33.1%)	114 (23.2%)	
ER expression			0.3734
Negative	148 (45.4%)	207 (42.2%)	
Positive	178 (54.6%)	283 (57.8%)	
PR expression			0.7134
Negative	203 (62.3%)	274 (55.9%)	
Positive	123 (37.7%)	216 (44.0%)	
Nottingham Combined Histologic Grade			< 0.001
1	6 (1.9%)	45 (9.6%)	
2	121 (38.4%)	220 (46.8%)	
3	188 (59.7%)	205 (43.6%)	
Missing values	11	20	
cT stage			0.01947
T0	2 (0.6%)	2 (0.4%)	
T1	28 (8.8%)	36 (7.4%)	
T2	188 (58.9%)	314 (64.3%)	
T3	60 (18.8%)	54 (11.2%)	
T4	41 (12.9%)	82 (16.8%)	
Missing	7	2	
cN stage			< 0.001

N0	148 (49.3%)	164 (33.8%)	
N1	139 (46.3%)	207 (42.7%)	
N2	5 (1.7%)	47 (9.7%)	
N3	8 (2.7%)	67 (13.8%)	
Missing	26	5	
cAJCC stage			0.00687
I	16 (5.2%)	13 (2.7%)	
II	214 (69.0%)	301 (61.2%)	
III	80 (25.8%)	171 (35.3%)	
Missing	16	5	
pT stage			< 0.001
T0	81 (26.2%)	101 (20.6%)	
T1	133 (43.0%)	262 (53.5%)	
T2	69 (22.3%)	102 (20.8%)	
T3	20 (6.5%)	23 (4.7%)	
T4	6 (1.9%)	2 (0.4%)	
Missing	17	0	
pN stage			0.2339
N0	197 (60.8%)	261 (53.8%)	
N1	77 (23.8%)	132 (27.2%)	
N2	37 (11.4%)	72 (14.8%)	
N3	13 (4.0%)	20 (4.1%)	
Missing	2	5	
pAJCC stage			0.7091

0	81 (25.3%)	111 (22.9%)	
I	82 (25.6%)	124 (25.6%)	
II	100 (31.3%)	149 (30.7%)	
III	57 (17.8%)	101 (20.8%)	
Missing	6	5	
pCR/RD status			0.01741
RD	245 (75.2%)	402 (82.0%)	
pCR	81 (24.8%)	88 (18.0%)	

Table 1. PRIMUNEO and CGFL breast cancer neoadjuvant dataset characteristics for pCR prediction using the initial biopsy.

NAC, neoadjuvant chemotherapy; HER2, human epidermal growth factor; HER2+, HER2-positive; HER2-, HER2-negative; ER, estrogen receptor; ER+, estrogen receptor positive; TN, Triple Negative; PR, progesterone receptor; AJCC, American Joint Committee on Cancer; c, clinical; p, pathological; T, tumour; N, node; RT, radiotherapy; pCR, pathological Complete Response.

Table 1 summarizes the clinical and pathological characteristics of all patients included in the study. We tested the differences in characteristics between the two datasets and found no statistically significant difference in menopausal status (P=0.9046), breast surgery type (P=0.6483), ER expression (P=0.3734), PR expression (P=0.7134) or pathologic node stage and pathologic tumour stage (pAJCC), with p-values of 0.2339 and 0.7091, respectively. However, we found statistically significant differences in clinical tumour stage (P=0.01947), clinical AJCC stage (P=0.00687), the Ki-67 percentage, Mitotic index, clinical node stage and pathologic tumour stage (P<0.001), Nottingham Combined Histologic Grade (P<0.001), tumour molecular subtype (P<0.001), NAC treatment protocol (P<0.001), and pCR status (P=0.01741). Notably, the PRIMUNEO dataset exhibited a higher incidence of patients achieving pCR compared to the CGFL dataset, at 24.8% and 18% respectively. Regarding treatment, most of the patients received a combination of anthracycline and taxane therapies, 85% in the PRIMUNEO cohort and 59.6% in the CGFL dataset ('Other' category in **Table 1** also including intensified treatments combining both anthracyclines and taxanes). All patients

with HER2 amplification were treated with tailored trastuzumab therapy along with standard chemotherapy. All patients in our training and validation cohorts were treated before 2022, i.e. before the widespread adoption in France of KEYNOTE-522-type regimens combining anthracycline–taxane chemotherapy with carboplatin and pembrolizumab in triple-negative breast cancer (TNBC), and before routine dual HER2 blockade (trastuzumab + pertuzumab) in HER2-positive disease. No patient received neoadjuvant carboplatin, pembrolizumab, or pertuzumab.

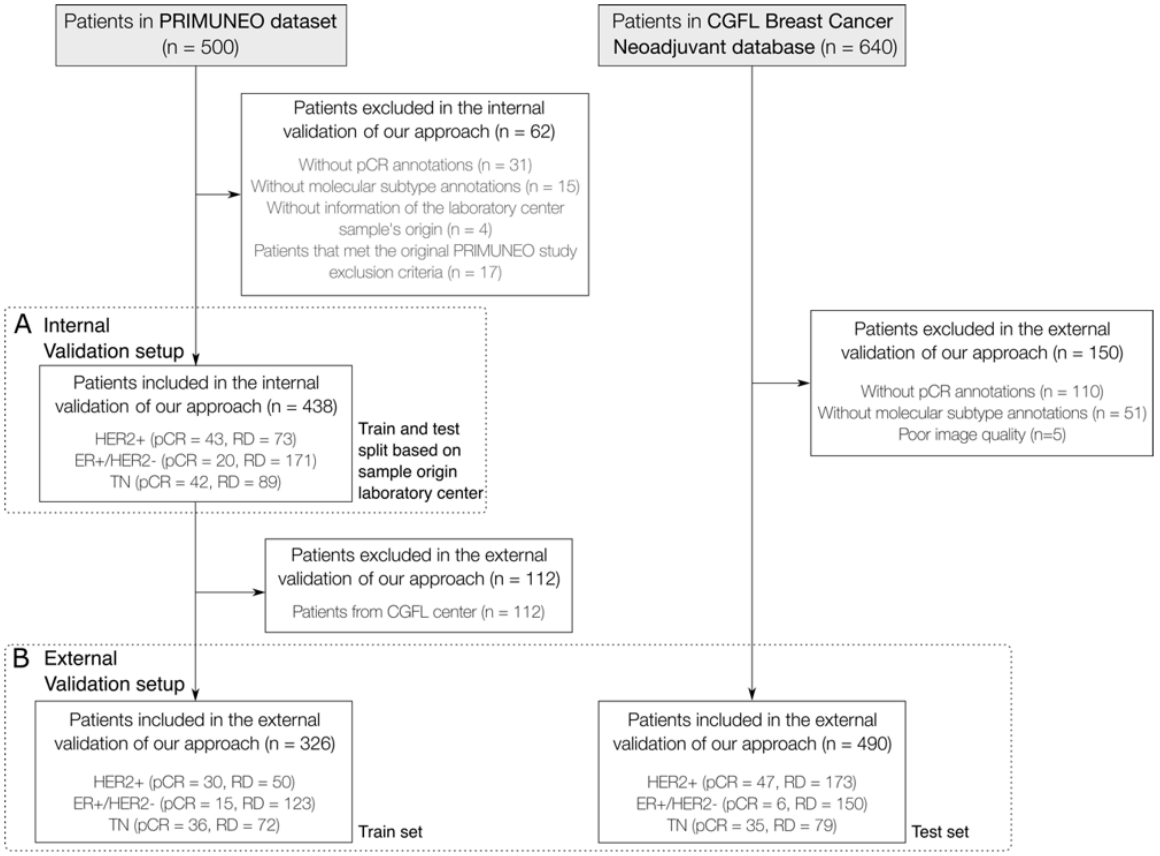


Figure 8. Workflow for patient inclusion and dataset partitioning in the PRIMUNEO and CGFL Breast Cancer Neoadjuvant cohorts. The flow diagram summarises patient selection, quality control, and dataset allocation for the internal and external validation setups. (A) Internal validation: patients from the PRIMUNEO cohort (n = 500) were screened for eligibility, excluding cases lacking pCR annotations, molecular subtype information, or adequate image quality. The resulting 438 patients were randomly partitioned into training and test subsets based on their laboratory of origin to prevent site-specific data leakage.
(B) External validation: to ensure full independence, patients from the Dijon centre (PRIMUNEO) were excluded from the training pool, and the external test set was composed

exclusively of patients from the CGFL Breast Cancer Neoadjuvant cohort, coming from Dijon (n = 490).

For each setup, the number of patients with pathological complete response (pCR) or residual disease (RD) is reported per molecular subtype (HER2+, ER+/HER2−, TNBC). This workflow ensures reproducible cohort curation, minimises cross-centre bias, and defines the foundation for the analytical validation presented in later sections.

Following the patient selection procedure to keep all patients with an initial biopsy slide and a surgical specimen with pCR information, as detailed in **Figure 8**, our study cohort was refined to 928 participants from the initial 1140. This included 438 from the PRIMUNEO dataset and 490 from the CGFL Breast Cancer Neoadjuvant database.

2.1.1 PRIMUNEO Database

A total of 500 patients who received NAC between May 2012 and February 2015 in 12 different cancer centers in France were enrolled in the PRIMUNEO study. Briefly, the main inclusion criteria were female patients between the ages of 18 and 80 years, with proven localized breast cancer, regardless of the histological type or molecular subtype (HER2-amplified, ER+/HER2-, TNBC); treated with standard NAC incorporating taxanes ± anthracyclines (treatment protocol at the physician's discretion, see **Table 1** for more information). The main exclusion criteria were metastatic breast cancer; neoadjuvant radiotherapy; patient not amenable to surgery; and ongoing therapy for any other type of cancer. A total of 438 hematoxylin and eosin (H&E)-stained slides were used from this dataset; 62 patients were excluded because they met the original PRIMUNEO study exclusion criteria (n=17) or had no available molecular subtype status (n=15), no available pathological response report (n=31) or no information of the laboratory origin (n=4) (**Figure 8a**). Some patients had more than one exclusion criteria.

Table 2. Describes the repartition of pCR and RD for each molecular subtype in the PRIMUNEO dataset. As known in the literature, ER+/HER2- tumours have significantly less pCR compared to the other type, although it is not as well linked with survival compared to TNBC or HER2+ tumours.

Center	Total patients	HER2+		ER+/HER2-		TN	
		pCR	RD	pCR	RD	pCR	RD
DIJON	112	13	23	5	48	6	17
STRASBOURG	69	10	3	4	24	9	19
REIMS	48	5	11	3	17	4	8
RENNES	23	2	4	0	6	2	9
NANCY	38	3	8	2	10	7	8
CAEN	34	3	6	0	14	3	8
NICE	26	2	4	4	8	2	6
PARIS	30	1	5	1	12	4	7
TOULOUSE	25	1	3	1	15	3	2
ST CLOUD	23	2	3	0	13	2	3
BORDEAUX	9	1	2	0	4	0	2
CLERMONT	1	0	1	0	0	0	0
TOTAL	**438**	**43**	**73**	**20**	**171**	**42**	**89**

Table 2. Description of the PRIMUNEO Database, including the number of patients exhibiting either pathological complete response (pCR) or residual disease (RD), categorized by the tumour's molecular subtype within each center.

Regarding survival outcomes, **Table 3** displays relapse and deaths events for each molecular subtype for the PRIMUNEO dataset.

Cancer Subtype	DFS (censored)	DFS (relapse)	Total Patients	OS (censored)	OS (death)	Total Patients
HER2+	92	24	116	108	8	116
HER2-/ RH+	150	41	191	167	24	191
TNBC	86	45	131	95	36	131

Table 3. Patients that present the event or are censored in the DFS and OS task stratified by molecular subtype in the PRIMUNEO dataset.

OS and DFS data distribution in the PRIMUNEO dataset are shown in **Figure 9** and **Figure 10**.

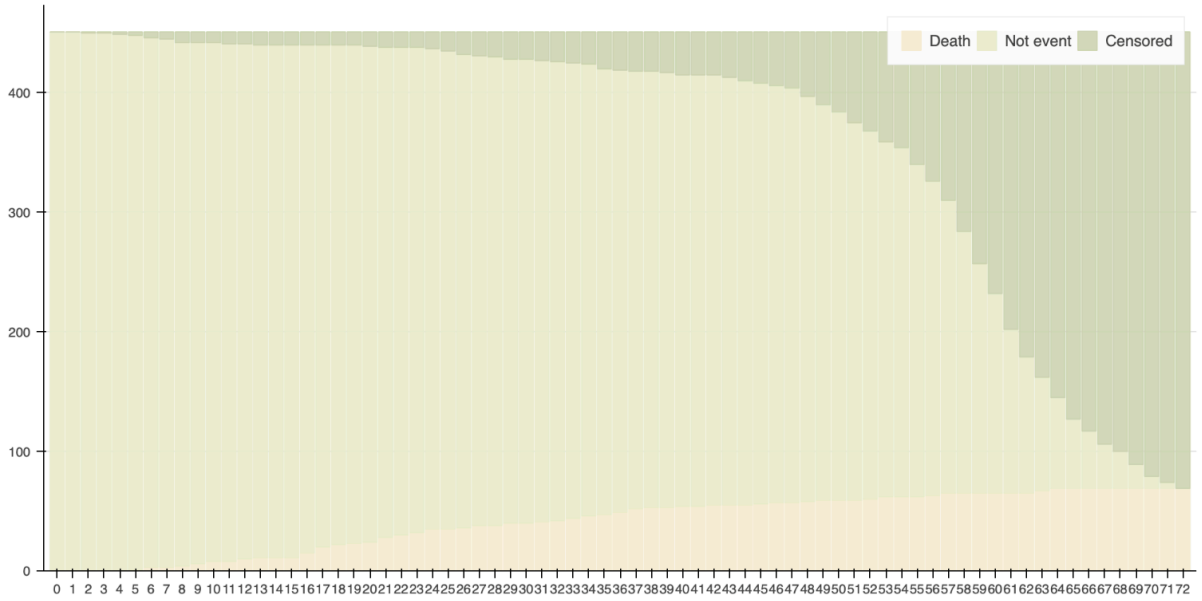


Figure 9. Distribution of survival events and censoring over time in the PRIMUNEO dataset. Stacked area plot showing the number of patients at risk, censored, or deceased at each monthly interval. "Death" corresponds to the occurrence of the event of interest, while "Not event" indicates patients still under observation, and "Censored" those lost to follow-up or whose observation ended before an event occurred. This visualisation illustrates the censoring pattern and follow-up duration used in subsequent survival analyses.

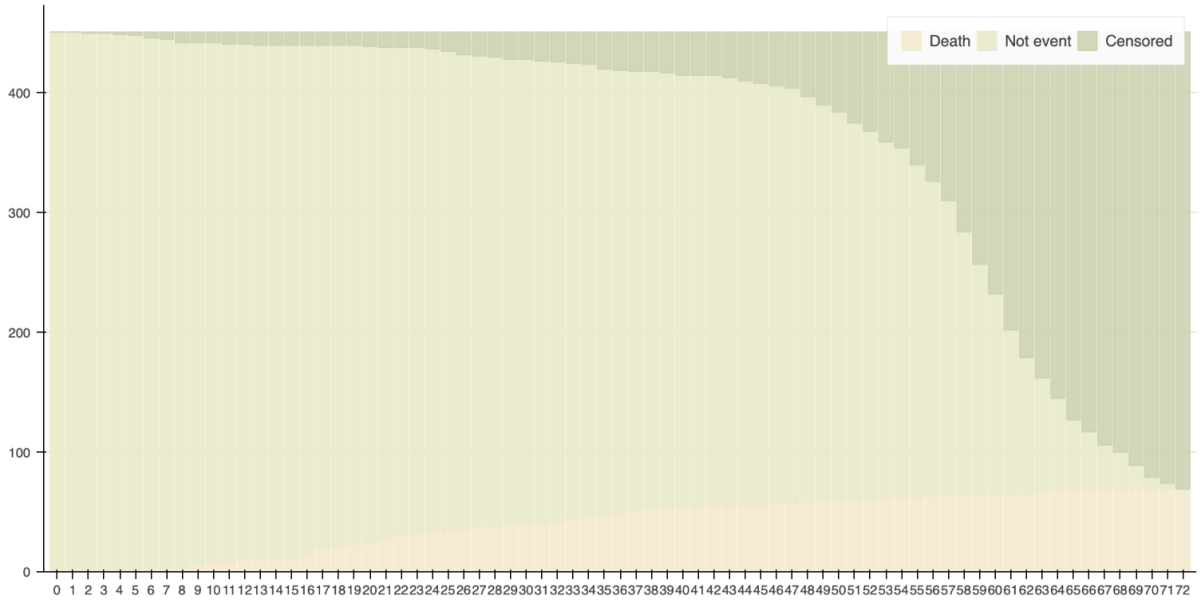


Figure 10. Distribution of relapse events and censoring over time in the PRIMUNEO dataset. Stacked area plot showing the number of patients at risk, censored, or presenting a relapse at each monthly interval. "Relapse" corresponds to the occurrence of the event of interest, while "Not event" indicates patients still under observation, and "Censored" those

lost to follow-up or whose observation ended before an event occurred. This visualisation illustrates the censoring pattern and follow-up duration used in subsequent survival analyses.

2.1.2 CGFL Breast Cancer Neoadjuvant dataset

The CGFL breast cancer neoadjuvant dataset comprises 1319 digitized WSI of initial diagnostic tumour biopsies obtained from 640 patients who received NAC between January 2000 and January 2022, some patients having more than 1 slide per tumour. Of these, 150 were excluded. Exclusion was for the following reasons: poor quality WSI (n=5), no molecular subtype status (n=51) or no available pathological response report (n=110) (**Figure 8b**). Some patients had more than one exclusion criteria. A more detailed description of this cohort is provided in **Table 1 and Table 4.**

Cancer subtype	Number of patients	pCR	RD	Prevalence
HER2+	220	47	173	0.214
ER+/HER2-	156	6	150	0.038
TN	114	35	79	0.307
Total	**490**	**88**	**402**	**0.180**

Table 4. CGFL breast cancer neoadjuvant database description according to tumour molecular subtype.

Regarding survival outcomes, **Table 5** displays relapse and death events for each molecular subtype for the CGFL Neoadj dataset.

Cancer Subtype	DFS (censored)	DFS (relapse)	Total Patients	OS (censored)	OS (death)	Total Patients
HER2+	161	59	220	202	18	220
HER2-/ RH+	97	59	156	131	25	156
TNBC	57	57	114	91	23	114

Table 5. Patients that present the event or are censored in the DFS and OS task stratified by molecular subtype in the CGFL Neoadj dataset.

OS and DFS data distribution in the CGFL Neoadj dataset are shown in **Figure 11** and **Figure 12**.

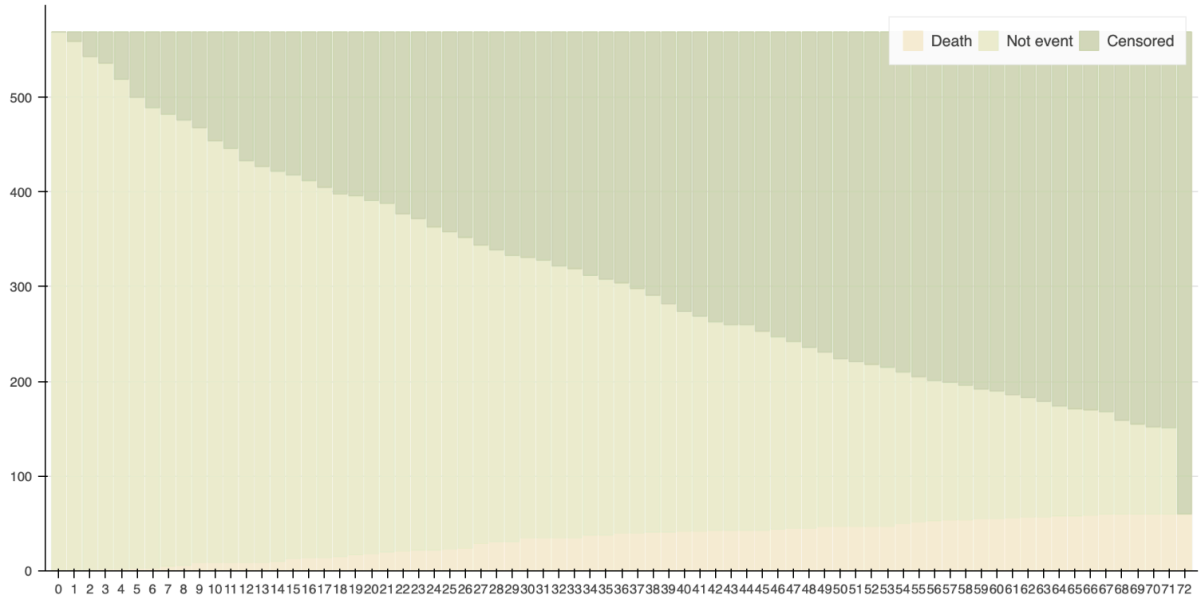


Figure 11. Distribution of survival events and censoring over time in the CGFL Neoadj dataset. Stacked area plot showing the number of patients at risk, censored, or deceased at each monthly interval. "Death" corresponds to the occurrence of the event of interest, while "Not event" indicates patients still under observation, and "Censored" those lost to follow-up or whose observation ended before an event occurred. This visualisation illustrates the censoring pattern and follow-up duration used in subsequent survival analyses.

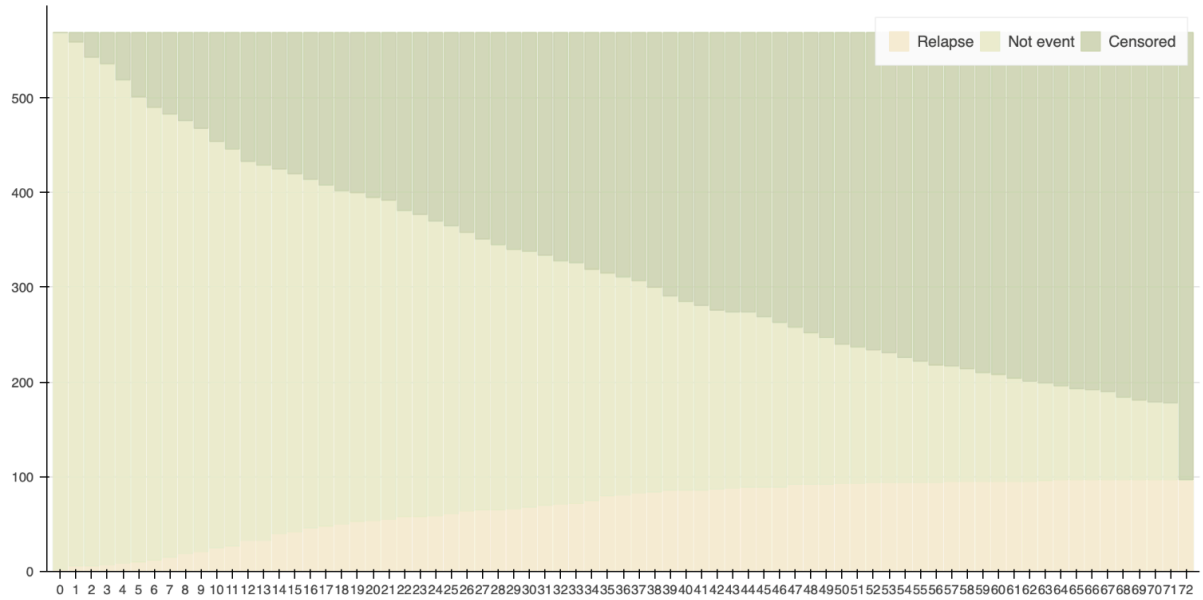


Figure 12. Distribution of relapse events and censoring over time in the CGFL Neoadj dataset. Stacked area plot showing the number of patients at risk, censored, or deceased at

each monthly interval. “Relapse” corresponds to the occurrence of the event of interest, while “Not event” indicates patients still under observation, and “Censored” those lost to follow-up or whose observation ended before an event occurred. This visualisation illustrates the censoring pattern and follow-up duration used in subsequent survival analyses.

When examining the distribution of survival events across the two datasets, notable differences emerge. In the PRIMUNEO dataset (**Figures 9 and 10**), the number of death events is relatively low and occurs more gradually over time, whereas the CGFL Neoadj dataset (**Figures 11 and 12**) shows a higher and more immediate incidence of such events, with more censored events. These differences reflect the nature of the cohorts: PRIMUNEO is derived from a controlled clinical trial environment, where patient selection criteria, follow-up protocols, and treatment regimens are standardised and closely monitored. In contrast, CGFL Neoadj represents real-world clinical practice, where patient populations are more heterogeneous and may include higher-risk or less tightly managed cases. The higher event rate observed in CGFL reflects the variability of clinical reality, including factors such as comorbidities, variable treatment adherence, and differential access to post-operative therapies. These differences must be considered when interpreting model generalisability and comparing performance across cohorts.

2.1.3 Internal and External validation study design

The complete flowchart of internal and external validation sets is provided in **Figure 8a - 8b**. In order to ensure that the site of origin was not biasing the prediction performance, as previously shown by Howard et al. (2021)[115], we used a stratified grouped cross-validation approach for the internal validation (**Figure 13**). We used 4 distinct training and test sets preserving similar pCR prevalence in each subset and without site overlap between a training set and its associated test set.

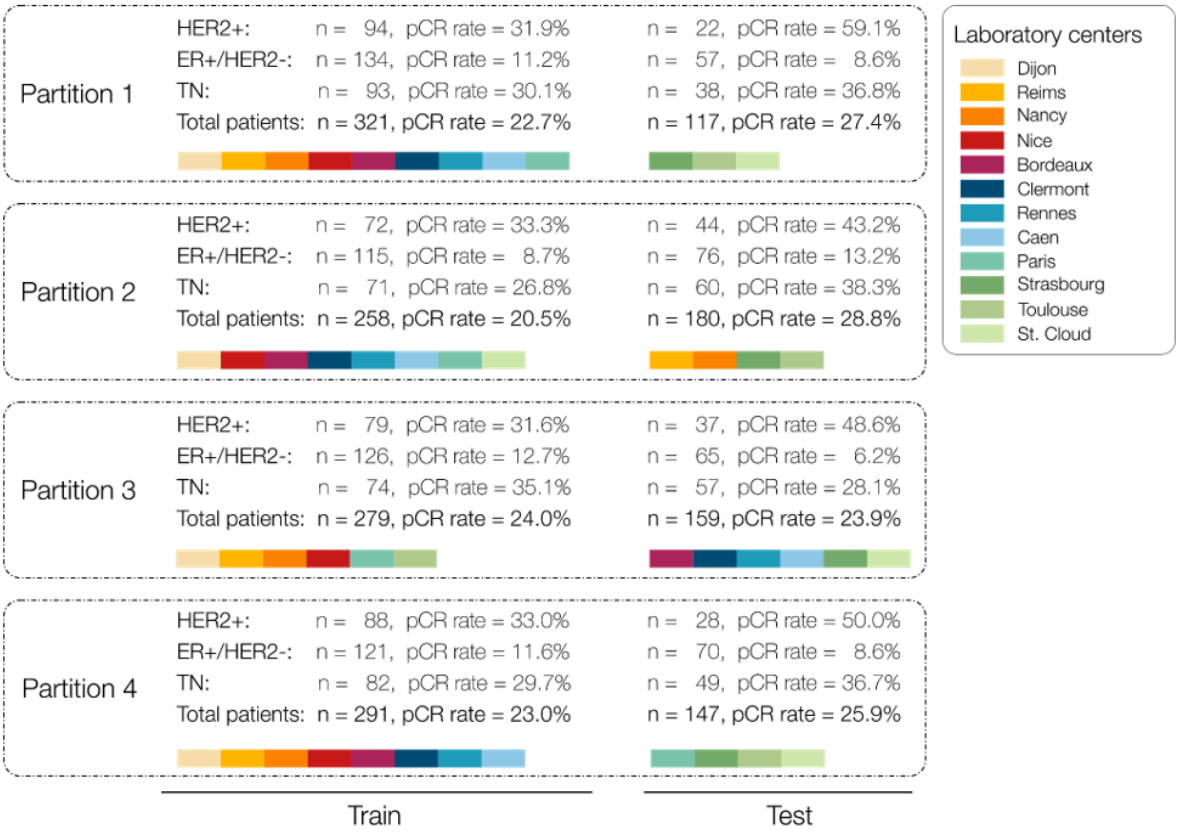


Figure 13. Overview of data partitions, patient distribution, and pathological complete response (pCR) rates in the internal validation study. Each panel represents one of the four cross-validation partitions used to train and test the models. Within each partition, patients are stratified by molecular subtype (HER2+, ER+/HER2−, and triple-negative), and their corresponding pCR rates (%) are reported for both the training and test subsets. Colours indicate the contributing pathology laboratories across France, ensuring balanced representation and mitigating centre-specific bias. This figure illustrates the consistency of subtype and institutional distribution across folds, supporting robust internal validation.

The PRIMUNEO dataset (**Figure 13**), comprising 438 patients categorized into HER2+ (n=116, with a 37.1% pCR rate), ER+/HER2- (n=191, 10.5% pCR rate), and TN breast cancer (n=131, 32.1% pCR rate).

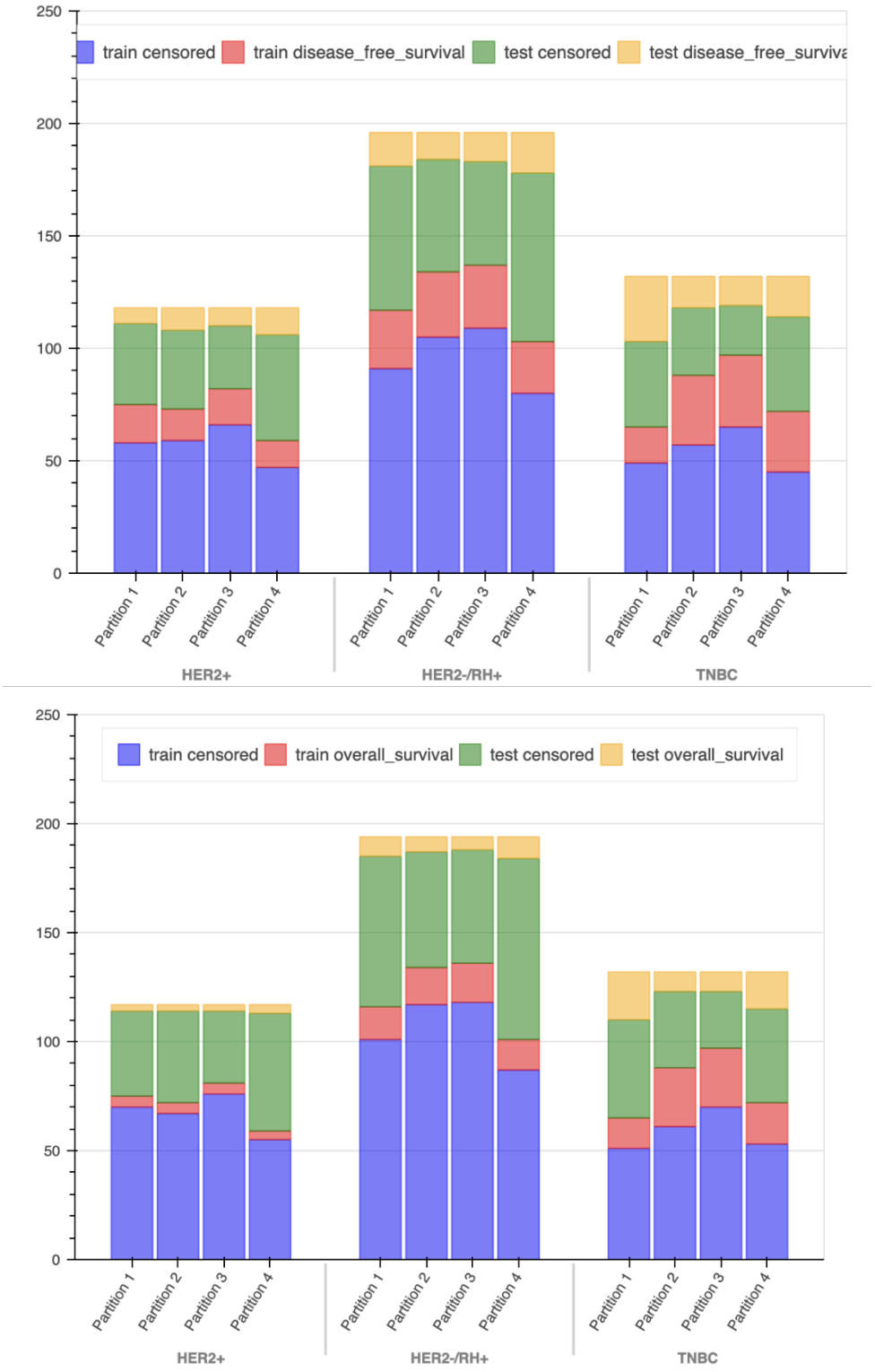


Figure 14. Distribution of survival events and censored cases across partitions in the internal validation study. (A) Disease-Free Survival (DFS) and (B) Overall Survival (OS) data distributions are shown for each molecular subtype (HER2+, ER+/HER2–, and TNBC) across the four training and testing partitions used for cross-validation. Each stacked bar represents the number of patients contributing to the survival analysis, distinguishing between censored cases (blue and orange) and patients with observed events (red and green). The

consistent height of bars across partitions demonstrates balanced sampling and equitable representation of censored and event cases between training and testing subsets, ensuring methodological robustness for downstream survival modelling.

Figure 14 illustrates the distribution of patient samples per partition across molecular subtypes (HER2+, HER2−/RH+, and TNBC) for the internal validation. The distributions show balanced partitioning across folds, both in terms of event occurrence and subtype representation, which supports the robustness of the cross-validation setup. HER2−/RH+ is the most represented subtype, followed by HER2+ and TNBC, reflecting their respective prevalence in the dataset.

For the external validation stage, we used the CGFL Breast Cancer Neoadjuvant database comprising 490 individuals characterized as follows (**Table 5**): HER2+ (n=220, 21.4% pCR rate), ER+/HER2- (n=156, 3.8% pCR rate), and TN breast cancer (n=114, 30.7% pCR rate).

Whole slide images of patients with a pCR were included in the training, but excluded in the analysis, serving as a data augmentation strategy.

2.2 Complementary datasets

The Cancer Genome Atlas (TCGA)[98] is one of the most comprehensive and widely used resources in cancer genomics, providing large-scale, multi-omics and clinical data across a broad range of tumour types. The dataset includes high-quality H&E-stained diagnostic slides from many different cancer types, extensive molecular characterisation (including mutations, gene expression, methylation, and copy number alterations), and annotated clinical outcomes, including overall survival. This combination makes TCGA particularly valuable for developing and benchmarking deep learning models that aim to uncover morpho-molecular correlates from histology and prognosis predictors. However, it is important to note that TCGA primarily reflects early-stage, treatment-naïve tumours collected at the time of diagnosis, and is not designed to capture treatment response or long-term real-world follow-up. As such, while TCGA is ideal for tasks like mutation prediction or molecular subtype inference from baseline histology, its utility for post-treatment survival modelling or chemo-sensitivity studies is limited. In this thesis, we leverage TCGA as a training ground for

inferring molecular profiles (e.g. HRD and PAM50) but rely on PRIMUNEO and CGFL Neoadj cohorts for modelling treatment response and survival outcomes in the neoadjuvant setting. Hence, they are used separately to train models which are then transferred into the prediction task on the PRIMUNEO or CGFL Neoadj datasets. For more detailed information, see **Chapter 4 Material and Methods section**.

2.3 Pathological evaluation

All tissue sections, initial biopsies, and surgical specimens after NAC were examined microscopically by experienced pathologists (LA, AB, FB). For each patient included in the study, a tumour block from the initial biopsy and a representative block of residual tumour was chosen by the reference pathologist in each investigating center. Depending on the quality of pathological response, this representative block could come from an area of complete tumoural regression (in case of pCR), an area of partial tumour regression, or an area of unmodified residual tumour (in case of non-pCR). Pathological complete response (pCR) was defined as the disappearance of an invasive tumour on the surgical specimen and in the lymph nodes after NAC (pT0 pN0). The pathological response was dichotomized as pCR vs residual disease (RD).

RD can be subdivided into two groups: "no response" and "partial response". "No response" was defined as cases where the AJCC stage either remained the same or progressed compared to the stage at diagnosis. "Partial response" referred to cases where the AJCC stage decreased but remained above stage 0 compared to the initial diagnosis. Tumour infiltrating lymphocyte (TIL) quantification, and Residual Cancer Burden (RCB) information were not available in these two cohorts. The estrogen receptor (ER), progesterone receptor (PR), and HER2 status were assessed by immunohistochemistry (IHC). HER2 status was defined as positive only when IHC (3+) or IHC (2+) and HER2 amplification by fluorescence in situ hybridization (FISH), while breast cancer with IHC (0/1+) or IHC (2+) without HER2 amplification by FISH were considered as HER2-negative disease. ER/PR positivity was defined as positive nucleus staining in more than 10% of tumour cells. The nuclear grade was assessed based on the Nottingham grading system, and clinical staging according to the American Joint Committee on Cancer (AJCC) classification.

2.4 Clinical Variables

For all analyses involving clinical information, we compiled a comprehensive set of patient-level variables describing demographics, tumour biology, pre- and post-treatment status, and pathological response. These features were used for all models relying on clinical data, including CoxPH, DeepHit, and image–clinical fusion frameworks. Categorical variables were one-hot encoded, continuous variables were standardised, and missing information was encoded as zero vectors.

The clinical variables included:

- **Age at surgery**, time from diagnosis to surgery (in months), and the interval between diagnosis and surgery.
- **Pre- and post-neoadjuvant chemotherapy molecular subtypes**: Binary variables indicating HER2+, HER2–/HR+ (hormone receptor-positive), and TNBC (triple-negative breast cancer) status both at diagnosis (subtype) and on the surgical specimen (psubtype).
- **Histological grade**: Scarff-Bloom-Richardson (SBR) grade before neoadjuvant chemotherapy and post-treatment SBR grade (PSBR) assessed on the surgical specimen.
- **Staging**: Pretreatment clinical stage (AJCC) and post-treatment pathological stage (PAJCC) according to the American Joint Committee on Cancer classification.
- **Pathological complete response (pCR):** Binary variable defined as no residual invasive disease in the breast and axilla (ypT0 ypN0) after neoadjuvant chemotherapy.

This unified clinical feature set was used consistently throughout the survival-prediction experiments whenever clinical information was included, allowing direct comparison across modelling strategies and facilitating integration with image-derived risk scores in later chapters.

2.5 Image processing

As for the surgical specimen, the H&E-stained WSIs of initial biopsies were obtained using a Hamamatsu Nanozoomer 2.0HT scanner at 40 × magnification.

2.6 Metrics

To assess the performance of our survival prediction models throughout this thesis, we employed a combination of discrimination-based metrics, including concordance indices and area under the curve (AUC), as well as aggregated metrics designed to account for censoring imbalance and temporal consistency. The following metrics were used across the various tasks, with evaluation intervals tailored to the clinical endpoint of interest: years 2–5 for Disease-Free Survival (DFS), and years 3–5 for Overall Survival (OS).

Harrell's Concordance Index (C-index Harrell)

The C-index, introduced by Harrell, is a widely used metric to evaluate the discriminative power of survival models. It measures the proportion of all comparable pairs of subjects whose predicted risk scores are correctly ordered in accordance with observed event times. The index ranges from 0.5 (random performance) to 1.0 (perfect concordance).

C-index Harrell is formally defined as:

$$C = \frac{\sum_{i<j} \mathbf{1}(T_j < T_i) \cdot \mathbf{1}(\hat{\eta}_j > \hat{\eta}_i) \cdot \delta_j}{\sum_{i<j} \mathbf{1}(T_j < T_i) \cdot \delta_j}$$

Equation 1

Where:

- T_i, T_j: observed time-to-event for patients i and j
- $\hat{\eta}_i, \hat{\eta}_j$: predicted risk scores
- δ_j: event indicator (1 if patient j had the event, 0 if censored)
- $\mathbf{1}(\cdot)$: indicator function

This metric evaluates whether, for any pair where one patient has experienced an event and the other has not, the model assigns a higher risk score to the event patient.

Uno's Concordance Index (C-index Uno)

C-index Uno is a modified version of the concordance index that adjusts for censoring using Inverse Probability of Censoring Weighting (IPCW). It is considered more appropriate when the censoring distribution is non-uniform or when the censoring rate is high.

It is computed via IPCW estimators as described in Uno et al. (2011). Although similar in interpretation to Harrell's C-index, it generally provides a more robust assessment of model performance in censored data.

Time-Dependent AUC (AUCt)

To assess model discrimination at specific time points, we computed the Area Under the Receiver Operating Characteristic Curve (AUC) for each year t, denoted as AUCt. It evaluates the model's ability to rank patients who will experience the event at time t against those who will not.

The overall performance is reported using the mean AUC across the clinically relevant time windows:

- Years 2–5 for DFS
- Years 3–5 for OS

The AUCs at each year are computed using IPCW-adjusted estimators to account for censoring.

Weighted Mean AUC

The Weighted AUC Mean is a summary measure that averages the yearly AUCs across multiple time points, assigning weights based on the number of uncensored patients (at risk) at each time t. This approach avoids bias introduced by imbalanced event occurrence across years.

$$\text{Weighted AUC} = \frac{\sum_t w_t \cdot AUC_t}{\sum_t w_t}$$

Where wt is the number of uncensored patients at risk at time t.

C-t Index (DeepHit)

The C-t Index is used to evaluate ranking performance over time in discrete-time survival models such as DeepHit. It computes concordance at each time point by comparing predicted survival scores for all patient pairs and aggregating over the full prediction window.

Similar to the C-index, a higher C-t Index indicates better temporal ranking of risk, particularly in multi-output neural networks.

Weighted Mean C-t Index

As with the AUC, the C-t Index can also be averaged over time points with weights proportional to the number of patients at risk. This allows a time-aware evaluation of the model's temporal consistency and ranking quality.

Risk Score

The Risk Score is the output of the survival model (typically the CoxPH log-risk or the final sigmoid activation in a neural survival model). It serves as a latent representation of predicted hazard and is used for:

- Stratifying patients into high- and low-risk groups
- Calculating pairwise concordance
- Generating Kaplan–Meier curves for clinical interpretation

These scores are continuous and uncalibrated, and require post-processing (e.g. thresholding) to produce clinically actionable stratifications.

Having defined the datasets, preprocessing workflows, pathological evaluation standards, and performance metrics to be used throughout this thesis, we now present the first set of experimental results. Chapter 3 focuses on the central clinical question of residual disease stratification after neoadjuvant chemotherapy. Using the PRIMUNEO cohort for development and the CGFL dataset for external validation, we systematically compare a wide range of survival modelling strategies from classical Cox proportional hazards models to advanced deep learning approaches applied to post-NAC whole-slide images.

Chapter 3: Breast cancer residual disease stratification.

Abstract

In **Chapter 3**, we systematically developed and evaluated a deep-learning pipeline for predicting overall survival (OS) and disease-free survival (DFS) from post-neoadjuvant chemotherapy (post-NAC) surgical specimens. We began with a baseline model using pre-computed embeddings and a Cox proportional-hazards framework to establish reference performance. We then explored a range of machine-learning-based survival models (e.g., Random Forests, Survival SVM) and compared them to more advanced deep-learning survival approaches, including DeepSurv and DeepHit, assessing their predictive accuracy and stability. To complement image-derived predictions, we investigated the predictive power of clinical variables alone, building a Clinical Model and analysing its contribution both independently and in combination with histological features. In parallel, we explored unsupervised pre-training strategies, including contrastive learning and self-supervised Vision Transformer (ViT) backbones, to enhance feature extraction. We further developed and tested an end-to-end survival-prediction model, directly mapping raw histological input to survival-risk scores, and benchmarked its performance against modular architectures. Finally, we conducted an external validation of our best-performing pipelines using an independent dataset (CGFL NeoAdj), evaluating the generalisability and robustness of our approach in predicting long-term outcomes from surgical histology.

The methodological benchmarking performed in this chapter revealed a recurring pattern: in controlled datasets such as TCGA, sophisticated deep architectures, multi-task survival networks and end-to-end whole-slide approaches, can achieve very high apparent performance, yet they overfit severely when applied to unseen, real-world data. In contrast, simpler pipelines, notably foundation-model feature extraction followed by Cox proportional-hazards regression, generalised more reliably. This finding tempers the optimism of much of the recent methodological literature: apparent improvements on public benchmarks often reflect data familiarity rather than generalisable learning. It highlights that model simplicity, transparency, and clinical plausibility may outweigh incremental AUC or C-index gains from highly parameterised architectures. For the field, this advocates a shift

from benchmark-driven performance inflation toward prospective, cross-site validation as the true gold standard for evaluating survival-prediction models.

3.1 Introduction

As extensively covered in the literature review (**Chapter 1**), breast cancer remains one of the most prevalent health concerns globally, with significant morbidity and mortality rates, impacting millions of persons each year. Despite advancements in screening, diagnostic tools, and targeted therapies that have contributed to better patient outcomes, challenges persist, particularly concerning overall survival (OS) and disease-free survival (DFS), which are crucial measures of therapeutic success and disease prognosis[99]. OS, defined as the duration a patient survives following a cancer diagnosis, and DFS, indicating the period during which a patient remains cancer-free post-treatment, are essential endpoints in oncology research and practice[100].

Accurately predicting OS and DFS can significantly help clinicians in crafting tailored treatment regimens, tracking patient responses, and designing personalised care plans. OS, as a gold-standard outcome in clinical trials, often serves as a benchmark for regulatory drug approvals. When OS data are challenging to obtain due to time constraints, surrogate endpoints like DFS or, in a neoadjuvant context, pathological complete response (pCR), are frequently employed to approximate long-term survival outcomes.

Traditional prognostic factors like tumour size, lymph node involvement, hormone receptor status, and histological grade have long been instrumental in OS and DFS prediction models[101,102].

Our objective in this chapter is to develop a solution, or a range of solutions, to assist clinicians in selecting the most appropriate treatment strategies. We specifically focus on the neoadjuvant setting, targeting high-risk tumours, including luminal B (which accounts for 30% of all luminal tumours, representing 70% of breast cancer cases), as well as triple-negative and HER2+ breast cancers, each constituting approximately 15% of all breast cancer cases. Consequently, our solutions aim to cover approximately 50% of all breast cancer cases, addressing a substantial portion of the patient population.

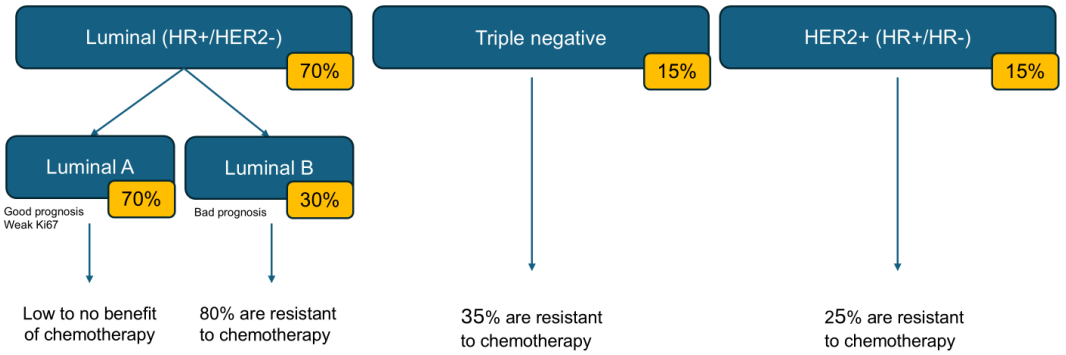


Figure 15. Molecular subtypes and therapeutic response patterns in early breast cancer. This schematic summarises the major intrinsic subtypes of early breast cancer: luminal (ER+/HER2−), triple-negative (TNBC), and HER2-positive (HER2+). Luminal tumours, representing approximately 70% of cases, are subdivided into Luminal A (good prognosis, limited benefit from chemotherapy) and Luminal B (poorer prognosis, ≈ 80% resistant to standard regimens). TNBC accounts for 15% of cases, with about 35% showing resistance to anthracycline–taxane chemotherapy, while HER2+ tumours (15%) exhibit resistance in roughly 25% of cases despite targeted therapy. These subtype-specific response profiles motivate the development of predictive models for chemosensitivity and residual-disease risk explored in subsequent chapters.

Each breast cancer subtype presents distinct therapeutic challenges. Recent advances in deep learning are revolutionizing predictive capabilities, enabling more refined and accurate survival prognostications by identifying intricate patterns within imaging data[103,104]. However, to develop the most effective solutions, it is essential to integrate these technological advancements within the broader clinical context and decision-making pipeline.

Treatment strategy for Luminal B Breast Cancer

For luminal B tumours, the standard treatment approach consists of neoadjuvant chemotherapy based on an anthracycline and/or taxane regimen, followed by surgery and adjuvant hormonotherapy (**Figure 16, Case 1**). Recently, CDK4/6 inhibitors have emerged as promising therapeutic agents for high-risk patients. In advanced ER-positive, HER2-negative breast cancer, these inhibitors have significantly improved outcomes, doubling progression-free survival (PFS) and yielding clinically meaningful overall survival (OS) benefits across both first-line and subsequent treatment settings[105].

However, their role in early breast cancer remains uncertain due to mixed results from clinical trials. The PALLAS and PENELOPE-B trials[26,106], which assessed the addition of palbociclib to standard adjuvant chemotherapy and endocrine therapy, failed to demonstrate an improvement in invasive disease-free survival (iDFS) compared to adjuvant endocrine therapy alone in patients with high-risk localized breast cancer[107,108]. Consequently, alternative strategies are under investigation, including neoadjuvant hormonotherapy combined with CDK4/6 inhibitors as a potential replacement for systemic chemotherapy, as explored in the NEOPAL and CORALLEEN trials.

Given that systemic chemotherapy is associated with medium- and long-term adverse effects, such as fatigue and cognitive impairment, reducing treatment toxicity could significantly enhance patients' quality of life. Another promising avenue under evaluation is the Neo-CheckRay study[109], which explores the integration of immunotherapy (anti-CD73 +/- anti-PD-L1) with standard systemic chemotherapy and stereotactic body radiation therapy (SBRT) to enhance the immunogenicity of luminal B breast cancer. Preliminary findings from this trial indicate an increase in pathological complete response (pCR) rates. Further analysis is required to determine the statistical significance of these results.

Thus, current strategies for managing luminal B breast cancer are evolving, incorporating novel targeted therapies and immunomodulatory approaches to optimize patient outcomes while minimizing treatment-related toxicity.

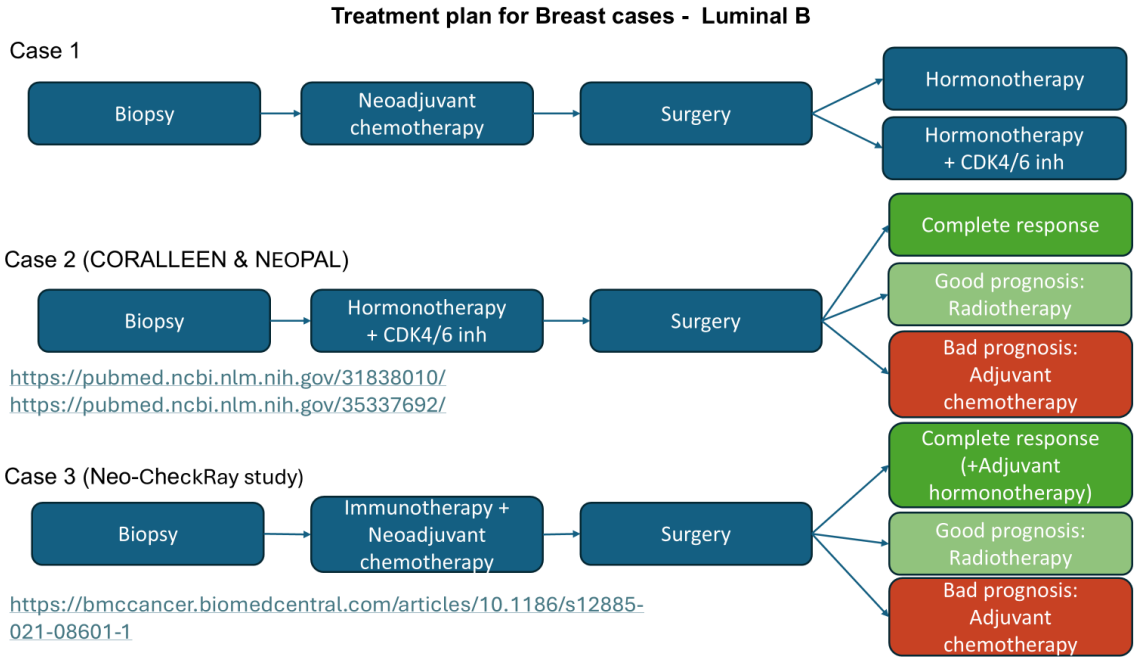


Figure 16. Therapeutic strategies and clinical outcomes in early Luminal B breast cancer. The diagram summarises current neoadjuvant and adjuvant treatment

strategies for Luminal B tumours, a biologically heterogeneous and moderately chemosensitive subtype. **Case 1** represents the standard approach using anthracycline–taxane neoadjuvant chemotherapy followed by surgery and adjuvant endocrine therapy, with or without CDK4/6 inhibition. **Case 2** (CORALLEEN & NEOPAL trials) illustrates an alternative chemo-free neoadjuvant strategy combining endocrine therapy with CDK4/6 inhibitors, achieving molecular responses comparable to chemotherapy in selected patients (Prat et al., 2020; Llombart-Cussac et al., 2022). **Case 3** (Neo-CheckRay study) explores an intensified regimen adding immunotherapy to neoadjuvant chemotherapy.

Each path leads to outcome-specific recommendations, complete response, good prognosis requiring only radiotherapy, or poor prognosis prompting adjuvant chemotherapy, highlighting the need for accurate baseline predictors of chemosensitivity to individualise treatment intensity.

In this chapter, we will stratify surgical specimens from patients who have undergone systemic chemotherapy to identify residual disease with either favorable or poor prognostic implications. The primary clinical application of this approach corresponds to Case 1 (**Figure 16**), which represents the treatment workflow in our cohorts. In this scenario, our tool would assist in distinguishing low-risk patients who may safely receive standard adjuvant hormonotherapy from those who could benefit from the addition of CDK4/6 inhibitors (**Figure 17.A**).

If our results demonstrate sufficient robustness, we aim to extend our analysis to a cohort of patients who received neoadjuvant hormonotherapy combined with CDK4/6 inhibitors or immunotherapy (Cases 2 and 3, **Figure 16**). In these settings, our algorithm could support clinical decision-making by guiding therapeutic escalation, helping physicians determine whether to introduce adjuvant chemotherapy or maintain a less aggressive treatment regimen (**Figure 17.B**).

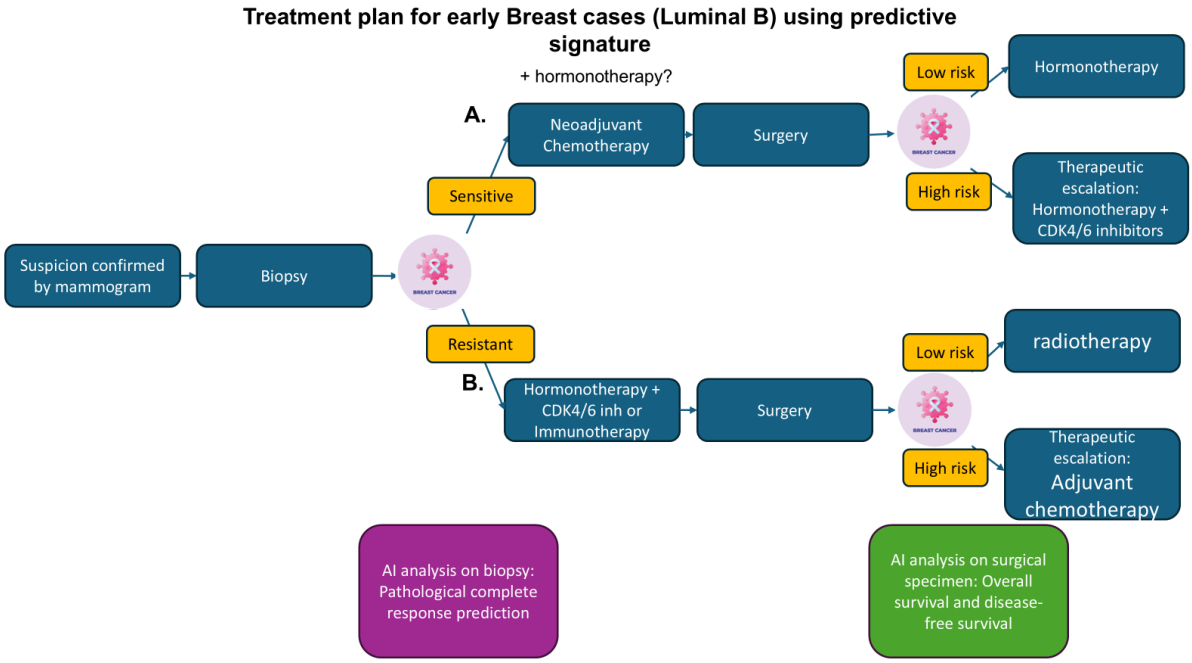


Figure 17. Integrative treatment pathway for early Luminal B breast cancer guided by AI-derived predictive signatures. The diagram illustrates how deep-learning–based histological analyses can inform two major decision points in the management of early Luminal B breast cancer. After biopsy confirmation, the first AI model (Chemo-prAIdict Breast) estimates the probability of pathological complete response (pCR) to neoadjuvant chemotherapy. Path A (Sensitive tumours): patients predicted as chemosensitive receive standard neoadjuvant chemotherapy, followed by surgery and adjuvant hormone therapy. Path B (Resistant tumours): patients predicted as chemoresistant may receive alternative regimens such as hormonotherapy combined with CDK4/6 inhibitors or immunotherapy prior to surgery.

Following surgical resection, a second AI model analyses the residual disease to predict long-term outcomes, including overall survival (OS) and disease-free survival (DFS).

This dual-stage framework enables patient stratification into high- and low-risk prognostic groups, guiding therapeutic escalation (adjuvant chemotherapy or CDK4/6 inhibitors) or de-escalation (hormone or radiotherapy alone). Together, these AI-driven analyses provide a biologically informed approach to personalise neoadjuvant and adjuvant treatment intensity using only routine histopathology slides.

Treatment strategy for Triple-Negative Breast Cancer (TNBC)

The overall treatment rationale for **TNBC** follows a similar framework to that of luminal B tumours, though the therapeutic regimen differs. According to the ESMO 2024 guidelines[52], neoadjuvant chemotherapy (NAC) is the standard of care for T1c/N0 or greater TNBC.

The baseline systemic chemotherapy regimen consists of anthracycline and taxane, with or without carboplatin. The inclusion of carboplatin has been shown to improve pathological complete response (pCR) rates and event-free survival (EFS); however, its impact on overall survival (OS) remains uncertain. As a result, not all patients receive carboplatin, and its benefit-risk ratio requires careful evaluation. For high-risk TNBC patients, pembrolizumab (immunotherapy) is added to the neoadjuvant regimen. This combination has demonstrated significant efficacy in improving pCR and long-term outcomes in patients with PD-L1-positive disease. Patients with germline BRCA1/2 mutations (gBRCA1/2m) generally exhibit strong responses to standard anthracycline-taxane-based chemotherapy, regardless of platinum use. However, PARP inhibitors are currently recommended only for patients with residual disease and a germline BRCA1/2 mutation, as supported by findings from the BRIGHTNESS study[110].

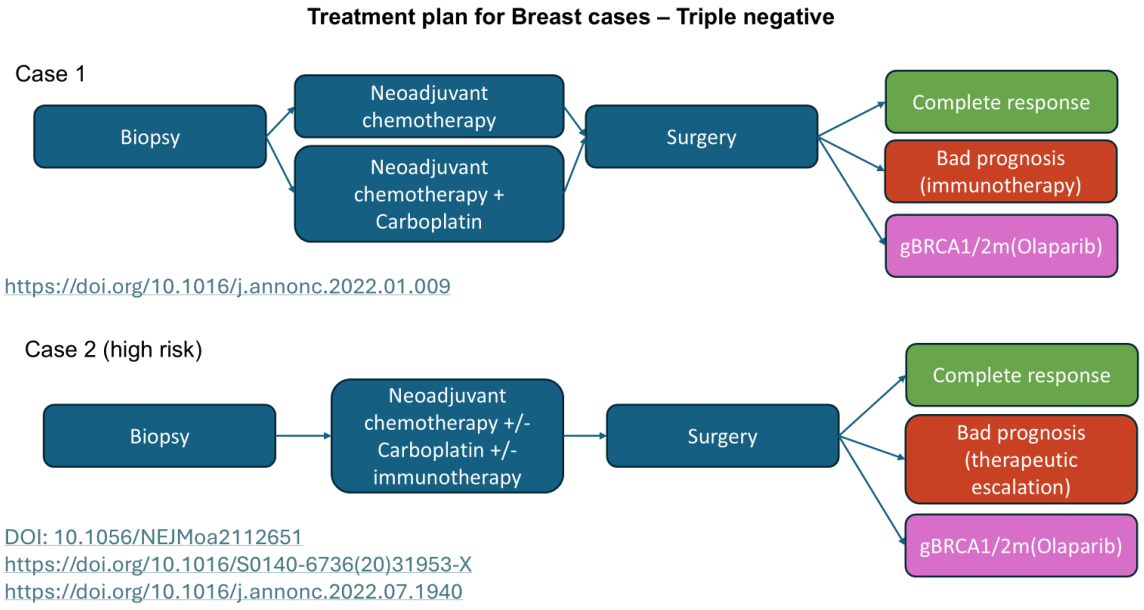


Figure 18. Evolving treatment algorithms for triple-negative breast cancer (TNBC). This diagram summarises two representative management strategies reflecting the shift toward biomarker-guided and risk-adapted therapy in TNBC.

Case 1 illustrates standard practice, where patients receive conventional anthracycline–taxane neoadjuvant chemotherapy or an intensified regimen including carboplatin, followed by surgery. Postoperative treatment is guided by the pathological response: patients achieving complete response (pCR) enter surveillance, while those with residual disease may receive adjuvant immunotherapy or the PARP inhibitor olaparib if carrying a germline BRCA1/2 mutation (gBRCA1/2m). **Case 2** represents a high-risk TNBC scenario, where baseline tumour or molecular features prompt an upfront intensified regimen combining carboplatin and/or immunotherapy before surgery. After resection, patients with persistent disease are candidates for further escalation (e.g., adjuvant olaparib or immune checkpoint blockade).
This figure highlights the increasing role of predictive biomarkers (e.g., BRCA1/2 status) and dynamic treatment adaptation based on tumour response and molecular context.

The two strategies outlined above (**Figure 18**) differ primarily in the timing of immunotherapy administration, either before or after surgery.

For TNBC patients, early identification of high-risk, chemoresistant tumours based on biopsy analysis is crucial. This enables the prompt integration of immunotherapy. Additionally, post-surgical stratification of residual disease is essential to determine whether therapeutic escalation is warranted (**Figure 19**), ensuring that patients with persistent high-risk disease receive optimized adjuvant treatment.

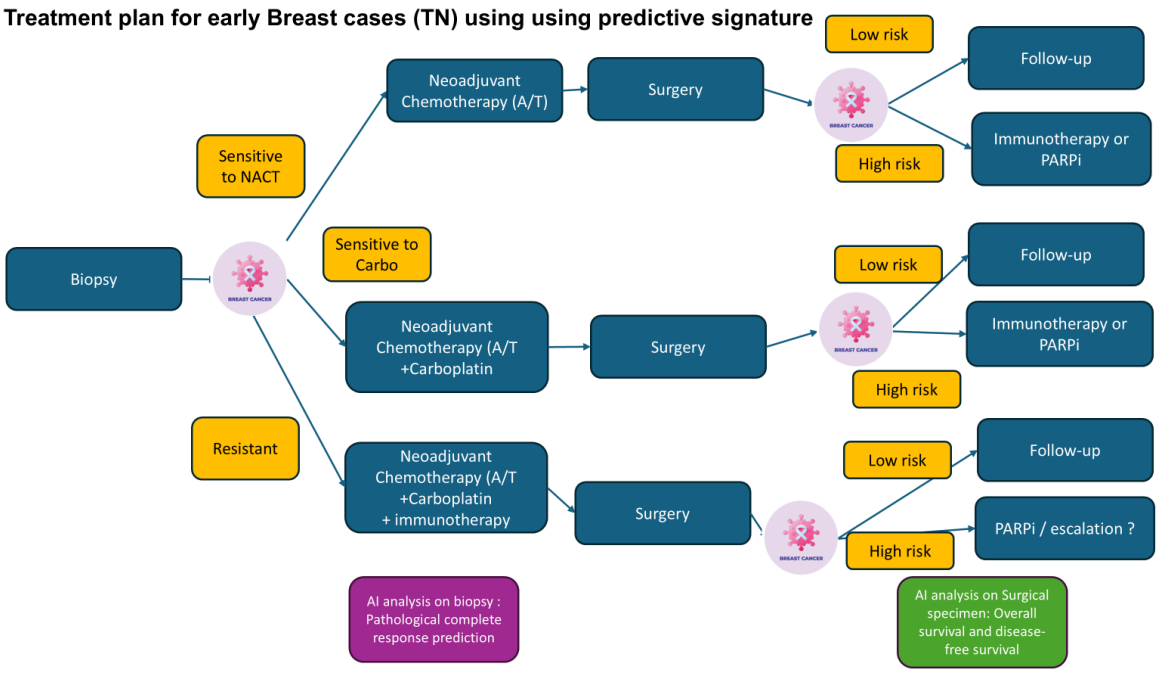


Figure 19. AI-assisted treatment pathway for early-stage triple-negative breast cancer (TNBC). This diagram illustrates a dual-stage decision framework guided by deep-learning–derived predictive signatures from histological slides. After biopsy confirmation, the first AI model (Chemo-prAIdict Breast) predicts the tumour's sensitivity to neoadjuvant chemotherapy (NACT). Sensitive to anthracycline/taxane (A/T): patients receive standard NACT followed by surgery. Sensitive to carboplatin: patients receive NACT augmented with carboplatin. Resistant to both: patients receive an intensified regimen combining A/T, carboplatin, and immunotherapy.

Following surgery, a second deep learning model analyses the residual tumour to estimate the risk of recurrence (overall survival and disease-free survival). Based on this post-NAC assessment, patients are stratified into low-risk (eligible for standard follow-up) and high-risk (eligible for adjuvant escalation such as immunotherapy, PARP inhibition, or clinical trial enrolment).

Together, these two AI models enable a fully data-driven workflow for TNBC, integrating baseline chemosensitivity prediction with post-treatment prognostication to personalise therapeutic intensity and improve long-term outcomes.

Treatment strategy for HER2 amplified Breast Cancer (HER2+)

For HER2-amplified tumours, we follow the same two-step approach as in other subtypes. For stage II–III HER2-positive disease, which typically benefits from a neoadjuvant protocol, the standard treatment regimen consists of an anthracycline-taxane-based combination with

HER2-targeted therapy[111] (trastuzumab-pertuzumab, HP). However, anthracycline-free regimens, incorporating carboplatin and taxanes, have been evaluated in several phase II (PREDIX HER2, TRAIN-2, TRYPHAENA) and phase III (BCIRG006) clinical trials[111–114]. These studies have reported comparable efficacy to anthracycline-containing regimens while demonstrating improved cardiac safety.

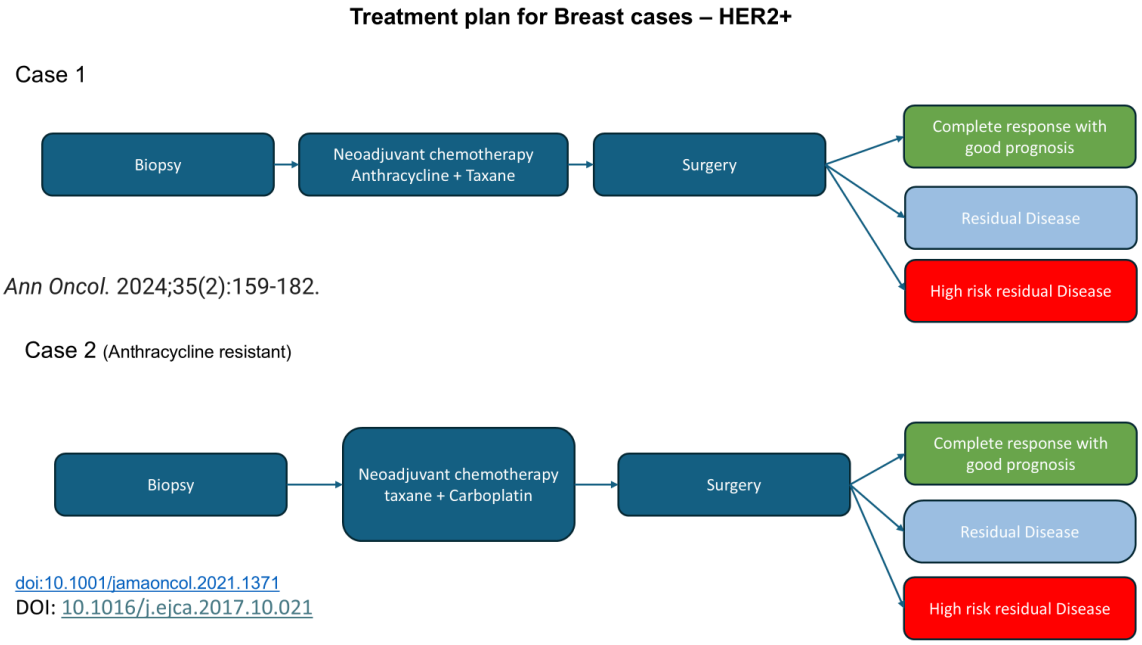


Figure 20. Treatment algorithms for HER2-positive (HER2$^+$) early breast cancer. This diagram summarises the main neoadjuvant strategies for HER2$^+$ breast cancer, structured according to anthracycline sensitivity. **Case 1** illustrates the standard approach, where patients receive anthracycline–taxane–based neoadjuvant chemotherapy followed by surgery. Post-surgical evaluation determines prognosis based on residual disease burden. Complete response (pCR) is associated with excellent outcomes and continuation of standard anti-HER2 therapy. Residual disease prompts consideration of adjuvant escalation, depending on the extent of persistence. High-risk residual disease indicates poor prognosis and eligibility for intensified adjuvant therapy, such as trastuzumab emtansine (T-DM1). **Case 2** represents anthracycline-resistant or contraindicated cases, where the neoadjuvant backbone is adapted to a taxane plus carboplatin combination before surgery. Postoperative management follows the same risk-stratified scheme based on residual disease status.

As with other tumour types treated in a neoadjuvant setting, the analysis of residual disease in the postoperative specimen is crucial for guiding further treatment decisions. Patients achieving pathological complete response (pCR) follow standard recommendations, which

include one year of trastuzumab-based therapy. However, even in cases of pCR, patients with a high initial tumour burden remain at an elevated risk of relapse. The presence of residual invasive disease in the breast or lymph nodes is associated with poorer outcomes. For these patients, clinicians prefer a trastuzumab-pertuzumab (HP) regimen to enhance therapeutic efficacy.

Current & Emerging Strategies for Residual Disease are the following :

- Standard Approach: Residual disease is currently managed based on established classifications, with antibody-drug conjugate (ADC) therapy, specifically T-DM1, serving as the standard of care.
- De-escalation Considerations: Some patients may benefit from HP ± endocrine therapy (ET) monitoring as a potential de-escalation strategy.
- Escalation for High-Risk tumours: For high-risk residual disease, treatment intensification strategies are being explored, including T-DM1 in combination with ET or radiotherapy (RT).

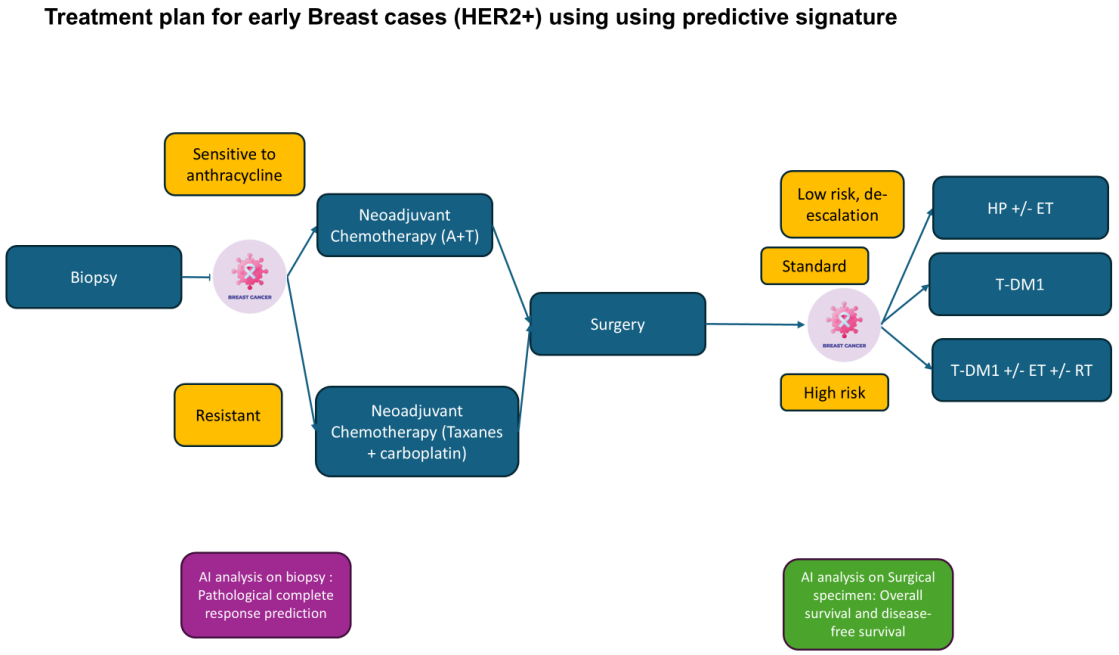


Figure 21. AI-guided treatment strategy for early-stage HER2-positive (HER2⁺) breast cancer. This figure outlines a proposed workflow integrating AI-derived predictive signatures at both the pre- and post-treatment stages. The process begins with a biopsy analysed by an AI model trained to predict anthracycline sensitivity. Patients predicted to be anthracycline-sensitive receive the standard anthracycline–taxane–based neoadjuvant

chemotherapy (A+T), while those predicted to be resistant are directed to an alternative regimen combining taxane and carboplatin, reflecting current strategies for anthracycline-intolerant or resistant disease.

After surgery, a second AI model analyses the residual tumour to estimate overall survival (OS) and disease-free survival (DFS), enabling post-neoadjuvant risk stratification. Low-risk patients may undergo treatment de-escalation, typically continuing trastuzumab (H) or trastuzumab + pertuzumab (HP), with or without endocrine therapy (ET), depending on hormone receptor status. Standard-risk patients follow conventional T-DM1-based adjuvant therapy, in line with the KATHERINE trial (von Minckwitz et al., NEJM 2019). High-risk patients may receive escalated adjuvant regimens combining T-DM1 with endocrine therapy and/or radiotherapy.

This approach reflects the international standard of care, where dual HER2 blockade (HP) is routinely used in the neoadjuvant setting, although in France, pertuzumab remains unreimbursed for neoadjuvant use. T-DM1, conversely, is specifically indicated for post-NAC escalation in patients with residual invasive disease, representing the cornerstone of adjuvant intensification in HER2-positive early breast cancer.

Conclusion

Each strategy has its own benefits and risks. It is therefore critical to identify patients first on the biopsy that will respond to each regimen, and stratify the surgical specimen afterwards to distinguish good from bad prognosis residual disease and adapt the treatment protocol with the best strategy.

To do so, we leverage Whole Slide Imaging (WSI) of post-Neo Adjuvant Chemotherapy (post-NAC) surgical specimens from early Breast Cancer (eBC) patients, utilising the PRIMUNEO and Neoadj datasets. Our objective is to predict OS and DFS as distinct tasks, examining a spectrum of methodological approaches:

- **Baseline Method**: Employing a pretrained EfficientNet model alongside a Cox Proportional Hazards (CoxPH) model for initial OS and DFS estimation.
- **Advanced Survival Analysis Method**s: Exploring machine learning-based survival models that directly predict OS and DFS by incorporating survival-specific frameworks.

- **Deep Learning Survival Approaches**: Applying survival neural networks optimised for high-dimensional imaging data, a cutting-edge method that may enhance predictive accuracy.
- **Clinical Integration**: Integrating patient-specific clinical data to further refine and augment survival predictions, acknowledging the critical role of clinical variables in survival outcomes.
- **Unsupervised Pre-Training**: Leveraging unsupervised pre-training techniques for feature extraction, improving the depth of the model's learned representations from the imaging data.
- **End-to-End Prediction Pipeline**: Developing a comprehensive pipeline to streamline the prediction of OS and DFS, aiming for seamless integration into clinical workflows.
- **Combined Approaches:** Identifying and synthesising the most effective techniques across experiments to deliver the highest predictive accuracy.

These methodologies will undergo testing on the internal PRIMUNEO validation cohort, with the highest-performing model validated on the independent NEOADJ dataset. This work represents a novel approach, aiming to utilise deep learning-derived insights from post-NAC histopathological specimens as an alternative source of prognostic information beyond conventional clinical data. To our knowledge, this project constitutes the first initiative to apply deep learning on such samples for survival prediction in breast cancer patients.

3.2 Material and Methods

In this section, we will use only the post neoadjuvant systemic chemotherapy (post-NAC) surgical specimens and their whole slide images. They come from 2 different datasets: the PRIMUNEO dataset, the CGFL breast cancer neoadjuvant dataset, and the TCGA. The cohorts and experimental setups are described in **Chapter 2 Neoadjuvant early Breast Cancer Datasets sections**. Specific adaptations are covered in the dedicated sections throughout this chapter.

3.3 Results

3.3.1 Baseline model for OS and DFS prediction using post-NAC surgical specimen

Overall Survival (OS) refers to the length of time a patient survives after their cancer diagnosis, while DFS (Disease Free Survival) represents the duration during which a patient remains cancer-free after completing treatment. Understanding and addressing the factors that impact OS and DFS in breast cancer patients is crucial for optimising treatment strategies and improving long-term outcomes, as it is the primary endpoint for deciding which treatment is the better option in clinical trials

In this section, we use Whole Slide Images (WSIs) of the post neoadjuvant systemic chemotherapy (post-NAC) surgical specimen in patients from the PRIMUNEO and Neoadj datasets (**see Chapter 2**) to predict the patient's OS and DFS.

This approach intends to use state-of-the-art deep learning methods to extract information from the operative piece from the patient, as an alternative source of features rather than the clinical information. Additionally, this project will explore for the first time the OS and DFS prediction in the PRIMUNEO and CGFL Breast Cancer Neoadjuvant databases.

Questions and objectives

The questions and objectives we would like to answer in this section are the following:

- How do we split the data for training and testing in both datasets (PRIMUNEO and CGFL Breast Cancer Neoadjuvant database) ?
- How do we stratify the data ?
- Is the molecular subtype relevant for this type? Do other studies stratify their patients by a specific molecular subtype? Do we train a model for each subtype?
- Do we part from the raw WSI?
- What models are suitable for extracting the features from the WSI without spending days in an end-to-end training?
- What hyperparameter values are the most adequate for having a high performance?
- To perform the DFS and OS prediction, what type of methods do we use for these tasks in survival analysis?

- What Clinical Information can be relevant for boosting the performance of the method?

List of propositions (based on the literature review)

Data splitting: We have two datasets (PRIMUNEO and CGFL Breast Cancer Neoadjuvant database). PRIMUNEO dataset is multicentric and we decided to use it as a training set, as it has a higher range of domains (colors, preparation protocols, patients, centers etc.) giving us a higher chance to better generalise. The CGFL Breast Cancer Neoadjuvant database, despite being monocentric, is bigger and can act as a good external validation setting as it will not be seen in the training phase by the deep learning algorithms. This setup, described in **Chapter 2**, allows a robust demonstration of the performances of the algorithm without any bias.

Data stratification: In order to ensure that the site of origin is not biasing the prediction performance, as previously shown by Howard et al.[115], we use a stratified grouped cross-validation approach (see **Chapter 2**).

Molecular subtype stratification: The molecular subtype is critical in our clinical setting, as the treatment strategies and underlying biology are very different from one another. Luminal, TNBC and HER2 amplified tumours should be treated separately from a clinical endpoint. However, as the more the merrier in AI, we could train a single deep learning solution with all the data points and stratify the results for each molecular subtype at the end.

Baseline pipeline: As shown by Laleh et al. (2022)[81], a basic EfficientNet B7 pretrained on Imagenet often yields the best results for a weakly supervised learning framework applied to histopathology images.

Advanced feature extractors: EfficientNet B7 pretrained on histology task, end-to-end prediction or unsupervised pretraining[116,117]. We shall compare them and find the best solution possible.

Training Strategies: As we need to predict survival, our training pipeline has two critical steps. The first one is to train the feature extractor and the second one to predict the survival or event. We can do it either by a direct approach with a deep learning fine-tuning task or

end-to-end training , or do first the feature extraction and then use traditional survival analysis methods (Cox Proportional Hazards) or existing DL methods (DeepHit, DeepSurvNet).

Clinical information: To ensure that our model is better than the current standard of care, we will compare our WSI model to a clinical model using critical information used in day to day practice by pathologists and clinicians.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

As described in the **List of proposition** section, we first implement a baseline model. A single slide was used for each patient for performance evaluation. We then extracted the foreground using an in-house trained U-net and tiled the images in non-overlapping patches of 600x600 pixels at a 5x resolution. The resulting patches were used as inputs of an EfficientNetB7 feature extractor pretrained on ImageNet, adding a global average pooling layer to produce an embedding vector of size 2560 (**Figure 22**).

The extracted embeddings are then fed into a top classifier, a two-layer Multi Layer Perceptron (MLP) of size 512 and 64. The goal of this task is to reduce the size of the embedding to feed it into an adequate survival method. To do so, we use a pretext task linked to our objective, by training the MLP on the binary task of survival (0)/death (1) or not relapse (0) or relapse (1). Next, the patch-features are averaged to have a slide-level feature survival vector. This feature survival vector is fed into a Cox Proportional Hazard method (**Figure 22.C**).

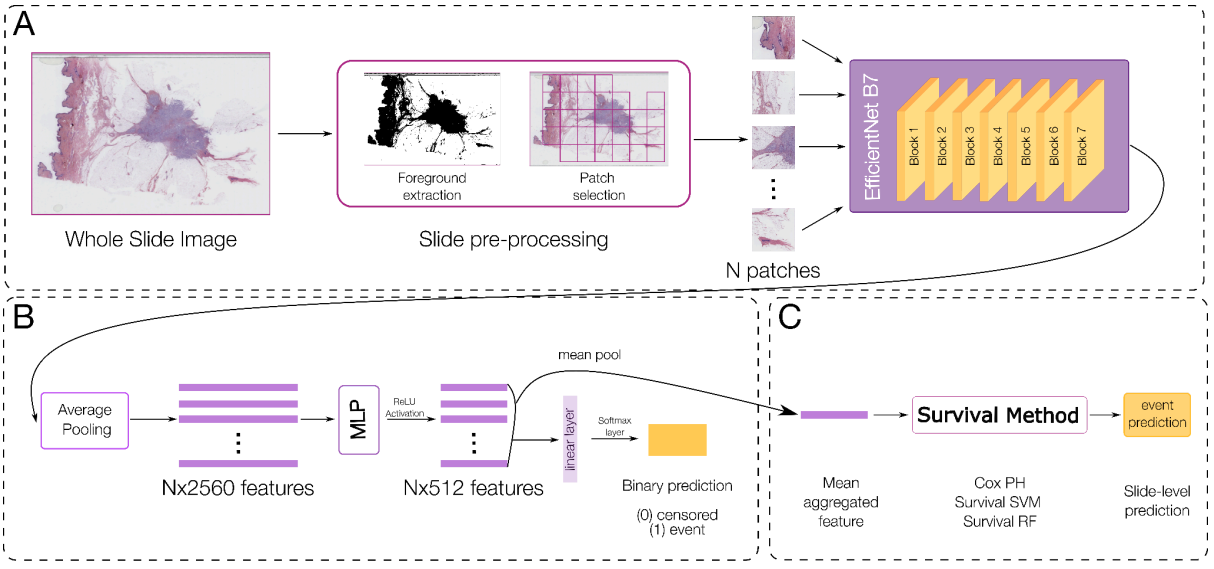


Figure 22. Overview of the baseline pipeline for Overall Survival (OS) and Disease-Free Survival (DFS) prediction from whole-slide images. (A) Whole-slide images (WSIs) are pre-processed through foreground extraction to remove background and artefacts, followed by patch selection to divide each slide into N non-overlapping image tiles. Each patch is then encoded using a pre-trained EfficientNet-B7 backbone, producing high-dimensional feature embeddings. (B) The extracted patch-level embeddings (N × 2560) are passed through a multi-layer perceptron (MLP) for dimensionality reduction (to 512 features) with non-linear activations (ReLU) and dropout regularisation. The resulting patch features are aggregated by mean pooling to obtain a slide-level representation summarising global histological patterns.
(C) This aggregated feature vector is then fed into survival modelling frameworks, including Cox proportional hazards (Cox-PH), survival support vector machines (Survival-SVM), and random survival forests (Survival-RF), to predict survival outcomes.
The models were trained using censoring-aware loss functions and evaluated for both OS and DFS tasks. This modular baseline establishes a transparent, interpretable pipeline linking histological morphology to time-to-event outcomes.

The model was trained during 5 epochs, with a batch size of 32, adam optimizer with a learning rate of 1e-4.

RESULTS

Our first approach consists of doing a mean aggregation of the patch features just after the MLP to generate the slide-level survival vector. Then, these WSI-level vectors are used for training the Cox-PH survival model. This first value achieves 58.4% AUC on average (mean of the 4 internal partitions, **Figure 23**). Hence, we only have one survival vector per patient in this method, used as an input feature for the CoxPH (setting Patches feature mean aggregation, **Figure 23**).

Disease Free Survival (DFS) prediction

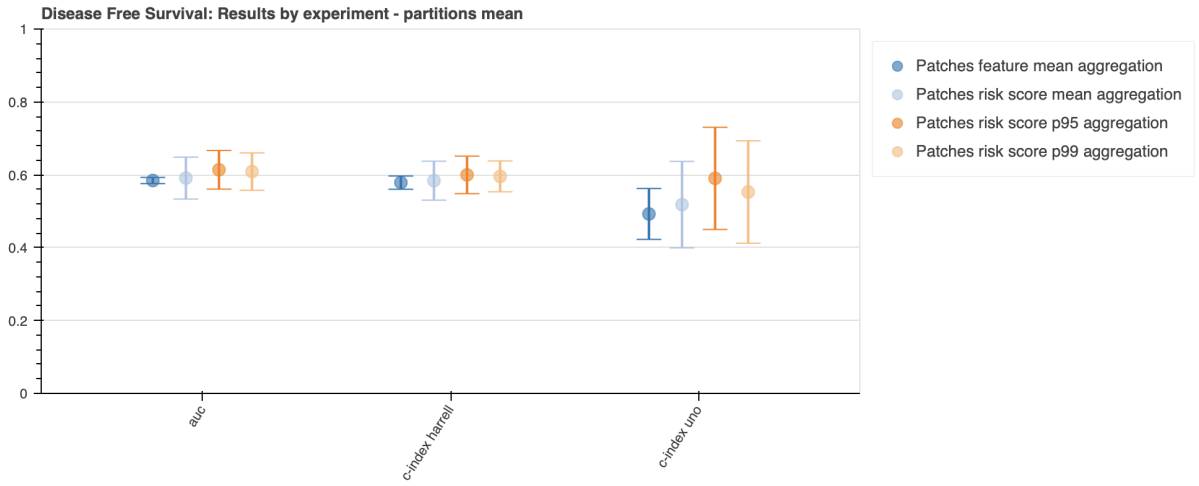


Figure 23. Comparison of patch-level aggregation strategies for Disease-Free Survival (DFS) prediction. This figure compares different strategies for aggregating patch-level representations into whole-slide predictions. The baseline model uses mean aggregation of patch features after the MLP layer (dark blue). Alternative methods aggregate risk scores computed at the patch level using the mean, 95th percentile (p95), and 99th percentile (p99) to capture highly predictive regions. Performance metrics are averaged over four internal validation partitions. The p95 risk-score aggregation achieved the best overall performance (AUC = 61.4%, Harrell's C = 60.0%, Uno's C = 59.0%), suggesting that weighting high-risk regions better preserves discriminative information than averaging feature embeddings.

In the first experiment, (patches risk score mean, **Figure 23**), we i**nput all the patches embedding in the Cox-PH** model, and we perform the aggregation in the patches risk scores to get the WSI-level score. The rationale behind this experiment is that doing the mean of the embedding vector will destroy the information generated contained in each individual patch embedding. Therefore, if we want to keep the information intact, we should convert these embedding into a risk score before doing the aggregation.

We show the results for the mean (setting Patches risk score mean aggregation), the percentile 95 (61.4% AUC, 60% C-index Harrell, 59% C-index Uno, setting Patches risk score p95 aggregation) and the percentile 99 (60.9% AUC, 59.6% C-index Harrell, 55.3% C-index Uno, setting Patches risk score p99 aggregation) aggregations, p95 being the best performant in the 3 metrics.

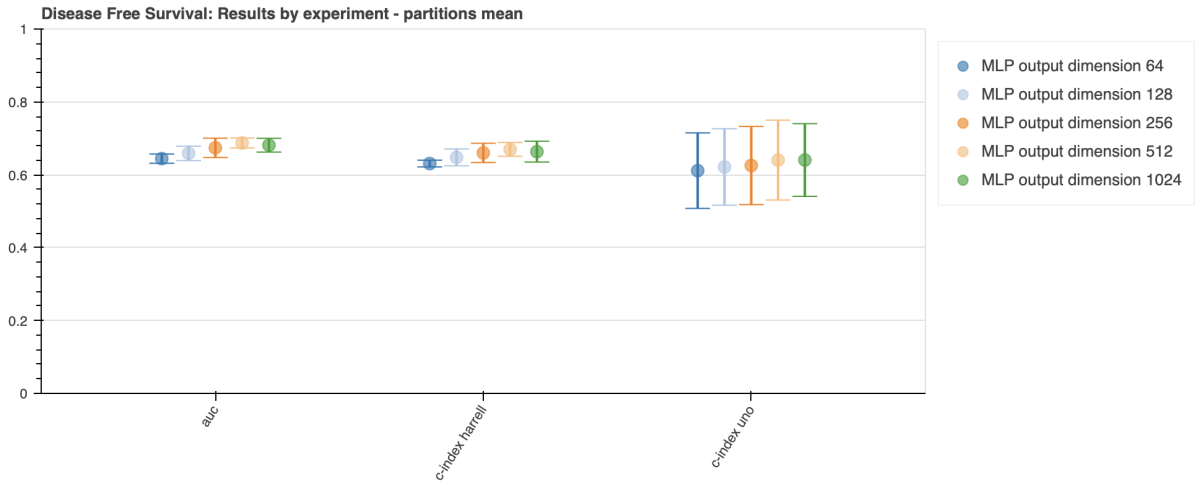


Figure 24. Effect of MLP output dimensionality on DFS prediction performance. This experiment evaluates how varying the output dimension of the MLP (64, 128, 256, 512, 1024) affects model accuracy and stability. Increasing dimensionality improves AUC and C-index up to 512 features, beyond which performance plateaus while computation time increases substantially. The configuration with 512 output features achieved the best balance between accuracy (AUC = 68.7%) and training efficiency, outperforming lower-dimensional settings while avoiding the computational cost of 1024-dimensional embeddings.

To determine the optimal hyperparameters, we first varied the output dimension of the MLP (**Figure 24**). Increasing the output dimension led to slight performance improvements; however, it also significantly increased processing time. Notably, while a dimension of 512 yielded better results compared to lower values, further increasing it to 1024 resulted in only a marginal improvement of 0.2% in AUC, while dramatically increasing training time to approximately one day. Given this trade-off, we selected an output dimension of 512, achieving an AUC of 68.7%, to balance performance and computational efficiency. Next, we experimented with batch sizes of 32, 64, and 128. A batch size of 64 produced the best results, reaching an AUC of 69.1%, along with C-index scores of 67.1% (Harrell) and 64.1% (Uno) (**Figure 25**). This configuration outperformed batch sizes of 32 and 128, yielding 0.4% and 0.1% higher AUC, respectively.

Additionally, we conducted further experiments by varying training epochs and learning rates, but no significant changes in performance were observed.

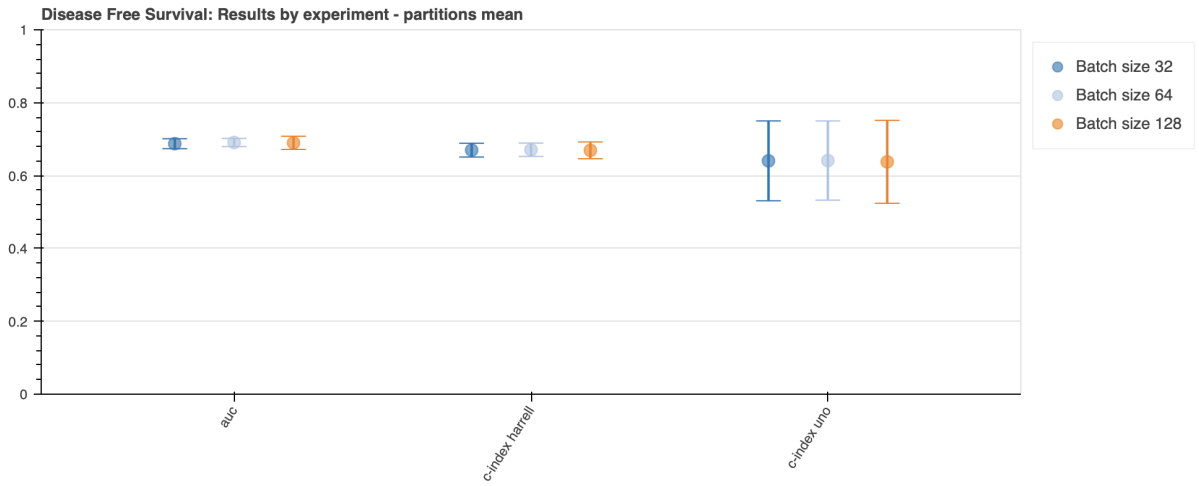


Figure 25. Impact of batch size on DFS task performance. Batch sizes of 32, 64, and 128 were compared to assess their influence on training stability and discrimination metrics. The model trained with a batch size of 64 achieved the highest mean AUC (69.1%) and the most consistent C-index values (Harrell = 67.1%, Uno = 64.1%), indicating a favourable trade-off between convergence speed and generalisation. Very small or large batches slightly reduced performance, possibly due to noisier gradients or poorer batch normalisation dynamics.

Overall Survival (OS) prediction

The same approach was used for the overall survival (OS) prediction pipeline. The aggregation on the patch embeddings achieved 59.9% AUC on average (mean of the 4 internal partitions, **Figure 26**), whereas the aggregation on the risk scores were the best for the percentile 99 aggregation. This method achieves 73.5% AUC, 71.2% in the C-index (Harrell) metric, and 63.0% in the C-index (Uno) metric (**Figure 27**).

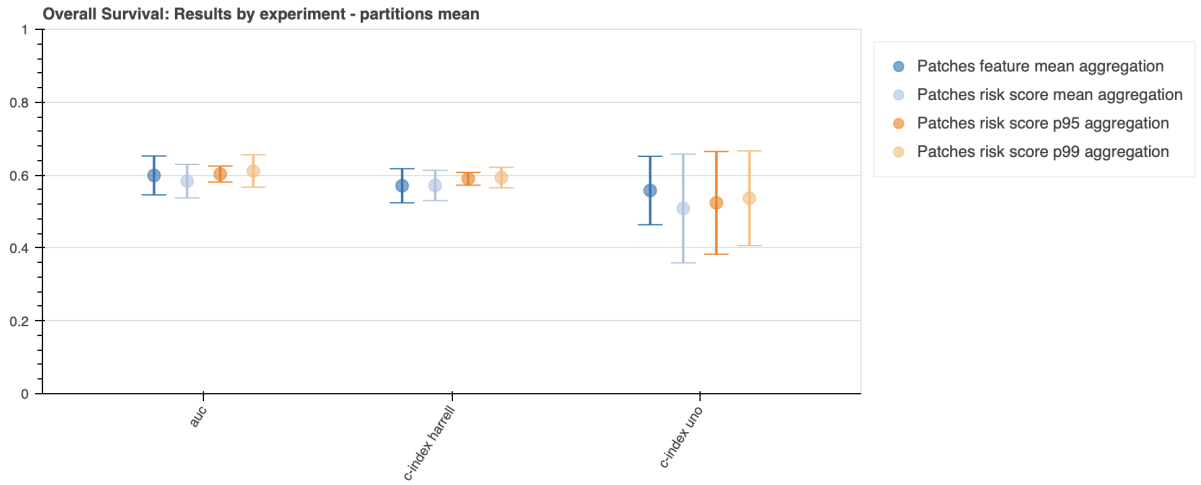


Figure 26. Comparison of patch-level aggregation strategies for Overall Survival (OS) prediction. This figure compares alternative approaches for aggregating patch-level representations into whole-slide survival predictions. The baseline model aggregates patch embeddings directly after the MLP layer (blue). In contrast, alternative methods compute patch-level risk scores and aggregate them by mean, 95th percentile (p95), or 99th percentile (p99) before slide-level prediction. Performance, averaged across the four internal validation partitions, shows that risk-score aggregation outperforms feature aggregation, with the p99 percentile achieving the best discrimination (AUC = 73.5%, Harrell's C = 71.2%, Uno's C = 63.0%), suggesting that extreme high-risk regions carry the strongest prognostic information for OS.

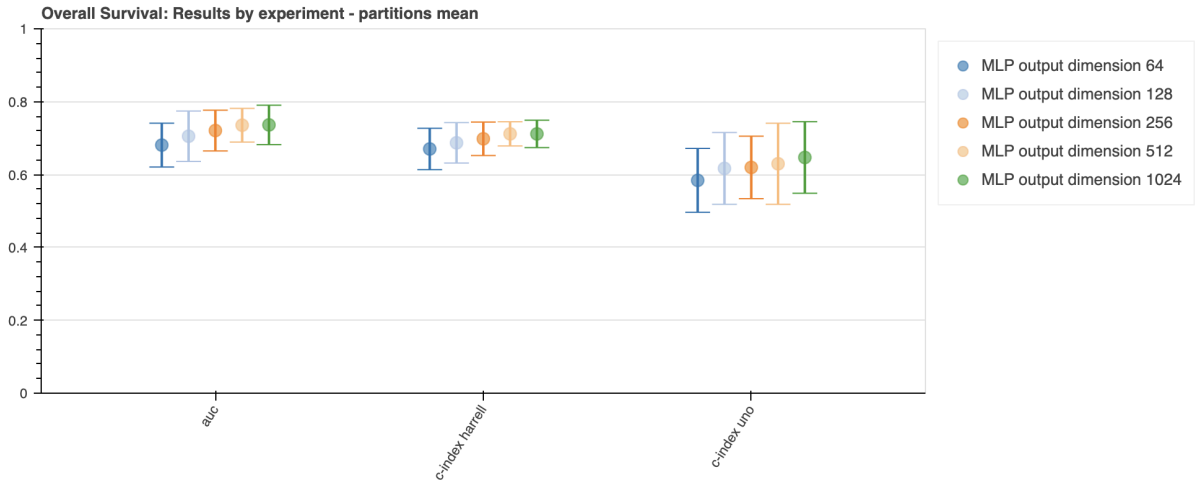


Figure 27. Effect of MLP output dimensionality on OS prediction performance. This experiment explores how varying the dimensionality of the MLP output layer (64, 128, 256, 512, 1024) impacts model accuracy and computational cost. Increasing the feature space improves overall performance up to 512 dimensions, beyond which results plateau while training time increases substantially. The 512-dimensional configuration yielded the best

trade-off between predictive power and computational efficiency, maintaining high consistency across internal validation partitions.

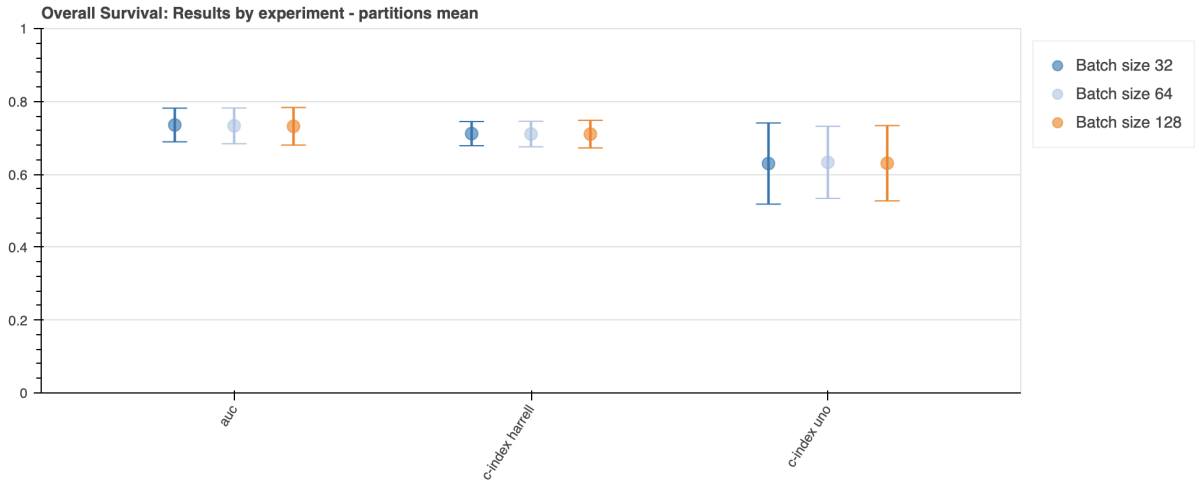


Figure 28. Impact of batch size on OS prediction performance. Performance in the OS task using different batch sizes. Batch sizes of 32, 64, and 128 were evaluated to assess their influence on convergence stability and discriminative power. In contrast with the DFS task, where a batch size of 64 was optimal, the OS prediction task showed very little improved results with a batch size of 32, suggesting better gradient stability and regularisation in this configuration. Larger batches did not improve AUC or C-index and occasionally degraded performance due to reduced gradient variance.

We keep the 32 batch size as this method shows very little performance improvements compared to a batch size of 64 (**Figure 28**).

3.3.2 Using Machine Learning survival Methods for predicting OS and DFS using post-NAC surgical specimen

In the previous approach, we used the Cox Proportional Hazards (Cox-PH) model to estimate the risk of an event occurring. Here we extend our exploration of survival analysis methods by incorporating traditional machine learning techniques, such as Support Vector Machines (SVM) and Random Forests (RF). The primary objective is to compare the performance of ML-based survival models with the traditional Cox-PH method, assessing their potential advantages and limitations in predicting survival outcomes.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

For this comparison, we use the same architecture as before, only the survival method differs, as described in **Figure 22, Chapter 3 Material and Methods**. Given that machine learning-based survival models require significantly more time to converge, we reduced the input feature dimensions to 64 as it was necessary to improve computational efficiency.

Neural network training, model selection

For the Random Survival Forest, we used the default parameters from the sksurv library, configuring it with 100 estimators and a random seed for reproducibility.
For the SVM, we implemented a Fast Survival SVM (linear SVM), training it for 1,000 iterations with a termination tolerance of 1e-5.

RESULTS

Disease Free Survival (DFS) prediction

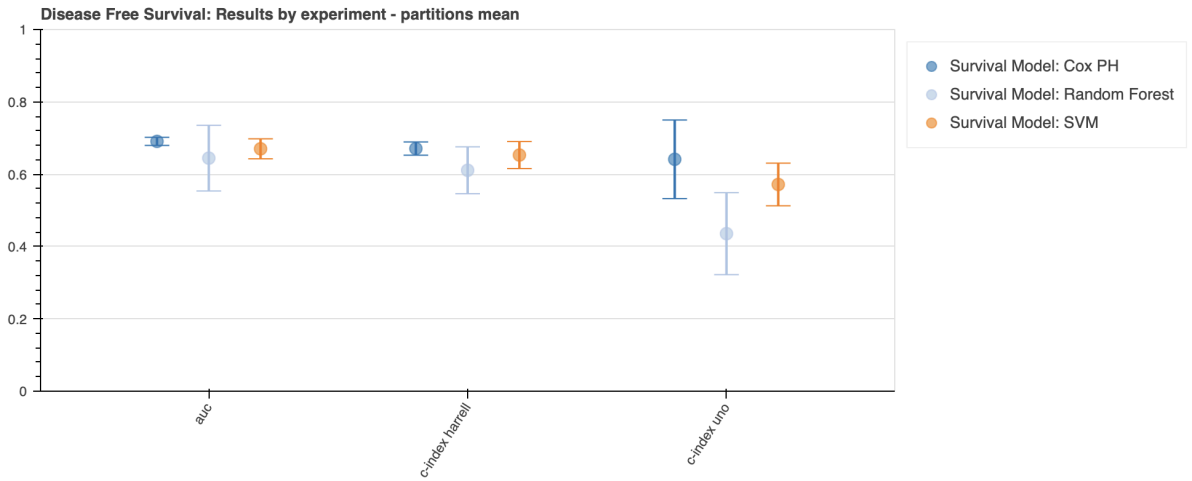


Figure 29. Comparison of machine learning survival modelling methods for Disease-Free Survival (DFS) prediction. This figure compares three survival modelling approaches applied to image-derived features: the traditional Cox proportional hazards (Cox-PH) model, Random Survival Forests (RF), and Survival Support Vector Machines (SVM). Performance is averaged across four internal validation partitions using AUC, Harrell's C-index, and Uno's C-index.

Figure 29 illustrates that the Cox-PH model consistently outperforms the ML-based approaches across all three evaluation metrics. In the AUC metric, Cox-PH achieves 69.1%, followed by SVM with 67.0% and RF with 64.4%. Similar trends are observed in the C-index (Harrell) metric. However, in the C-index (Uno), performance differences become more pronounced, with Cox-PH achieving 64.1%, while SVM and RF lag behind at 57.2% and 43.6%, respectively.

Overall Survival (OS) prediction

In the overall survival (OS) task, the Survival SVM model outperforms both the Cox-PH and Random Forest (RF) models, as illustrated in **Figure 30**. While the AUC difference between SVM and Cox-PH is relatively small (75.3% vs. 73.5%, a 1.8-point difference), the gap is more pronounced in the C-index (Uno) metric, where SVM achieves 71.3%, outperforming Cox-PH (62.9%) by 8.4 percentage points.

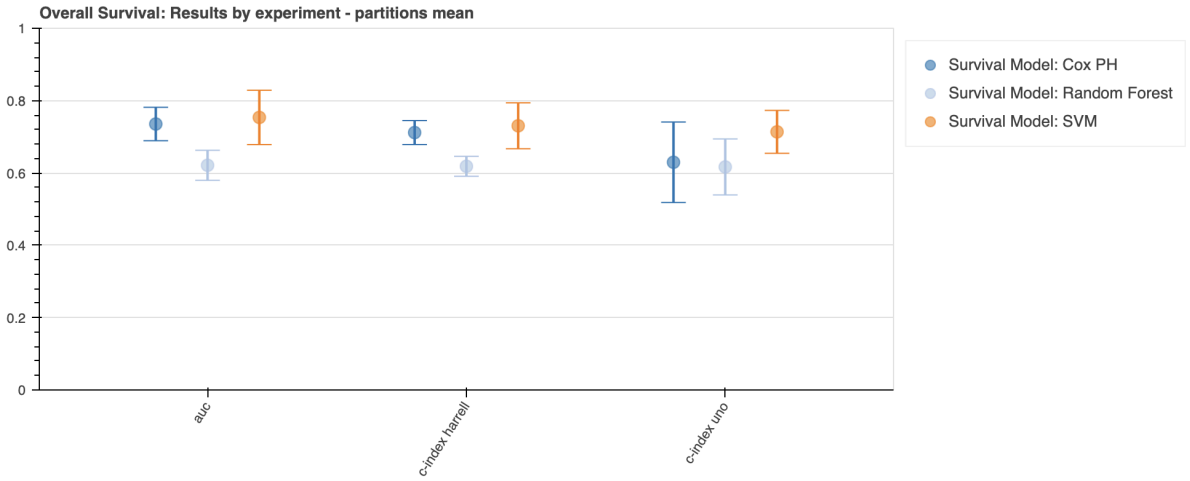


Figure 30. Comparison of machine learning survival modelling methods for Overall Survival (OS) prediction. This figure compares three survival modelling approaches applied to image-derived features: the traditional Cox proportional hazards (Cox-PH) model, Random Survival Forests (RF), and Survival Support Vector Machines (SVM). Performance is averaged across four internal validation partitions using AUC, Harrell's C-index, and Uno's C-index.

3.3.3 Using Deep Learning survival Methods for predicting OS and DFS using post-NAC surgical specimen

In this approach, we employ deep learning models specifically designed for survival prediction, namely DeepSurvNet and DeepHit. The primary objective is to evaluate their predictive performance and compare their effectiveness against the traditional Cox Proportional Hazards (Cox-PH) model.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

To train these deep learning survival models, we explored two approaches utilizing EfficientNet-B7 embeddings pretrained on ImageNet:

1. **Dimensionality Reduction Approach**: We reduced the feature dimensions of the embeddings, following the same procedure as in **Part 3.3.1**, **Figure 31** here.
2. **Direct Embedding Approach**: We used the raw EfficientNet-B7 embeddings as direct inputs for the deep learning survival models, as depicted in **Figure 32.**

To generate OS and DFS predictions, we first obtained risk scores for all WSI patches and applied a p99 aggregation strategy. The MLP model for dimensionality reduction was trained for five epochs using the Adam optimizer with a learning rate of 1e-4. The batch size was set to 32 for OS prediction and 64 for DFS prediction. The MLP produced 64-dimensional embeddings, which were then used as inputs for the deep learning-based survival models.

For the DeepSurvNet and DeepHit models, we set a batch size of 256 and trained for 100 epochs, incorporating early stopping to optimize performance. This was feasible due to the reduced input feature dimensions and the efficient implementation provided by the PyCox library.

As DeepHit is a discrete survival model, it generates a separate survival prediction for each evaluation time point. To aggregate patch-level predictions, we applied a percentile 99 (p99) strategy (**Figure 32C**). This resulted in 72 survival probability outputs, corresponding to the first five years (72 months) of follow-up for each patient.

To fairly evaluate the DeepHit model, we computed the mean of the time predictions, using Harrell's C-index and AUC as performance metrics.

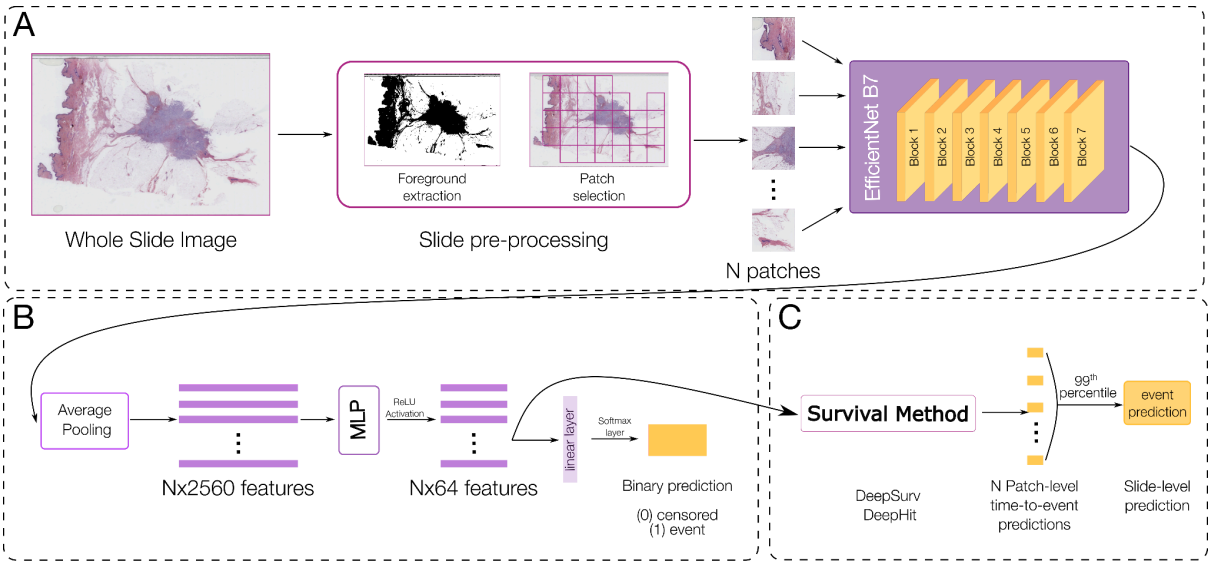


Figure 31. Overview of the baseline pipeline for Overall Survival (OS) and Disease-Free Survival (DFS) prediction from whole-slide images. (A) Whole-slide images (WSIs) are pre-processed through foreground extraction to remove background and artefacts, followed by patch selection to divide each slide into N non-overlapping image tiles. Each patch is then encoded using a pre-trained EfficientNet-B7 backbone, producing high-dimensional feature embeddings. (B) The extracted patch-level embeddings (N × 2560) are passed through a multi-layer perceptron (MLP) for dimensionality reduction (to 512 features) with non-linear activations (ReLU) and dropout regularisation. The resulting patch features are aggregated by mean pooling to obtain a slide-level representation summarising global histological patterns.

(C) This aggregated feature vector is then fed into survival modelling frameworks, including Cox proportional hazards (Cox-PH), survival support vector machines (Survival-SVM), and random survival forests (Survival-RF), to predict survival outcomes.

The models were trained using censoring-aware loss functions and evaluated for both OS and DFS tasks. This modular baseline establishes a transparent, interpretable pipeline linking histological morphology to time-to-event outcomes.

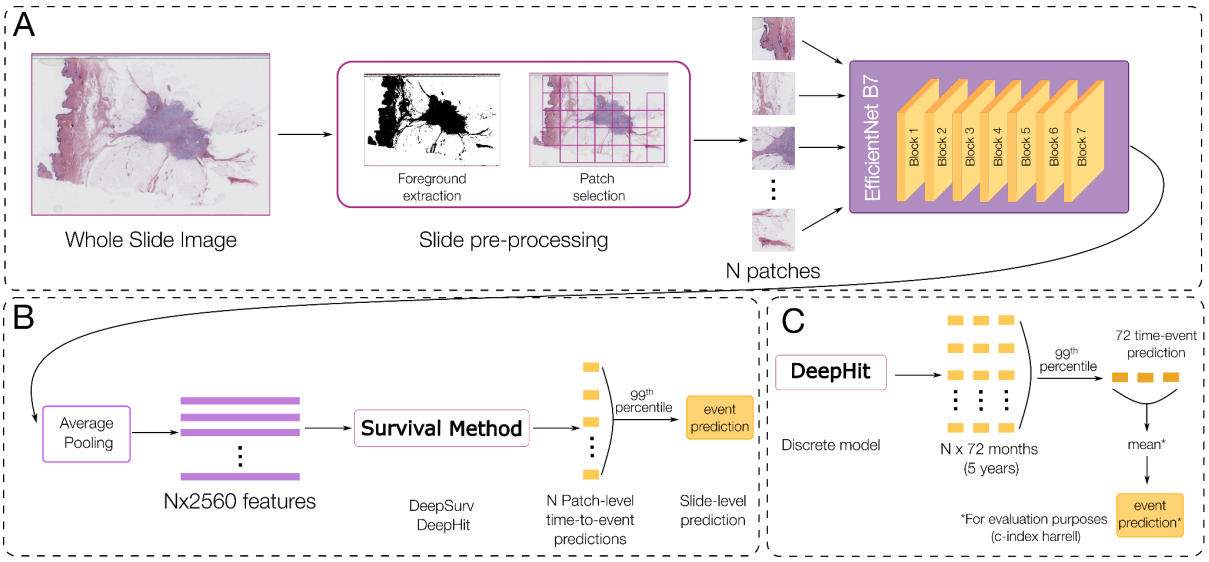


Figure 32. Overview of the deep-learning survival framework for Overall Survival (OS) and Disease-Free Survival (DFS). This architecture extends the baseline CoxPH pipeline by replacing the linear survival layer with non-linear deep survival networks (DeepSurv and DeepHit). (A) Whole-slide images are processed through EfficientNet-B7 to extract patch-level embeddings. (B) These embeddings are aggregated and used to train the DeepSurv model (continuous hazard formulation) and DeepHit (discrete time-to-event formulation).

(C) The DeepHit outputs are ensembled over 72 monthly time intervals to produce slide-level survival predictions, allowing evaluation with C-index metrics.

Compared with the baseline CoxPH approach (Figure 1), this framework captures non-linear interactions between image features and survival risk and can model discrete time-to-event probabilities, improving flexibility at the cost of interpretability.

RESULTS

Disease Free Survival (DFS) prediction

As described in **Figure 33**, in the DFS task coxPH method's superiority is only observed in the AUC (0.667) and Harrell's concordance index (0.649), whereas for the C-index Uno, it performs the worst, with a value of 0.496.

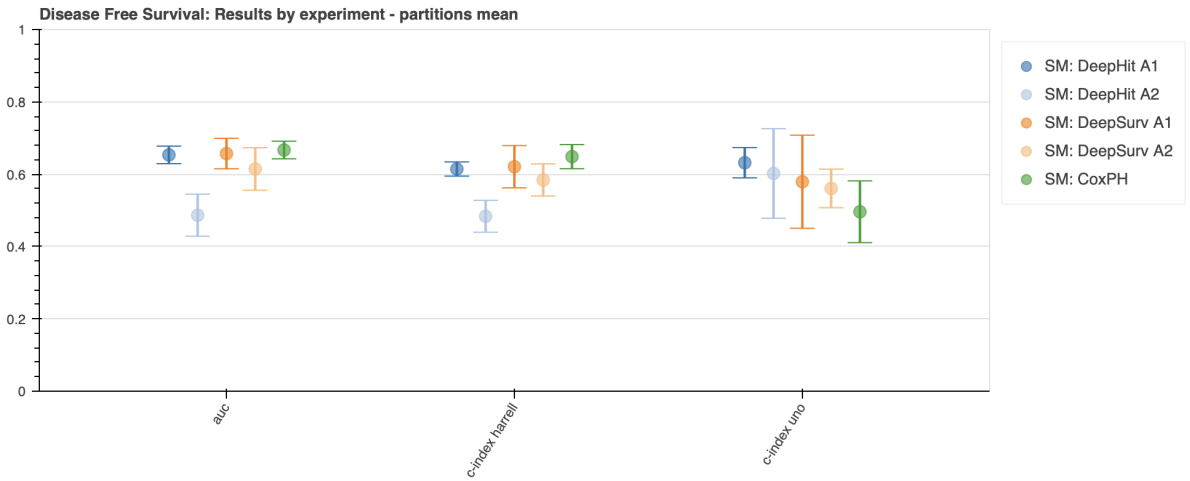


Figure 33. Comparison of Survival Models (SM) using traditional CoxPH and DL-based methods for Disease-Free Survival. This figure compares the performance of traditional Cox proportional hazards (Cox-PH) and deep learning–based survival models (DeepSurv, DeepHit) across internal validation partitions. Two configurations are shown for each deep model: **A1**, using dimensionally reduced ("condensed") feature vectors; and **A2**, using the full EfficientNet-B7 feature vector of size 2560.

Among the alternative models, the next best performer in AUC and Harrell's C-index is DeepHit trained on dimensionally reduced embeddings, achieving an AUC of 0.653, a Harrell's C-index of 0.614 and a Uno's C-index of 0.632. DeepSurv delivers comparable performance but exhibits higher standard deviation, indicating greater variability across runs.

Overall Survival (OS) prediction

Figure 34 shows that the Cox-PH model achieves the highest performance across all three evaluation metrics, with an AUC of 0.745, a Harrell's C-index of 0.722, and a Uno's C-index of 0.716.

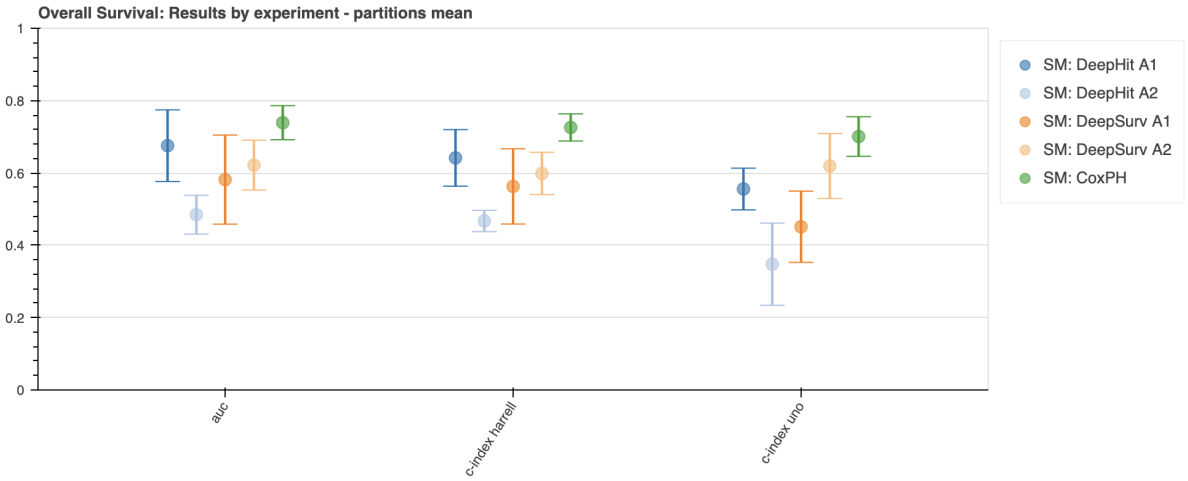


Figure 34. Comparison of Survival Models (SM) using traditional CoxPH and DL-based methods for Overall Survival (OS) prediction. This figure compares the performance of traditional Cox proportional hazards (Cox-PH) and deep learning–based survival models (DeepSurv, DeepHit) across internal validation partitions. Two configurations are shown for each deep model: **A1**, using dimensionally reduced ("condensed") feature vectors; and **A2**, using the full EfficientNet-B7 feature vector of size 2560. For DeepHit and Deepsurv we compare two different approaches (A#). A1: using the features of the condensed vector, A2: use the efnb7 feature vector of size 2560.

3.3.4 Using clinical information for predicting OS and DFS using post-NAC surgical specimen

In this section, we aim to assess whether our deep learning approach outperforms standard clinical information alone and to quantify the potential clinical gain achieved by integrating a deep learning image-based approach with traditional pathology data.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

The clinical model was trained using the clinical feature set described in **Chapter 2** (**Section Clinical variables**). All variables were one-hot encoded following the preprocessing pipeline introduced earlier. In this chapter, we focus on evaluating the prognostic information contained in these clinical features alone and comparing their performance to image-based and hybrid models.

Given the results from the previous sections, we decided to keep the two best survival methods for our evaluation: CoxPH and DeepHit.

To evaluate the added value of image-based features, we also combined the outputs of the survival models (i.e. Cox-PH and DeepHit) using two complementary data sources:

- Image-based features extracted from the EfficientNet-B7 model, referred to as the **Whole Slide Image Model (WSIM)**
- Clinical information encoded as a one-hot vector, referred to as the **Clinical Model (CM)**

For the combined prediction, we performed a simple average of the model scores at the whole-slide level.While exploring the possibility of averaging scores at the patch level, it is important to note that clinical information is identical for all patches from the same patient. Consequently, averaging at the patch level would yield the same result as averaging at the whole-slide level.

Prediction normalisation for WSIM and CM combination

In survival analysis tasks, particularly when combining outputs from different models such as those based on image features (e.g., WSIM) and clinical data (e.g., CM), score normalisation plays a crucial role in ensuring fair and effective ensembling.

One key reason for normalising the scores is the inherent difference in their value ranges. The output scores from models like DeepHit-CM were observed to be an order of magnitude higher than those from DeepHit-WSIM. Without normalisation, averaging such disparate values would result in the model with larger outputs disproportionately influencing the final prediction, regardless of its actual predictive accuracy.

Models like Cox-PH generate risk scores, while models like DeepHit predict time-dependent survival probabilities. Normalising these scores aligns their behavior, ensuring the ensemble reflects meaningful survival patterns rather than amplifying differences due to score distribution alone.

To this end, we choose two normalisation methods: Min-Max scaling and the sigmoid function. Min-Max scaling is intuitive and aligns scores proportionally within the observed range, making it a straightforward choice for standardising model outputs. However, this method is dependent on the score distribution of the specific test set, which may vary significantly. The sigmoid function, on the other hand, offers greater stability, making it less sensitive to extreme values and better suited for cases where score distributions are inconsistent.

RESULTS

Disease Free Survival (DFS) prediction

In the DFS task (**Figure 35),** the simple averaging of the CM and WSIM-CoxPH models led to slight improvements in both the AUC (0.691) and Uno's C-index (0.7402) metrics.

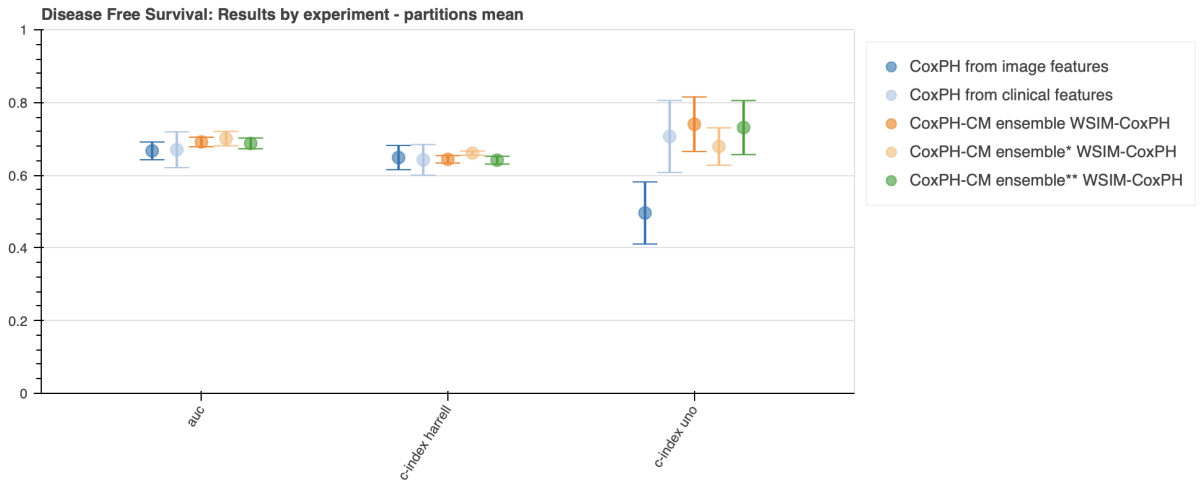


Figure 35. Comparison of different ensembling methods for the Cox PH model from image features (WSIM-CoxPH) and the CoxPH model from clinical features in the DFS task. The slide-level scores of this model were averaged after normalizing with *the minimum and maximum values of the test scores, and with a **sigmoid function.

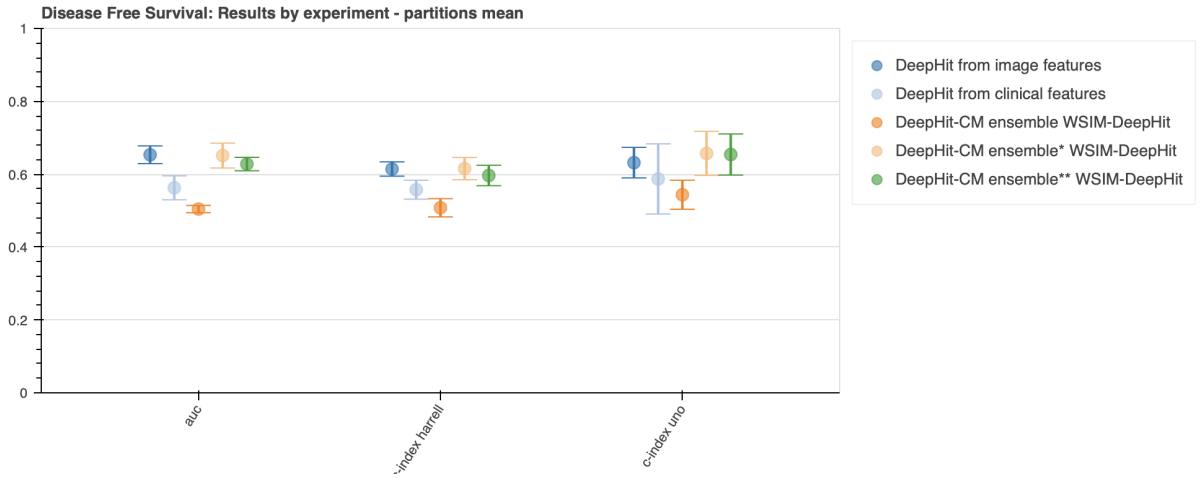


Figure 36. comparison of different ensembling methods for the DeepHit model from image features (WSIM-DeepHit) and the DeepHit model from clinical features in the DFS task. The slide-level scores of this model were averaged after normalizing with *the minimum and maximum values of the test scores, and with a **sigmoid function.

Conversely, **Figure 36** shows that ensembling the DeepHit model with clinical features was detrimental to performance in the DFS task without normalisation, achieving a notably low AUC of 0.577. This highlights that for DeepHit, ensembling with clinical features in this

context could be counterproductive, or is at least not giving additional information about the DFS.

Overall Survival (OS) prediction

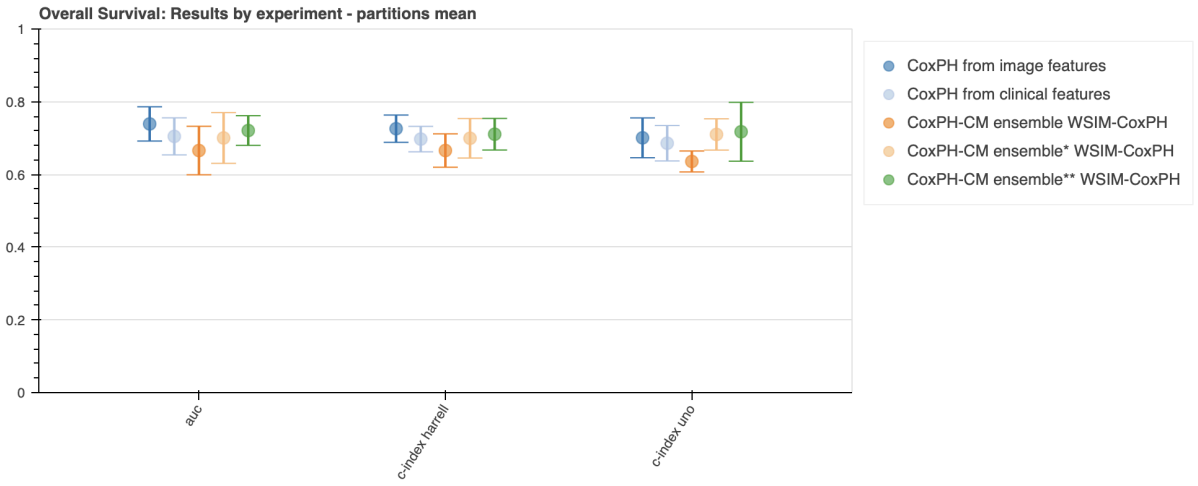


Figure 37. Comparison of different ensembling methods for the Cox PH model from image features (WSIM-CoxPH) and the CoxPH model from clinical features in the OS task. The slide-level scores of this model were averaged after normalizing with *the minimum and maximum values of the test scores, and with a **sigmoid function.

For the OS task, the simple ensembling approach without normalisation underperformed compared to the individual WSIM and CM models when using both Cox-PH (**Figure 37**) and DeepHit (**Figure 38**), underlying the need for normalisation when combining clinical and image-based information. Both models demonstrated improved performance with the sigmoid normalisation (but not a statistically significant improvement).

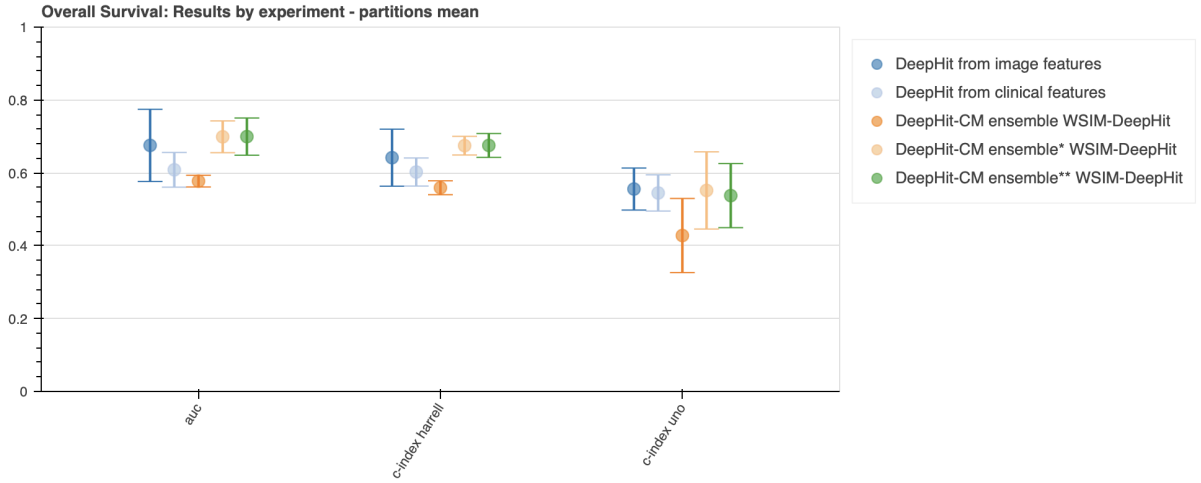


Figure 38. comparison of different ensembling methods for the DeepHit model from image features (WSIM-DeepHit) and the DeepHit model from clinical features in the OS task. The slide-level scores of this model were averaged after normalizing with *the minimum and maximum values of the test scores, and with a **sigmoid function.

Notably, none of the ensemble Cox-PH models surpassed the performance of the WSIM-CoxPH alone (**Figure 37**).

3.3.5 Exploring unsupervised pre-training for better feature extraction

Deep learning models in histopathology commonly rely on supervised pretraining on datasets like ImageNet to generate feature embeddings. While effective, these embeddings may overlook subtle domain-specific patterns essential for survival prediction.

Contrastive learning approaches such as SimSiam, MoCoV2, and SwAV have shown strong results by encouraging models to associate multiple views of the same image while distinguishing unrelated images. More recently, transformer-based frameworks like DINO have extended these successes by leveraging Vision Transformers (ViT), offering enhanced flexibility and performance in visual representation learning.

Motivated by these advances, we investigate whether unsupervised embeddings can improve survival prediction over conventional ImageNet-pretrained models like EfficientNet-B7.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

We investigate an alternative feature extraction strategy based on SimSiam[117], a self-supervised learning method that leverages Siamese networks and contrastive learning to generate robust image representations without labelled data. Specifically, we use a ResNet50 backbone pretrained with SimSiam to extract embeddings from histological patches.

SimSiam trains the network to maximise cosine similarity between embeddings from two augmented views of the same image, encouraging the model to learn invariant and generalisable morphological features. We fine trained the SimSiam model on the PRIMUNEO dataset for 100 epochs and replaced the previous EfficientNet-B7 embeddings in our survival pipeline with those obtained from the SimSiam-pretrained ResNet50. All other methodological aspects remained unchanged, including the use of Cox-PH and DeepHit models for survival prediction.

To evaluate the impact of this feature extraction method, we used the same partition splits as in earlier experiments and assessed model performance using Harrell's and Uno's C-index, as well as AUC at years 2–5 for DFS and years 3–5 for OS. For context, we also include performance results from prior experiments using supervised embeddings, allowing for a direct comparison and clearer insight into the benefits of self-supervised pretraining.

RESULTS

Disease Free Survival (DFS) prediction

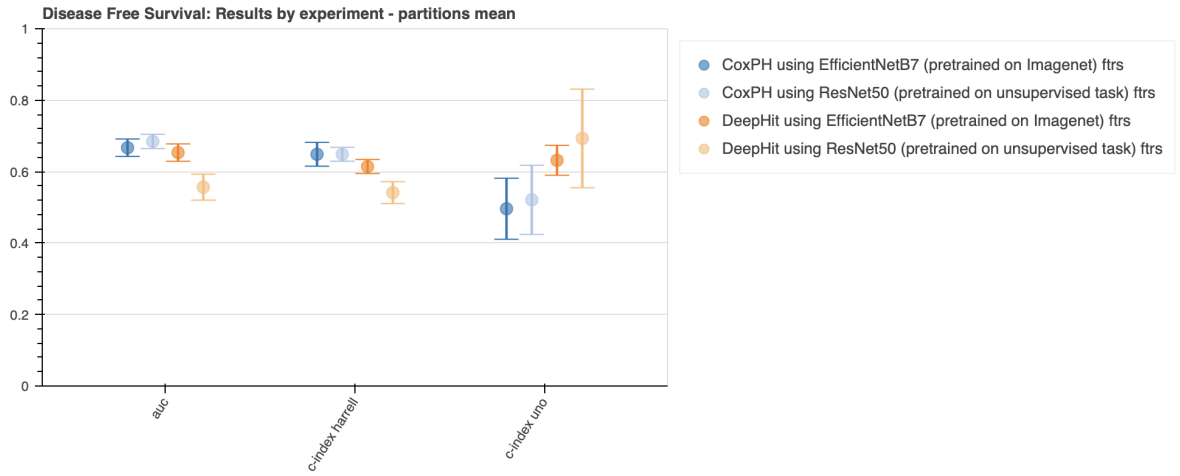


Figure 39. Comparison of feature extractors and survival modeling approaches for the DFS task. This figure compares the performance of CoxPH and DeepHit survival models when trained on embeddings extracted from two architectures: EfficientNet-B7 pretrained on ImageNet (supervised learning) and ResNet-50 pretrained with SimSiam (self-supervised learning). The results highlight the impact of backbone choice and pretraining strategy on disease-free survival prediction.

As shown in **Figure 39**, incorporating the ResNet50 embeddings into the Cox-PH model resulted in improved performance across multiple metrics. The model achieved an AUC of 0.685, a C-index (Harrell) of 0.649, and a C-index (Uno) of 0.521, representing a notable improvement compared to previous results.

However, the same ResNet50 embeddings proved less beneficial for the DeepHit model. One possible explanation for this outcome is that the ResNet50 features may already encode a substantial amount of the information that DeepHit is designed to capture.

Overall Survival (OS) prediction

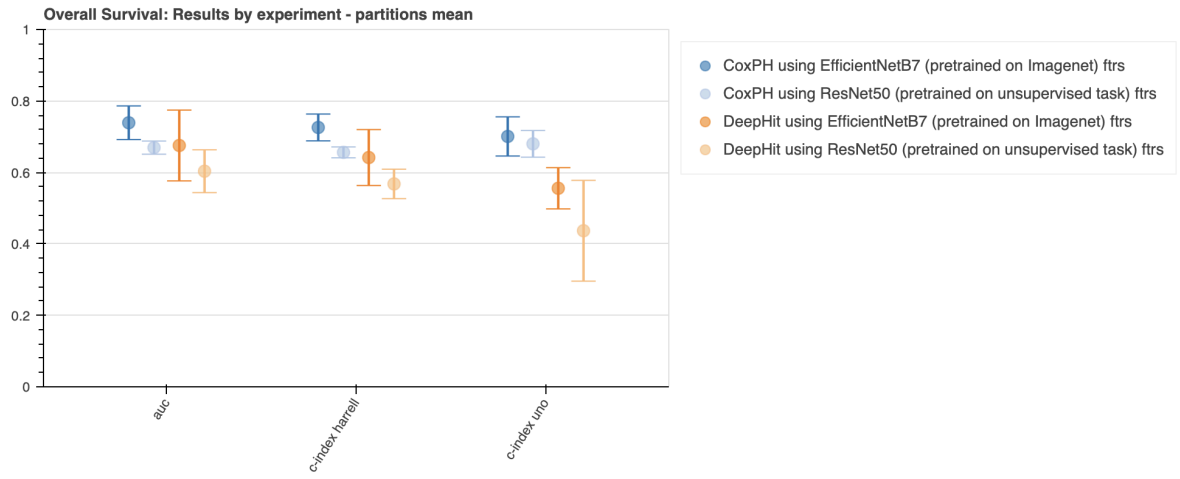


Figure 40. Comparison of feature extractors and survival modeling approaches for the OStask. This figure compares the performance of CoxPH and DeepHit survival models when trained on embeddings extracted from two architectures: EfficientNet-B7 pretrained on ImageNet (supervised learning) and ResNet-50 pretrained with SimSiam (self-supervised learning). The results highlight the impact of backbone choice and pretraining strategy on disease-free survival prediction.

In the OS task, neither the Cox-PH nor the DeepHit model demonstrated improved performance when using ResNet50 embeddings (**Figure 40**). These results suggest that ResNet50 embeddings may be less optimal for the OS prediction task compared to those generated by EfficientNet-B7, despite both models being pretrained on datasets within the histopathological image domain.

3.3.5bis Exploring more unsupervised pre-training for better feature extraction

In this section, our objective is to explore alternative pretrained architectures as feature extractors to enhance the performance of the best-performing survival prediction model. This exploration aims to determine whether different architectures and training strategies can generate richer, more informative feature representations, particularly for challenging prediction tasks such as DFS and OS.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

We evaluate the impact of different pretrained feature extractors:

- EfficientNet-B7 (ImageNet-pretrained, fined-tuned on PRIMUNEO)
- EfficientNet-V2 M (ImageNet-pretrained, fined-tuned on PRIMUNEO)
- ResNet50 (ImageNet-pretrained and additional self-supervised versions fine-tuned with MoCoV2, BT, and SwAV, fined-tuned on PRIMUNEO)
- Vision Transformer (ViT) models with 8-patch and 16-patch configurations, trained with the DINO self-supervised method on Lunit's dataset (already pretrained, fine-tuned on PRIMUNEO).

The selection of self-supervised learning (SSL) methods in this study was driven by their demonstrated success in computer vision and their suitability for the specific demands of histopathological image analysis in survival prediction tasks. We focused on four state-of-the-art models: DINO[118], SwAV[85], MoCoV2[87] , and Barlow Twins[117–120], each offering complementary strengths aligned with our dataset and objectives.

Given the limited availability of large-scale annotated histology datasets, we prioritised SSL methods capable of learning rich, discriminative visual representations from unlabeled data. DINO and SwAV were selected for their ability to capture complex tissue architecture and long-range spatial dependencies, particularly when combined with Vision Transformers (ViTs), which are well-suited for modelling the tumour microenvironment and global tissue context.

For convolutional architectures like ResNet50, we employed SwAV, MoCoV2, and BT, which combine local feature learning with robust contrastive frameworks. We excluded alternatives like SimCLR and BYOL[85,121] due to their reliance on large batch sizes or compute-intensive setups, which are less feasible with whole slide images (WSIs). Likewise, although promising, methods such as VICReg were deprioritised due to comparatively limited performance benchmarks in medical imaging contexts[122].

All models were evaluated using the Cox-PH and DeepHit survival frameworks. The only adjustment involved adapting the top classifier's input dimensions to the embedding size of each architecture.

RESULTS

Disease Free Survival (DFS) prediction

From **Figure 41** where the survival method used is the CoxPH, we observe that the ViT/Small/16 patches (ViT/S16) model trained with the DINO self-supervised method achieves the highest performance across multiple metrics, with an AUC of 70.7%, a C-index (Harrell) of 68.3%, and a C-t index of 69.4%. This result demonstrates the strong potential of self-supervised learning methods like DINO when applied to ViT architectures for survival prediction tasks. The second-best performing model was the ViT/Small/8 patches (ViT/S8), further emphasizing the potential of transformer-based models in this context.

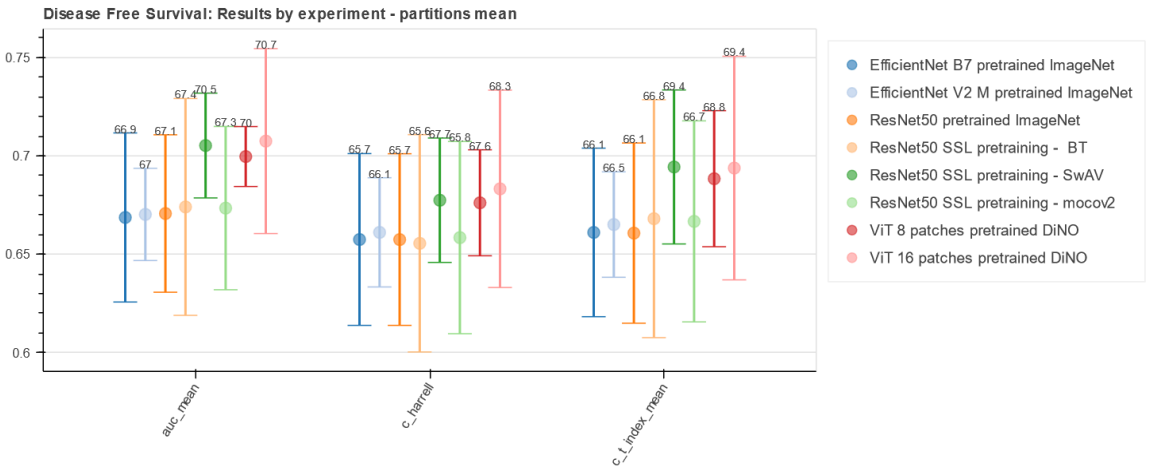


Figure 41. Comparison of pretrained feature extractors for CoxPH-based survival prediction in the DFS task. Performance of Cox proportional hazards (CoxPH) models trained on embeddings extracted from multiple pretrained networks: EfficientNet-B7 and EfficientNet-V2M (supervised on ImageNet), ResNet-50 (supervised or self-supervised with

BYOL, SwAV, MoCo v2), and Vision Transformers (ViT, pretrained with DINO on 8×8 and 16×16 patches). Bars indicate AUC, Harrell's C-index, and Uno's C-index performance across internal validation partitions.

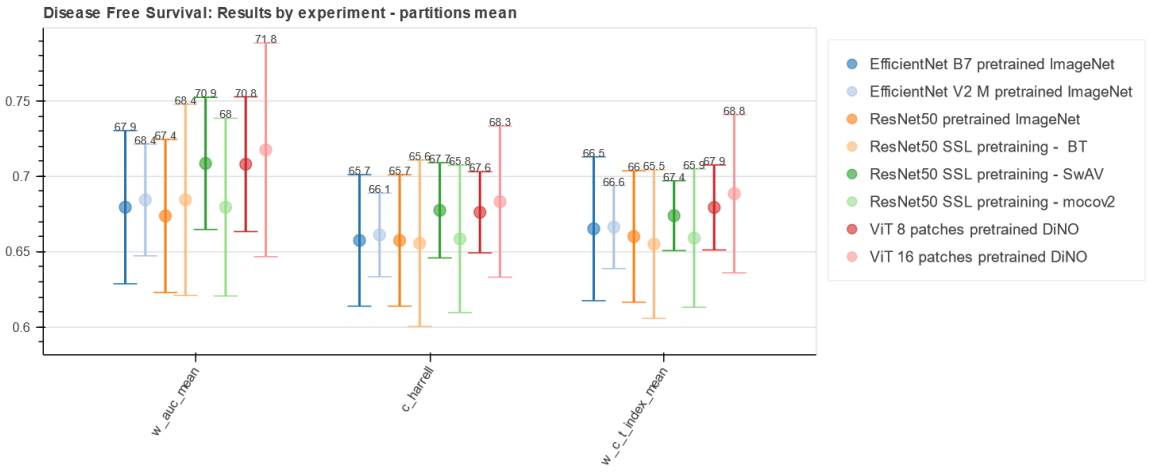


Figure 42. Comparison of pretrained feature extractors for DeepHit-based survival prediction in the DFS task. Performance of DeepHit survival models trained on the same set of pretrained image embeddings as in Figure 1. Weighted AUC, Harrell's C-index, and Uno's C-index are shown for each backbone, illustrating the impact of supervised versus self-supervised pretraining and convolutional versus transformer architectures on disease-free survival prediction.

In **Figure 42** where we use the DeepHit survival method, the ViT/S16 model once again achieved the best performance in the weighted AUC (71.8%) and weighted Ct-index (68.8%), further confirming its robustness in the DFS task. Notably, between the ViT/S8 and ResNet50 SSL (SwAV) models, both demonstrated competitive performance, consistently ranking as the second-best methods across all evaluated metrics.

Stratified Analysis by Molecular Subtype

Figure 43 presents the weighted mean AUC for OS prediction using embeddings from the ViT/S16 model pretrained with DINO, stratified by molecular subtype and dataset partition.

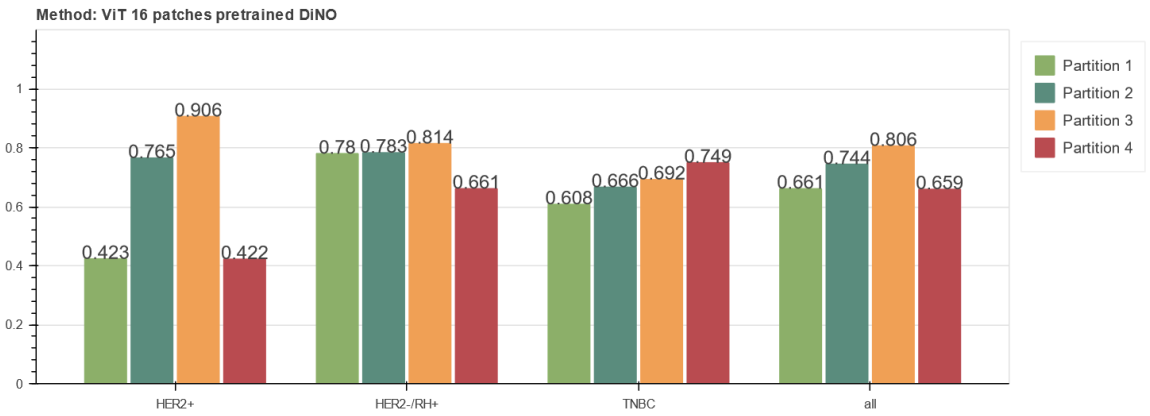


Figure 43. Subtype-specific performance of the ViT-S/16 model pretrained with DINO for DFS prediction. Mean weighted AUC values obtained using embeddings from the self-supervised Vision Transformer (ViT-S/16) pretrained with DINO, stratified by molecular subtype (HER2+, ER+/HER2−, and TNBC) and internal validation partition. The results highlight variability across partitions and molecular subtypes, with notably higher performance observed in HER2+ tumours.

Performance varied significantly across subtypes and partitions. For HER2+ tumours, the AUC ranged widely, from 0.422 in partition 4 to 0.906 in partition 3, indicating substantial variability. This may be due to clinical heterogeneity, especially since HR+ and HR– cases were pooled to increase statistical power, potentially introducing noise.

For HER2–/RH+ patients, results were more stable, with AUCs ranging from 0.661 to 0.814, and highest performance again seen in partition 3. Similarly, TNBC patients showed consistent AUCs between 0.608 and 0.749, suggesting the model captured key patterns across partitions despite lower absolute performance.

When aggregating all subtypes, the model performed best in partition 3 (AUC 0.806) and worst in partition 4 (0.659), reflecting the same partition-dependent trend.

These findings support the robustness of ViT/S16 for HER2–/RH+ and TNBC, while highlighting variability and potential heterogeneity in HER2+ patients. This underscores the need for refined stratification and possibly subtype-specific models to improve generalisation.

Overall Survival (OS) prediction

Figures 44 and **45** illustrate the performance of the proposed pipeline in the OS prediction task using embeddings generated from different architectures and model weights for the coxPH and DeepHit methods respectively.

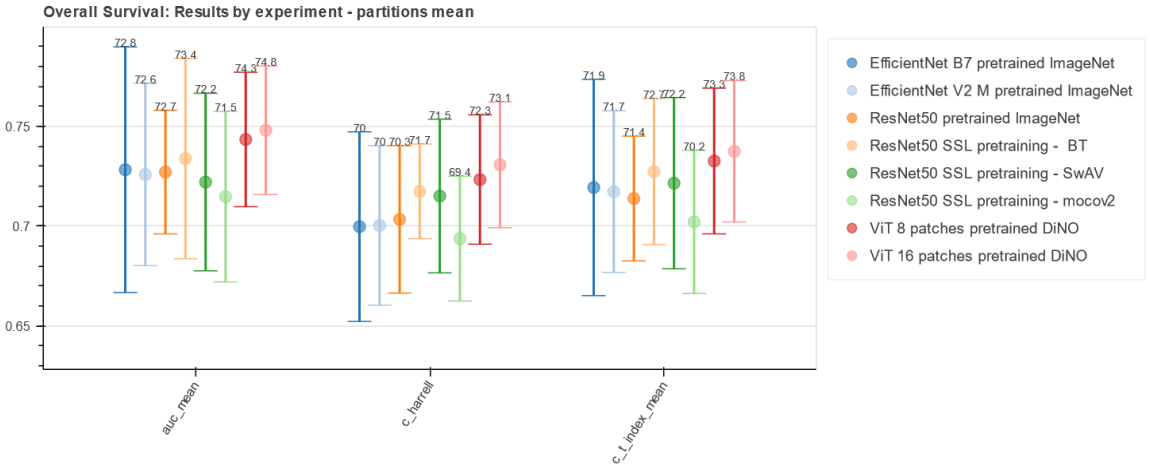


Figure 44. Comparison of pretrained feature extractors for CoxPH-based survival prediction in the OS task. Performance of Cox proportional hazards (CoxPH) models trained on embeddings extracted from multiple pretrained networks: EfficientNet-B7 and EfficientNet-V2M (supervised on ImageNet), ResNet-50 (supervised or self-supervised with BYOL, SwAV, MoCo v2), and Vision Transformers (ViT, pretrained with DINO on 8×8 and 16×16 patches). Bars indicate AUC, Harrell's C-index, and Uno's C-index performance across internal validation partitions.

The ViT/Small/16 patches (ViT/S16) model pretrained with the DINO self-supervised method achieved the highest performance across key metrics. Specifically, it reached an AUC of 74.8%, a C-index (Harrell) of 73.1%, and a C-t index of 73.8%, outperforming all other tested models. The ViT/S8 model ranked second in overall performance, confirming the strong potential of Vision Transformers trained with DINO for survival prediction tasks.

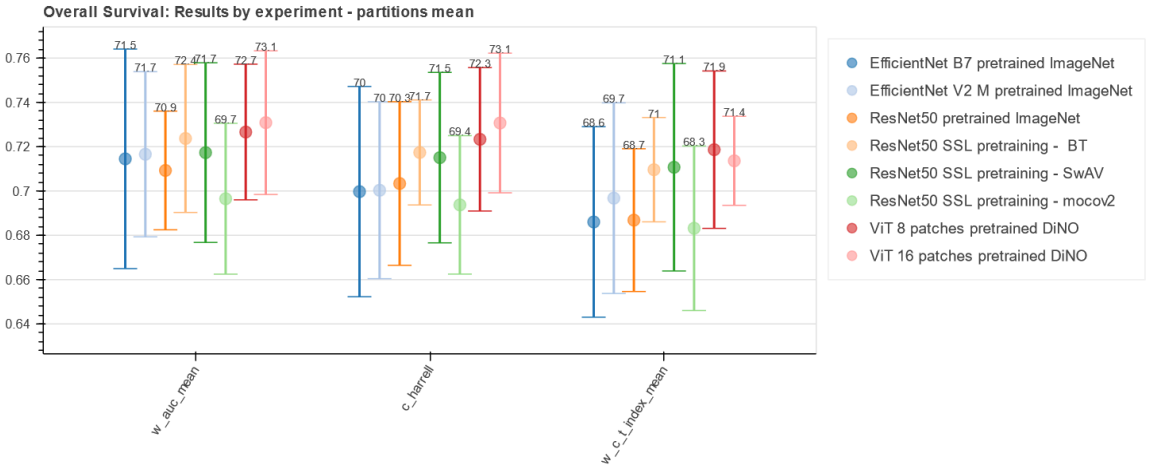


Figure 45. Comparison of pretrained feature extractors for DeepHit-based survival prediction in the OS task. Performance of DeepHit survival models trained on the same set of pretrained image embeddings as in Figure 1. Weighted AUC, Harrell's C-index, and Uno's C-index are shown for each backbone, illustrating the impact of supervised versus self-supervised pretraining and convolutional versus transformer architectures on disease-free survival prediction.

In **Figure 45**, the ViT/S16 model achieved the best performance in weighted AUC (73.1%).

Stratified Analysis by Molecular Subtype

Figure 46 presents the weighted mean AUC results for the ViT/S16 with DINO model, stratified by molecular subtype and partition.

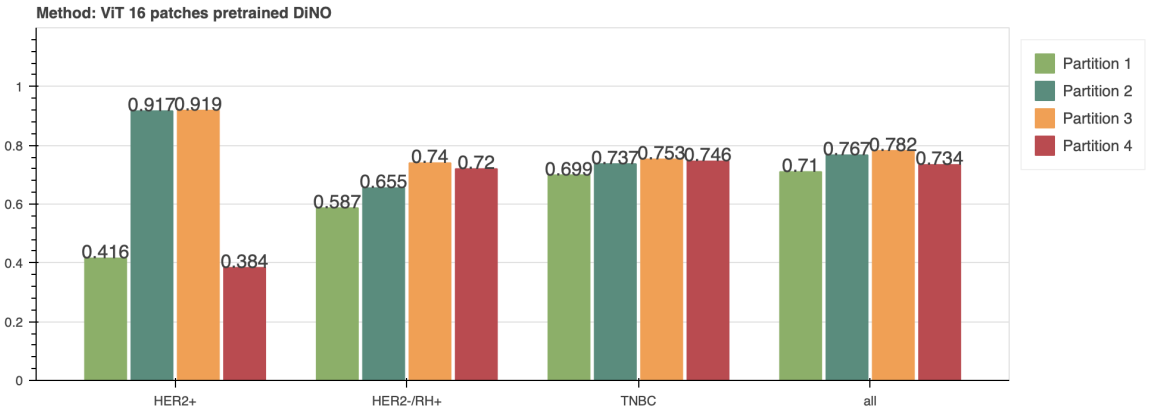


Figure 46. Subtype-specific performance of the ViT-S/16 model pretrained with DINO for OS prediction. Mean weighted AUC values obtained using embeddings from the self-supervised Vision Transformer (ViT-S/16) pretrained with DINO, stratified by molecular subtype (HER2+, ER+/HER2−, and TNBC) and internal validation partition. The results

highlight variability across partitions and molecular subtypes, with notably higher performance observed in HER2+ tumours.

For patients with HER2+ tumours performance varied considerably across partitions, ranging from a minimum AUC of 0.384 in partition 4 to a maximum AUC of 0.919 in partition 3. This wide performance gap aligns with the previously observed variability in the HER2+ subgroup, which may result from clinical heterogeneity, data imbalance, or confounding factors introduced by pooling HR+ and HR- patients.

For the HER2-/RH+ subgroup, performance showed more stability, with AUC values ranging between 0.587 and 0.74, suggesting that the model effectively captured relevant prognostic patterns within this group.

For the TNBC subgroup, the model achieved relatively stable results as well, with AUC values between 0.699 and 0.753, indicating strong generalization despite TNBC's known clinical complexity.

Time-Dependent AUC for DFS Prediction Stratified by Partition and Subtype

Figure 47 displays the time-dependent AUC from year 3 to 5 post-surgery across the four dataset partitions, stratified by molecular subtype (HER2+, HER2−/HR+, TNBC, and All patients). These results are based on the best-performing model using ViT/S16 embeddings and CoxPH. P-values above each point indicate the statistical significance of the AUCs.

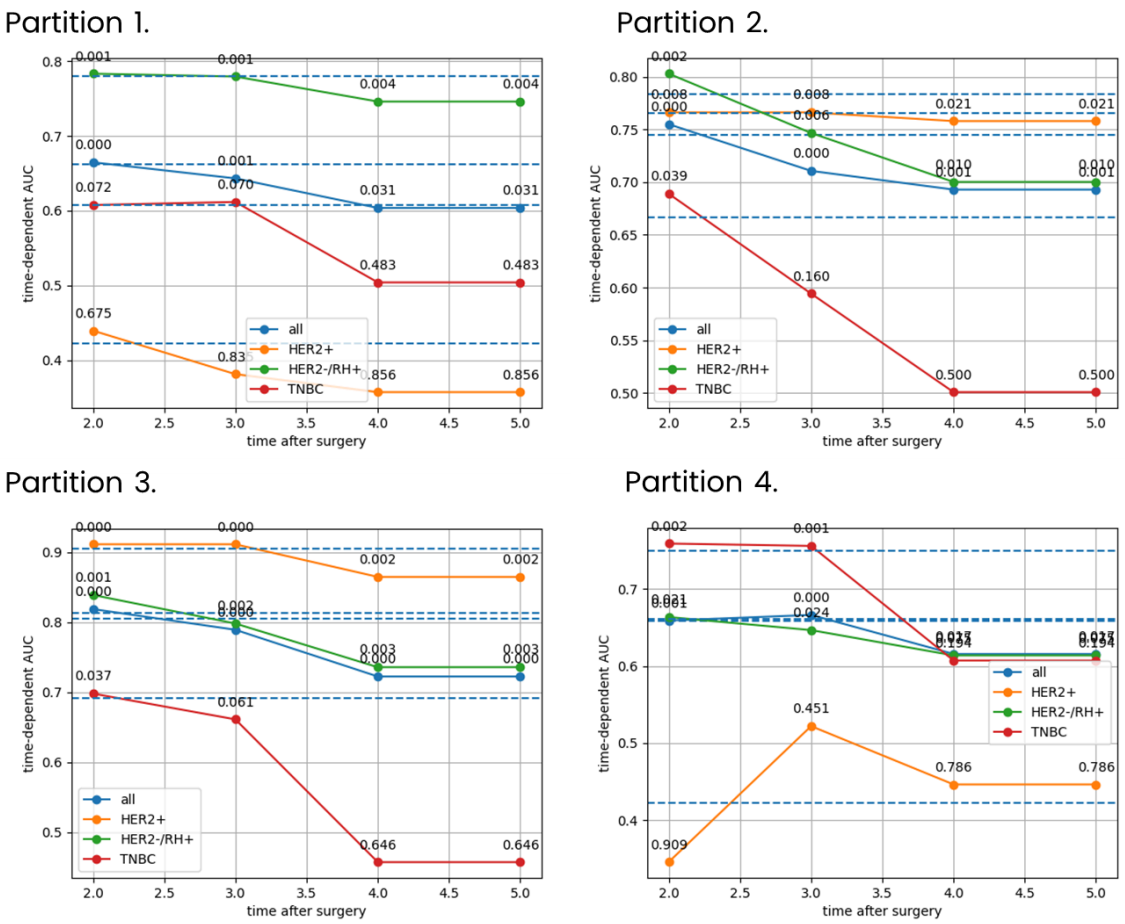


Figure 47. Temporal evolution of weighted AUC across molecular subtypes and internal partitions using the SSL ViT-S/16 model. Time-dependent AUC values computed yearly after surgery for disease-free survival prediction using embeddings from the self-supervised ViT-S/16 model pretrained with DINO. Each subplot corresponds to one internal validation partition, and curves represent molecular subtypes (HER2+, ER+/HER2−, and TNBC), along with the global cohort ("all"). The results illustrate the temporal stability of performance across subtypes, with HER2+ and luminal tumours showing more consistent predictive power over time compared to TNBC.

HER2−/HR+ patients showed the most consistent and robust performance across partitions and years, suggesting that the model captures reliable prognostic patterns in this group. In contrast, HER2+ patients exhibited high variability, with good early performance followed by drops in later years. This may stem from pooling HR+ and HR− cases, which introduced heterogeneity.

TNBC patients showed more stable results in partitions 3 and 4, though with lower performance and significance in later years. This reflects the challenge of modelling this aggressive and heterogeneous subtype.

The overall cohort ("All patients") achieved stable, significant performance, suggesting that common morphological signatures may be shared across subtypes.

Notably, AUCs were higher at year 3 across all groups, likely due to the pooling of early events (years 1–3) to address data sparsity. This also hints that early relapsing patients may share distinctive morphological traits, aligning with clinical knowledge about early-aggressive disease.

3.3.6 Exploring the Clinical Model

In this section, our primary objective is to compute and analyze the coefficients of the Cox Proportional Hazards (Cox-PH) model when trained using clinical information. By examining these coefficients, we aim to better understand the relationship between specific clinical features and the predicted outcomes for overall survival (OS) and disease-free survival (DFS). Identifying these associations may provide valuable insights into the prognostic value of various clinicopathological variables.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

The clinical model was trained using the clinical feature set described in **Chapter 2** (**Section Clinical variables**). All variables were one-hot encoded following the preprocessing pipeline introduced earlier.

RESULTS

Disease Free Survival (DFS) prediction

In the DFS task, the Clinical Model Cox-PH achieved a weighted AUC of 67.3%, a C-index (Harrell) of 66.5%, and a weighted C-t index of 66.3%. An analysis of the computed coefficients across folds and partitions shows that the values obtained from the sksurv and lifelines libraries are closely aligned (**Table 6**). This consistency suggests that both libraries effectively capture the same underlying relationships between clinical variables and DFS outcomes.

Table 6. Clinical information available in the PRIMUNEO dataset with all the possible values.

Clinical Information	Coefficient (sksurv)	Coefficient (lifelines)	Hazard ratio (exp(coefficient))
AJCC	0.164440	0.154167	1.166685
PAJCC	0.251513	0.269167	1.308873

PSBR	0.242779	0.244167	1.276557
SBR	0.056438	0.063333	1.065382
Time to surgery since diagnosis	0.203736	0.205000	1.227525
age_surgery	-0.013094	-0.010833	0.989225
pCR	-0.008762	-0.010833	0.989225
psubtype HER2+	-0.177413	-0.195000	0.822835
psubtype HER2-/RH+	-0.220763	-0.207500	0.812613
psubtype TNBC	0.114348	0.178333	1.195224
subtype HER2+	-0.231608	-0.236667	0.789254
subtype HER2-/RH+	-0.151099	-0.175000	0.839457
subtype TNBC	0.382707	0.392500	1.480678

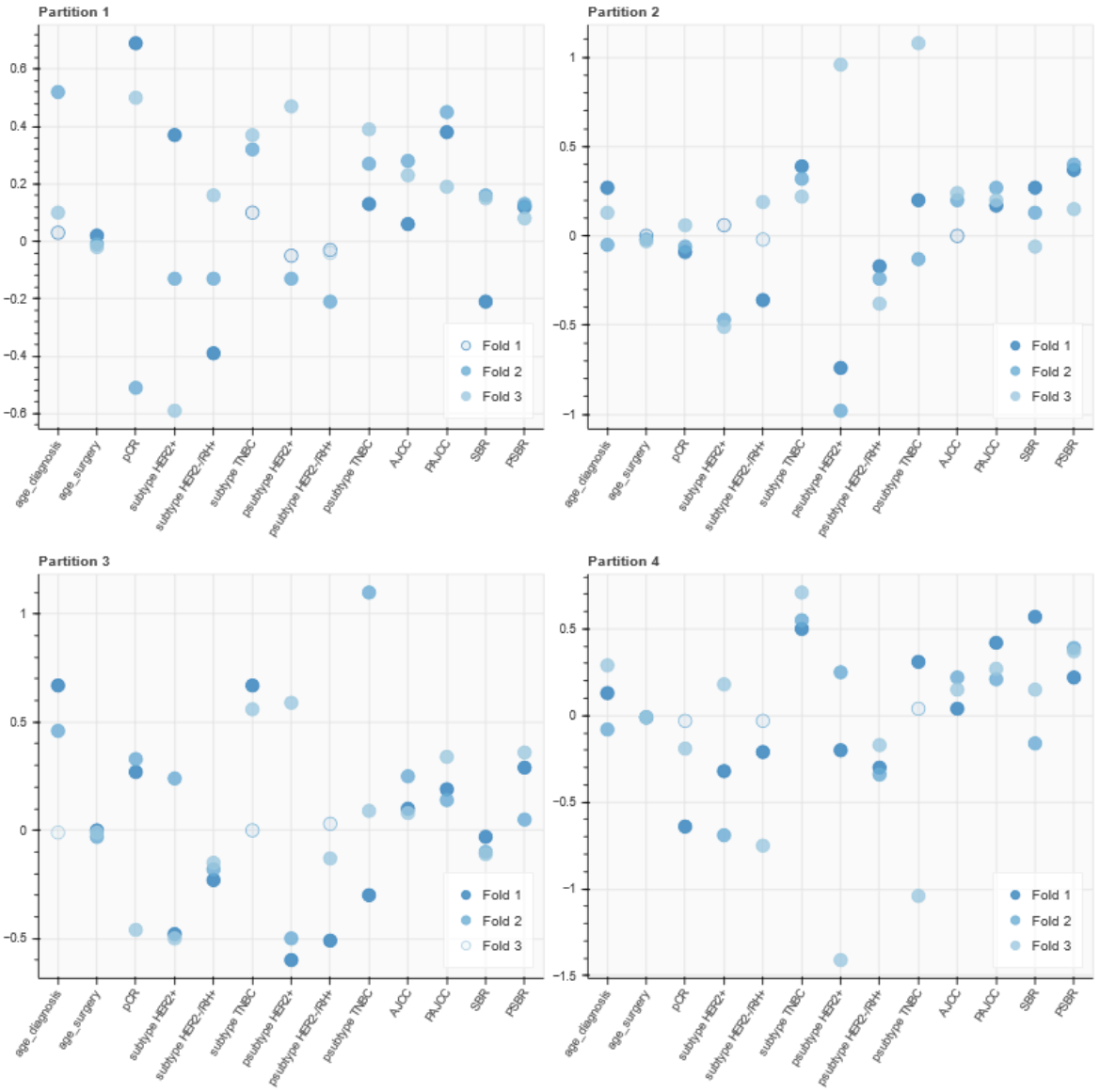


Figure 48. Significance and stability of clinical covariate coefficients across CoxPH models for the DFS prediction. Computed coefficients of the Clinical Model using the Cox proportional hazards (CoxPH) survival framework (lifelines library). Each panel corresponds to one of the four internal validation partitions, and dots represent the estimated coefficients across the three cross-validation folds. Fully coloured circles indicate statistically significant coefficients ($p < 0.005$), while empty circles denote non-significant effects. Positive coefficients correspond to covariates associated with increased risk, whereas negative values represent protective factors. This visualisation highlights both the stability and significance of clinical predictors, such as tumour size, nodal involvement, and Ki67, across partitions.

Most hazard ratios in our DFS analysis were near 1, indicating limited impact. However, higher HRs were observed for AJCC (1.166), pAJCC (1.308), and pSBR (1.276), reaffirming their roles as poor prognostic indicators. A longer interval between diagnosis and surgery (HR = 1.227) was also linked to worse DFS, potentially reflecting treatment delays in aggressive tumours rather than timing alone. TNBC status showed elevated risk in both biopsy (HR = 1.195) and surgical specimen (HR = 1.481), consistent with its known aggressiveness. In contrast, HER2+ and HER2-/HR+ subtypes were associated with better DFS (HRs < 1), aligning with their typically more favourable treatment responses.

Overall Survival (OS) prediction

In the OS task, the Clinical Model Cox-PH achieved a weighted AUC of 71.6%, a C-index (Harrell) of 70.0%, and a weighted C-t index of 70.2%. An analysis of the computed coefficients reveals that the values obtained from the sksurv and lifelines libraries are generally consistent, with the exception of differences observed for the pCR variable and the subtypes. This suggests that, despite minor discrepancies in implementation details between the libraries, both effectively capture the broader patterns of association between clinical variables and OS outcomes.

Table 7. Clinical information available in the PRIMUNEO dataset with all the possible values.

Clinical Information	Coefficient (sksurv)	Coefficient (lifelines)	Hazard ratio (exp(coefficient))
AJCC	0.152250	0.134167	1.143583
PAJCC	0.273470	0.270000	1.309964
PSBR	0.318205	0.300833	1.350984
SBR	0.349616	0.326667	1.386339
age_diagnosis	0.157818	0.149167	1.160866
age_surgery	-0.015878	-0.014167	0.985933
pCR	-0.110919	-0.015833	0.984291

psubtype HER2+	0.062838	-0.134167	0.874444
psubtype HER2-/RH+	-0.470715	-0.415000	0.660340
psubtype TNBC	0.059253	0.109167	1.115348
subtype HER2+	-1.386016	-0.866667	0.420350
subtype HER2-/RH+	0.299343	-0.044167	0.956794
subtype TNBC	1.086673	0.795000	2.214441

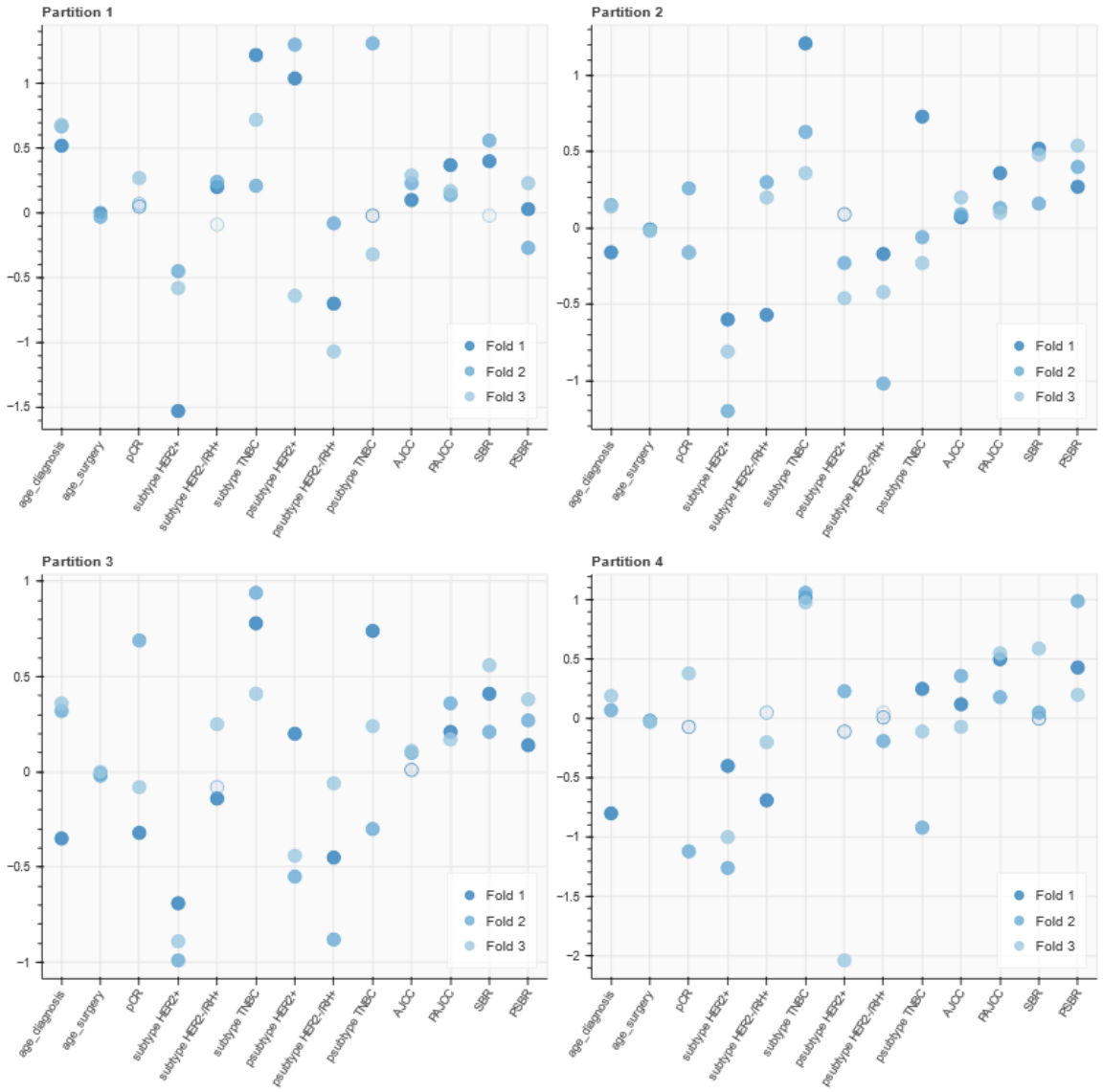


Figure 49. Significance and stability of clinical covariate coefficients across CoxPH models for the OS prediction. Computed coefficients of the Clinical Model using the Cox proportional hazards (CoxPH) survival framework (lifelines library). Each panel corresponds

to one of the four internal validation partitions, and dots represent the estimated coefficients across the three cross-validation folds. Fully coloured circles indicate statistically significant coefficients ($p < 0.005$), while empty circles denote non-significant effects. Positive coefficients correspond to covariates associated with increased risk, whereas negative values represent protective factors.

Examining hazard ratios (HR) reveals key clinical predictors of survival. TNBC (from biopsy) shows the highest risk, with HR = 2.21, indicating more than twice the risk of death compared to other subtypes, which is consistent with its aggressive nature. Conversely, the HER2+ subtype appears protective, with HRs of 0.42 (biopsy) and 0.87 (surgical), reflecting known treatment benefits. Higher AJCC/pAJCC stages and SBR/pSBR grades also correlate with worse outcomes (HR > 1), as expected. Age at surgery shows minimal effect (HR ≈ 0.985), suggesting limited prognostic value in this cohort.

3.3.7 Predicting Overall Survival (OS) and Disease Free Survival (DFS) with an end to end model

In the previous chapter, we introduced a three-step pipeline to predict Overall Survival (OS) and Disease-Free Survival (DFS) using the PRIMUNEO dataset. This approach combined: (1) feature extraction via EfficientNet-B7 pretrained on ImageNet or with SSL methods; (2) an intermediate top classifier for dimensionality reduction and binary event prediction; and (3) survival modeling using CoxPH or DeepHit. While effective, this modular design can introduce limitations, particularly the disjointed learning of features and survival outcomes, which may hinder the model's ability to capture fine-grained specific prognostic signals.

To address this, we now propose an end-to-end deep learning approach that eliminates precomputed embeddings and integrates survival prediction directly from Whole Slide Images (WSIs). By training the model to predict time-dependent risk scores from raw image data, we allow it to learn task-specific representations optimised for survival. This method is supposed to draw on the strengths of recent end-to-end frameworks in digital pathology[123,124] and the flexibility of DeepHit, which avoids CoxPH's proportional hazards assumption.

This unified architecture should enable the model to capture subtle morphological features and global spatial patterns often lost in patch-based pipelines. It also holds particular promise for aggressive subtypes like TNBC or early relapsing tumours, where high-resolution spatial context may be essential. By aligning representation learning directly with outcome prediction, our goal is to improve both accuracy and clinical relevance in survival modelling.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

In this new approach, we train an EfficientNet-B7 model directly from WSIs to predict survival or relapse within the 5 years following surgery for patients treated with neoadjuvant therapy. This formulation is treated as a 6-class classification task (including year 0 for patients who relapse immediately after surgery), where each class corresponds to the patient's event risk score for that year.

To aggregate the patch-level risk scores and generate slide-level predictions, we apply a percentile 99 aggregation strategy, selecting the highest predicted risk score among patches to ensure that high-risk regions are effectively captured (**Figure 50**).

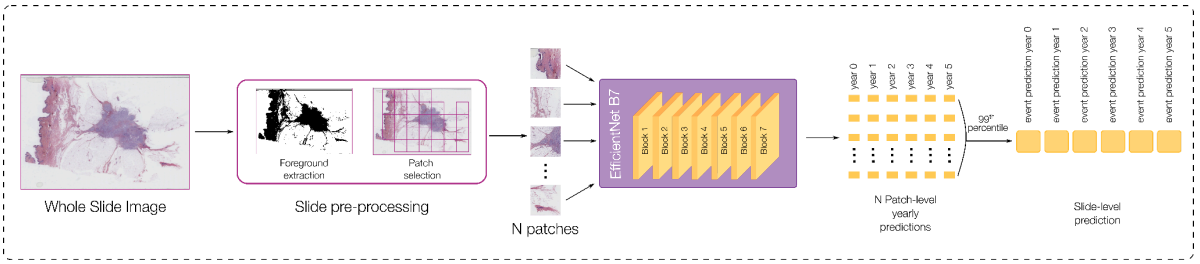


Figure 50. Overview of the end-to-end approach for overall survival (OS) and disease-free survival (DFS) prediction. The pipeline illustrates the end-to-end workflow used to train and evaluate the deep learning models. Whole-slide images (WSIs) are first pre-processed by foreground extraction and tiling into fixed-size patches. Each patch is then encoded using an EfficientNet-B7 backbone to extract high-dimensional visual features. These patch-level embeddings are passed through survival-specific network heads that output yearly risk predictions. The predictions are subsequently aggregated at the slide level to produce patient-specific survival risk estimates for OS and DFS tasks.

The EfficientNet-B7 model is initialized with ImageNet-pretrained weights. We conduct experiments with two distinct loss functions:

- Binary Cross-Entropy (BCE) loss applied independently to each year.
- The DeepHit loss is designed to optimize survival score prediction in a time-dependent framework.

To control model complexity and promote stable training, we adopt a freeze ratio of 0.75, meaning that 75% of the EfficientNet-B7 layers are frozen during training. The model is trained for 5 epochs using a batch size of 64, an Adam optimizer, and a learning rate of 5e-5.

To assess the performance of the proposed end-to-end approach, we retained the four previously defined partitions of the PRIMUNEO dataset as established in the first chapter.

Evaluation Metrics

To evaluate model performance, we employ the same metrics as before, but replacing the C-index Uno by the C-t index. It is the metric used in the original DeepHit implementation, designed to measure the model's ability to predict time-to-event outcomes in a discrete time

setting. While this metric was initially included in previous experiments, it demonstrated high instability and produced inconsistent results. Additionally, no other studies in the literature review reported using this metric for breast cancer survival prediction, further justifying its exclusion.

Focus on the OS Task

Due to the time-intensive nature of the end-to-end training approach, this set of experiments will focus exclusively on the OS task. Training the model for the DFS task would require additional experimentation time, as each full experiment takes approximately one day per partition .

RESULTS

Overall Survival (OS) prediction

Figure 51 presents the performance outcomes for various end-to-end models trained with different configurations, including:

- Loss functions: BCE loss and Single DeepHit loss.
- Training durations: 5 epochs and 20 epochs.
- Data sampling strategies: Random Patch Selection (RPS) using 500 randomly selected patches.

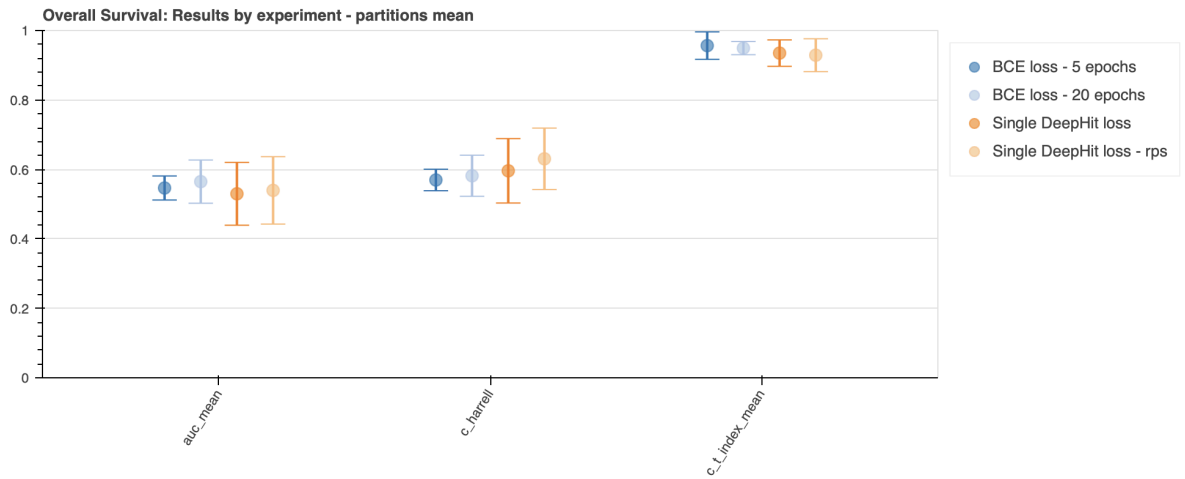


Figure 51. AUC, C-index Harrell and DeepHit's C-t index performance of end-to-end models trained with different hyperparameters. A comparison of the BCE loss trained during 5 and 20 epochs is done. We also show a performance comparison of BCE loss with

the Single DeepHit loss, the latter trained during 5 epochs, and training with 500 patches selected randomly (random patch selection, rps).

Overall, the end-to-end models underperformed compared to our previous 3-component pipeline, which achieved an AUC of 0.745 and a Harrell's C-index of 0.722 in the OS task using pretrained embeddings and CoxPH. In contrast, the best end-to-end model reached only an AUC of 0.564 (BCE loss, 20 epochs) and a C-index of 0.631 (DeepHit loss with random patch selection), highlighting a clear drop in predictive performance.

Models trained with the DeepHit loss yielded higher C-index values than those trained with BCE, suggesting better risk ranking. However, this improvement did not extend to AUC or C-t index metrics, indicating that DeepHit may enhance ranking without significantly improving overall classification performance.

Longer training also proved beneficial: extending BCE training from 5 to 20 epochs led to a marked increase in AUC, implying the original model was underfitted. The DeepHit model, trained for only 5 epochs, may similarly benefit from further optimisation.

Finally, random patch selection slightly improved both AUC (+0.1) and C-index (+0.4), suggesting that increased patch-level variability could support better model generalisation. Nonetheless, the end-to-end models still lag behind the pretrained embedding-based pipeline, indicating a need for further refinement.

Table 8. AUC performance in years 3 to 5 of the end-to-end model trained with the Single DeepHit loss with random patch selection. P-value is shown for statistical significance.

Yearly Performance	AUC	p-value
Year 3	0.534	0.384
Year 4	0.584	0.227
Year 5	0.519	0.458

These results confirm the low predictive power of this method across all evaluated time points. Moreover, none of the p-values indicate statistical significance, further demonstrating the model's limited reliability in effectively predicting survival outcomes.

3.3.8 External validation of the Overall Survival (OS) and Disease Free Survival (DFS) prediction pipeline

To evaluate the generalisability of our survival models, we externally tested the best-performing configuration (ViT/S16 pretrained with DINO pretrained by Lunit, fined-tuned on PRIMUNEO) on the CGFL Neoadj dataset, composed of breast cancer patients treated with NAC at a different institution.

This model, trained exclusively on the PRIMUNEO cohort (excluding CGFL samples), was selected based on its superior internal performance across both OS and DFS tasks. We retained all training settings and assessed its ability to generalise to unseen data, reflecting real-world clinical applicability.

In this external validation, we also explored the impact of magnification scale. PRIMUNEO WSIs were available at both x20 and x40 magnifications, allowing us to evaluate multi-scale training strategies, while CGFL slides were scanned only at x40. This setup helped assess model robustness to resolution shifts and potential benefits of multi-scale learning. This experiment offers a preliminary but essential test of model reliability across datasets, supporting its future clinical integration.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Our evaluation pipeline follows a structured process to ensure a robust assessment of the model's performance on the CGFL Neoadj dataset. The methodology consists of three key stages: Feature Extraction, MLP Training, and Survival Prediction, selected as the best performing model in the internal validation phase.

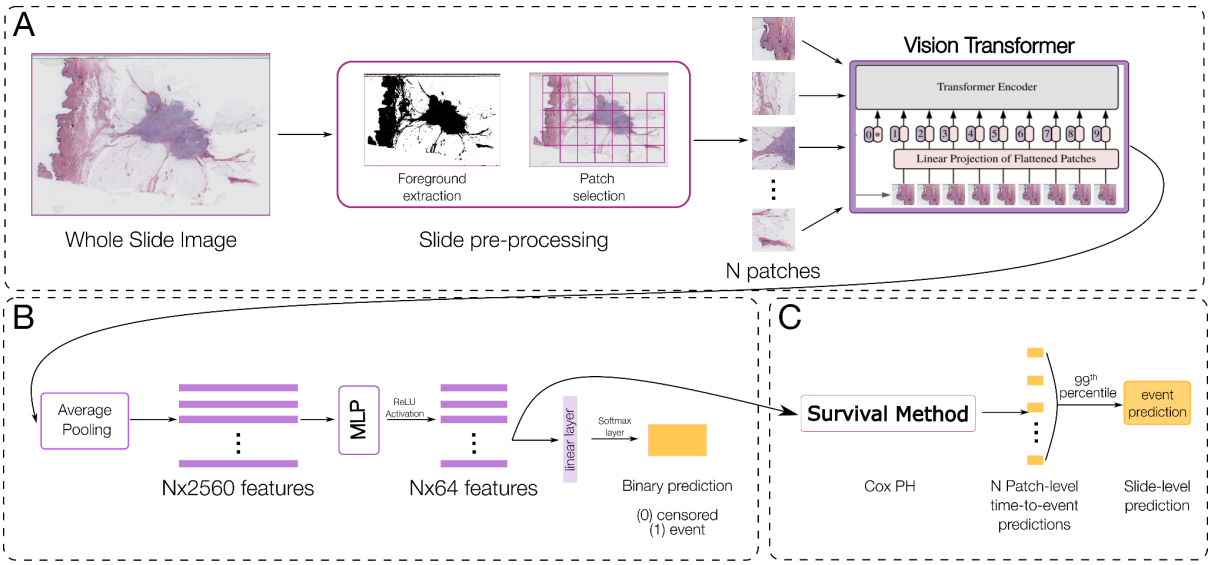


Figure 52. Overview of the best-performing method in the internal validation setup. This figure summarises the workflow of the optimal configuration identified during internal validation. **(A)** Whole-slide images (WSIs) are pre-processed through foreground extraction and tiling into fixed-size patches, which are then encoded using a Vision Transformer (ViT) pretrained on a self-supervised learning task to obtain patch-level embeddings. **(B)** These embeddings are processed through a multilayer perceptron (MLP) that reduces their dimensionality and outputs patch-level risk logits. **(C)** Patch-level predictions are aggregated into slide-level survival estimates using a percentile-based pooling strategy (99th percentile) and fitted to survival models such as Cox proportional hazards (CoxPH). This end-to-end configuration achieved the best overall performance for both disease-free survival (DFS) and overall survival (OS) tasks during internal validation.

Feature Extraction (Figure 52.A)

We used the ViT/S16 model pretrained with DINO (self-supervised learning) to extract features from WSIs. The PRIMUNEO dataset (excluding CGFL patients) was used for training, while both PRIMUNEO and CGFL datasets were encoded to ensure consistency.

Dimensionality Reduction (Figure 52.B)

A Multi-Layer Perceptron (MLP) with 512 and 64 output nodes was trained on a binary event prediction task (death/relapse vs. no event), using Binary Cross-Entropy (BCE) loss. This step reduces the high-dimensional ViT embeddings into compact, survival-relevant representations.

Survival Prediction (Figure 52.C)

The 64-dimensional embeddings from the MLP were used as input for a Cox Proportional Hazards (CoxPH) model, trained separately for OS and DFS. Patch-level risk scores were aggregated using the 99th percentile to generate slide-level predictions.

Training Details

Models were trained using 3-fold cross-validation, 5 epochs per fold, a batch size of 32, and the Adam optimizer (learning rate 5e-5). This setup promotes generalisability and robustness across institutions.

Exploring Alternative MLP Training Strategies

To improve performance, we tested three variations of the MLP training:

1. **Summed BCE Loss per Year**: The MLP was trained using BCE loss summed across 5 years.
   Rationale: Encourages the model to capture both early and late event risks.
2. **Single Model with Six Outputs:** A single MLP predicted binary event outcomes for each year (0–5) as six independent outputs.
   Rationale: Allows the model to specialise in year-specific survival intervals.
3. **Year-Specific MLPs with Embedding Concatenation**: Five separate MLPs were trained for each year, and their intermediate embeddings were concatenated before CoxPH modelling.
   Rationale: Captures time-specific features while preserving shared patterns.

RESULTS

Disease Free Survival (DFS) prediction

Figure 53 illustrates the importance of aligning image magnification between training and external validation datasets. Training on PRIMUNEO slides scanned at x40, the same resolution as the CGFL Neoadj dataset, significantly improved performance (weighted AUC = 58.1%) compared to using x20 scans (AUC = 49.5%). This 10-point increase underscores how consistent image detail and cellular resolution enhance feature extraction and prediction accuracy.

Alternative training strategies did not outperform the baseline model trained on x40 data, confirming this setup as the optimal configuration for external validation.

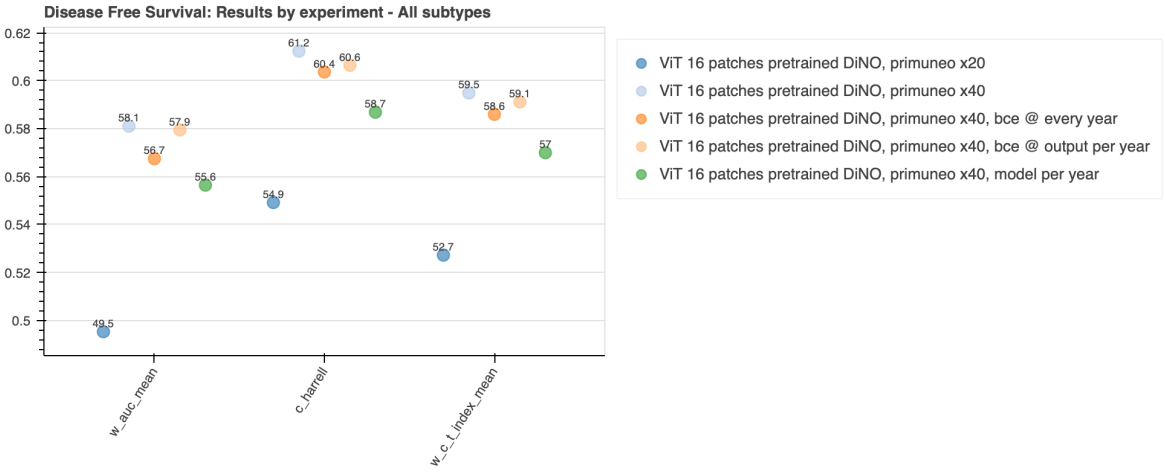


Figure 53. Comparison of weighted performance metrics for the disease-free survival (DFS) task in external validation. The figure presents the weighted AUC, Harrell's C-index, and Uno's C-index obtained in the external validation cohort for various configurations of the Vision Transformer (ViT) model pretrained with DINO. Each colour represents a different training strategy: varying the zoom level (×20 vs. ×40) and loss-function settings (binary cross-entropy applied at each yearly output or per-year model training). Across configurations, the model trained with ViT-16/DINO at ×40 magnification and yearly loss weighting achieved the highest and most stable results, confirming the robustness of temporal risk modelling in survival prediction.

Figure 54 presents model performance by molecular subtype in the DFS task. TNBC achieved the best results (AUC = 64.5%) with statistically significant predictions across all years, suggesting strong ability to detect aggressive disease patterns. HER2-/RH+ (Luminal) reached an AUC of 60.1%, with significant results at years 2 and 3, indicating some predictive capacity for early relapse. HER2+ showed the weakest performance (AUC = 51.5%) with no significant yearly results, highlighting challenges in capturing relevant prognostic signals for this group. These findings validate the model's generalisability for TNBC and luminal tumours, while pointing to the need for further optimisation in HER2+ subtypes.

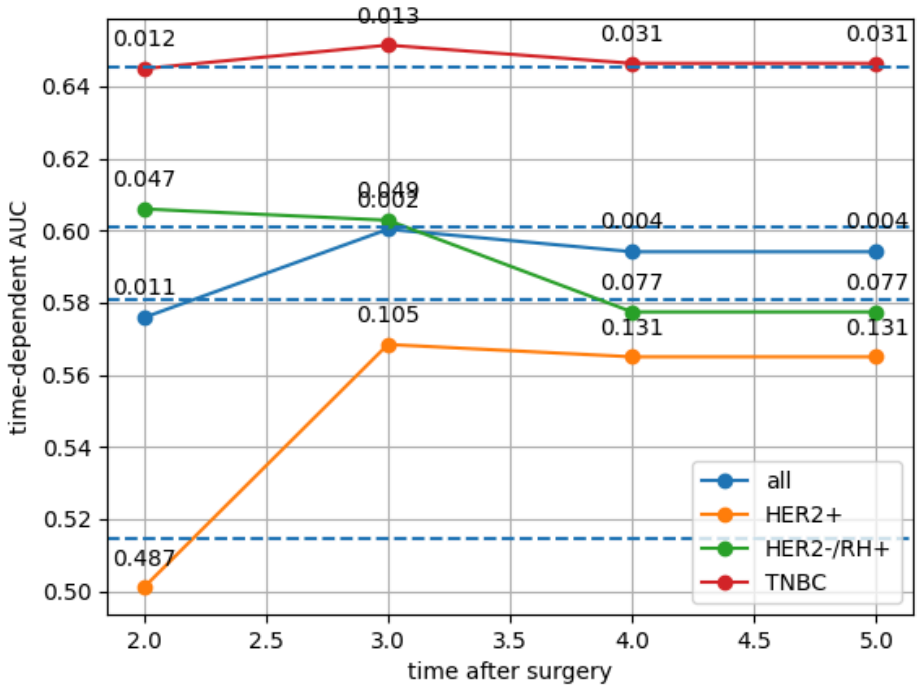


Figure 54. Time-dependent weighted AUC performance by molecular subtype for disease-free survival (DFS) prediction in external validation. The figure presents the temporal evolution of weighted AUCs across five years after surgery for the main molecular subtypes (HER2+, ER+/HER2–, and triple-negative breast cancer [TNBC]) as well as the overall population ("all"). The model was trained on the PRIMUNEO dataset (scanned at ×40 magnification) and evaluated on the independent CGFL cohort.

Overall Survival (OS) prediction

Figure 55 confirms the benefit of matching image magnification between training and test datasets. Training on PRIMUNEO slides scanned at x40 improved weighted AUC to 63.4%, compared to 56.9% when using x20 scans. As in the DFS task, none of the alternative training strategies outperformed this x40 baseline, underscoring the importance of magnification consistency and the robustness of the original MLP pipeline.

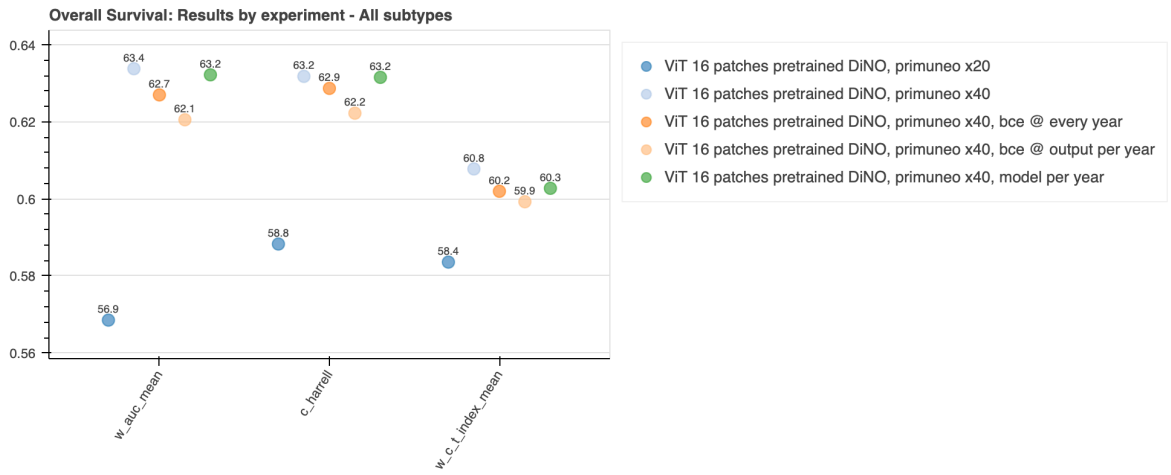


Figure 55. Comparison of weighted performance metrics for the overall survival (OS) task in external validation. The figure presents the weighted AUC, Harrell's C-index, and Uno's C-index obtained in the external validation cohort for various configurations of the Vision Transformer (ViT) model pretrained with DINO. Each colour represents a different training strategy: varying the zoom level (×20 vs. ×40) and loss-function settings (binary cross-entropy applied at each yearly output or per-year model training). Across configurations, the model trained with ViT-16/DINO at ×40 magnification and yearly loss weighting achieved the highest and most stable results, confirming the robustness of temporal risk modelling in survival prediction.

Figure 56 shows the model's performance by molecular subtype. TNBC yielded the best results (AUC = 69.3%), with statistically significant predictions across all years, confirming its value in high-risk groups. HER2+ achieved moderate performance (AUC = 61.8%), with significant prediction at year 3, suggesting potential for mid-term outcome prediction. HER2-/RH+ (Luminal) showed the weakest performance (AUC = 56.7%), with no significant results, indicating that alternative strategies may be needed for this more heterogeneous subgroup.

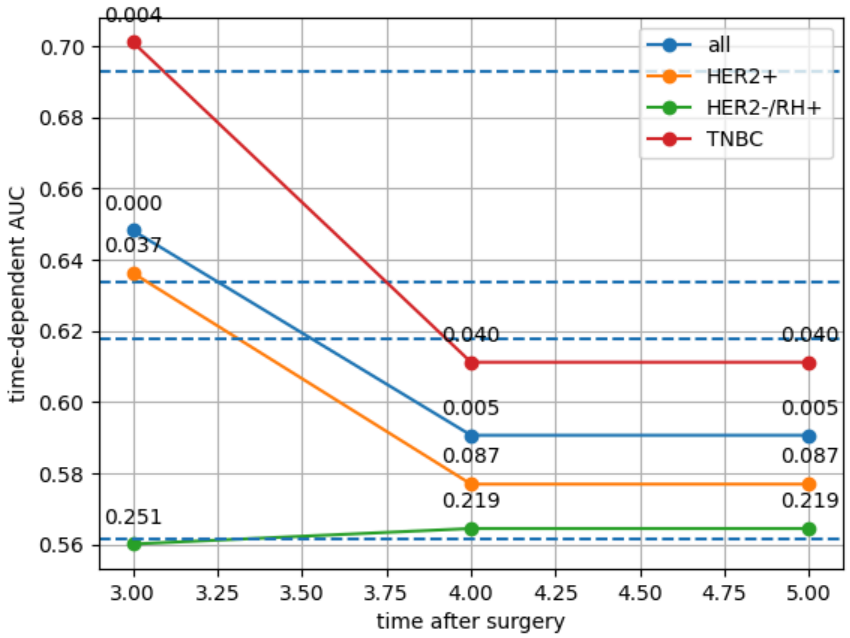


Figure 56. Time-dependent weighted AUC performance by molecular subtype for overall survival survival (OS) prediction in external validation. The figure presents the temporal evolution of weighted AUCs across five years after surgery for the main molecular subtypes (HER2+, ER+/HER2–, and triple-negative breast cancer [TNBC]) as well as the overall population ("all"). The model was trained on the PRIMUNEO dataset (scanned at ×40 magnification) and evaluated on the independent CGFL cohort.

To better account for the known interaction between molecular subtype and pCR status, we re-ran the multivariable clinical Cox model separately within each molecular subtype. As expected, the prognostic value of pCR was highly subtype-dependent (**Table 9).** In luminal (HR+/HER2−) tumours, achieving pCR was associated with a hazard ratio close to 1 (HR = 0.955), confirming its very limited additional prognostic information in this subtype. In contrast, pCR remained strongly protective in HER2+ (HR = 0.276) and TNBC (HR = 0.249) tumours.

Covariate	Luminal (HR+/HER2-)	HER2+	TNBC
AJCC (clinical)	1.125688	1.143855	**1.353554**
Pathological AJCC	1.189752	1.289754	1.274587
Post-NAC SBR	1.274563	**1.457815**	**2.174589**
Pre-NAC SBR	**1.887956**	**1.843254**	**1.998453**
Age at diagnosis	1.245687	1.189750	1.168795
Age at surgery	0.989564	1.024502	0.887562
pCR status	0.954875	**0.275846**	**0.248756**

Table 9. Hazard ratios from the subtype-stratified clinical multivariable Cox model

Beyond pCR, other clinical variables showed subtype-specific patterns. Clinical and pathological AJCC staging had a stronger prognostic impact in TNBC (HR = 1.354) compared to luminal and HER2+ disease. Post-NAC SBR grade was markedly more prognostic in HER2+ (HR = 1.458) and especially TNBC (HR = 2.175) than in luminal tumours (HR = 1.275). Interestingly, Pre-NAC SBR grade retained a consistent prognostic value across all three subtypes (HR ranging from 1.843 to 1.998), supporting the continued relevance of grade-based scores such as the Neo-Bioscore even in the post-neoadjuvant setting.

Overall, these results highlight that, even after proper stratification by subtype, classical clinical variables provide only modest discrimination. This is particularly true in luminal disease, where residual disease remains highly heterogeneous.

3.4 Discussion

In this chapter, we systematically benchmarked multiple survival-prediction strategies using post-neoadjuvant (post-NAC) surgical whole-slide images (WSIs) in early breast cancer, progressively evaluating classical, machine-learning, and deep-learning pipelines. Our findings reveal that, despite the growing methodological complexity in the literature, simpler and more transparent models often remain the most reliable and generalisable.

1. Baseline and traditional models

The baseline pipeline combining EfficientNet-B7 features, a multi-layer perceptron (MLP), and a Cox proportional-hazards (Cox-PH) model achieved moderate discrimination for both disease-free survival (DFS) and overall survival (OS). Aggregation strategy and MLP output dimension were key determinants of performance. Although absolute AUCs remained below 80 %, this framework provided a stable and interpretable reference.

2. Classical machine-learning approaches

Support-vector machines (SVM) and random forests (RF) did not surpass the Cox-PH model in DFS prediction. SVM performed best for OS but remains limited by its ranking-based formulation, lack of explicit hazard estimation, and high inference cost. These drawbacks make it unsuitable for clinical deployment.

3. Deep-learning survival models

Among deep architectures, DeepSurv underperformed across all metrics, while DeepHit approached Cox-PH results, particularly when paired with improved feature extraction. DeepHit remains a promising direction given its ability to model non-proportional hazards, though interpretability and calibration remain challenges.

4. Integration of clinical data

Incorporating clinical variables through ensembling modestly improved some metrics, especially after score normalisation (sigmoid transformation), but the image-based Cox-PH alone remained the top performer for OS. This underscores the dominant prognostic value of morphological information encoded in post-NAC tissue.

5. Self-supervised and unsupervised pre-training

Replacing ImageNet pretraining with unsupervised strategies on in-domain data enhanced DFS prediction and highlighted the promise of self-supervised ViT (DINO) and ResNet-SwAV backbones. These representations capture domain-specific structure more effectively than generic features, suggesting that future pipelines should leverage foundation models adapted to histopathology.

6. Clinical Cox-PH model

Analysis of clinical-only Cox-PH coefficients confirmed expected trends (TNBC status as the strongest prognostic factor, limited value of age, and variable effects of stage and grade) but demonstrated lower predictive power compared to WSI-based models. This reinforces the complementary, rather than substitutive, role of clinical features.

To better account for the known interaction between molecular subtype and pCR status, we additionally re-ran the multivariable clinical Cox model separately within each molecular subtype in the external validation setup (**Chapter 3.3.8**). As expected, the prognostic value of pCR was highly subtype-dependent (**Table 9**). In luminal (HR+/HER2−) tumours, achieving pCR was associated with a hazard ratio close to 1 (HR = 0.955), confirming its very limited additional prognostic information in this subtype. In contrast, pCR remained strongly protective in HER2+ (HR = 0.276) and TNBC (HR = 0.249) tumours.

Beyond pCR, other clinical variables showed clear subtype-specific patterns. Clinical and pathological AJCC staging had a stronger prognostic impact in TNBC, while Post-NAC SBR grade was markedly more prognostic in HER2+ and especially TNBC than in luminal tumours. Interestingly, Pre-NAC SBR grade retained consistent prognostic value across all three subtypes, supporting the continued relevance of grade-based scores such as the Neo-Bioscore even in the post-neoadjuvant setting.

Overall, these results highlight that, even after proper stratification by subtype, classical clinical variables provide only modest discrimination. This is particularly true in luminal disease, where residual disease remains highly heterogeneous.

7. End-to-end architectures

End-to-end survival networks integrating feature extraction and hazard prediction proved computationally demanding and prone to overfitting, offering no gain over modular pipelines.

Random patch selection improved generalisation marginally, but at substantial computational cost, indicating that optimisation and scalability remain open challenges.

8. External validation

Validation on the CGFL-Neoadj dataset confirmed that magnification consistency and subtype context critically affect generalisation. Models trained on 40× slides achieved markedly higher AUCs than those trained at 20×, demonstrating the need for harmonised scanning protocols. The ViT-based Cox-PH model remained the most robust and reproducible across folds, particularly for TNBC patients.

Overall synthesis

Across all experiments, the Cox-PH model built on foundation-model embeddings emerged as the best balance between interpretability, robustness, and clinical plausibility. While advanced deep-learning methods offered incremental improvements under controlled conditions, they frequently overfitted and failed to generalise externally.

This chapter thus defines a clear methodological lesson: in real-world survival prediction, transparent, biologically interpretable, and computationally tractable approaches outperform purely end-to-end black-box networks.

Future work should prioritise:

- integrating biologically meaningful representations (e.g., HRD, PAM50) to enhance prognostic depth,
- combining biopsy and surgical data to capture tumour evolution across treatment, and
- expanding external, multi-scanner validation to achieve regulatory-grade reproducibility.

These directions lay the foundation for Chapter 4, which extends the analysis to morpho-molecular prediction and cross-cancer generalisation.

Chapter 4: Morpho-molecular correlate analysis.

Abstract

Recent advancements in deep learning for medical image analysis have significantly improved our ability to predict molecular characteristics, such as tumour mutational status, directly from histology images[66]. In this chapter, we introduce a virtual molecular biology framework that leverages neural networks to uncover and analyze morpho-molecular correlates combining observed histological features with molecular data to infer underlying genetic mutations.

Our primary objective is to harness these morpho-molecular correlations not only for mutation prediction but also to enhance the stratification of residual disease in early breast cancer patients post-neoadjuvant chemotherapy. By applying this framework in **Chapter 5**, we aim to distinguish between high-risk patients, who are more likely to experience relapse, and those with a favorable prognosis, ultimately improving personalised treatment strategies.

Building on our foundational work[66], we will employ our mutation prediction deep learning framework to predict two critical molecular signatures in breast cancer:

1. **PAM50 Intrinsic Subtypes**, a well-established classifier that stratifies breast cancers into key molecular subtypes with distinct clinical behaviors.
2. **Homologous Recombination Deficiency (HRD) Scores**, a crucial biomarker reflecting genomic instability and predictive of response to certain therapies.

Both signatures will be predicted using training data from the TCGA BRCA cohort. By successfully applying our framework to predict these signatures, we aim to develop a model that can then be transferred and adapted to the problem of residual disease stratification in early breast cancer patients. This transfer learning strategy will allow us to leverage the rich molecular insights gained from TCGA data to improve risk stratification in the post-neoadjuvant chemotherapy setting.

But the research developed in this present chapter goes beyond predicting gene-level mutations. Building on our previously developed framework, we have designed MultiVarNet, a neural network capable of decoding complex genomic patterns and accurately predicting

specific mutation variants (e.g., p.G12C or p.G12D in non-small cell lung cancers (NSCLCs)). This enhanced precision is critical for supporting clinical applications such as the selection of targeted therapies based on precise genetic profiles.

For the first time, our method has successfully identified over 20 distinct mutation variants across key oncogenes. Beyond variant prediction, our study also emphasises the importance of integrating the molecular biology of tumours into deep learning frameworks. By merging histological data with molecular insights, our approach not only improves prediction accuracy but also introduces a new paradigm in digital pathology , one that significantly enhances the performance and clinical applicability of deep learning tools in cancer diagnostics.

We believe that this study establishes a foundation for future research, particularly in the context of residual disease stratification and the development of personalised therapeutic strategies. This chapter extends our findings from our recent work published in Scientific Reports and at the MICCAI 2024 main conference, with additional experiments and expanded explanations that further demonstrate the clinical value and versatility of our approach.

4.1 Introduction / Background work

Targeted therapies and molecularly guided treatment have transformed oncology, but they rely on accurate, timely characterisation of tumour genetics and transcriptomic programmes[125,126]. In routine practice, this information is obtained by DNA sequencing, RNA-based assays, or methylation profiling[127,128]. However, these tests remain constrained by cost, turnaround time, tissue requirements, and access inequality[58,60,129]. For instance, Gondos et al. reported that nearly a quarter of patients with newly diagnosed advanced non-small cell lung cancer (NSCLC) in the USA (Medicare data) do not receive gold-standard genomic testing for key therapeutic targets (ALK, BRAF, EGFR, and ROS1) before beginning treatment, due to these constraints. In France (**Figure 57**), the testing rate in 2024 for patients with a NSCLC is 53% for EGFR, 46% for ALK, 34% for ROS1, 38% for BRAF, 12% for MET, 45% for KRAS, 30% for HER2 where in an ideal world it should be 100%[130,131].

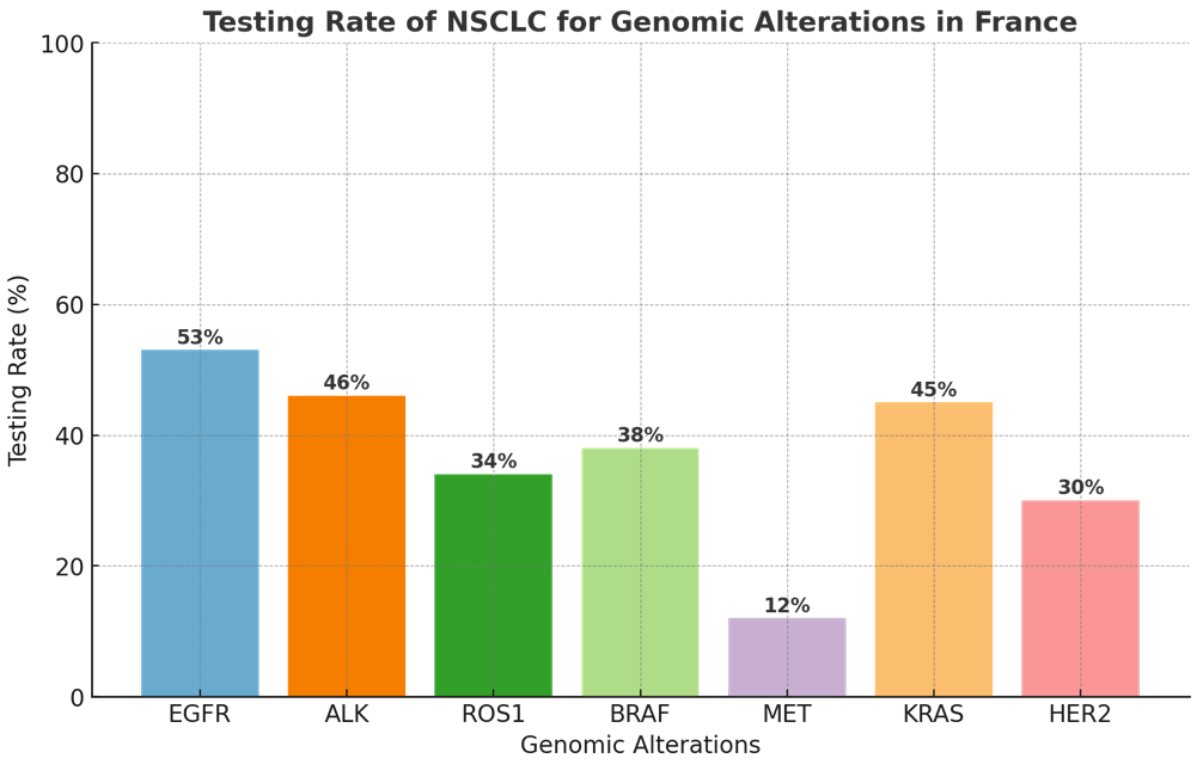


Figure 57: testing rates for key genomic alterations in non-small cell lung cancer (NSCLC) in France, as reported by De Jager et al. (2024) and Kerr et al. (2021). The bar chart highlights substantial variability in testing frequency across clinically relevant biomarkers. EGFR is the most frequently tested alteration, with a rate of 53%, followed by ALK at 46%, and KRAS at 45%. Testing rates for ROS1 (34%) and BRAF (38%) are more low, while HER2 (30%) and MET (12%) are significantly under-tested.

In parallel, there is now robust evidence that deep learning applied to routine H&E histopathology can recover part of this “missing molecular information”. Multiple groups have shown that convolutional and transformer-based networks can predict single-gene mutations, copy-number alterations, expression signatures, and prognosis directly from whole-slide images[132–134]. The working hypothesis is that genetic and transcriptional alterations leave subtle, yet reproducible, footprints on tissue architecture and cytology, **morpho-molecular correlates**, that neural networks can detect even when they are not obvious to the human eye.

In our previous work (Morel et al., 2023, Scientific Reports)[66], we used relatively simple CNNs on TCGA slides to predict **gene-level mutational status** and explored how such models might be used to triage patients for sequencing. We formalised three screening strategies (save-all, fixed-capacity, and prioritisation) and showed that a deep-learning prescreening step could substantially increase the number of mutation-positive patients identified under a fixed sequencing budget, effectively “stretching” limited testing capacity. However, that work suffered from two major limitations:

- predictions were made at the gene level only, without distinguishing clinically relevant protein variants within the same gene ;
- We treated tumours largely within a single-cancer context and did not test whether learned morpho-molecular signatures could generalise across tissues.

In practice, precision oncology is almost always variant-centred rather than gene-centred. KRAS p.G12C and KRAS p.G12D belong to the same gene but have very different therapeutic implications: p.G12C can be targeted with sotorasib[135] or adagrasib[136], whereas p.G12D currently cannot. Similar distinctions exist for BRAF (V600E versus non-V600), PIK3CA (E545K/H1047R versus other variants) and IDH1 (R132H versus alternative substitutions)[137]. Moreover, not all functionally relevant alterations are visible to DNA sequencing: epigenetic events such as BRCA1 promoter hypermethylation can produce a homologous recombination–deficient phenotype identical to that caused by a truncating mutation, yet remain invisible to standard NGS[138]. This means that the observable

morphology is shaped by a functional context (driver, passenger, co-mutations, epigenetic silencing) rather than by a single mutation in isolation.

These considerations lead to two key gaps that this chapter addresses:

1. **Protein-specific variant-level prediction and label design**. Most histology-based mutation-prediction studies treat "mutated versus wild-type" as a binary label at the gene level, ignoring the diverse functional consequences of distinct variants. This limits both clinical usefulness (no direct link to variant-specific drugs) and biological interpretability.

2. **Generalisation of morpho-molecular signatures across cancers**. It is unknown whether the morphological footprint of a given protein-specific variant (e.g. KRAS p.G12C) is conserved across tissues, or whether it is fundamentally tissue-program–specific, shaped by lineage, microenvironment, and local treatment patterns. If signatures are highly context-dependent, the field is inherently heading towards a companion-diagnostic paradigm rather than universal pan-cancer models.

In the first part of this chapter, we introduce MultiVarNet, a biologically informed, variant-aware framework that decomposes gene-level mutation labels into protein-specific variants. Instead of training a network to detect "KRAS-mutated versus wild-type", we explicitly supervise on pairs of variants (e.g. p.G12C vs p.G12D, p.G12V vs p.G13D) across 11 TCGA cancer types. We show that this label engineering, without architectural complexity, improves prediction performance and reveals finer-grained morpho-molecular signatures that align more closely with therapeutic actionability.

In the second part, we probe the transferability of these variant-level signatures across tissues. We perform pairwise cross-cancer experiments (train on cancer A, test on cancer B) and then adopt a multi-domain adversarial strategy inspired by our previous work[139], extending classical DANN to multi-tissue settings. The model is trained to predict the variant while being penalised if it can recover the tissue of origin, encouraging tissue-invariant representations when possible. This allows us to ask a principled question: *are some morpho-molecular signatures genuinely pan-cancer, or are they almost entirely tissue-bound?*

In the third part, we shift from single mutations to clinically used molecular signatures in breast cancer. Using TCGA-BRCA, we train networks to predict the PAM50[140] intrinsic subtypes and homologous recombination Deficiency[141] (HRD) score directly from histology. PAM50 (Luminal A, Luminal B, HER2-enriched, Basal-like) is the basis of the Prosigna assay and guides chemotherapy and endocrine decisions[142]. HRD captures the functional state of DNA repair and predicts sensitivity to platinum and PARP inhibitors[143–145]. We selected these two signatures because they are widely used in guidelines and because they form a natural bridge to Chapter 5, where PAM50/HRD-like "virtual molecular signatures" derived from post-NAC resections are integrated into survival models for residual disease.

Together, this chapter moves beyond gene-level mutation prediction to a variant-centred, tissue-aware, and clinically grounded view of morpho-molecular signatures. It establishes how far morphology can take us towards variant-level precision oncology, where it breaks down across tissues, and how molecular signatures learned in one setting can be reused to stratify residual disease in another.

4.2 Material and Methods

4.2.1 Dataset and population description

This study is based upon a retrospective analysis, employing de-identified scanned Whole Slide Images (WSIs) procured from TCGA. TCGA dataset spans a diverse range of cancer types from multiple centers. Comprehensive information regarding the TCGA dataset and patient-related particulars can be found in the literature[146,147]. In order to assess the robustness of our methodology across a diverse spectrum of cancer subtypes, we selected 11 distinct TCGA datasets out of the 21 cancer datasets, as given in **Table 10**. This selection was made by the availability of WSIs, DNA, protein-specific variant and transcriptomic information. This selection criterion resulted in keeping only 11 datasets. While many criticisms arise from using TCGA data[148], it is still used as an international benchmark that allows other teams to replicate and improve their results.

Dataset	Cancer type	Slides Analysed	Protein Alterations
BLCA	Bladder cancer	457	4
BRCA	Breast cancer	719	3
COAD	Colorectal cancer	459	4
GBM	Brain cancer	860	3
HNSC	Head & neck squamous cell carcinoma	472	2
LGG	brain tumor	844	2
LUAD	Lung adenocarcinoma	571	4
LUSC	Lung squamous cell carcinoma	512	2
SKCM	Skin cancer	475	5
THCA	Thyroid cancer	517	2
UCEC	Uterine cancer	566	4

Table 10. TCGA dataset list used in Chapter IV.

4.3 Results

4.3.1 Gene Mutations Status Prediction

As discussed in the introduction of **Chapter IV**, deep learning approaches have shown promise in predicting gene mutations directly from histology images. However, these methods often focus on predicting gene-level mutations without addressing the complex molecular signatures that distinguish different protein-specific variant mutations. Variants at the protein level are often more clinically relevant, particularly for guiding targeted therapies like Sotorasib for KRAS p.G12C mutations.

The MultiVarNet architecture is designed to leverage protein-specific gene variant morphological patterns. Importantly, the two-branch architecture presented here is limited to the two most recurrent protein-specific variants selected within each gene, rather than modelling all possible variants. This design choice was made to ensure adequate statistical power while focusing on the most biologically and clinically relevant alterations.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset

To develop our MultiVarNet method and test it, we used the 11 TCGA dataset described in **Chapter IV Material and Method, Table 10**. Here our aim is to predict not only genomic mutations, but also their respective variants. To do so, and to ensure adequate statistical power for quantifying significant effects, we restricted our analysis to protein variants occurring in at least 15 patients. Moreover, we did not take into account the fast frozen for slides selection, as they are often of poor quality. Consequently, this criterion yielded a final set of 20 gene mutations and 35 protein alteration/pathology pairs across all datasets, as detailed in **Table 11** and **Table 12**.

Gene Mutation	Dataset	Number of cases
PIK3CA	blca	87/386
TP53	blca	192/386
PIK3CA	brca	209/687
KRAS	coad	173/451
PIK3CA	coad	119/451
TP53	coad	229/451
EGFR	gbm	66/389
IDH1	gbm	16/389
CDKN2A	hnsc	101/450
PIK3CA	hnsc	75/450
IDH1	lgg	374/491
EGFR	luad	69/478
KRAS	luad	137/478
PIK3CA	lusc	58/478
TP53	lusc	406/478
BRAF	skcm	233/433
NRAS	skcm	115/433
PIK3CA	ucec	252/505
BRAF	thca	294/504
NRAS	thca	39/504

Table 11. Gene list selected with more than 15 mutated patients.

Variant	gene	dataset	Number of cases
p.E545K	PIK3CA	blca	27/386
p.E542K	PIK3CA	blca	16/386
p.E545K	PIK3CA	brca	41/687
p.E542K	PIK3CA	brca	27/687
p.H1047R	PIK3CA	brca	75/687
p.E545K	PIK3CA	coad	31/451
p.E542K	PIK3CA	hnsc	15/450
p.E545K	PIK3CA	lusc	15/478
p.R88Q	PIK3CA	ucec	36/505
p.E542K	PIK3CA	ucec	15/505
p.G118D	PIK3CA	ucec	9/505
p.E545K	PIK3CA	ucec	15/505
p.V600E	BRAF	skcm	196/433
p.V600M	BRAF	skcm	39/433
p.V600E	BRAF	thca	286/504
p.A289V	EGFR	gbm	10/389
p.G598V	EGFR	gbm	7/389
p.L858R	EGFR	luad	19/478
p.E746_A750del	EGFR	luad	14/478
p.Q61R	NRAS	thca	31/504
p.Q61L	NRAS	skcm	17/433
p.Q61K	NRAS	skcm	37/433
p.Q61R	NRAS	skcm	49/433
p.R280T	TP53	blca	11/386
p.R248Q	TP53	blca	15/386
p.R175H	TP53	coad	26/451
p.R158L	TP53	lusc	14/478
p.R80*	CDKN2A	hnsc	20/450
p.R132H	IDH1	gbm	13/389
p.R132C	IDH1	lgg	15/491
p.R132G	IDH1	lgg	10/491
p.G12D	KRAS	luad	18/478
p.G12C	KRAS	luad	55/478
.p.G12V	KRAS	coad	33/451
p.G12D	KRAS	coad	48/451

Table 12. protein alteration/pathology pairs across all datasets.

Proposed pipeline

Our proposed method, MultiVarNet, introduces a multi-pathway architecture designed to predict gene-level mutations and the two most recurrent protein-specific variants within each

gene. The pipeline integrates efficient feature extraction, specialised MLP pathways, and a combination of high-level features. The overall framework is illustrated in **Figure 58**.

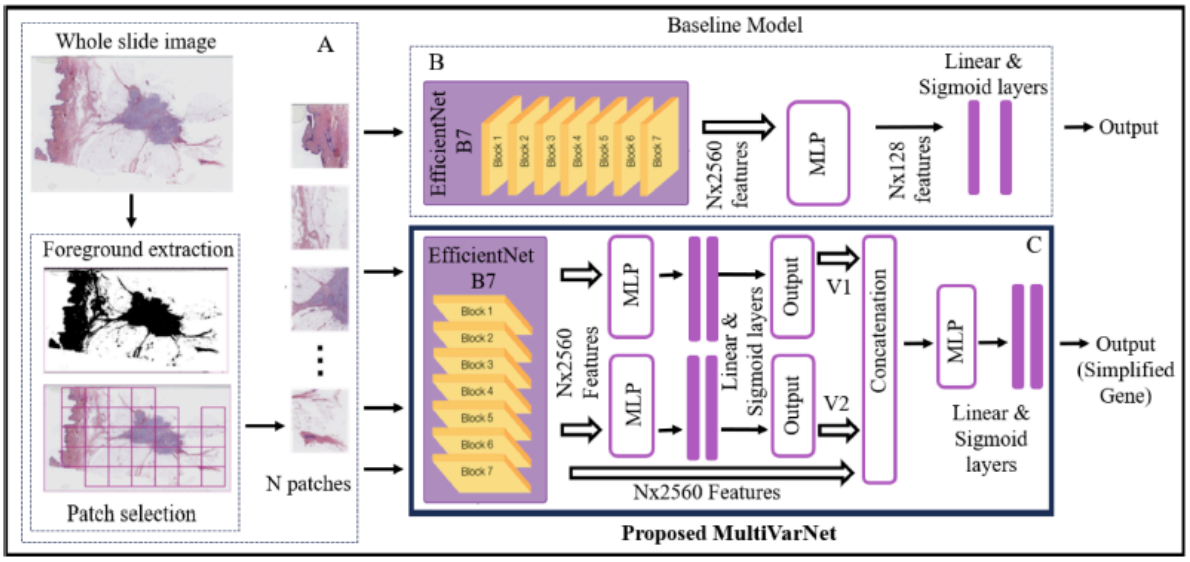


Figure 58. Architecture of the MultiVarNet framework for predicting gene mutations, protein alterations, and protein-specific variant morphological signatures. The figure illustrates the workflow and architectural differences between the baseline model (top) and the proposed MultiVarNet framework (bottom). **(A)** Whole-slide images (WSIs) are divided into non-overlapping tiles after foreground extraction to remove empty regions, generating *N* image patches per slide. **(B)** In the baseline approach, each patch is encoded using a pretrained EfficientNet-B7 backbone to extract 2560-dimensional features, which are aggregated through a multi-layer perceptron (MLP) and linear–sigmoid layers to predict simplified gene-level mutation labels (mutated vs. wild-type). **(C)** The proposed MultiVarNet extends this architecture to handle variant-aware supervision. Instead of a single binary output per gene, the model jointly learns to predict protein-specific variant sublabels (e.g., KRAS p.G12C, p.G12D, p.G12V) using parallel classification branches. Each variant head outputs an independent probability score, which is then concatenated into a shared representation and re-aggregated by an MLP to produce both variant-level and gene-level predictions.

This multi-task structure allows MultiVarNet to capture morphological cues that are unique to each functionally distinct protein alteration while preserving global gene-level context. The framework thus bridges molecular granularity (from gene to variant) with morphological representation, improving interpretability and performance in variant-specific mutation prediction across cancers.

The core concept behind the MultiVarNet method is to enhance the model's ability to predict gene mutations by leveraging protein-specific variant-specific morpho-molecular signatures. Instead of directly predicting the overall gene mutation status, where a gene is considered mutated if any of its variants are present, MultiVarNet decomposes the prediction task by individually predicting the mutational status of each variant. This strategy is grounded in the hypothesis that each variant exhibits distinct morphological patterns in histological slides, reflecting unique underlying molecular alterations. By training the network to recognize these finer, variant-specific signatures, the model can better identify the subtle histological features that correlate with specific mutations, ultimately improving its predictive accuracy and biological interpretability.

1. Data Preprocessing and Patch Extraction (Figure 58.A)

We begin by processing diagnostic slides stored in Aperio SVS files from 11 TCGA datasets (see **Table 10, Chapter IV.2.1)**, identified by the 'DX' label in their filenames. To isolate tissue regions from the background, we employ an in-house trained U-Net for foreground extraction. This step ensures that only tissue structures are analysed, improving computational efficiency and model performance.

Next, the slides are divided into non-overlapping image patches of 600 × 600 pixels at 5x magnification. This patch size and resolution balance the need for capturing detailed morphological features while maintaining a manageable computational load.

2. Baseline Model (Figure 58.B)

For baseline comparison, we employ a pretrained EfficientNet-B7 model initialised with ImageNet weights. After feature extraction through its average pooling layers, the model generates N × 2560 dimensional feature embeddings for each patch.

These embeddings are passed through a 128-dimensional MLP with dropout for regularisation, followed by linear and sigmoid layers to predict the presence of specific genetic protein-specific variants. This baseline model provides a simple yet effective framework for evaluating our proposed MultiVarNet architecture and has the best generalisation results.

3. Proposed MultiVarNet Architecture (Figure 58.C)

The MultiVarNet architecture expands on the baseline model by introducing a multi-stream structure.

Image patches are processed through a pretrained EfficientNet-B7 network, generating N × 2560 feature embeddings. These embeddings are divided and directed into two distinct MLP branches, each designed to extract high-level features for variant prediction. Each branch generates unique linear and sigmoid outputs for predicting distinct variant mutations (V1 and V2). This enables the model to differentiate fine morphological patterns unique to specific variants.

Then the outputs from both V1 and V2 branches are concatenated with the extracted feature vector generated for the patch by the pretrained EfficientNet-B7 network, generating N × 2562 feature embeddings.

This combined representation is then processed by a third MLP with a 128-dimensional hidden layer followed by linear and sigmoid layers to predict simplified gene mutations.

Label generation

To guide our analysis and improve mutation prediction accuracy, we generated three distinct types of labels: genetic mutation labels, protein variant labels, and simplified gene labels. Each label type serves a specific purpose in capturing molecular complexity and supporting the training of our MultiVarNet architecture.

1. Genetic Mutation Labels

Genetic mutation labels were derived from the Mutation Annotation Format (MAF) file, which was processed using the MuTect2 algorithm[125]. Key features extracted from the MAF file included:

- ‘IMPACT’ : An indicator of pathogenicity, where mutations categorized as HIGH or MODERATE were selected, as they suggest significant functional changes with potential clinical relevance.
- ‘tumour_Sample_Barcode’ : Identifies individual tumour samples, ensuring each mutation is correctly assigned to its corresponding patient.

- ‘Hugo_Symbol’ : Specifies the gene symbols, enabling gene-level mutation tracking.
- ‘HGVSp_Short’ : Describes somatic mutations at the protein level, capturing amino acid changes that can alter protein function.

To generate the genetic mutation labels, we encoded the presence of pathogenic mutations as '1' and their absence as '0'. This binary labeling system ensures a clear distinction between mutated and non-mutated samples.

2. Protein Variant Labels

For a more granular analysis, we constructed protein variant labels by combining gene symbols with detailed mutation information extracted from the ‘HGVSp_Short’ field. This approach allowed us to track mutations at the variant-specific level, focusing on individual amino acid changes that may have distinct morphological and functional impacts. By associating these mutations directly with protein structure alterations, the network is trained to capture the unique histological features linked to each variant.

This level of detail is crucial, as different variants within the same gene can produce distinct molecular and clinical outcomes, often with differing prognostic and therapeutic implications.

3. Simplified Gene Labels

To further refine the analysis, we introduced the concept of simplified gene labels. This strategy was designed to reduce complexity while preserving the most informative variant data. Each simplified gene label is constructed using only the two most prevalent variants within a given gene.

A simplified gene is labeled as mutated if either of these two variants is detected. This approach simplifies the learning task for the network, ensuring the model focuses on the most biologically relevant and statistically significant alterations while minimizing noise from rare or less impactful variants.

By focusing on two dominant variants per gene, we aimed to determine whether improved prediction performance arises from the model architecture itself and not from the reduction in complexity by excluding less common variants. This strategy enhances the interpretability of the model’s predictions while maintaining robust mutation detection.

RESULTS

To predict gene mutations and precise protein variants from WSIs, we used the established baseline deep learning setup (**Figure 58.B)**. We derived slide-level predictions by aggregating tile-level predictions, focusing on the 99th percentile of these values as an indicator of mutations. We categorised clinically actionable genes (KRAS, EGFR, IDH1, BRAF, NRAS, presented in **Table 13**) separately from others relevant but yet less clinically critical (PIK3CA, TP53, CDKN2A in **Table 14**) to guide the reader to the most relevant results. The model successfully predicted 15 out of 20 gene mutations Among these mutations, the IDH1 gene mutation showed robust discrimination in GBM (AUC=81.17%, P=1.07E-12) and LGG (AUC=80.04%, P=1.72E-34) **(see Table 13)**. Similarly, for the NRAS gene mutation, the baseline model achieved an AUC of 56.72% (P=0.026) and an AUC of 77.58% (P=1.00E-08) for THCA for SKCM datasets. However, the model exhibited reduced discriminative capacity for TP53 in LUSC (AUC=57.25%, P=0.055), as detailed **in Table 14**. This variability underscores the intricate nature of morphological signatures associated with mutations, emphasising the imperative for our ongoing research to refine and enhance predictive methodologies.

Protein-specific gene variants are predictable using histology

Separating gene mutation by their protein-specific variant drastically reduces the proportion of positive samples in the datasets and therefore can only be done for few genes and few variants. In our datasets, some specific variant mutations were distinctly identifiable and achieved higher AUC scores compared to the overall gene mutations AUC scores. For instance, BRAF p.V600E (AUC=87.76%,P=6.54E-49) in THCA, as observed in **Table 13,** exhibited enhanced discernibility compared to the gene's overall mutation (AUC=82.81%, P=5.36E-37). However, not all variant mutations showed this pattern. Some, like p.Q61R for NRAS in THCA (AUC=70.88%, P=9.60E-05) or p.V600M for BRAF in SKCM (AUC= 54.48%, P=0.34), were less predictable than their overall gene mutation counterparts. Interestingly, the predictability of the same protein variant, such as PIK3CA p.E545K (refer to **Table 14**), varied significantly across different types of cancer, suggesting distinct morphological signatures specific to each variant.

Table 13 : Genes and protein-specific gene variants prediction results with the baseline prediction framework for current clinically actionable genes.

dataset	Gene	Variant	Mean AUC (%)	Mean P value
COAD	KRAS	All	66.31	4.26E-09
		p.G12V	65.29	0.003
		p.G12D	58.98	0.034
GBM	EGFR	All	65.00	4.02E-09
		p.A289V	67.33	0.0055
		p.G598V	65.08	0.0386
	IDH1	All	81.17	1.07E-12
		p.R132H	73.05	9.47E-06
LGG	IDH1	All	80.04	1.72E-34
		p.R132C	74.33	8.34E-06
		p.R132G	86.29	1.50E-06
LUAD	EGFR	All	62.823	0.0002
		p.L858R	75.67	1.57E-06
		p.E746_A750del	70.58	0.005
	KRAS	All	62.07	1.76E-05
		p.G12D	63.53	0.045
		p.G12C	55.12	0.2025
SKCM	BRAF	All	58.77	0.001
		p.V600E	64.64	3.67E-08
		p.V600M	54.48	0.3431
	NRAS	All	56.72	0.0258
		p.Q61L	58.62	0.2276
		p.Q61K	70.76	3.67E-06
		p.Q61R	63.46	0.0018
THCA	BRAF	All	82.81	5.36E-37
		p.V600E	87.76	6.54E-49
	NRAS	All	77.58	1.00E-08
		p.Q61R	70.88	9.60E-05

dataset	Gene	Variant	Mean AUC (%)	Mean P value
BLCA	PIK3CA	All	64.61	4.39E-06
		p.E545K	70.95	1.50E-05
		p.E542K	49.31	0.9255
	TP53	All	70.33	6.20E-14
		p.R280T	76.04	0.0002
		p.R248Q	55.29	0.486
BRCA	PIK3CA	All	55.28	0.0236
		p.E545K	68.63	5.01E-05
		p.E542K	61.83	0.0337
		p.H1047R	50.04	0.9913
COAD	PIK3CA	All	57.59	0.0129
		p.E545K	63.34	0.0096
	TP53	All	67.67	5.77E-11
		p.R175H	59.36	0.1088
HNSC	CDKN2A	All	53.89	0.22153
		p.R80*	72.60	0.0003
HNSC	PIK3CA	All	45.97	0.2612
		p.E542K	40.47	0.1953
LUSC	PIK3CA	All	54.21	0.2488
		p.E545K	52.42	0.721
	TP53	All	57.25	0.0554
		p.R158L	58.34	0.2291
UCEC	PIK3CA	All	57.97	0.001
		p.R88Q	60.45	0.0294
		p.E542K	72.29	0.00095
		p.G118D	67.15	0.0234
		p.E545K	50.36	0.963

Table 14 : Genes and protein-specific gene variants prediction results with the baseline prediction framework for relevant genes but not yet targeted in routine practice.

Tissue type is a factor of predictability

We thus investigated in more depth the variability of the results (a.k.a predictability or difference of AUC) across genes and cancers types which could help us find performance improvements. To do so, we analysed the relation between the prediction performances and other variables such as the number of slides or the proportion of mutated patients in the sample. We analysed the AUC correlation with the proportion of mutated samples or number of samples. The AUC was positively correlated with the number of samples (r = 0.24, p-value = 0.0020, Pearson R test) and the proportion of mutated samples (r = 0.31, p-value = 8.6e-05, Pearson R test). Using a naive linear model, we found a coefficient of 0.134 for the proportion of mutated samples which means that an increase of 10% prevalence gives an increase of 1.34% in AUC. We also found a coefficient of 0.0075 for the number of samples which means that adding 132 samples was associated with a 1% increase in AUC. These results suggest

that the enrichment of the mutated population and increasing the number of slides directly improves the AUC for this specific task.

Next we explored whether some tissues were associated with higher predictability than others. We first started by plotting the mutation predictability distribution **(Figure 59)** for all the genes available in the datasets (see **section IV.2.1**) without excluding any alteration, keeping even the low prevalence ones. **Figure 59** shows that the majority of the DNA mutations can be predicted with an AUC over 55% across all datasets. This 55% AUC is usually the significant threshold for a p-value < 0.05.

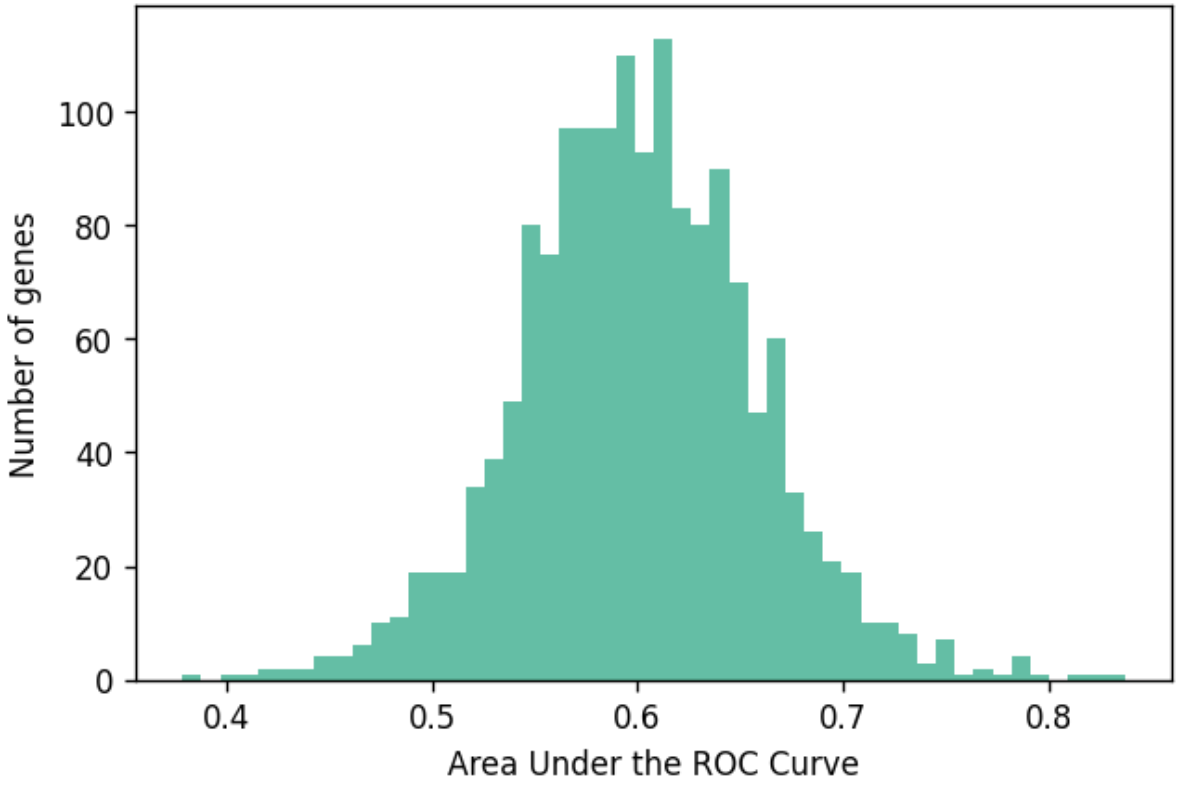


Figure 59 : Distribution of the genes' predictability using the baseline prediction pipeline across 11 TCGA cancer datasets.

We then analysed the scores for each datasets. **Figure 60** displays the mean predictability for LUSC, BLCA, LUAD, HNSC, SKCM, COAD, LGG, GMB, UCEC and THCA datasets. The first thing to notice is that different organs display statistically different predictability, as for uterine cancer (UCEC) or thyroid cancer (THCA) and lung cancer, either LUSC or LUAD. More interestingly, within the same organ, predictability can vary significantly. Literature shows that lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC) have different causes and molecular profiles[149–151], and here it is interesting to note that they also

share different predictability. We observed that gene mutational status was significantly more predictable in uterine corpus endometrial carcinoma (UCEC) and thyroid carcinoma (THCA) compared to other tumour types. In particular, UCEC and THCA consistently achieved the highest prediction AUCs using our deep learning pipeline, whereas lung squamous cell carcinoma (LUSC) presented the lowest performance, with a difference of approximately 7 percentage points in AUC compared to THCA.

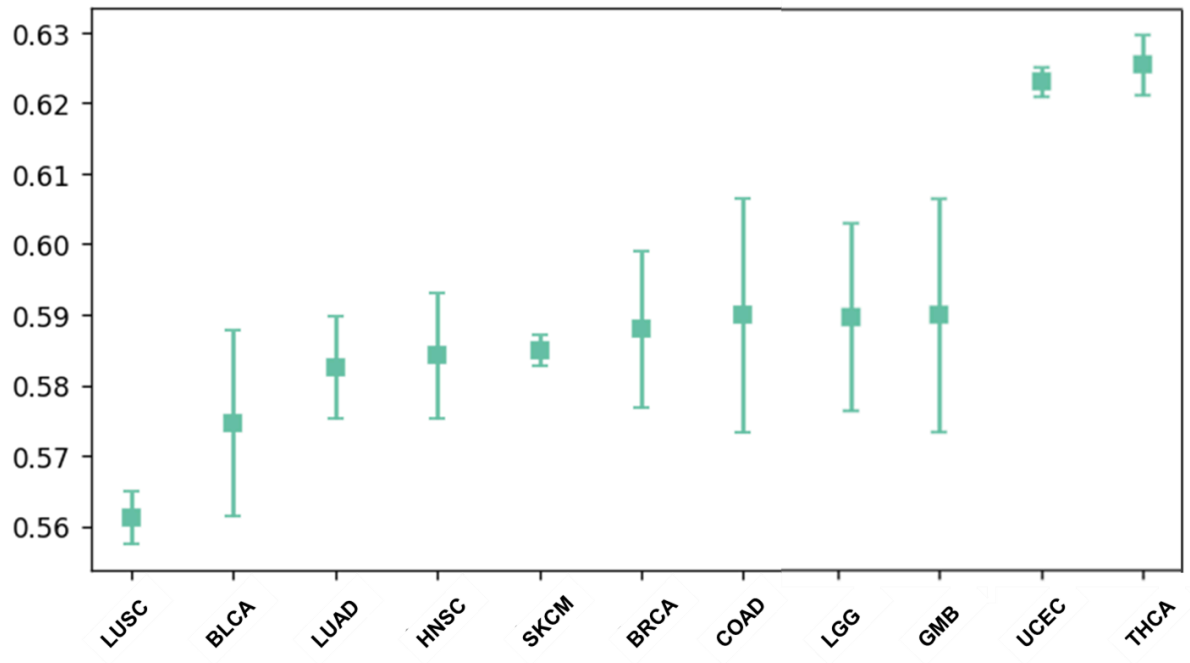


Figure 60 : Comparison of the morpho-molecular correlates' predictability across 11 TCGA cancer datasets.

This finding aligns with recent studies showing that UCEC and THCA display stronger morphological correlates of genomic alterations. For UCEC, the presence of well-defined molecular subtypes, such as POLE-ultramutated, MSI-high, and copy-number low/high, which correlates with distinct histopathological features, that could make them easier to classify and predict using histological images[152,153]. Similarly, in THCA, mutations such as BRAF p.V600E are associated with highly recognisable morphological features, such as papillary structures and characteristic nuclear changes, which can be reliably detected by deep learning models[154]. In contrast, LUSC exhibits a more homogeneous histological appearance, with fewer visual correlates of specific mutations. As shown by Coudray et al. (2018)[155], mutation prediction in LUSC from histopathology remains challenging, likely due to the lack of a strong relationship between gene alterations and morphologic presentation. This supports our empirical observation that LUSC mutation prediction is significantly less accurate than in

UCEC or THCA. Other than this literature, no clear clinical or histological rationale has yet been established to fully explain why UCEC and THCA exhibit such high predictability, especially when compared to other cancers with similarly high mutation burdens The pathologists interviewed following these observations confirmed there was no trivial explanation for this. Our findings suggest that these tumours may possess subtle but reproducible morpho-molecular signatures that remain under-characterised in the existing literature.

Genes predictability is not consistent across tissues

We next investigated whether there was a gene effect on the predictability across cancer tissues. Assuming that predictability is a consequence of the pattern's discriminability, a consistency in the predictability (i.e. easiness of prediction) for a specific gene across different cancer types and organs could suggest that these genes could share similar patterns between the different tissues. In other words, if some genes are more predictable than others consistently across tissues, this would suggest that there might be something in common in their morphological signature that would explain this effect.

To test if genes had intrinsic predictability that was consistent over multiple tissues, we ran a 1-way ANOVA on the genes occurring at least in five datasets, which led to 55 genes without any selection criterion (we kept even the genes with a very low prevalence). We found no statistical difference between mean gene predictabilities (statistic=1.10, p-value=0.300), even after correction for the dataset of origin (statistic = 1.15, p-value = 0.230). As an example, BRAF is highly predictable in the THCA dataset (AUC = 0.83) but modestly detectable in SKCM (AUC = 0.59) as shown in **Table 13**.

Protein-specific gene variants predictability is not consistent across tissues

To investigate whether protein-specific gene variant morphological signatures are consistent across tissues, we analysed the predictive performance of several recurrent protein mutations across multiple cancer types. The analysis reveals that the same variant can have highly variable AUCs depending on the tissue of origin, suggesting that morphological features

associated with certain mutations are not uniformly expressed across tumour types. For this analysis, we selected within our protein-specific variant list variants observed in at least two cancer types (**Table 15**).

Gene	Variant	Dataset	Mean AUC (%)	Mean P value	Std Dev (AUC)
KRAS	p.G12D	COAD	58.98	4.26E-09	3.22
		LUAD	63.53	0.045	
BRAF	p.V600E	SKCM	64.64	3.68E-08	16.33
		THCA	87.76	6.54E-49	
NRAS	p.Q61R	SKCM	63.46	0.0018	5.24
		THCA	70.88	9.60E-05	
PIK3CA	p.E545K	BLCA	70.95	1.50E-05	8.36
		BRCA	68.63	5.01E-05	
		COAD	63.34	0.0096	
		LUSC	52.42	0.721	
		UCEC	50.36	0.963	
	p.E542K	BLCA	49.31	0.9255	12.93
		BRCA	61.83	0.0337	
		HNSC	40.47	0.1953	
		UCEC	72.29	0.00095	

Table 15. Predictive performance (AUC and p-value) of selected protein variants across different tissues.

For example, the BRAF p.V600E variant shows a high AUC of 87.76% in thyroid cancer (THCA) but only 64.64% in skin melanoma (SKCM), despite being a well-known driver mutation in both cancers. Similarly, PIK3CA p.E545K, one of the most frequent alterations, shows strong predictability in bladder cancer (BLCA, AUC = 70.95%) and breast cancer (BRCA, AUC = 68.63%), but drops considerably in lung squamous carcinoma (LUSC, AUC = 52.42%) and uterine cancer (UCEC, AUC = 50.36%). Even more striking, PIK3CA p.E542K performs poorly in head and neck squamous cell carcinoma (HNSC, AUC = 40.47%), yet reaches 72.29% in UCEC, showing a 30-point difference across tissues. This variability suggests that the morphological manifestation of these alterations differs significantly depending on tissue context.

To quantify this inconsistency, we measured the variance in AUC across tissues for each protein-specific gene variant. Protein-specific variants like PIK3CA p.E545K and p.E542K exhibited high variance (standard deviation ≈ 8–12 points), reflecting heterogeneous morphological expression. In contrast, variants with more stable AUCs across tissues (e.g.,

KRAS p.G12D: 58.98% in COAD, 63.53% in LUAD) showed lower variance, but their overall predictability remained modest.

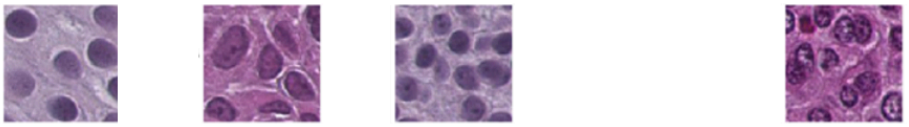
These findings support the hypothesis that protein-specific variant predictability is not an intrinsic property of the mutation alone but is highly dependent on tissue-specific morphological context. This could help explain why gene-level prediction models often underperform or appear inconsistent: if variant composition varies, so too will the morphological features they induce, making gene-level signatures more diffuse. This highlights the importance of tissue-aware variant modeling and justifies the need for approaches that can explicitly account for variant-specific morphology in context.

Protein-specific gene variants exhibit specific, gene-independent morphological signatures

To better understand the structure of the morphological signals learned by our models, we asked the following question: do protein-specific gene variants from the same gene share a common morphological pattern, or is the morphological signature of each variant mostly independent, even within the same gene and cancer type?

From a clinical and biological perspective, this question is important. In practice, gene mutations are currently always modeled as a single label in predictive algorithms, ignoring the internal heterogeneity of their protein-specific variants (as our work was the first one to predict variants). However, if different variants of the same gene have distinct morphological patterns, this assumption may weaken performance and interpretability. Conversely, if variants from a given gene consistently share similar features, then aggregating them into a gene-level label may be justified.

To address this, we define the *morphological signature* of a variant as the vector of predicted values across all test patches, obtained from a model trained specifically to detect that variant. This signature captures the spatial expression of the variant as learned by the model and serves as a compact representation of its morphological footprint (**Figure 61**).



	Tile 1 Case 1	Tile 2 Case 1	Tile 3 Case 1	...	Tile N Case M
PIK3CA	0.61	0.12	0.89	...	0.67
TP53	0.14	0.09	0.34	...	0.64
KRAS	0.42	0.23	0.27	...	0.91
...	...	...	...	...	...
IDH1	0.55	0.61	0.07	...	0.87

Figure 61 : concept of a morphological signature for each protein-specific variant. For each protein-specific variant, a dedicated variant-specific model is trained. The morphological signature of a given variant is defined as the vector of prediction scores generated by this dedicated model across all individual image tiles from the test set. Each row of the resulting matrix corresponds to one protein-specific variant (e.g., PIK3CA p.E545K, TP53 p.R175H, KRAS p.G12D), and each column represents the prediction score assigned to a particular image tile. This matrix encodes the distribution and strength of the morphological signal learned specifically for that variant across all cases and tiles. Pairwise correlations between these signatures can then be computed to investigate relationships between variants, genes, and tissue types.

We then computed the pairwise correlations between these morphological signatures within each dataset. In doing so, we compared the similarity across each datasets between:

- intra-gene pairs: protein-specific variants belonging to the same gene, and
- inter-gene pairs: protein-specific variants from different genes.

If a gene has a "true" morphological identity, we would expect its variants to be more correlated with each other (i.e., higher intra-gene correlations) than with variants from unrelated genes (lower inter-gene correlations). This idea is illustrated conceptually in **Figure 62.**

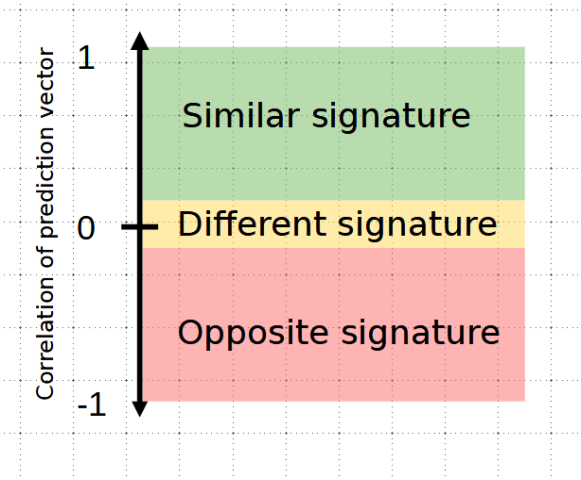


Figure 62 : Statistical definition of morphological signature using Deep Learning models. A higher intra-gene correlation would indicate that the gene has a consistent morphological identity across its variants.

By comparing these two types of correlations across datasets, we aim to understand whether gene-level labels are morphologically meaningful, or whether each variant should be treated as an independent unit of analysis. **Figure 63** displays the distributions of these correlation values across cancer types.

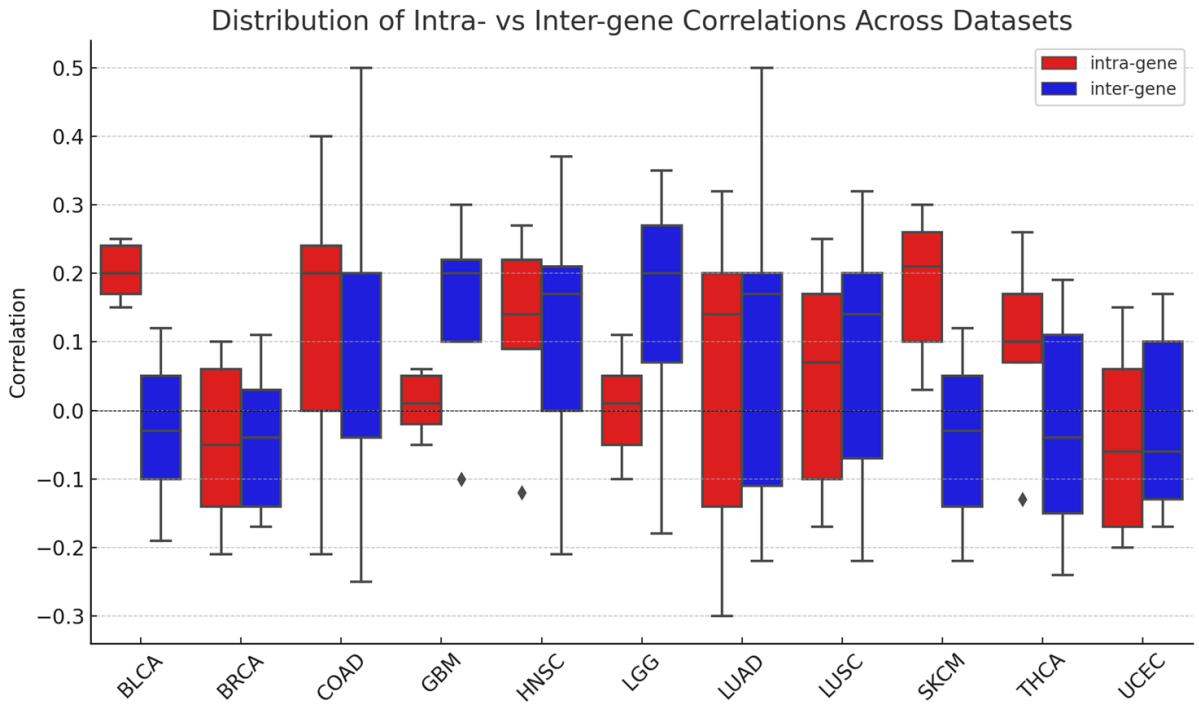


Figure 63 : Correlation between morphological signatures of protein-specific variants within and across genes (*intra-gene* in red, *inter-gene* in blue). In most datasets, correlations are similar or higher across genes than within genes, suggesting the absence of a strong gene-level morphological structure.

Interestingly, we found that *intra-gene correlations are not consistently higher* than inter-gene correlations. In many datasets such as COAD, GBM, UCEC, and THCA, the distributions of inter-gene correlations were similar to or even exceeded those of intra-gene pairs. This suggests that variant-level morphological signals are not necessarily grouped by gene identity. The only dataset showing a statistically significant difference was BLCA, where intra-gene correlations (mean = 0.188) were significantly higher than inter-gene ones (mean = -0.02), with a p-value of 9.03e-05 (Student's t-test). Notably, this dataset only includes two major genes (TP53 and PIK3CA), which might reduce heterogeneity and explain the observed signal. However, when we isolated only TP53 and PIK3CA in other datasets (e.g., COAD or UCEC), this intra-gene advantage did not replicate, further supporting that protein-specific variant predictability is primarily *variant-specific* rather than gene-dependent.

These results support the idea that the morphological signal learned by deep learning models is more closely tied to individual protein-specific variant-level alterations, rather than broader gene identity. This has important implications: it suggests that modeling gene mutations as a single label might dilute signals when variants behave differently, and that variant-specific modeling may offer a more powerful route to improving gene-level mutation prediction.

MultiVarNet: Leveraging protein-specific variant specific signature to enhance predictive performance

The results from the previous section highlight a key insight: protein-specific variant prediction from histology is not only possible but often more robust than prediction of overall gene mutation status. In many cases, individual variants within a gene exhibited stronger and more consistent morphological signals than the gene-level label, especially when evaluated across multiple tissue types. These findings suggest that variant-specific signatures are a meaningful biological and computational target for modeling, and that aggregating all proteic variants under a single gene label may dilute these learnable signals.

To further investigate this, we conducted an analysis comparing the AUC of individual protein-specific variants to that of their combined gene-level mutation status. Specifically, we assessed whether two protein-specific variants of a gene were each more or less predictable than their logical OR combination (a proxy for the overall gene mutation status). In many cases, both proteic variants were either easier or harder to predict than the gene-level label (**Table 13, Table 14**). This raised a fundamental question: if individual proteic variants provide stronger morphological cues than their aggregated gene mutation label, could modeling these protein-specific variants independently lead to better gene-level predictions?

This observation forms the basis of our MultiVarNet hypothesis: by leveraging the distinct morphological features associated with individual gene variants (and training a model to recognize these variant-specific patterns), we can build a more effective representation of the gene's mutational landscape. In other words, training on variant-level signals may enhance gene-level mutation prediction by better capturing the heterogeneity of morpho-molecular expression.

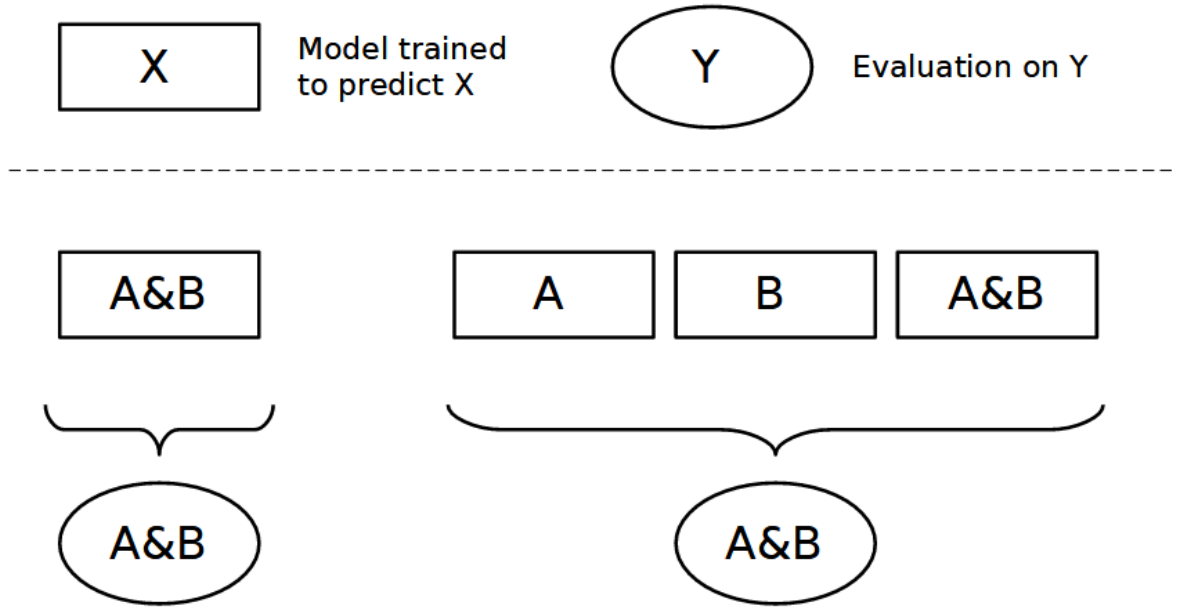


Figure 64. Label engineering principle for protein-specific variant- and gene-level ensemble prediction. This schematic illustrates how MultiVarNet integrates models trained on different supervision targets, from fine-grained protein-specific variant-level labels to broader gene-level labels, to form a unified ensemble for mutational status prediction. In this setup, individual models are trained to predict either variant-specific alterations (e.g., KRAS p.G12C, p.G12D), other functionally related proteic variants, or (A&B) their joint occurrence

within the same gene. During inference, outputs from these separate models are combined into an aggregated prediction for the higher-level target (gene mutation status).

This "label engineering" strategy enables cross-resolution learning, where models specialised on protein-specific variant morphology inform and regularise broader gene-level predictions, improving robustness and interpretability across molecular contexts.

To test this hypothesis, we introduce MultiVarNet, a proof-of-concept architecture that integrates predictions from multiple proteic variant-specific streams to enhance overall gene mutation prediction. As illustrated in **Figure 64** and extensively described in the **METHODOLOGY section**, MultiVarNet combines protein-specific variant-level and gene-level predictions through a label engineering strategy that ensembles models trained on distinct but related targets (individual protein variants and their parent gene). This approach contrasts with standard weakly supervised frameworks in the literature, which typically rely on a single binary label for gene mutation status. By explicitly incorporating proteic variant heterogeneity, MultiVarNet aims to align the learning process more closely with the biological structure of mutation expression.

We benchmark MultiVarNet against conventional architectures using two widely accepted aggregation strategies for whole-slide inference: mean pooling and the 99th percentile of tile predictions, both of which have been shown to be effective in recent molecular prediction studies[81]. As shown by our results displayed in **Tables 16 and 17**, MultiVarNet consistently outperformed the baseline models across various datasets and cancer types, revealing a significant improvement in predictive accuracy as measured by the Area Under the Curve (AUC).

Dataset	Gene Mutation	Number of cases	Proteins Pair	Baseline Model [14] mean AUC (%)/ Mean P value	MultiVarNet mean AUC (%)/ Mean P value
BLCA	PIK3CA	43/386	p.E545K, p.E542K	69.14/4.10E-06	**71.47/2.39E-07**
BRCA	PIK3CA	68/687	p.E545K, p.E542K	56.872/0.0587	**57.46/0.0403**
COAD	KRAS	81/451	p.G12D, p.G12V	66.17/3.99E-06	**67.21/9.10E-07**
GBM	EGFR	17/389	p.A289V, p.G598V	60.08/0.0355	**61.20/0.0195**
LGG	IDH1	25/491	p.R132C, p.R132G	84.10/2.47E-14	**84.48/1.28E-14**
LUAD	KRAS	72/478	p.G12C, p.G12D	57.67/0.0319	**60.30/0.004**
SKCM	NRAS	83/433	p.Q61K, p.Q61R	62.72/0.0001	**64.52/1.39E-05**
UCEC	PIK3CA	24/505	p.E542K, p.G118D	**65.11/0.0031**	64.76/0.0039

Table 16: Performance of baseline model and MultiVarNet for simplified gene mutations (using the 99th percentile of their tile prediction values).

Dataset	Gene Mutation	Number of cases	Proteins Pair	Baseline Model [9] mean AUC (%)/ Mean P value	MultiVarNet mean AUC (%)/ Mean P value
BLCA	PIK3CA	43/386	p.E545K, p.E542K	71.94/1.30E-07	**72.09/1.05E-07**
BRCA	PIK3CA	68/687	p.E545K, p.E542K	59.24/0.0110	**59.58/0.0084**
COAD	KRAS	81/451	p.G12D, p.G12V	65.75/7.05E-06	**66.21/3.75E-06**
GBM	EGFR	17/389	p.A289V, p.G598V	68.61/0.0001	**69.29/5.72E-05**
LGG	IDH1	25/491	p.R132C, p.R132G	85.62/1.68E-15	**86.59/2.84E-16**
LUAD	KRAS	72/478	p.G12C, p.G12D	**61.45/0.0013**	60.94/0.0022
SKCM	NRAS	83/433	p.Q61K, p.Q61R	62.99/0.0001	**63.15/8.34E-05**
UCEC	PIK3CA	24/505	p.E542K, p.G118D	69.96/9.41E-05	**70.48/6.17E-05**

Table 17: Performance of baseline model and MultiVarNet for simplified gene mutations (using the mean of their tile prediction values).

For example, in the bladder cancer (BLCA) dataset, MultiVarNet using 99th percentile aggregator achieved a higher mean AUC of 71.47% (P=2.39E-07) for the PIK3CA simplified gene mutation compared to the baseline's 69.14% (P=4.10E-06) as given in **Table 16**. Enhanced performance was observed across multiple datasets, including COAD, SKCM, and LUAD, underscoring the relevance of our approach in capturing the nuanced morphological features associated with these mutations. However, in the UCEC dataset for PIK3CA simplified gene mutation, the baseline model marginally outperformed MultiVarNet with the mean AUC of 65.12% (P=0.0031) compared to MultiVarNet's mean AUC of 64.75% (P=0.0039).

Consistent with the 99th percentile aggregator, the MultiVarNet also outperformed the baseline model using the mean aggregation (**Table 17**) for simplified genes defined in various cancer subtypes. For instance, in BLCA (72.09%, P=1.05E-07) compared to 71.94%, P=1.30E-07), BRCA (59.58%, P=0.0084 vs 59.24%, P=0.0110), COAD (66.21%, P=3.75E-06 vs 65.75%, P=7.05E-06), and LGG (86.59%, P= 2.84E-16 vs 85.62%, P=1.68E-15), respectively.

To quantitatively assess whether MultiVarNet provided a statistically significant improvement over the baseline mutation-prediction model, we performed a paired t-test across all evaluated gene–dataset pairs (n = 16, combining both 99th percentile and mean aggregation strategies).

MultiVarNet achieved a higher mean AUC (69.0 ± 9.3%) compared to the baseline model (68.1 ± 9.2%), corresponding to an average gain of +0.86 AUC points. The difference was statistically significant (paired $t = 2.44$, $p = 0.027$), a finding confirmed by a non-parametric Wilcoxon signed-rank test ($p = 0.031$). These results indicate that the performance improvements observed with MultiVarNet are moderate but consistent across datasets and aggregation strategies. The gain arises primarily from better protein-specific variant-level signal integration rather than architectural complexity, suggesting that biologically structured supervision contributes measurable but moderate enhancements to mutation-prediction accuracy.

In summary, the MultiVarNet approach demonstrates that explicitly modeling protein-specific variant-level morphological signals not only aligns more closely with the biological complexity of tumour genomics but also yields tangible improvements in gene-level mutation prediction across multiple cancer types. By leveraging distinct protein-specific variant signatures and incorporating them through a label engineering framework, MultiVarNet effectively captures heterogeneity that standard models tend to overlook. These findings validate our initial hypothesis and position MultiVarNet as a biologically informed alternative to traditional weakly supervised methods, opening new avenues for enhancing molecular prediction in computational pathology.

4.3.2 Pan cancer gene mutation status prediction

In **Section IV.3.1**, we demonstrated that morphological prediction of gene mutations is highly dependent on both the tissue of origin and the protein-specific variant composition of each gene. We observed that even identical protein variants, such as PIK3CA p.E545K, could be strongly predictable in one cancer (e.g. BLCA, AUC = 70.9%) yet nearly random in another (e.g. LUSC, AUC ≈ 52%), suggesting that the visual correlates of mutations are modulated by tissue context. This raised a central question: can deep learning models trained to recognise mutation-associated morphology in one cancer generalise to another? In other words, do certain driver mutations exhibit transferable, pan-cancer morphological signatures, or are these signals fundamentally tissue-specific?

To further address this question, we designed a pairwise cross-cancer prediction framework, focusing on genes occurring in at least 15 mutated patients across two or more TCGA cancer types. For each eligible gene, we trained a model on one cancer (source domain) and tested it on another (target domain), then repeated the process in reverse to evaluate transfer symmetry. This design isolates the capacity of models to generalise across histogenetically distinct but genomically comparable tissues. It also provides an interpretable metric of "morphological portability" for individual mutations. Representative results are reported in **Table 19**.

Beyond this pairwise setting, two genes (PIK3CA and TP53) were sufficiently prevalent across multiple cancers (≥3 datasets) to enable multi-domain training. For these, we implemented a Domain-Adversarial Network (DANN) approach to test whether invariant, mutation-informative representations can be learned despite inter-tissue variability. DANNs have been successfully applied in computational pathology to reduce site and staining bias by penalising features predictive of the data source [156,157]. Our architecture extends this principle with two output heads: one optimised for mutation prediction and a second adversarial head predicting cancer type through a gradient reversal layer. By maximising performance on the mutation branch while minimising tissue classification accuracy, the network is encouraged to encode features orthogonal to organ identity.

Together, this section aims to quantify the degree and directionality of morphological transferability across cancers and to evaluate whether adversarial strategies can mitigate tissue dependence in histology-based mutation prediction. These analyses complement **Section**

IV.3.1 by moving from protein-specific variant-aware intra-tissue modelling to cross-tissue generalisation, ultimately probing the limits of pan-cancer morphological learning.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset

We used the 11 TCGA cohorts described in **Section IV.2.1**. For the cross-cancer experiments, we selected genes occurring in at least 15 mutated patients in two or more cancer types. For each eligible gene, we constructed *pairwise tasks* in which a model was trained on one cancer (source) and tested on another (target), and vice versa. This ensured balanced transfer evaluation while preserving independence between training and testing slides.

For the Domain Adversarial training part, we only kept PIK3CA and TP53 as the requirement is to be represented n>15 in at least 3 cancer types.

All slides were processed through the same pipeline as in **Section IV.3.1**, including tissue segmentation with a U-Net, 600×600 pixel non-overlapping tile extraction at 5× magnification, and feature encoding with EfficientNet-B7 pretrained on ImageNet.

Domain Adversarial Training

For genes detected in three or more cancer types (*PIK3CA* and *TP53*), we implemented a Domain-Adversarial Neural Network (DANN) architecture following [157,158]. The network comprises a shared feature extractor feeding two output heads (**Figure 65**):
(1) a mutation-prediction branch minimising binary cross-entropy for the mutation status, and (2) a cancer-type branch connected through a gradient reversal layer (GRL) penalising the network when tissue identity can be inferred from its latent representation.

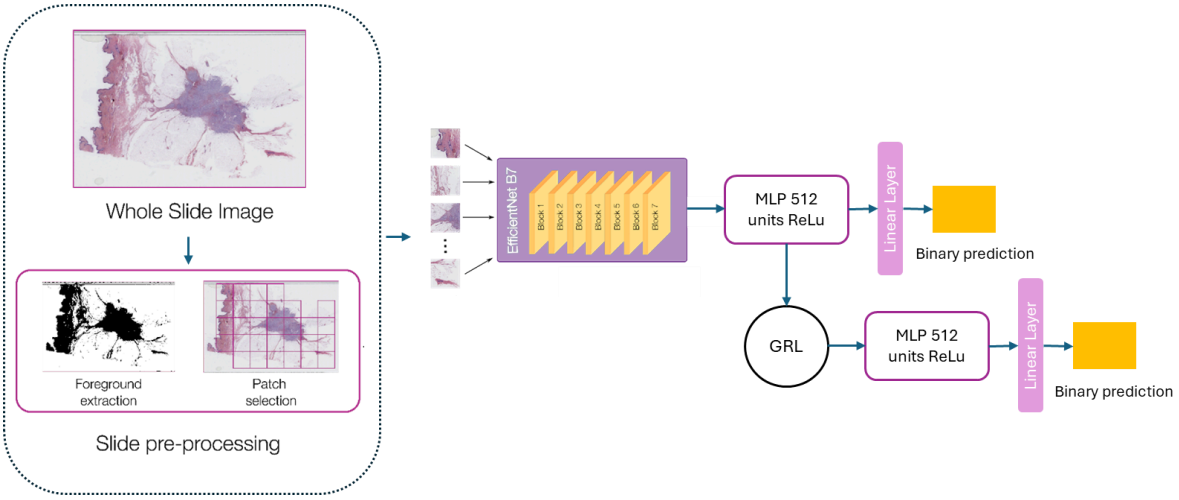


Figure 65. Overview of the domain-adversarial mutation prediction pipeline. Whole-slide images are first pre-processed by foreground extraction and tiling into 600 × 600 px patches at 5× magnification. Patches are encoded using a pretrained EfficientNet-B7 feature extractor, producing latent representations subsequently fed into a dual-branch network. The main branch (top) predicts the mutation status via a multilayer perceptron (MLP, 512 units, ReLU activation), while the auxiliary branch (bottom) predicts the cancer type through a Gradient Reversal Layer (GRL), enforcing domain invariance by penalising features informative of tissue origin. The joint optimisation of mutation and adversarial losses encourages the model to learn representations discriminative for genomic alterations yet robust to inter-tissue variability.

The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{mut} + \lambda \mathcal{L}_{DA}$$

where L_{mut} is the main mutation loss, L_{DA} the categorical cross-entropy for domain classification, and λ the adversarial coefficient (empirically set to 0.1). Training and domain-adversarial phases were run simultaneously with mixed batches from all available tissues for each gene.

Training Configuration and evaluation

Each pairwise task was trained for 10 epochs with early stopping based on validation AUC. Binary labels were defined at slide level using the presence of at least one pathogenic

mutation. Models were optimised with Adam (lr = 1e-4, weight decay = 1e-5, batch size = 64). Prediction scores were aggregated at slide level using the 99th percentile of tile scores, as in prior work. Reported metrics include mean AUC and corresponding one-sided Mann–Whitney *U* test p-values averaged across 3 random initialisations.

To systematically evaluate cross-cancer generalization, we adopted a leave-one-cancer-out (LOCO) design. For each gene, one cancer type was held out as an independent test set, while the model was trained on the remaining available cancers. This procedure was repeated such that each cancer type served once as the test domain. This resulted is 9 non-oriented combination :

Gene	Test cancer type	Training cancer types
PIK3CA	UCEC	BLCA, BRCA, COAD, HNSC, LUSC
PIK3CA	LUSC	BLCA, BRCA, COAD, HNSC, UCEC
PIK3CA	HNSC	BLCA, BRCA, COAD, LUSC, UCEC
PIK3CA	COAD	BLCA, BRCA, HNSC, LUSC, UCEC
PIK3CA	BRCA	BLCA, COAD, HNSC, LUSC, UCEC
PIK3CA	BLCA	BRCA, COAD, HNSC, LUSC, UCEC
TP53	LUSC	BLCA, COAD
TP53	COAD	BLCA, LUSC
TP53	BLCA	COAD, LUSC

Table 18. Cross-cancer combinations used for the leave-one-cancer-out (LOCO) experiments on PIK3CA and TP53. For each configuration, the model was trained on all remaining tissues and tested on the held-out cancer type.

The models then followed the same training procedure and parameters as for the pairwise comparison.

All experiments were repeated with three random seeds, and results were averaged. Performance was reported on held-out test slides unseen during training, ensuring no patient overlap.

RESULTS

Pair-wise comparison

Table 19 summarises the cross-cancer pairwise prediction results obtained by training on one tumour type and testing on another for the same mutation. Across all evaluated genes, AUC values generally remained close to random, ranging from 41.3% to 55.2%, except for *IDH1*, which displayed strong bidirectional transferability between glioma subtypes (GBM ↔ LGG; AUC = 71.1% and 74.2%, $p < 1.5\times10^{-5}$). This observation likely reflects the strong morphological continuity between low-grade and high-grade gliomas, where IDH1-driven molecular phenotypes are preserved across grades.

Task	Mean AUC (%)	Mean P-value
KRAS - COAD → LUAD	52.91	0.641
KRAS - LUAD → COAD	45.27	0.4102
EGFR - GBM → LUAD	52.13	0.674
EGFR - LUAD → GBM	51.03	0.7405
IDH1 - GBM → LGG	**71.08**	**1.50E-5**
IDH1 - LGG → GBM	**74.22**	**1.72E-6**
BRAF - SKCM → THCA	49.87	0.8752
BRAF - THCA → SKCM	50.82	0.8425
NRAS - SKCM → THCA	41.27	0.0258
NRAS - THCA → SKCM	55.17	0.057

Table 19. Cross-cancer mutation status prediction results.

By contrast, other driver mutations such as *KRAS*, *EGFR*, *BRAF* and *NRAS* showed limited cross-tissue predictability, with AUCs typically below 55% and no significant *p*-values after correction. Notably, *NRAS* displayed marginally significant transfer from SKCM to THCA (AUC = 41.3%, $p = 0.026$), suggesting weak, unidirectional generalisation.

Domain adversarial training

To evaluate whether domain adaptation could mitigate the strong tissue dependency observed above, we implemented a domain-adversarial setup using a leave-one-cancer-out (LOCO) design for *PIK3CA* and *TP53* (**Table 20**). Models were trained on all available cancers for each gene except one, which was reserved for testing.

Gene	Test cancer type	Mean AUC (%)	Mean P-value
PIK3CA	UCEC	51.58	0.7410
PIK3CA	LUSC	55.48	0.3200
PIK3CA	HNSC	57.94	0.0740
PIK3CA	COAD	47.84	0.7604
PIK3CA	BRCA	51.02	0.8147
PIK3CA	BLCA	54.78	0.2578
TP53	LUSC	48.47	0.5800
TP53	COAD	53.36	0.5740
TP53	BLCA	52.98	0.6561

Table 20. Cross-cancer leave-one-cancer-out (LOCO) prediction results for PIK3CA and TP53. Models were trained on all remaining tissues and evaluated on the held-out cancer type. Mean AUC and p-values are averaged across three random initialisations.

For *PIK3CA*, cross-cancer generalisation remained modest, with AUC values ranging between 47.8% and 57.9%. The highest performance was observed when testing on HNSC (AUC = 57.9%, p = 0.074), suggesting partial transfer when epithelial morphologies are preserved across training domains. For *TP53*, results were similarly limited (AUC 48.5–53.4%), indicating that even with domain-adversarial regularisation, mutation prediction remains highly sensitive to histological context.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

Overall, these results suggest that morphological correlates of mutation status are largely tissue-dependent, and that transferable morpho-genomic signatures are the exception rather than the rule.

For Domain adversarial training strategies, no configuration achieved significance after correction, and improvements compared to the pairwise baseline were minor. These findings indicate that while domain-adversarial training reduces overfitting to specific tissues, it does not yield fully invariant representations of mutation-related morphology. Instead, the data support a **context-specific paradigm**, where each gene–cancer pair exhibits distinct morpho-molecular relationships that cannot be universally transferred across organs.

4.3.3 Predicting Homologous Recombination Deficiency (HRD) and The Prediction Analysis of Microarray 50 (PAM50) in TCGA-BRCA

The HRD score and PAM50 classification offer complementary perspectives on tumour biology. The HRD score highlights genetic vulnerabilities, while the PAM50 classification delineates molecular subtypes that drive tumour behaviour. Being able to include these features in our current OS and DFS pipeline has the potential to increase our current performance in both tasks and give us a molecular backbone for explaining why specific patients might respond better to systemic chemotherapy.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset

For this specific chapter, we use the TCGA/BRCA dataset, which contains HRD annotations for 493 patients (402 labeled 0, and 91 labeled 1). For PAM50, the distribution of the annotations are shown in **Table 21**.

PAM50 Label	Number Patients
None	188
0 - Luminal B	73
1- HER2-enriched	35
2 - Luminal A	143
3 - Basal Like	53
4 - Normal Like	1

Table 21. PAM50 annotation distribution in the TCGA dataset.

Methodology

To predict the HRD score and PAM50 intrinsic subtypes, we developed two distinct models, tailored to address the differences in available data and prediction objectives. As the datasets differ, with 188 patients annotated for HRD scores but lacking PAM50 information, we implemented separate models for each task while maintaining a common feature extraction strategy.

Feature Extraction with ViT/S16 trained by Lunit, fine-tuned on PRIMUNEO

For both models, we leveraged precomputed embeddings obtained using the ViT/S16 model at x40 magnification. This ViT-based model has demonstrated robust performance in previous experiments and serves as the foundational feature extractor. By building on these embeddings, we aimed to capitalise on the model's strong representation of histological features.

HRD Prediction Model

To predict the Homologous Recombination Deficiency (HRD) score, we employed a Multi-Layer Perceptron (MLP) designed to classify patients as HRD-positive or HRD-negative based on established clinical thresholds. Specifically:

- The MLP architecture includes output layers of 128, 64, 32, 16, and 1 nodes. This progressive reduction in feature dimensions facilitates both robust feature refinement and effective binary classification.
- We employed a Binary Cross-Entropy (BCE) loss to optimize the model for the binary task.
- HRD status was defined using the established cutoff:
    - 0 — Negative if HRD score < 42
    - 1 — Positive if HRD score ≥ 42

This model is designed to capture key morphological patterns linked to genomic instability, a crucial factor in breast cancer prognosis and treatment response.

PAM50 Prediction Model

To predict the PAM50 intrinsic subtypes, we adapted the pipeline to classify patients into one of four breast cancer subtypes:

- 0 — Luminal B
- 1 — HER2-enriched
- 2 — Luminal A
- 3 — Basal-like

The MLP architecture was modified to produce 4 output nodes, one for each PAM50 class. To train the model, we employed a Cross-Entropy loss, which effectively optimises for multi-class classification. This design aims to capture the morphological patterns that distinguish these key breast cancer subtypes, which have distinct clinical outcomes and treatment pathways.

Training Configuration

Both models were trained using a 3-fold cross-validation strategy to ensure robust evaluation and reduce the risk of overfitting. Each fold was split into two sets to ensure balanced sampling. We used a batch size of 32 and a learning rate of 1e-4

This setup allows for comprehensive assessment across different data partitions while ensuring consistent optimization for both tasks.

Evaluation

To obtain WSI-level predictions, we employed two aggregation strategies to combine the patch-level predictions: mean aggregation and percentile-99 aggregation. The mean aggregation method computes the average prediction score across all patches within a WSI, providing a balanced representation of the overall risk distribution. Conversely, the percentile-99 aggregation method selects the 99th percentile score from the patch predictions, emphasising the most aggressive or highest-risk regions within the slide. This latter strategy ensures that critical morphological patterns associated with poor prognosis are effectively captured. By combining these two aggregation techniques, we aim to capture both global tumour behavior and localized high-risk features, which are essential for improving prediction reliability.

For performance evaluation, we employed the Area Under the Curve (AUC) metric. For the HRD prediction task, we report the mean AUC to summarize the model's overall performance. To assess the statistical significance of the results, we performed a one-sided

Mann-Whitney U test, which evaluates whether the model's score distributions for HRD-positive and HRD-negative cases are statistically distinct.

In the PAM50 prediction task, we expanded the evaluation to provide both class-specific AUC scores and a macro-average AUC. The class-specific AUC allows us to examine performance for each PAM50 subtype independently (Luminal B, HER2-enriched, Luminal A, and Basal-like), offering insight into the model's strengths and weaknesses for each class. The macro-average AUC provides a single performance measure that calculates the AUC for each label, followed by their unweighted mean, ensuring that no individual class dominates the overall score, an important adjustment given the potential class imbalance in the dataset.

To ensure robust performance evaluation, we adopted a 3-fold cross-validation strategy. Results are first reported by fold to highlight potential performance variability across different data splits. Additionally, to provide a single consolidated result, we concatenate the predictions from all 3 folds to compute performance metrics on the combined dataset, ensuring a comprehensive assessment of the model's predictive capacity.

RESULTS

HRD prediction

The performance of the HRD prediction model on the TCGA/BRCA dataset is summarized in **Table 22** and **Figure 66**. Overall, the model achieved strong and consistent results across different folds and aggregation methods, demonstrating reliable predictive capabilities.

Patch ensemble method	Mean	Weighted Mean	Percentile 99
Fold 1	0.818 (1.2e-8)	0.821 (5.65e-8)	**0.825 (2.28e-7)**
Fold 2	0.752 (1.02e-4)	0.749 (1.3e-4)	**0.754 (4.73e-5)**
Fold 3	**0.81 (1.03e-6)**	0.805 (9.23e-7)	0.795 (7.97e-7)
Mean	**0.793**	0.792	0.791

Table 22. Pipeline's performance in the AUC (P-value) in each fold of the cross-validation using different functions for ensembling the patch-level predictions.

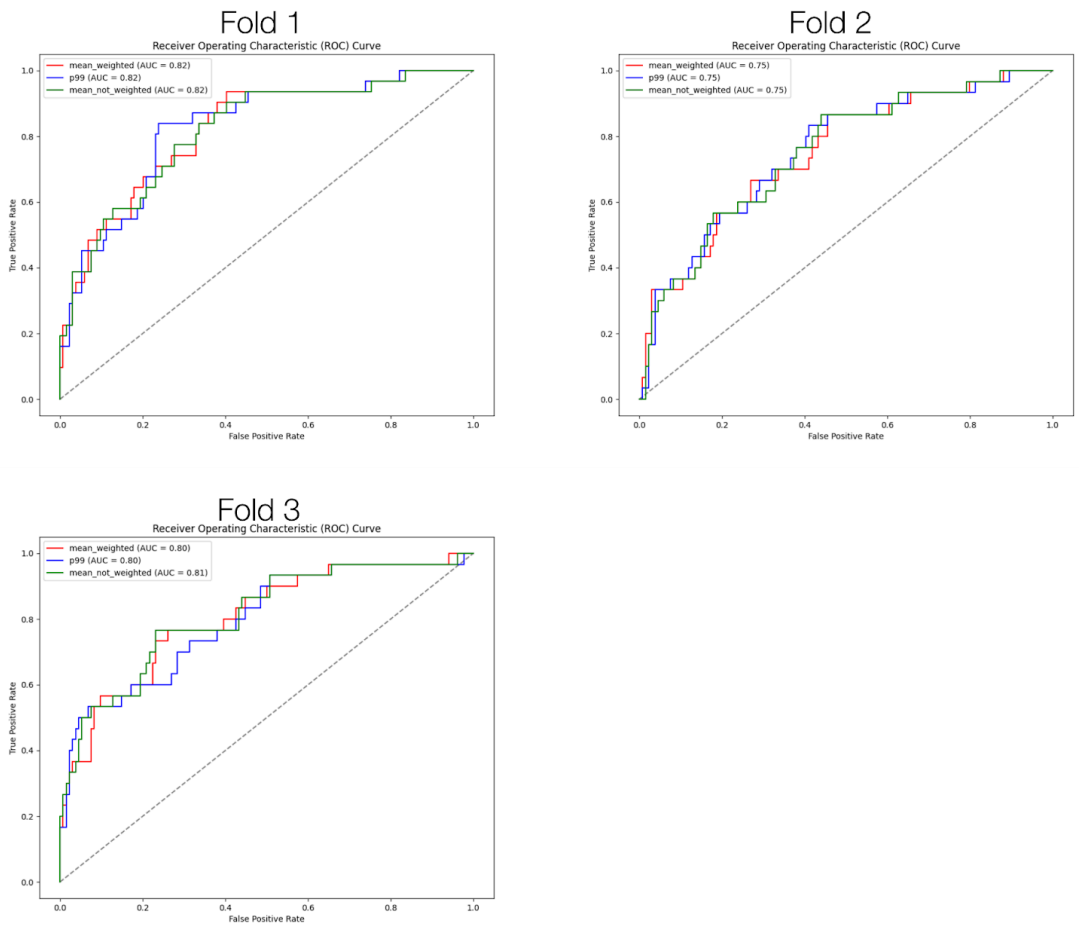


Figure 66. AUC performance in the HRD prediction task ensembling the patch scores with the weighted mean, percentile 99 and mean functions for the 3-fold cross-validation.

On average, the HRD prediction pipeline achieved an AUC of 79.3% when aggregating patch-level predictions using a mean ensemble strategy. This approach proved to be the most stable and effective method, maintaining robust performance across all folds. The model's overall performance remained close to 80% AUC across all evaluated aggregation methods, reinforcing the method's reliability.

Performance analysis by fold reveals some variability. While Fold 1 achieved the highest performance with an AUC of 0.818 (mean aggregation) and 0.825 (percentile-99 aggregation), the model's performance was notably lower in Fold 2, with AUC values ranging between 0.749 and 0.754. This drop may reflect differences in data distribution or sample complexity within this fold. In Fold 3, results improved again, with AUC values consistently above 0.79 across all aggregation strategies.

When comparing different aggregation methods, the percentile-99 aggregation method achieved the highest score in Fold 1 (AUC = 0.825) and Fold 2 (AUC = 0.754), while the mean aggregation method showed slightly superior consistency across the folds. The weighted mean method delivered comparable results, with an average AUC of 0.792, reinforcing the robustness of all three approaches.

Despite slight variations between folds, the overall model performance indicates that the proposed method is effective in predicting HRD status in the TCGA/BRCA dataset. The strong average AUC scores across all aggregation strategies, combined with highly significant p-values, underscore the model's capacity to identify key morphological patterns linked to HRD-positive tumours.

PAM50 prediction

The PAM50 prediction model achieved strong and consistent performance across all folds and aggregation methods. The model's overall performance reached 81.6% AUC when using the mean aggregation method, slightly outperforming the percentile-99 strategy, which achieved 80.9% AUC.

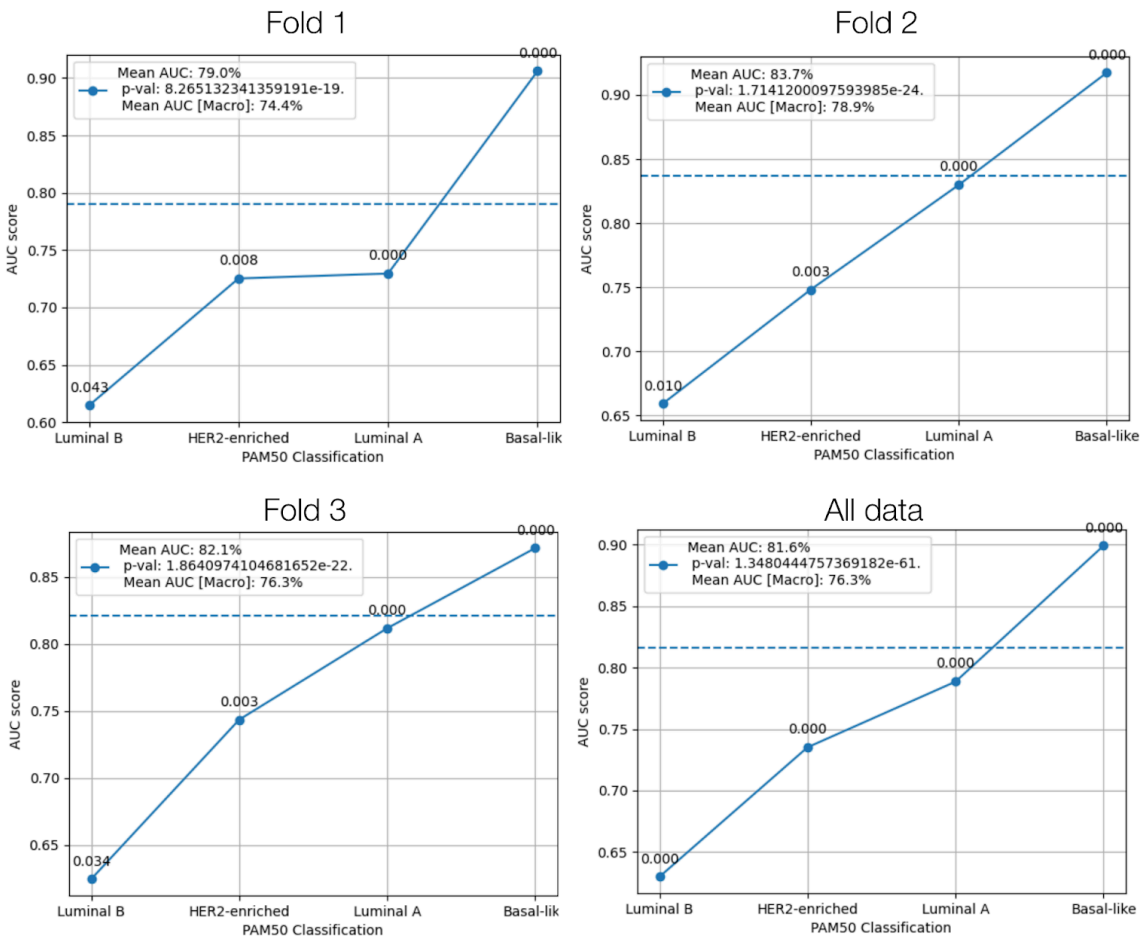


Figure 67. AUC by class and average AUC performance in the PAM50 task ensembling the patch scores with the mean function.

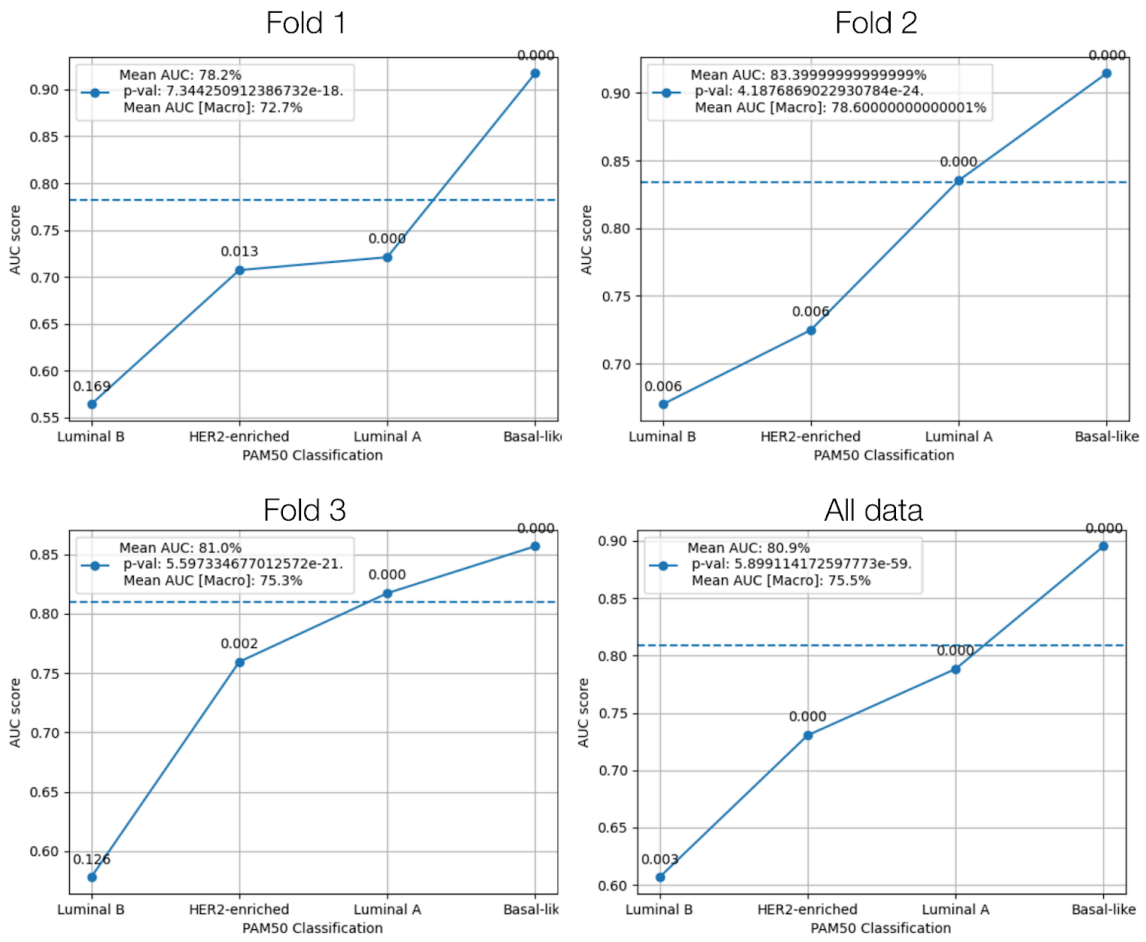


Figure 68. AUC by class and average AUC performance in the PAM50 task ensembling the patch scores with the percentile-99 function.

When examining individual folds, the mean aggregation method achieved 79.0% in Fold 1, 83.7% in Fold 2, and 82.1% in Fold 3, with all results showing strong statistical significance. In comparison, the percentile-99 method showed slightly lower performance, achieving 78.2%, 83.4%, and 81.0% in the same respective folds. The consistently superior results obtained with the mean aggregation suggest that averaging patch-level predictions is more effective for this multi-class classification task.

Despite these encouraging results, performance for the Luminal B subtype was noticeably weaker, with AUC scores consistently falling below 65% in most folds. Further analysis revealed that many patients labeled as Luminal B were frequently misclassified as Luminal A with high confidence. This confusion likely stems from the morphological similarities between these two subtypes, which are known to overlap in histological appearance.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

The proposed pipeline for predicting HRD status and PAM50 subtypes demonstrated strong performance, achieving AUC scores close to 80% across all three folds, with p-values < 0.05 confirming statistical significance. These results highlight the model's robustness and reliability in identifying key molecular features from histological data.

The model showed promising generalization capabilities when applied to breast cancer data. The next steps will involve deploying the trained models on the PRIMUNEO and CGFL datasets to predict both HRD status and PAM50 subtypes in these cohorts. These predicted molecular features will then be integrated into the Cox Proportional Hazards (CoxPH) model to enhance the performance of survival prediction tasks. This step aims to explore the potential of combining histological and molecular insights to improve risk stratification in breast cancer patients.

4.4 Discussion

Precision oncology aims to tailor treatment to each patient's tumour biology, particularly by identifying actionable mutations. In this context, deep learning applied to histology offers a cost-effective alternative to genomic testing. Our study advances this field by showing that deep learning can predict not only gene-level mutations but also protein-specific variants directly from WSIs, uncovering distinct morpho-molecular signatures with clinical relevance.

Through MultiVarNet, a biologically informed, protein-specific variant-aware model, we demonstrate that protein-specific variant-level training enables the detection of subtle morphological differences within the same gene. This granularity improves prediction accuracy compared to traditional gene-level approaches, which often mask heterogeneity. MultiVarNet achieves modest but significant gains across tumour types and datasets, outperforming conventional baselines. The shift from generic mutation detection to biologically grounded, protein-specific variant-specific inference represents a paradigm change in computational pathology, enabling the development of interpretable digital biomarkers. Importantly, we show that these gains arise not from complex architectures, but from informed label design. This highlights the power of biologically structured supervision. While further validation is needed, this study provides the first systematic demonstration of protein-specific variant-level histological prediction. Further explorations will be made to improve performance gains through this new paradigm.

We also predicted clinically used molecular signatures, HRD status and PAM50, from histology. These molecular labels, used routinely to guide treatment in breast cancer, were predicted with strong performance, confirming that WSIs can capture higher-order molecular patterns.

Together, protein-specific variant-level predictions and signature-level classification form a robust virtual molecular biology framework. Looking forward, we will apply these tools to predict outcomes in patients with residual disease after neoadjuvant chemotherapy. By integrating biopsy and surgical slide analysis, we aim to guide post-treatment decisions such as escalation or de-escalation of therapy.

Despite these encouraging results, several critical limitations must be acknowledged. First, our cross-cancer analyses (Section IV.3.2) revealed that morphological correlates of genomic alterations are largely non-transferable across tissues, with only closely related entities such as gliomas showing meaningful generalisation. This finding underscores the strong contextual dependency of histopathological features, shaped not only by tumour lineage and microenvironment but also by site-specific pre-analytical factors such as fixation, staining, slide digitisation or mutation clonality and subclonality. Clonal variants appear to generate more consistent and detectable morphological patterns, whereas subclonal heterogeneity may dilute or obscure these signals, contributing to the limited pan-cancer generalisability observed even with domain-adversarial approaches. Although domain-adversarial learning (multi-DA) partially mitigated such biases by enforcing domain-invariant representations, it did not fully overcome the biological specificity inherent to each tissue. This suggests that future progress in computational pathology will depend less on achieving pan-cancer universality and more on context-aware, biologically constrained models capable of disentangling phenotype from artefact.

Second, while protein-specific variant-level prediction advances interpretability by aligning machine learning outputs with molecular mechanisms, the clinical translation of such digital biomarkers will require large-scale, prospective validation across institutions, scanners, and treatment settings. As noted by Echle et al.[159], 2021; Kather et al., 2020[160], reproducibility and robustness remain major barriers to clinical adoption. Furthermore, the apparent morphological predictability of some proteic variants may reflect sampling biases or confounding correlations (e.g., tumour grade, necrosis, or stromal content) rather than direct genotype–phenotype causality.

Finally, the ambition of precision oncology should not be reduced to algorithmic performance. As highlighted by Adam et al., 20201[161], the path from molecular correlation to actionable insight requires interpretability, transparency, and biological validation. Our results thus argue for a measured optimism: deep learning can reveal meaningful morpho-molecular links, but these must be rigorously contextualised within tumour biology and verified through integrative, multi-omic approaches. In that sense, this work positions deep learning not as a replacement for molecular diagnostics, but as a complementary, hypothesis-generating tool that bridges morphology, genomics, and treatment response.

Chapter 5: Understanding prognosis and risk of relapse in breast cancer post neoadjuvant systemic chemotherapy.

Abstract

This chapter develops an integrative deep-learning framework connecting pre- and post-treatment histology to therapeutic response and survival in early breast cancer (eBC). Building on previous chapters, we introduce a series of neural-network models trained on whole-slide images (WSIs) from diagnostic biopsies and post-neoadjuvant (post-NAC) surgical specimens to predict both pathological complete response (pCR) and overall survival (OS) / disease-free survival (DFS).

The first model, Chemo-prAIdict Breast, predicts intrinsic chemosensitivity to anthracycline–taxane backbones from baseline biopsies. Its output probability (the "pCR score") stratifies patients across subtypes (TNBC, HER2+, and ER+/HER2−) and correlates with recognised histopathological features of response. The pCR score was then combined with post-NAC histological analysis and virtual molecular signatures (PAM50 and HRD predicted from WSIs) to refine outcome prediction. Using Cox proportional-hazards modelling and Kaplan–Meier analyses, we show that integrating pre-treatment chemosensitivity and post-treatment residual-disease morphology yields stronger prognostic separation than any single marker alone.

A second network trained on post-NAC slides directly predicts OS and DFS and recapitulates established morphological regression patterns (high-risk profiles displaying residual invasive carcinoma, necrosis, and nuclear pleomorphism, and low-risk profiles characterised by fibrosis and in-situ remnants). Pathologist-guided review of high-attention regions confirmed biological plausibility of the learned features.

Together, these findings propose a continuum of AI-assisted decision support linking baseline chemosensitivity, pathological response, and long-term outcome. While performance remains moderate and external validation monocentric, this work provides the first analytical demonstration that digital histopathology can integrate pre- and post-treatment morphology to

predict both response and survival, laying the foundation for context-aware, clinically interpretable deep-learning tools in precision oncology.

5.1 Introduction

As described in Chapter I, neoadjuvant chemotherapy (NAC) is increasingly used in the management of early breast cancer, to shrink tumours prior to surgery and to assess the biological behavior of the disease. Following NAC, the evaluation of the post-surgical specimen, particularly in identifying residual disease (RD), has become essential in guiding adjuvant treatment decisions. Patients who achieve a pathological complete response (pCR) generally have an excellent prognosis, while those with residual disease face a more heterogeneous outcome. However, not all residual disease carries the same prognostic weight, and current clinicopathological tools remain insufficient to capture this complexity.

One of the most urgent challenges in precision oncology today is the stratification of residual disease: distinguishing patients whose residual tumours are likely to lead to relapse from those whose remaining disease is indolent or biologically controlled. This is particularly important as therapeutic options post-NAC continue to expand, such as extended hormonotherapy, CDK4/6 inhibitors, T-DM1, and immunotherapy, but must be deployed selectively due to toxicity, cost, and long-term impact on quality of life.

Building on the findings of the previous chapters, where we demonstrated the capacity of deep learning models to predict OS and DFS using the post-NAC surgical specimen and to decode variant-level and molecular signature-level information from histology, we now aim to better predict and understand the stratification of the residual disease. In this chapter, we will leverage both biopsy and post-NAC surgical specimens to predict survival outcomes (Overall Survival, OS; and Disease-Free Survival, DFS) with higher granularity, but also use the morpho-molecular approach described in Chapter IV applied to our dataset. While prior chapters established that post-NAC surgical specimens contain prognostic information that can be harnessed using deep learning to predict overall survival (OS) and disease-free survival (DFS), this section explores whether integrating information from the pre-treatment biopsy and inferred molecular phenotypes can further enhance prognostic accuracy. The central hypothesis is that integrating multi-temporal (biopsy and surgery) and multi-layered (morphological and inferred molecular) information will lead to a more granular, biologically informed stratification of residual disease.

The first part of this work investigates whether survival outcomes (OS and DFS) can be directly predicted from the initial biopsy. Although the biopsy is collected before systemic treatment, it may contain morphological markers of aggressive biology that remain

prognostically relevant. If such features are preserved, combining biopsy-derived predictions with post-surgical information could improve long-term outcome prediction. This sub-task tests the additive value of the baseline tumour phenotype in a survival-focused setting.

Building on this, we then examine whether the biopsy can be used to predict the tumour's intrinsic sensitivity to chemotherapy, which is defined as the likelihood of achieving a pCR. This information, while not directly prognostic, may be an important latent variable influencing survival. This work, published in the European Journal of Cancer[170] demonstrated that chemo-sensitivity can be accurately predicted from baseline biopsy histology using deep learning. In this chapter, we assess whether incorporating this predicted response into the survival pipeline enhances the ability to distinguish between high- and low-risk residual disease. This experiment tests whether the biological behavior inferred from the biopsy (rather than its morphological outcome alone) provides added prognostic value, as a tumour with intrinsic chemo sensitivity could lead to a residual disease of different prognosis from one with intrinsic chemoresistant features.

The second major objective of this chapter is to integrate morpho-molecular correlates (HRD status and PAM50 subtypes) into the survival modeling framework. While ground-truth molecular data are not available in the post-NAC setting, we previously developed models capable of predicting these signatures from histology. By applying these models to the post-treatment specimen, we aim to infer the molecular phenotype of the residual disease and evaluate its association with OS and DFS outcomes. This approach provides a first step toward exploring the biological underpinnings of prognostic variability in RD, potentially illuminating why certain residual tumours are more dangerous than others.

The third objective of this chapter is to transform our OS/DFS prediction models into a clinically interpretable survival analysis framework by stratifying patients into good and poor prognosis groups and generating Kaplan–Meier survival curves. While models such as WSIM (Whole Slide Image Model) and CM (Clinical Model) output continuous risk scores, these scores do not directly indicate how to categorise patients into actionable clinical groups. For real world application, clinicians require discrete classifications that can guide treatment decisions, such as distinguishing patients who may benefit from therapeutic escalation versus those who could be safely spared.

Finally, to move beyond blackbox prediction, we propose a patch-level analysis of the histological slides to identify novel prognostic biomarkers in collaboration with expert

pathologists. This analysis will focus on the regions that most strongly influence the model's predictions, offering a pathway to interpretability and biological discovery. By highlighting visual features associated with good or poor prognosis, we aim to uncover new morpho-clinical patterns that could be further validated and eventually integrated into clinical workflows.

Together, these complementary efforts seek to redefine the way residual disease is assessed after neoadjuvant therapy.

5.2 Material and Methods

In this section, we will use two main data types as inputs for our prediction tasks: biopsy and surgical specimens whole slide images. They come from 2 different datasets: the PRIMUNEO dataset, the CGFL breast cancer neoadjuvant dataset, and the TCGA. The cohorts and experimental setups are described in **Chapter 2: Clinical Cohorts and Data Acquisition**. Specific adaptations are covered in the dedicated sections throughout this chapter.

5.3 Results

5.3.1 Combining biopsy and post-NAC surgical specimen to improve OS and DFS prediction pipeline.

5.3.1.1 Predict OS and DFS using the diagnosis biopsy and integrate the info in the OS DFS prediction pipeline

In this section, we investigate whether combining pre-treatment biopsy and post-treatment surgical WSIs improves long-term survival prediction (OS and DFS) in breast cancer. Leveraging the unique structure of the PRIMUNEO and CGFL Neoadj datasets which include both biopsy and surgical slides for the same patients, we explore how tumour evolution across the neoadjuvant chemotherapy timeline can inform prognosis.

The diagnostic biopsy reflects the tumour's baseline biology, while the surgical specimen captures the treatment response. We hypothesize that these two timepoints provide complementary prognostic information: intrinsic aggressiveness may be visible in the biopsy, while treatment resistance or sensitivity is more evident in the surgical sample.

We begin by evaluating each source independently to establish baseline performance and understand their individual predictive value. This helps assess whether prognosis can be inferred from the biopsy alone or if integrating post-NAC data adds meaningful value.

This approach also addresses a key clinical question: how much does initial phenotype vs. treatment response contribute to prognosis? If biopsy-only models underperform, but improve when combined with surgical data, or with inferred variables like chemosensitivity, it would underscore the importance of modeling tumour evolution.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset and study design

For this experiment, we included all patients having a surgical specimen as both the diagnostic biopsy and the corresponding post-NAC surgical specimen Whole Slide Images

(WSIs) were available. All patients were selected from the PRIMUNEO and CGFL Neoadj cohorts. A comprehensive description of the dataset composition, preprocessing steps, and experimental setup is provided in the **Chapter 2 early Breast Cancer dataset section**.

Methodology

For both biopsy and surgical specimen inputs, we trained separate models using an identical pipeline. This pipeline corresponds to the best-performing method developed so far, as described in **Chapter 3**. It was originally optimised using the post-NAC surgical specimens.

As described in **Figure 69**, we begin by using the precomputed feature embeddings generated by the ViT/S16 model, pretrained with self-supervised DINO weights. For the training phase, we use the PRIMUNEO dataset, excluding patients from the CGFL centre in Dijon to ensure independence from the external validation cohort.
A Multi-Layer Perceptron (MLP) is then trained on these embeddings, with two output layers of 128 and 64 nodes, respectively. This MLP is optimised for a binary classification task, predicting event occurrence (labelled as 1) or non-occurrence (labelled as 0) at each year from 0 to 5, using a Binary Cross-Entropy (BCE) loss. The MLP serves to reduce the dimensionality of the ViT embeddings while learning a discriminative, time-aware representation of the input features.
The resulting 64-dimensional output from the MLP is used as input to a Cox Proportional Hazards (CoxPH) model, trained separately for the OS and DFS tasks. Risk scores are computed at the patch level and then aggregated at the WSI level using the 99th percentile, which emphasises the most high-risk regions of the slide.
The full pipeline is trained using a 3-fold cross-validation strategy. Each fold is trained for 5 epochs, with a batch size of 32, using the Adam optimiser with a learning rate of 5e-5. For additional technical details, including rationale for the architectural choices and validation scheme, refer to the **Chapter 2 early Breast Cancer dataset section**.

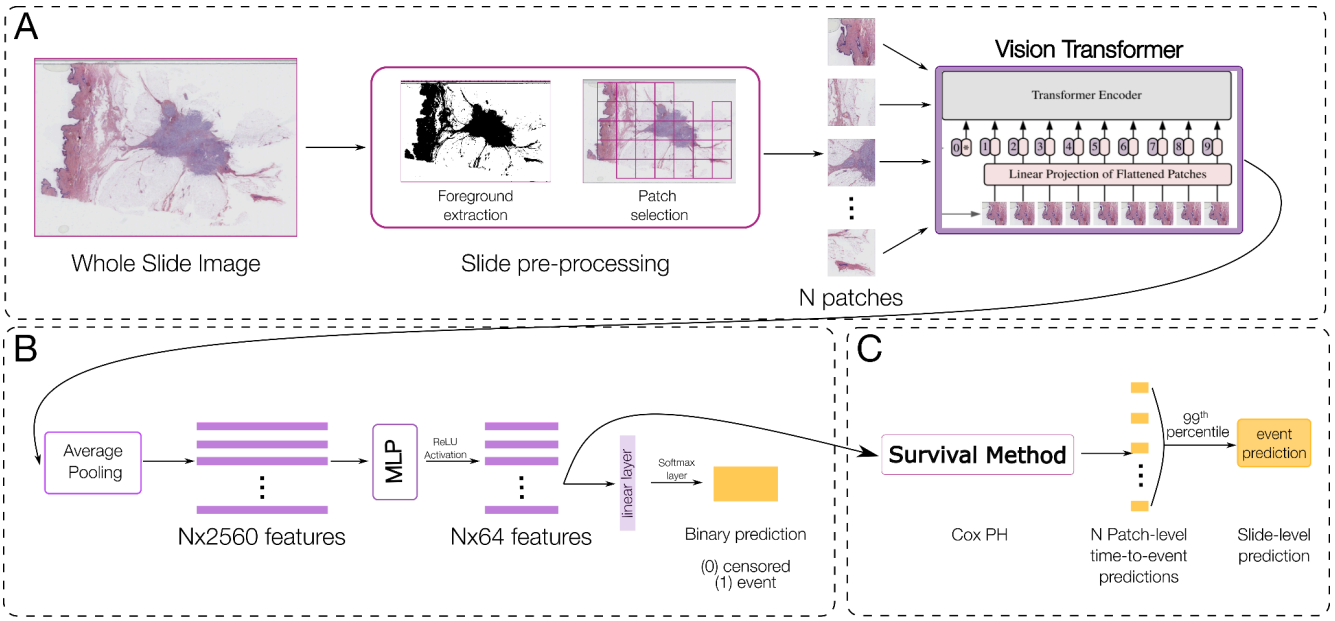


Figure 69. External validation' best performance method overview from Chapter 3. This figure summarises the workflow of the optimal configuration identified during internal validation. **(A)** Whole-slide images (WSIs) are pre-processed through foreground extraction and tiling into fixed-size patches, which are then encoded using a Vision Transformer (ViT) pretrained on a self-supervised learning task to obtain patch-level embeddings. **(B)** These embeddings are processed through a multilayer perceptron (MLP) that reduces their dimensionality and outputs patch-level risk logits. **(C)** Patch-level predictions are aggregated into slide-level survival estimates using a percentile-based pooling strategy (99th percentile) and fitted to survival models such as Cox proportional hazards (CoxPH). This end-to-end configuration achieved the best overall performance for both disease-free survival (DFS) and overall survival (OS) tasks during internal validation.

Evaluation

We aggregate the final WSI-level predictions by averaging the outputs of the three models trained across the folds during cross-validation. Model performance is then assessed using the mean AUC, Harrell's C-index, and the C-t index (as introduced in DeepHit), along with their weighted variants to account for class imbalance across timepoints.

For the OS task, evaluation is conducted from years 3 to 5, and for the DFS task, from years 2 to 5. We report the average performance across these time intervals to provide a summary metric that reflects long-term predictive accuracy.

In addition, we compute the p-value associated with the AUC and weighted AUC at each evaluated year using a one-sided Mann–Whitney U test, in order to assess the statistical significance of the model's discriminative ability.

RESULTS

Figures 70 to 73 summarise the performance of our models for predicting disease-free survival (DFS) and overall survival (OS), using either diagnostic biopsy or post-NAC surgical specimen WSIs as input. Results are reported for both internal and external validation cohorts.

Disease Free Survival (DFS) prediction

Figures 70 and 71 present the performance of the models for predicting DFS using either biopsy or surgical specimen WSIs, evaluated through both internal and external validation. Across all experimental settings, models trained on surgical specimens significantly outperformed those trained on biopsies.

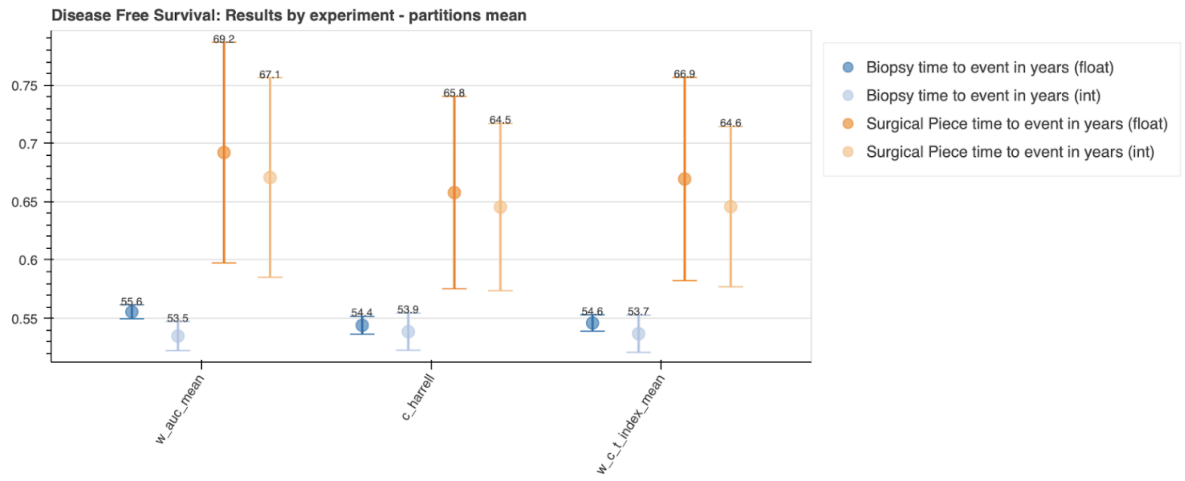


Figure 70. Internal validation – DFS prediction performance. Comparison of model performance for disease-free survival (DFS) prediction using whole-slide images (WSIs) from diagnostic biopsies (blue) and post-neoadjuvant surgical specimens (orange). Results are averaged across four internal partitions.

In the internal validation (**Figure 70**), surgical specimens consistently yielded higher weighted mean AUCs, ranging from 67.1% to 69.2%, whereas biopsy-based models achieved

lower scores between 53.5% and 55.6%. This performance gap was maintained across different stratification approaches and label formats, suggesting that the post-NAC histology contains stronger prognostic signals.

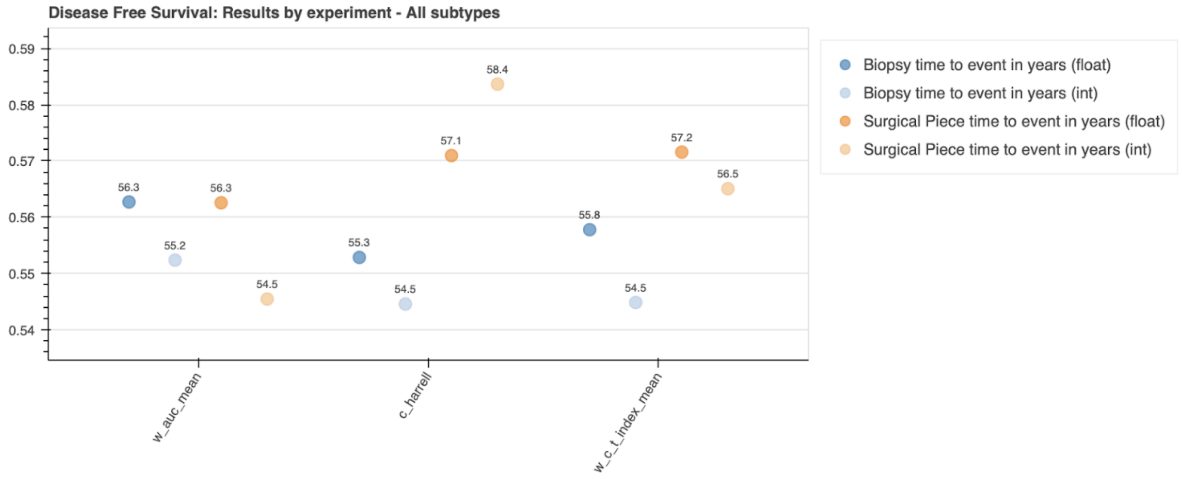


Figure 71. External validation – DFS prediction performance. Comparison of model performance for disease-free survival (DFS) prediction using whole-slide images (WSIs) from diagnostic biopsies (blue) and post-neoadjuvant surgical specimens (orange). The superior predictive value of post-NAC surgical slides is confirmed externally.

In the external validation (**Figure 71**), we again observe a similar pattern: biopsy-based models produced AUCs around 54–56%, while surgical specimen-based models reached 57–58%. These results demonstrate that the predictive advantage of surgical specimens generalises better across datasets and tumour subtypes.

Additionally, we compared performance based on the time-to-event label format. In most configurations, using continuous (float) labels for DFS prediction led to higher performance in weighted AUC metrics compared to discrete (integer) labels. This effect was particularly pronounced in the surgical specimen-based models, indicating that retaining the temporal resolution of the outcome improves discriminative power.

Overall Survival (OS) prediction

Figures 72 and **73** display OS prediction results. As with DFS, models trained on surgical specimens consistently outperformed those trained on biopsies, in both internal and external settings.

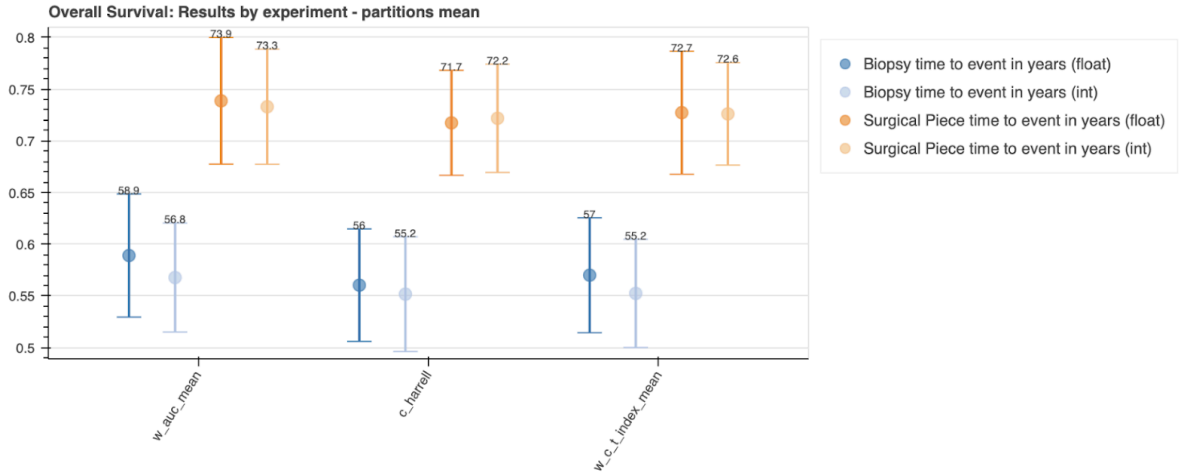


Figure 72. Internal validation – OS prediction performance. Comparison of model performance for overall survival (OS) prediction using whole-slide images (WSIs) from diagnostic biopsies (blue) and post-neoadjuvant surgical specimens (orange). Results are averaged across four internal partitions.

In the internal validation (**Figure 72**), surgical specimen-based models achieved weighted mean AUCs of up to 73.9%, while biopsy-based models did not exceed 58.9%, with most scoring around 55–57%. The superior performance of surgical specimens held across all stratification strategies.

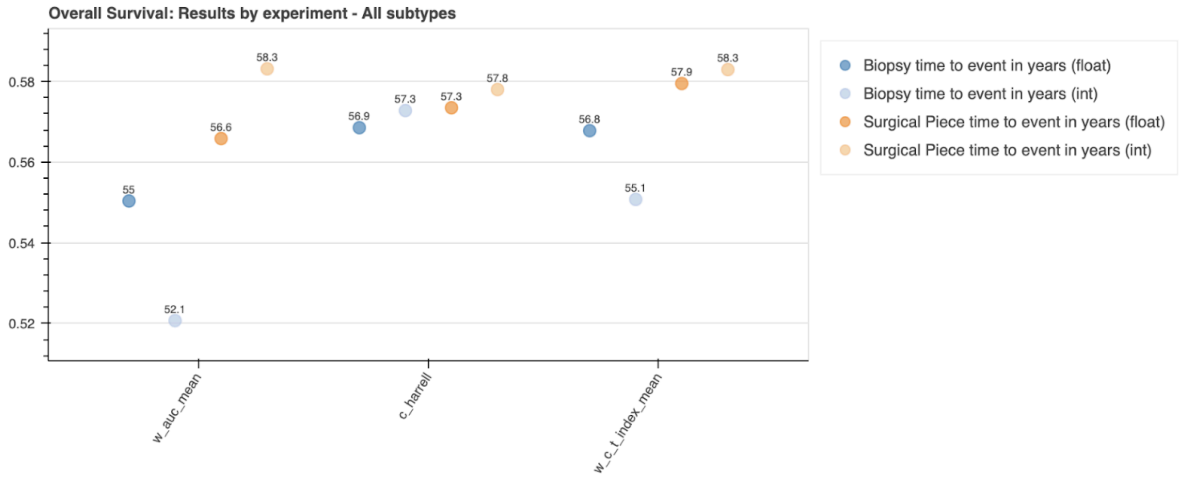


Figure 73. External validation – OS prediction performance. Comparison of model performance for overall survival (OS) prediction using whole-slide images (WSIs) from diagnostic biopsies (blue) and post-neoadjuvant surgical specimens (orange). The superior predictive value of post-NAC surgical slides is confirmed externally.

In the external validation (**Figure 73**), the trend persisted: surgical specimens achieved AUCs up to 58.3%, while biopsy-based models reached only 55% at best. Although the absolute values were lower than in internal validation, the relative benefit of using post-NAC data remained evident.

Again, the use of continuous time labels proved beneficial, particularly in the surgical specimen models, where it improved both weighted and standard AUCs, reinforcing the importance of precise temporal modelling for OS prediction.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

Overall, our results show that while biopsy WSIs contain useful prognostic signals, they underperform compared to surgical specimens for DFS and OS prediction. This likely reflects the added value of treatment response and residual disease captured post-NAC, as well as the pipeline's original optimization for surgical slides.

We also find that using continuous time-to-event labels improves model performance, particularly with surgical data.

In sum, this analysis provides a baseline and points to two key improvements: optimising the pipeline for biopsy data and adopting continuous time modeling for more accurate outcome prediction.

5.3.1.2 Combining Biopsy and Surgical specimen for OS and DFS prediction

The previous experiment (**Section 5.3.1.1**) demonstrated a significant drop in predictive performance when using diagnostic biopsy whole slide images (WSIs) alone to estimate overall survival (OS) and disease-free survival (DFS). This result highlights the limitations of relying solely on pre-treatment histology for long-term outcome modelling, likely due to the absence of treatment response information and the inherent biological variability captured only after neoadjuvant chemotherapy (NAC).

In this section, we aim to address two key questions: First, can we identify a more suitable deep learning architecture for feature extraction from biopsy slides to improve standalone predictive performance? Second, does combining biopsy-based and surgical specimen-based models, each representing a distinct clinical timepoint, enhance OS and DFS prediction when compared to using either input in isolation?

To do so, we evaluate multiple backbone architectures trained on biopsy data and assess their performance individually. We then explore different model ensembling strategies, integrating biopsy- and surgery-based predictions to test whether their information is complementary.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset and study design

For this experiment, we included all patients having a surgical specimen as both the diagnostic biopsy and the corresponding post-NAC surgical specimen Whole Slide Images (WSIs) were available. All patients were selected from the PRIMUNEO and CGFL Neoadj cohorts. A comprehensive description of the dataset composition, preprocessing steps, and experimental setup is provided in the **Chapter 2 early Breast Cancer dataset section**.

Methodology

To improve feature extraction from biopsy WSIs, we evaluated alternative deep learning backbones by replacing the original encoder with two state-of-the-art architectures:

Kaiko-ViT-B8 and Phikon[162,163]. These models were selected based on their strong performance in prior histopathological and molecular prediction tasks. The goal was to determine whether a more suitable representation of pre-treatment tumour morphology could enhance survival prediction performance when using biopsy slides alone.

The rest of the pipeline remained consistent with the approach described in **Section 5.3.1.1**, including the use of a multi-layer perceptron (MLP) for dimensionality reduction and a Cox Proportional Hazards (CoxPH) model for OS and DFS risk estimation.

For the ensemble model combining biopsy and surgical specimen data, we aggregated the WSI-level risk scores produced by each model (one from the biopsy, one from the surgical specimen). These scores were combined to generate a single, unified prediction for each patient.

RESULTS

Disease Free Survival (DFS) prediction

Figures 74 and 75 summarise DFS prediction performance across internal and external validation settings. Among the architectures tested, the ViT-S/16 backbone continued to provide the most robust results for biopsy-based models. In the internal validation, combining biopsy-derived and surgical specimen-derived risk scores led to modest improvements in predictive performance. This suggests a degree of complementarity between the two sources when evaluated on the training domain.

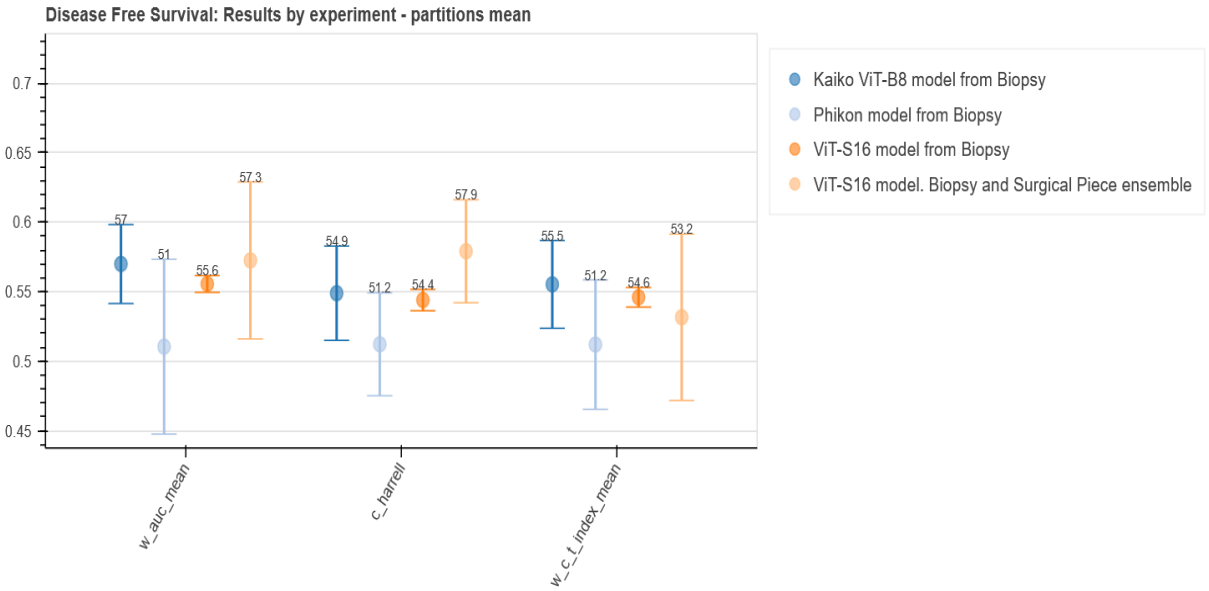


Figure 74. Disease-Free Survival (DFS) prediction performance in the internal validation. Comparison of multiple transformer-based architectures (ViT-S/16, ViT-B/8, Phikon) trained on biopsy-derived whole-slide images (WSIs). Results include both standalone biopsy models and ensembles combining biopsy- and surgery-based predictions.

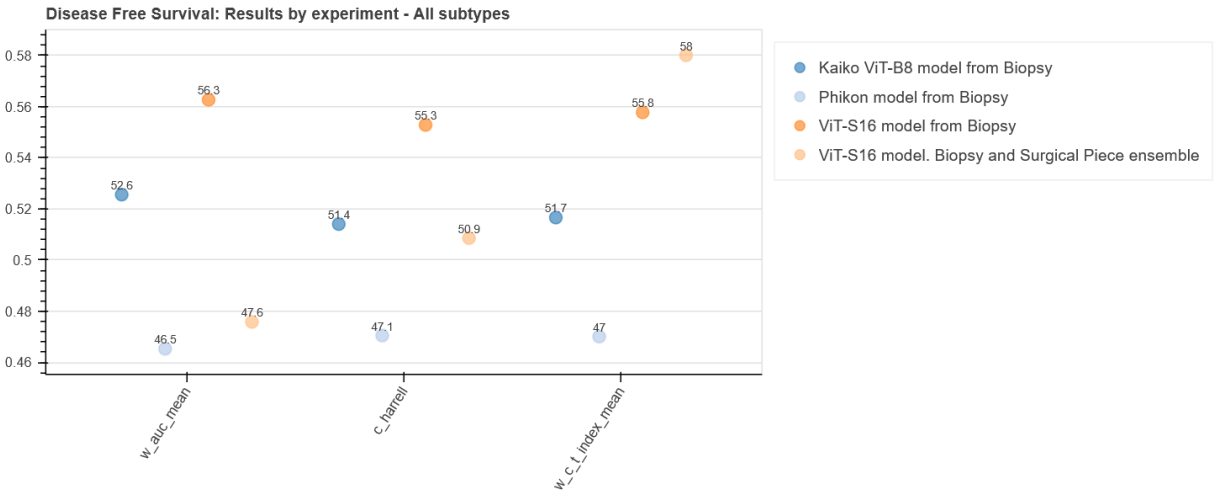


Figure 75. Disease-Free Survival (DFS) prediction performance in the external validation. Evaluation of the same transformer-based models on the independent external cohort, comparing biopsy-only and biopsy+surgery ensemble strategies for generalisation assessment.

However, this benefit did not translate to the external validation cohort. As shown in **Figure 75**, ensemble models performed consistently lower than models trained on surgical specimens alone, with ensemble AUC scores falling nearly 10 percentage points below the best

surgery-only model reported in **Section 53.1.1**. Despite additional experiments, such as architectural substitutions and hyperparameter tuning, the biopsy-based model remained limited in its standalone and ensemble contributions to DFS prediction.

Overall Survival (OS) prediction

Similarly, **Figures 76 and 77** present OS prediction performance. In line with the DFS findings, the ViT-S/16 model again produced the highest predictive scores among biopsy-based backbones. In the internal validation, ensembling biopsy and surgical specimen predictions led to minor gains, but these were insufficient to close the gap with surgery-only models. In the external validation, biopsy-derived features offered no added benefit, and the ensemble models remained consistently inferior to those based solely on surgical data.

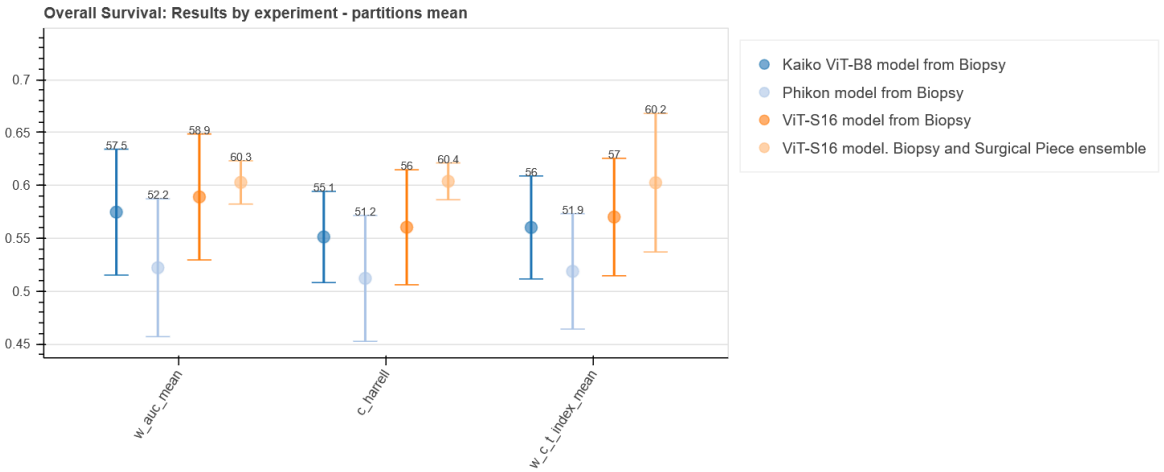


Figure 76. Overall Survival (OS) prediction performance in the internal validation. Performance comparison of transformer-based backbones for OS prediction using biopsy WSIs, with and without ensembling with surgical-specimen models. Internal validation results are shown across the three main survival metrics.

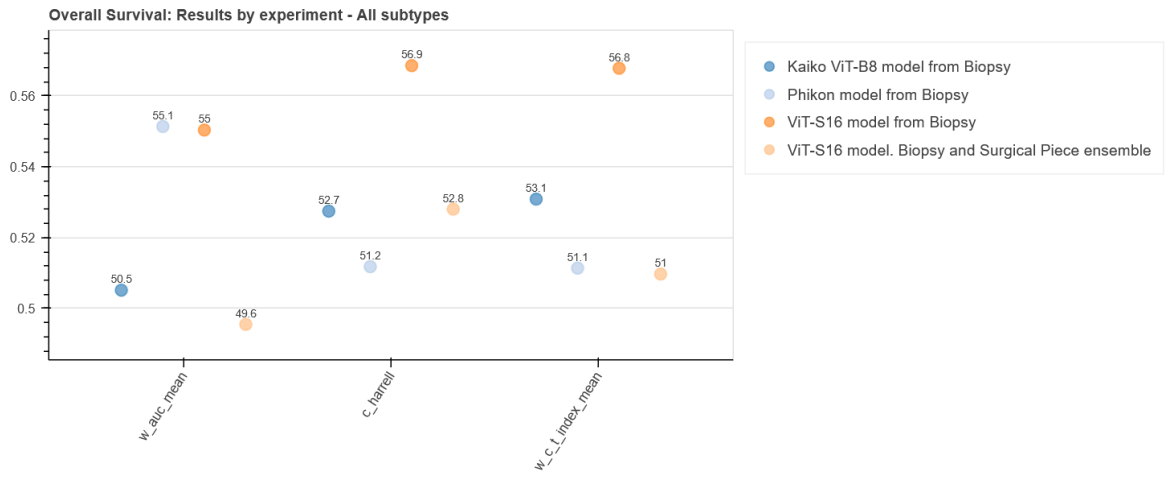


Figure 77. Overall Survival (OS) prediction performance in the external validation. External validation of biopsy-based transformer models and their ensembles with surgical counterparts for OS prediction, illustrating the cross-cohort consistency of the observed trends.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

These results underscore a key limitation of the current pipeline: under the existing framework, biopsy-derived features do not meaningfully enhance survival prediction compared to models based solely on post-NAC surgical specimens. Although ensembling biopsy and surgical predictions yielded modest improvements in internal validation, these gains did not generalise, as evidenced by weaker performance in the external validation cohort.

This suggests that the biopsy models fail to capture prognostically relevant features in a form that complements post-treatment data. Further work is needed to optimise feature extraction from diagnostic biopsies, potentially through task-specific fine-tuning, domain adaptation, or the integration of richer biological information, such as inferred molecular signatures or treatment response proxies.

5.3.2 Predict the chances of having a Residual Disease using the diagnostic biopsy

Although post-NAC surgical specimens reflect treatment response and offer prognostic value, they don't fully capture tumour biology. For instance, two patients with similar RD volumes may have different outcomes due to varying intrinsic chemosensitivity.

To address this, we introduce a deep learning-derived score estimating chemosensitivity from diagnostic biopsies. While biopsies alone have limited predictive power for OS/DFS (**see Section 5.3.1**), they offer crucial insight into tumour biology before treatment begins.

Pathological complete response (pCR), defined as the absence of invasive cancer in the breast and lymph nodes, is a known surrogate for improved survival, especially in HER2+ and TNBC subtypes. In these cases, patients with RD often benefit from adjuvant intensification (e.g., KATHERINE, CREATE-X trials[5,33]). For ER+/HER2− patients, pCR is less predictive, but identifying poor responders early could help avoid unnecessary toxicity, an idea explored in trials like CORALLEEN and NEOPAL.

We developed chemo prAIdict Breast, a deep learning model that predicts pCR probability from pre-treatment H&E biopsy slides. Trained to detect subtle, morphology-based markers of chemosensitivity, it provides a foundational tool for incorporating biological response into survival modeling. This section details the model's architecture, training, and performance, as described in our *European Journal of Cancer* paper[170].

Moreover, this work sets up the next section, where we test whether adding the biopsy-derived pCR score to post-NAC OS/DFS models improves prediction, potentially revealing intrinsic resistance or sensitivity not captured in surgical specimens alone.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset and study design

All patients were selected from the PRIMUNEO and CGFL Neoadj cohorts. A comprehensive description of the dataset composition, preprocessing steps, and experimental setup is provided in the **Chapter 2 early Breast Cancer dataset section**.

Pipeline

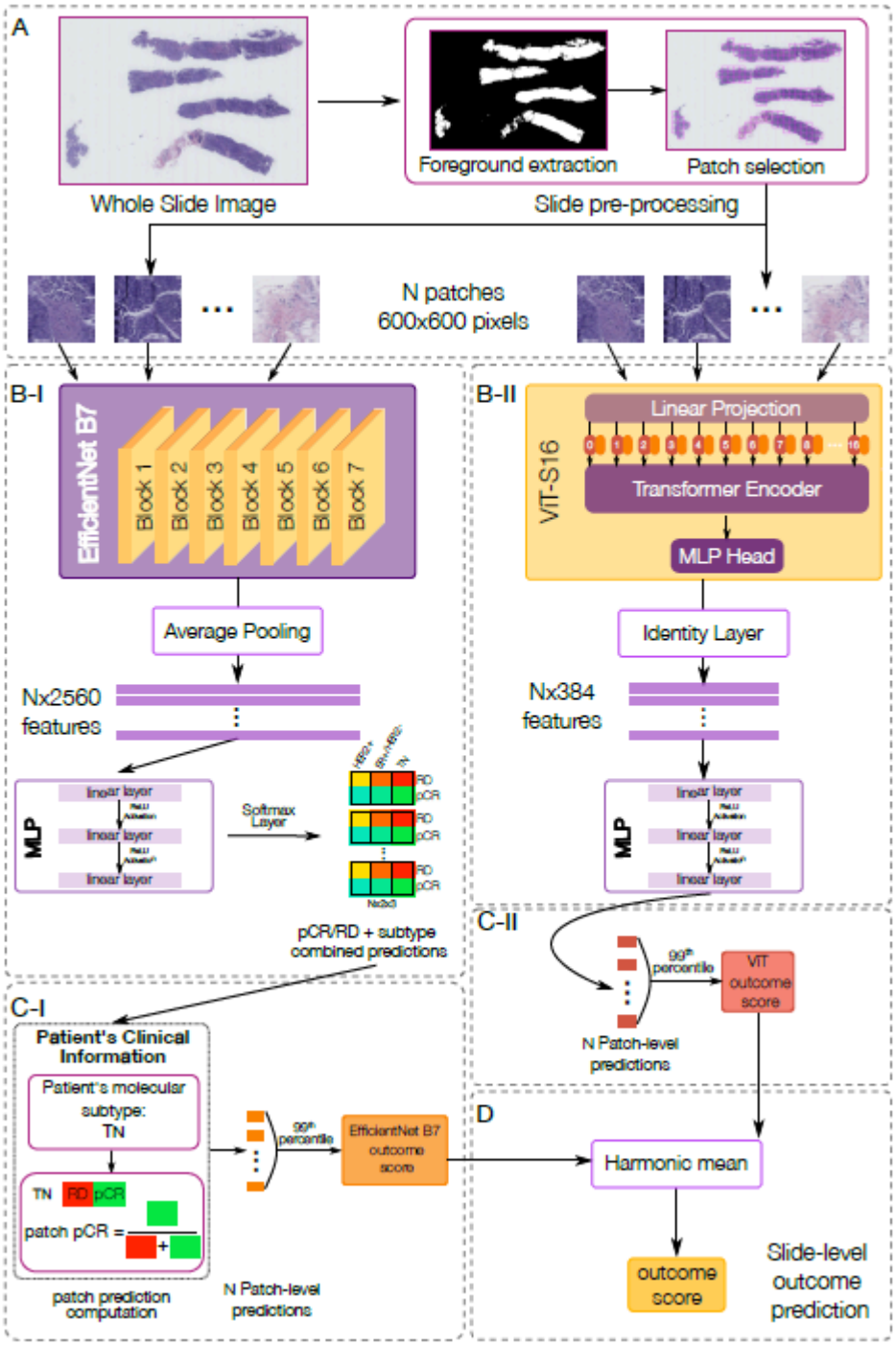


Figure 78. Pipeline Overview. Chemo-prAIdict Breast predicts patient response to neoadjuvant therapy using Whole Slide Images (WSIs). Our approach: a) Extracting the foreground from WSIs and dividing the tissue into non-overlapping patches. b-i) These patches are processed by an EfficientNetB7 neural network, pre-trained on ImageNet, to generate 2560-dimensional feature embeddings. c-i) A Multilayer Perceptron (MLP) analyzes these features, incorporating the patient's molecular subtype. Patch-level predictions for pathological complete response (pCR) or residual disease (RD), along with the molecular subtype, are made using a SoftMax layer. The combined score is normalized, ensuring proper weighting based on clinical information. b-ii) Simultaneously, the same patches are input into a Vision Transformer (ViT) neural network, pre-trained with Dino, to compute 384-dimensional feature embeddings. c-ii) An MLP then predicts the patch-level score. The final slide-level predictions are derived by taking the 99th percentile of the patch-level scores for each model. d) The results from both the EfficientNet and ViT models are then combined using the harmonic mean to generate the overall patient outcome (pCR/RD) prediction.

- Image preprocessing:

Hamamatsu NDPI files of H&E diagnostic biopsy slides from the PRIMUNEO and CGFL databases were first selected. A single slide was used for each patient for performance evaluation. We then extracted the foreground using an in-house trained U-net and tiled the images in non-overlapping patches of 600x600 pixels at a 5x resolution. The resulting patches were used as inputs of two separate neural network architectures (**Figure 78a**). For the first architecture, we employed EfficientNetB7 with ImageNet-pretrained weights, adding a global average pooling layer to produce an embedding vector of size 2560 (**Figure 78b-i**). For the second architecture, we used the Vision Transformer Small with 16 patches (ViT-S/16) with weights from the Self Supervised Learning (SSL) DINO method pre-trained on the TCGA dataset **(Figure 78b-ii**).

- Label processing:

These patches were associated with the label 0 if RD was observed on the surgical specimen, or 1 if pCR was described.

- Neural network training, model selection:

We developed a deep learning model based on two architectures to predict RD from the WSI of biopsies.

Architecture 1: EfficientNet B7-Based Model. The first architecture employs the EfficientNet B7 embeddings as inputs for a multi-layer perceptron (MLP). This MLP consists of two fully connected layers with output dimensions of 64 and 16, each followed by a ReLU activation layer. Then, we used a single SoftMax layer to jointly predict the patient's pCR or RD status and molecular subtype (**Figure 78b-i**). Lastly, the slide-level prediction was calculated using the 99th percentile of the patch-level prediction values (**Figure 78c-i**). A Multiple Instance Learning (MIL) aggregation was also tried but did not provide competitive results.

Architecture 2: Vision Transformer-Based Model. The second architecture utilizes ViT-S/16 patch embeddings as inputs for an MLP, also consisting of two fully connected layers with output dimensions of 64 and 16, each followed by a ReLU activation layer. A linear layer predicts the pCR or RD status. The slide-level prediction is calculated using the 99th percentile of the patch-level prediction values (**Figure 78c-ii**). A Multiple Instance Learning (MIL) aggregation was also tried but did not provide competitive results.

Models from both architectures were trained using a nested three-fold cross-validation approach. Each training set was divided into three subsets: two subsets were used for training a model, while the third subset served as the validation set. This process was repeated three times, each time using a different subset for validation, resulting in three trained models per architecture. We used the validation subset to select the training epoch with the best performance. The final slide-level pCR prediction was obtained by first averaging the predictions from the three models within each architecture. Next, we computed the harmonic mean of the ensemble predictions across both architectures (**Figure 78d and Figure 79)**.

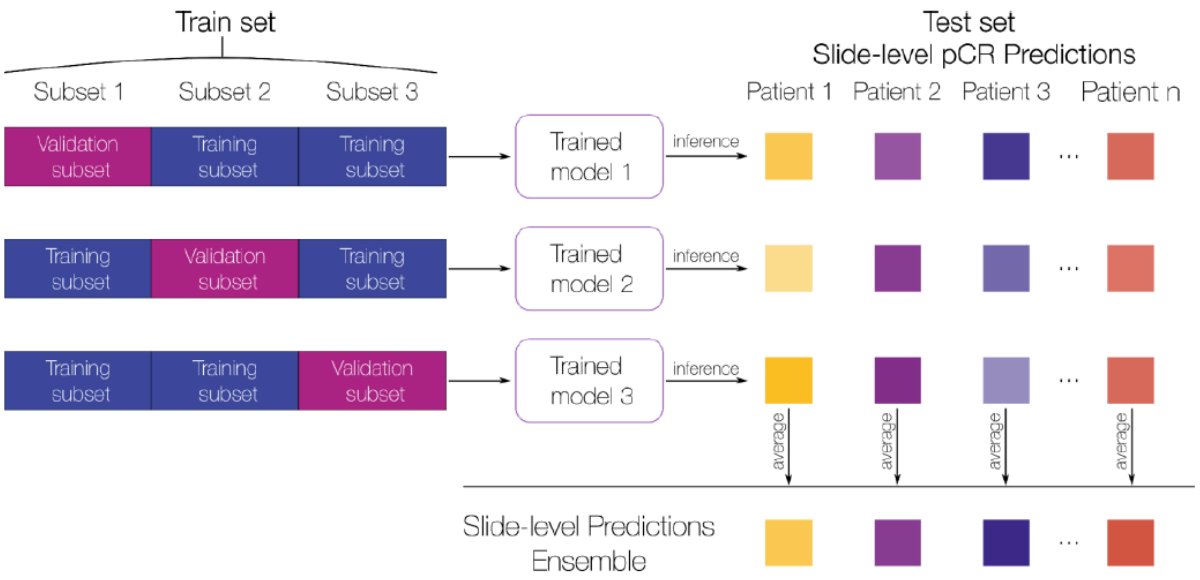


Figure 79. Three-Fold Cross-Validation training strategy. Schematic representation of the three-fold cross-validation setup used for model training and evaluation. In each fold, one subset of the training cohort is used for validation while the remaining two are used for training. Each trained model then performs inference on the independent test set, generating slide-level predictions per patient. Final ensemble predictions are obtained by averaging the outputs from the three folds to improve robustness and reduce variance.

Regarding further specifications, we trained each model for 5 epochs and a batch size of 32. For the EfficientNet B7-Based model we employed a cross-entropy loss and SAM wrapper with an Adam optimizer as the underlying optimizer and a base learning rate of 1e−3. For the Vision Transformer-Based Model, we employed a binary cross-entropy loss with an Adam optimizer and a learning rate of 1e−3.

- Clinical Model:

To provide a baseline for comparison with our deep learning model from WSIs, we developed a Clinical Model (CM) based on the pathological and clinical information. This model utilized the following data: age at surgery (ranging from 22 to 88 years), the time difference between biopsy and surgery, the menopausal status (2 classes: premenopausal or postmenopausal), the Ki67 (separated in 3 classes: ≤14%, 15% to 29%, and ≥30%), the mitotic index (separated in 3 classes: 0 to 6 mitoses, 7 to 12 mitoses, and > 12 mitoses), AJCC staging (6 classes: I, IIA, IIB, IIIA, IIIB, IIIC), the TNM tumour stage (5 classes: T0, T1, T2, T3, T4), the TNM node stage (4 classes: N0, N1, N2, N3), Nottingham Combined Histologic Grade (3 classes: 1, 2, 3) information, the molecular subtype (4 classes: ER+/HER2+,

ER+/HER2-, ER-/HER2+, and TN breast tumours (ER-/HER2-)), the HER2 IHC results (4 classes: 0, 1+, 2+, 3+), the estrogen receptors expression (2 classes: negative or positive), and the progesterone receptors expression (2 classes: negative or positive). If the data was available for the patient, it was encoded in a one-hot vector; otherwise, a vector of zeros was used. The resulting 40-feature vector was used as an input for an MLP with two hidden layers of output dimensions 64 and 16. We then used the same approach as described earlier to predict pCR/RD. To ensure a fair comparison between the models, the CM employed the same implementation details and training curriculum as detailed above.

- Output binarisation:

We binarized the predictions using different thresholds for each molecular subtype corresponding to the highest median of the positive labeled predictions in the validation subsets of the internal study.

- Hardware and software specifications:

Experiments were run with a NVIDIA RTX A4000 graphic card and the following libraries: PyTorch v1.12.1, CUDA 11.5.

Statistics and Metrics

Clinical and pathological characteristics were compared between cohorts using the Chi-squared test or two-sided Fisher’s exact test. Discriminatory power was measured using the area under the receiver operating characteristic curve (AUC) and statistical significance was assessed using a one-sided Mann–Whitney U test. To measure the strength of the association between the binarized predictions and actual pCR/RD, we used the odds ratio (OR) with the Haldane-Anscombe correction, and a two-sided Fisher's Exact test for statistical analysis. The models were systematically evaluated separately on each molecular subtype to avoid biological bias.

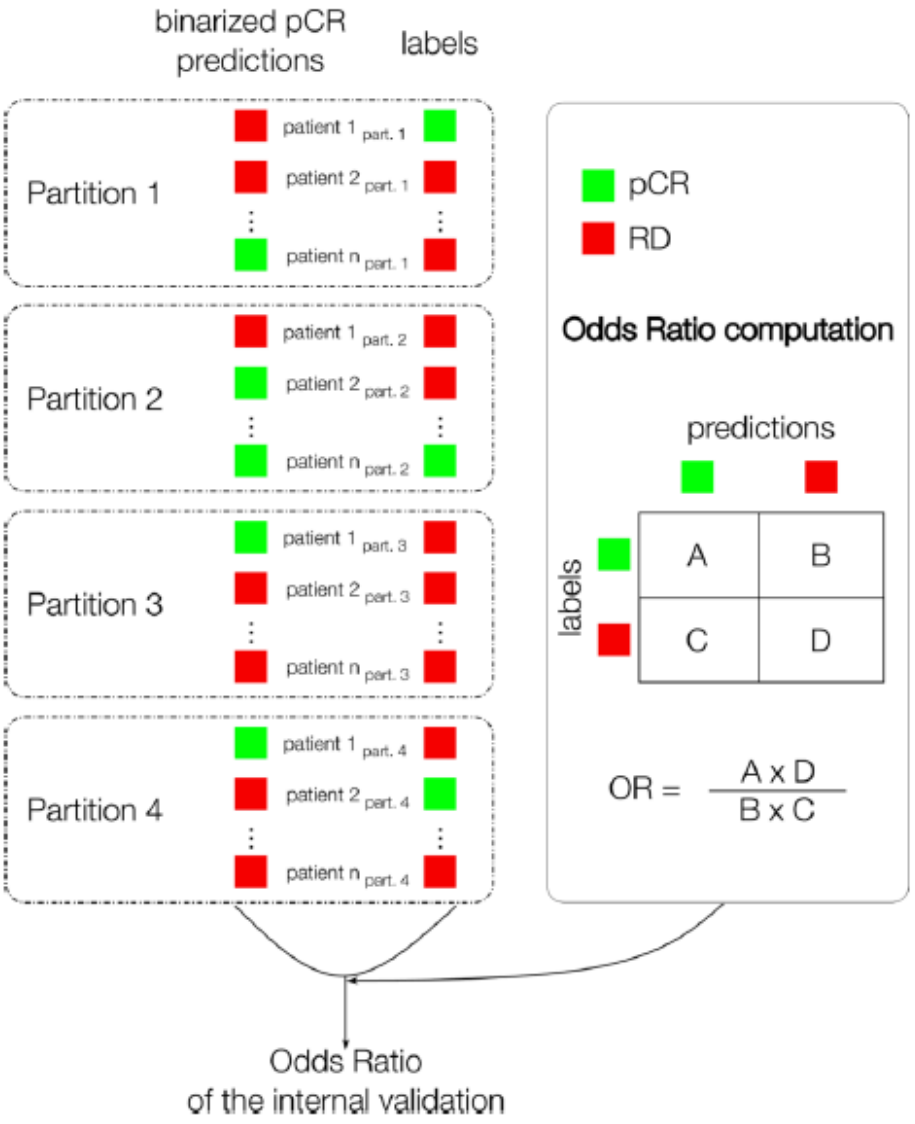


Figure 80. Ensemble prediction pipeline for internal validation and Odds Ratio (OR) computation. Each model trained on one internal partition generates binarized patient-level predictions for pathological complete response (pCR) and residual disease (RD). Predictions from all partitions are aggregated to form the full internal validation ensemble. The Odds Ratio (OR) metric is then computed from the aggregated contingency table, quantifying the association between predicted and true response outcomes.

For the internal study, we computed the AUC for each partition, while for the OR, we concatenated the predictions of each partition to obtain a single value (**Figure 80**). This approach was taken to mitigate the potential bias in the OR estimates due to the small sample size and low prevalence rates for certain molecular subtypes.

RESULTS

Pathological Complete Response is predictable by a Deep Learning System using WSI on HER2+, ER+/HER2- and TN Breast initial diagnostic cancer biopsies.

Internal validation study (PRIMUNEO cohort)

We introduce Chemo-prAIdict Breast, a deep learning model that uses WSI as input to predict pCR. The predictive capacity of Chemo-prAIdict Breast (see **METHODOLOGY section**) was systematically compared to a clinical model (CM) based on clinicopathological information (menopausal status, Ki67, mitotic index, tumour molecular subtype, AJCC staging, and Nottingham Combined Histologic Grade information). Internal validation was performed with cross-validation on the PRIMUNEO dataset. Comparison of the performance of the Chemo-prAIdict Breast and CM in terms of AUC are shown in **Figure 81** and for OR in **Table 23**.

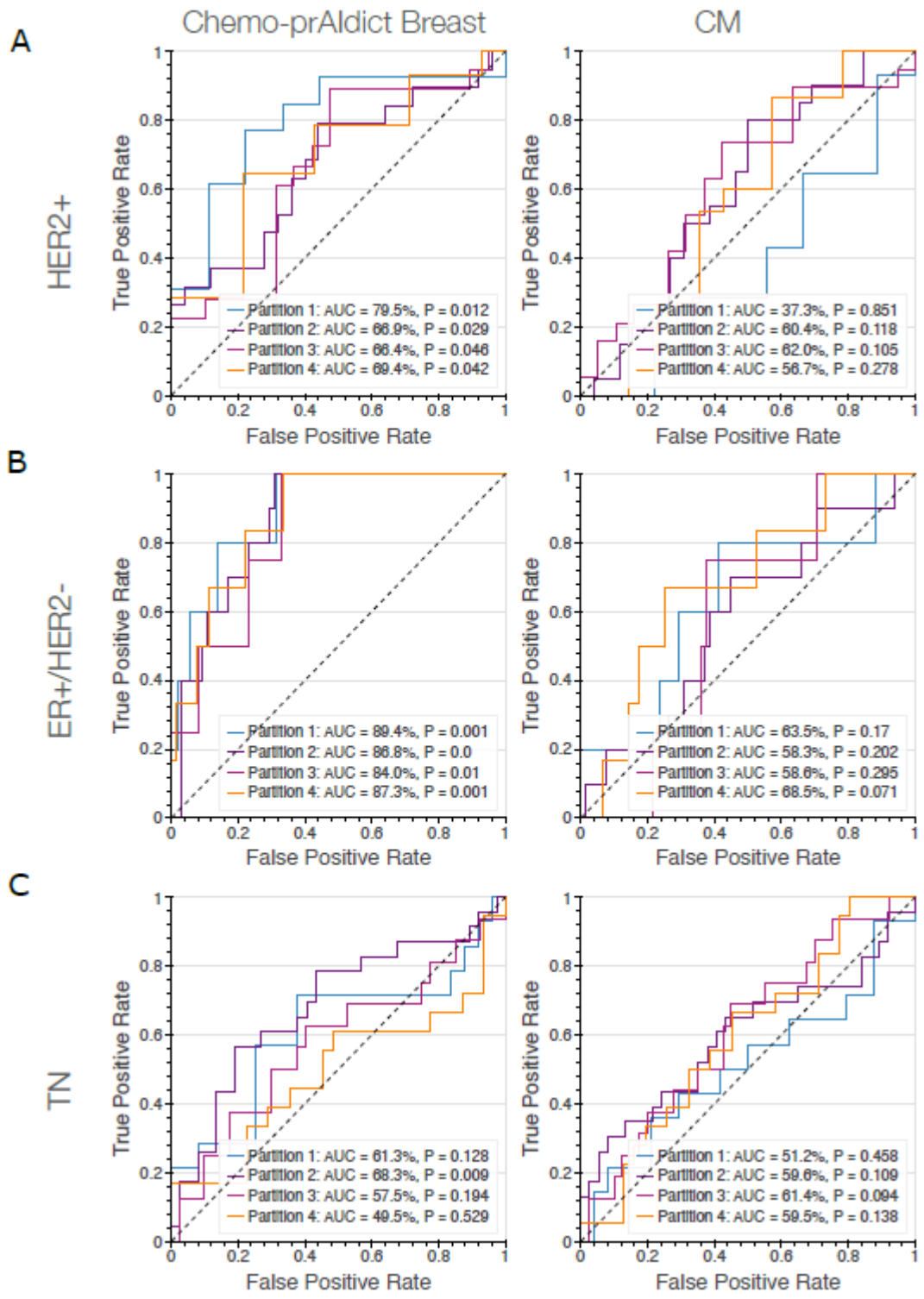


Figure 81. Comparison of AUC between Chemo-prAIdict Breast and Clinical model (CM) in internal validation. Receiver operating characteristic (ROC) curves illustrate the performance of Chemo-prAIdict Breast (left) and CM (right) across various partitions for a) HER2+, b) ER+/HER2- and c) TN breast cancer molecular subtypes. Legends in the figures display the Area Under the Curve (AUC) and p-value (P).

Validation Setup	Methods	Chemo-prAIdict Breast		Clinical Model (CM)	
		OR (95 % CI)	*P*	OR (95 % CI)	*P*
Internal	HER2+	4.44 (2.11 - 9.38)	8.8e-05	1.15 (0.55-2.41)	0.851
	ER+/HER2-	72.53 (4.36-1205.43)	1.36e-09	2.16 (0.90-5.20)	0.0942
	TN	2.22 (1.23-4.00)	0.00831	2.09 (1.11-3.91)	0.023
External	HER2+	2.70 (1.08-6.76)	0.0358	2.04 (0.85-4.89)	0.131
	ER+/HER2-	20.56 (1.14-371.74)	0.00413	2.41 (0.43-13.57)	0.416
	TN	3.02 (1.18-7.74)	0.0206	1.38 (0.52-3.66)	0.608

Table 23. Comparison of ORs between Chemo-prAIdict Breast and Clinical model (CM).

- Prediction of pCR in HER2+ subtype:

Figure 81a shows the results for patients with HER2+ tumours, where Chemo-prAIdict Breast achieved AUCs from 66.4% to 79.5% (worse to best partition). The CM achieved AUCs from 37.3% to 62.0%. Using the positivity threshold defined in the Methods section, Chemo-prAIdict Breast achieved an OR of 4.44 (95% CI 2.11 - 9.38, p<0.0001), whereas the CM achieved an OR of 1.15 (95% CI 0.55 - 2.41, P=0.851) (**Table 23**). Note that Chemo-prAIdict Breast achieved an AUC from 0.60 to 0.89 on ER-/HER2+ subgroup, and an AUC from 0.59 to 0.73 on ER+/HER2+ subgroup.

- Prediction of pCR in ER+/HER2- subtype:

As shown in **Figure 81b** for the luminal breast cancer subtype (ER+/HER2-), Chemo-prAIdict Breast achieved AUCs from 84.0% to 89.4%. In contrast, the CM achieved AUCs from 58.3% to 68.5%. Chemo-prAIdict Breast achieved an OR of 72.53 (95% CI 4.36 - 1205.43, P <0.0001), whereas the CM achieved an OR of 2.16 (95% CI 0.90 - 5.20, P=0.0942) (**Table 23**).

- Prediction of pCR in the Triple Negative Breast cancer subtype:

For patients with TN tumours (**Figure 81c**), Chemo-prAIdict Breast achieved AUCs from 49.5% to 68.3%. The CM achieved AUCs from 51.2% to 61.4%. However, considering OR instead of AUC, Chemo-prAIdict Breast achieved 2.22 (95% CI 1.23 - 4.00, P=0.008) compared to 2.09 (95% CI 1.11 - 3.91, P=0.023) for the CM (**Table 23**).

Overall, the Chemo-prAIdict Breast showed better performance in terms of both AUC and OR than the Clinical Model results, demonstrating the relevance of our approach to predict pCR versus RD in HER2+, ER+/HER2- and TN molecular subtypes.

External validation of Chemo-prAIdict Breast on the CGFL Breast Cancer Neoadjuvant database

To validate the prediction performance of Chemo-prAIdict Breast, we conducted an external validation on the CGFL breast cancer neoadjuvant database for a total of 490 patients. We used these thresholds for binarization (HER2+ = 0.38, ER+/HER2- = 0.41, TN = 0.50), following the binarization procedure described in the **METHODOLOGY section**.

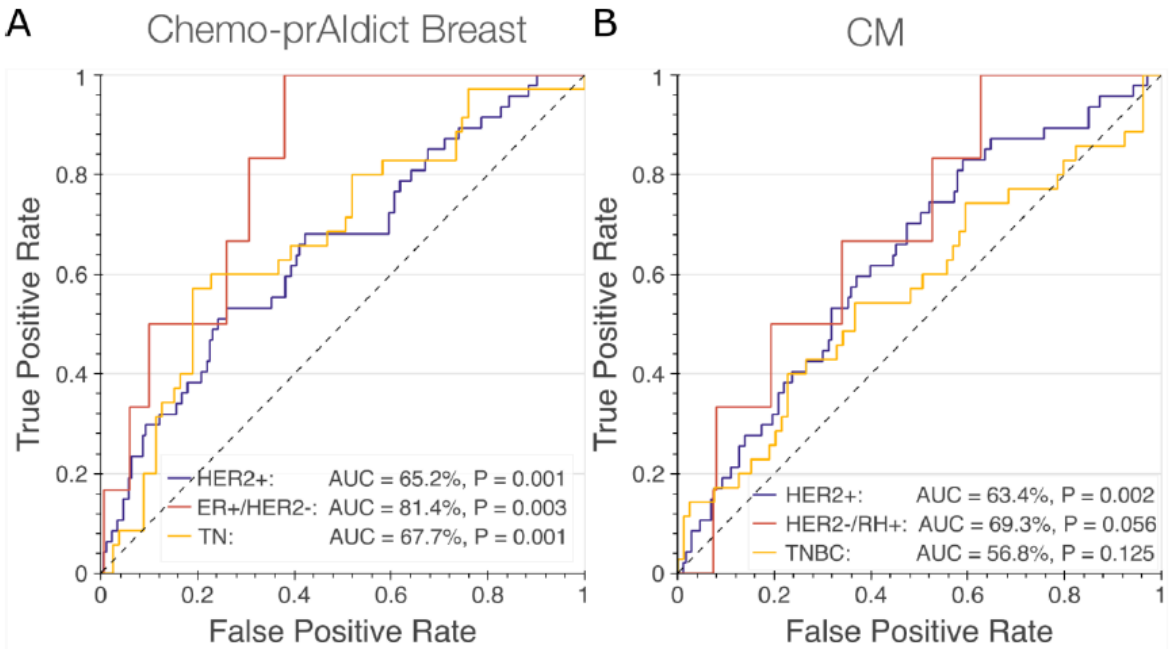


Figure 82. Chemo-prAIdict Breast vs Clinical model (CM) AUC performance comparison in the external validation. The ROC curves illustrate the performance of a) Chemo-prAIdict Breast and b) CM across the different molecular subtypes of tumours.

Legends in the figures provide details on the Area Under the Curve (AUC) and associated p-values (P).

Chemo-prAIdict Breast (**Figure 82a**) demonstrated high performance in predicting pCR with an AUC of 65.2% (P=0.001) and an OR of 2.70 (95% CI 1.08-6.76, P=0.0358, **Table 23**) for HER2+ tumours. Note that taken separately, ER-/HER2+ obtained an AUC of 0.58 (P=0.126) and ER+/HER2+ an AUC of 0.67 (P=0.013), suggesting that most of the discriminative effect is independent of the ER status for HER2+ tumours. Chemo-prAIdict Breast obtained an AUC of 81.4% (P=0.003) and an OR of 20.56 (95% CI 1.14-371.74, P=0.00413, **Table 24**) for ER+/HER2- tumours, and an AUC of 67.7% (P=0.001) and an OR of 3.02 (95% CI 1.18-7.74, P=0.0206) for TN tumours (**Figure 82a, Table 23**). Although the number of pCR in the ER+/HER2- subgroup is only 6, a p-value of 0.004 suggests that the effect is likely to be strong. Chemo-prAIdict Breast predicts well the RD with a PPV of 0.829, 0.891 and 1.00 for TN, HER2+ and ER+/HER2-,respectively. Other metrics such as sensitivity, specificity, NPV are provided in **Table 24**.

Metrics	Specificity	Sensitivity	NPV	PPV
HER2+	0.872	0.283	0.248	0.891
ER+/HER2-	1	0.567	0.084	1
TN	0.800	0.430	0.383	0.829

Table 24. Chemo-prAIdict Breast RD prediction performance is measured with specificity, sensitivity (recall), NPV negative predictive value and PPV positive predictive value (precision) on external validation for HER2+, ER+/HER2- and TN breast cancer molecular subtypes.

As shown in **Figure 82b** and **Table 23**, the CM achieved an AUC of 63.4% (P=0.002) and an OR of 2.04 (95% CI 0.85 - 4.89, P=0.131) for HER2+ tumours, an AUC of 69.3% (P=0.056) and an OR of 2.41 (95% CI 0.43 - 13.57, P=0.416) for ER+/HER2- tumours, and an AUC of 56.8% (P=0.125) and an OR of 1.38 (95% CI 0.52 - 3.66, P=0.608) for TN subtypes. These results demonstrate that Chemo-prAIdict Breast, with a very strong PPV, is very effective in

identifying the most chemoresistant tumours in patients for all molecular subtypes, which will result in the presence of RD after chemotherapy. As patient from the external validation has varying chemotherapy regimen (i.e. Taxanes, Anthracyclines or both Taxanes + Anthracyclines), we also evaluated the predictive performance for each regimen when it was possible (**Figure 83**). We found no major effect of the chemotherapy regimen on the pCR/RD predictability.

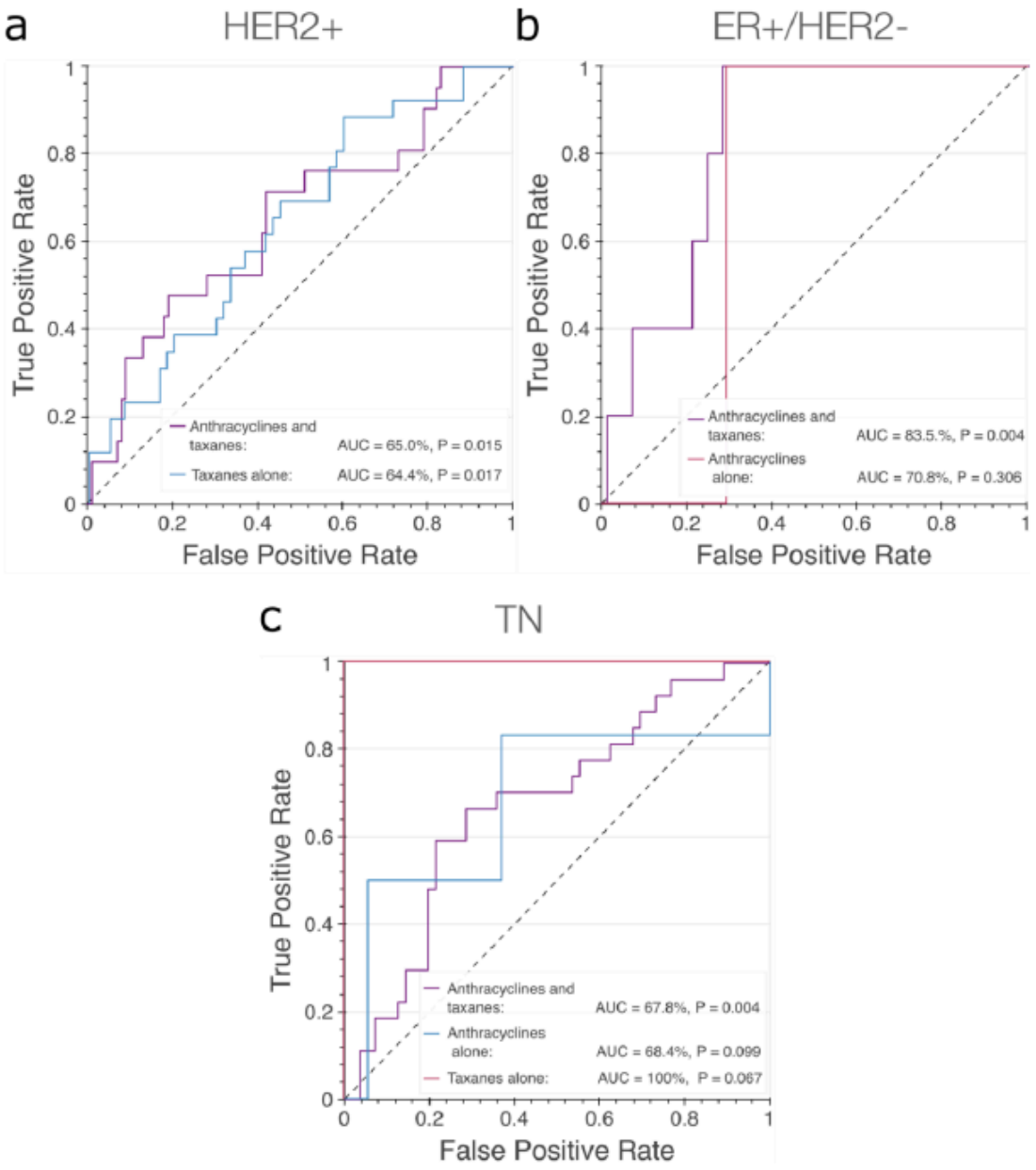


Figure 83. External validation AUC stratified by regimen treatment. AUC in patients with A) HER2+, B) ER+/HER2- and C) TN tumours in the external validation study, stratifying patients by the received treatment (Taxanes, Anthracyclines, or both).

Additionally, we highlight the advantages of combining the EfficientNet B7-based model with the ViT-S/16-based model. We present the results of the individual models in the external validation in **Table 25**, showing that averaging the predictions from both models enhances performance across all molecular subtypes.

Methods	EfficientNet B7-based model		ViT-S/16-based model		Chemo-prAIdict Breast	
	AUC	*P*	AUC	*P*	AUC	*P*
HER2+	0.632	0.003	0.599	0.019	0.652	0.001
ER+/HER2-	0.776	0.010	0.746	0.02	0.814	0.003
TN	0.661	0.003	0.647	0.006	0.677	0.001

Table 25. Ablation study on the ensemble of the EfficientNet B7-based model and the ViT-S/16-based model in the external validation. AUC in patients with HER2+, ER+/HER2-, and TN tumours is shown.

Aggregation of the Chemo-prAIdict Breast Model and the Clinical Model predictions show inconsistent improvement for all tumours.

We next tested whether combining Chemo-prAIdict Breast and CM by averaging their slide-level prediction could lead to performance improvement, hereafter referred to as Chemo-prAIdict Breast+CM. We compared the Chemo-prAIdict Breast+CM model to the Chemo-prAIdict Breast alone and found consistent improvement for the TN molecular subtype only in the internal validation.

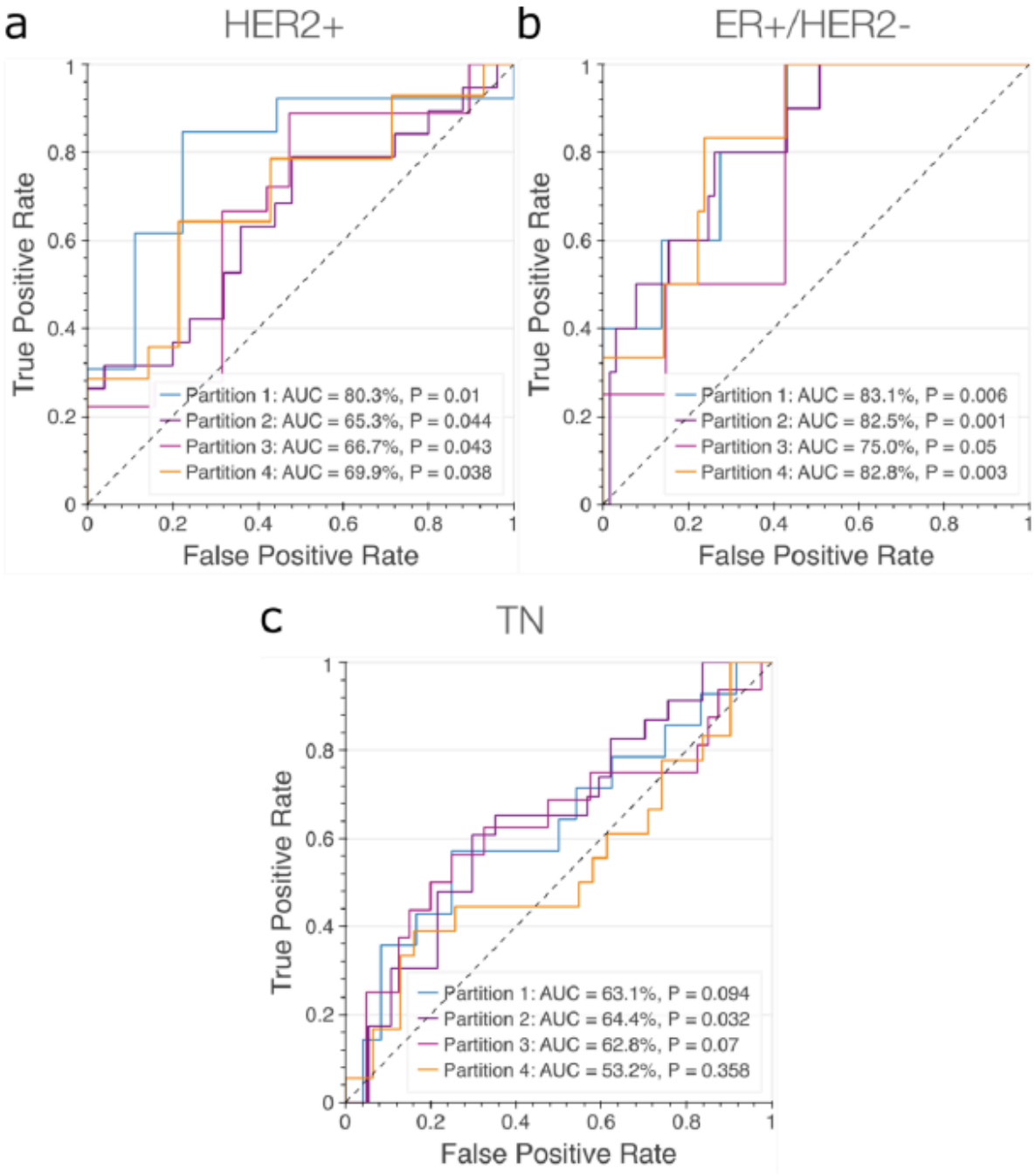


Figure 84. WSIM + CM AUC in patients with a) HER2+, b) ER+/HER2- and c) TN tumours in the internal validation study. The patients' outcome predictions are obtained by averaging Chemo-prAIdict Breast and CM predictions.

When applied to the mentioned strategy, the Chemo-prAIdict Breast+CM showed improvement in 3 out of 4 partitions over the Chemo-prAIdict Breast alone (**Figure 84 vs Figure 81**). On the external validation, the performance decreased in 3.1 points (AUC of 64.6%, P=0.005) compared to the Chemo-prAIdict Breast alone (AUC of 67.7%, P=0.001) (**Figure 85**).

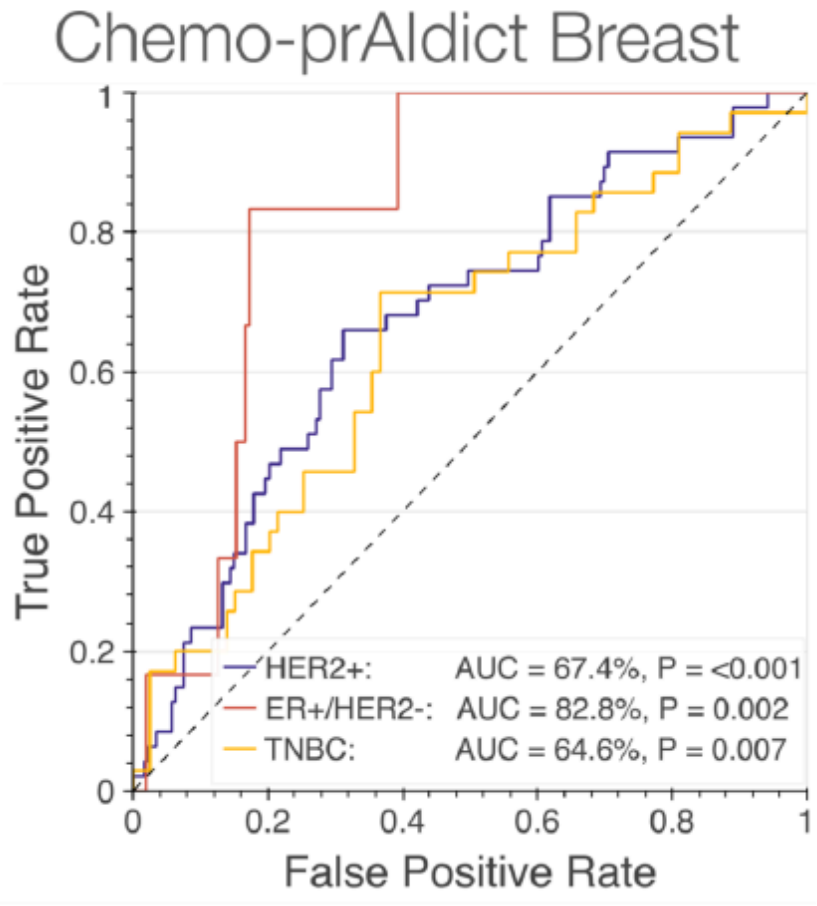


Figure 85. WSIM + CM AUC in patients with HER2+, ER+/HER2-, and TN tumours in the external validation study. The patients' outcome predictions are obtained by averaging Chemo-prAIdict Breast and CM predictions.

However, results are mixed for HER2+ and ER+/HER2- subtypes where internal validation tends to show no improvement (**Figure 84**), whereas external validation shows 2.2 and 1.4 points increase when compared with the Chemo-prAIdict Breast alone for HER2+ and ER+/HER2- subtypes respectively (**Figure 85**). These findings suggest that while the Chemo-prAIdict Breast Model is effective using only Whole Slide Image information for the evaluated molecular subtypes, incorporating clinical information might not always benefit the patient's outcome prediction.

Deciphering Intrinsic Chemo-Sensitivity with Chemo-prAIdict Breast

To evaluate whether the predictions generated by Chemo-prAIdict Breast reflect true biological chemo-sensitivity, as opposed to simply reproducing a binary pCR label, we performed a more granular analysis of tumour response across multiple clinical categories. While pCR is formally defined as the absence of residual invasive cancer (pT0N0), this binary classification fails to account for degrees of partial response and includes biologically heterogeneous cases. In real-world cohorts, many patients do not achieve complete response yet still experience substantial tumour regression and long-term benefit from NAC. Conversely, some patients may achieve only minimal response, reflecting intrinsic chemoresistance. Disentangling this biological variability is essential if pCR scores are to serve as meaningful surrogates for integration into survival prediction models.

To address this, we stratified patients into three groups based on pathological response: non-responders, partial responders, and complete responders (as detailed in the **Chapter 5. METHODS, Pathological evaluation** section). We then analysed the distribution of Chemo-prAIdict Breast scores across these categories in the external validation cohort, as a post analytical study.

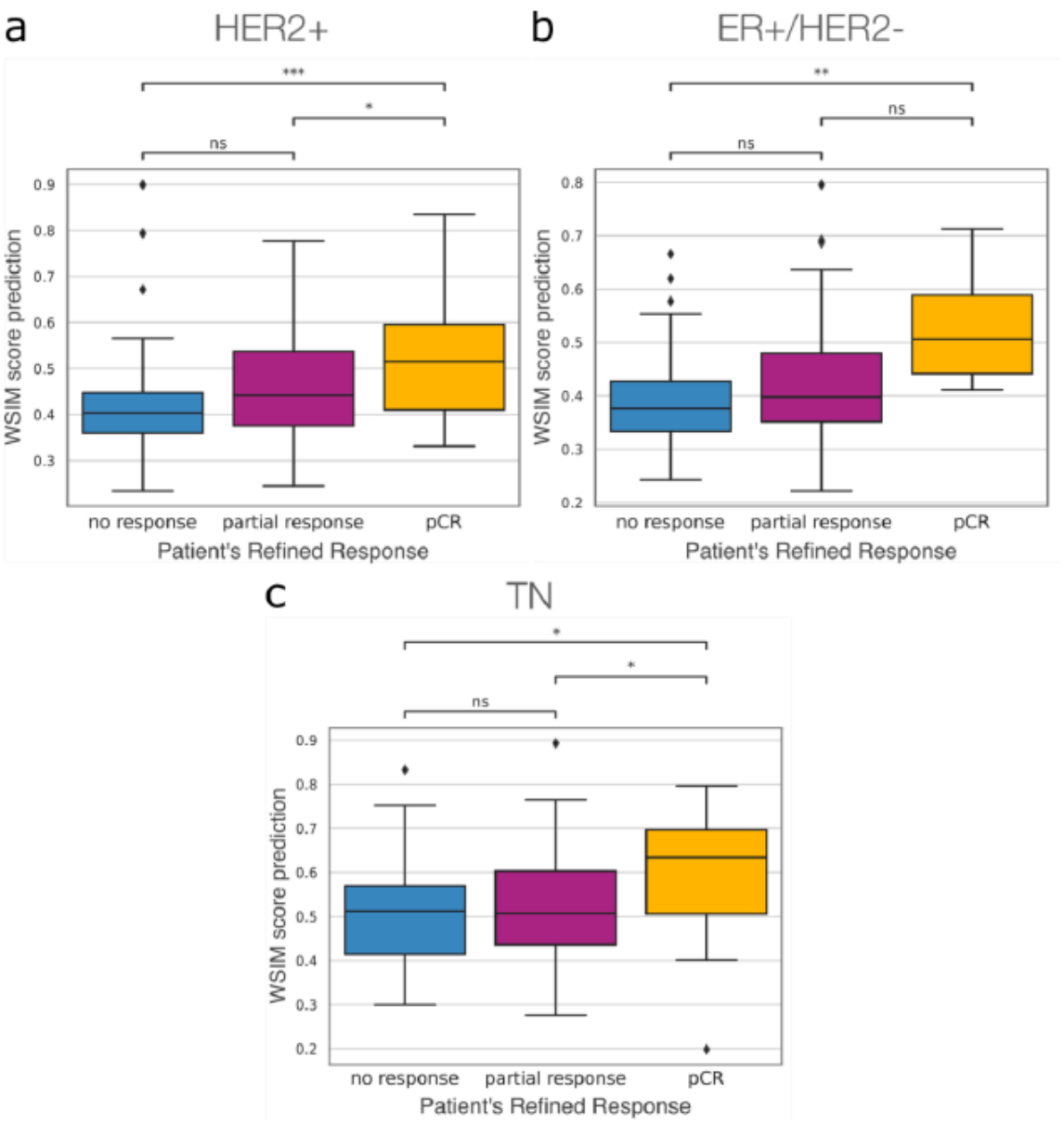


Figure 86. Comparison of Chemo-prAIdict Breast score prediction distribution within non-responder, partial responder, and complete responder patients. a) HER2+, b) ER+/HER2- and c) TN tumours. The one-sided Mann-Whitney-Wilcoxon test was used to compare the response categories. ns: 0.05 < p-value <= 1; *: 0.01 < p-value <= 0.05; **: 0.001 < p-value <= 0.01, ***: 0.0001 < p <= 0.001.

As shown in **Figure 86**, Chemo-prAIdict Breast scores were significantly associated with the degree of clinical response across all three molecular subtypes: HER2+, ER+/HER2−, and TNBC. Specifically, patients with complete response had significantly higher prediction scores than those with partial response, and partial responders had higher scores than non-responders, although this latter difference was more variable. The trend suggests that the

model is not only learning to distinguish pCR in a binary sense, but also capturing a biologically continuous representation of chemo-sensitivity, embedded in tumour morphology..

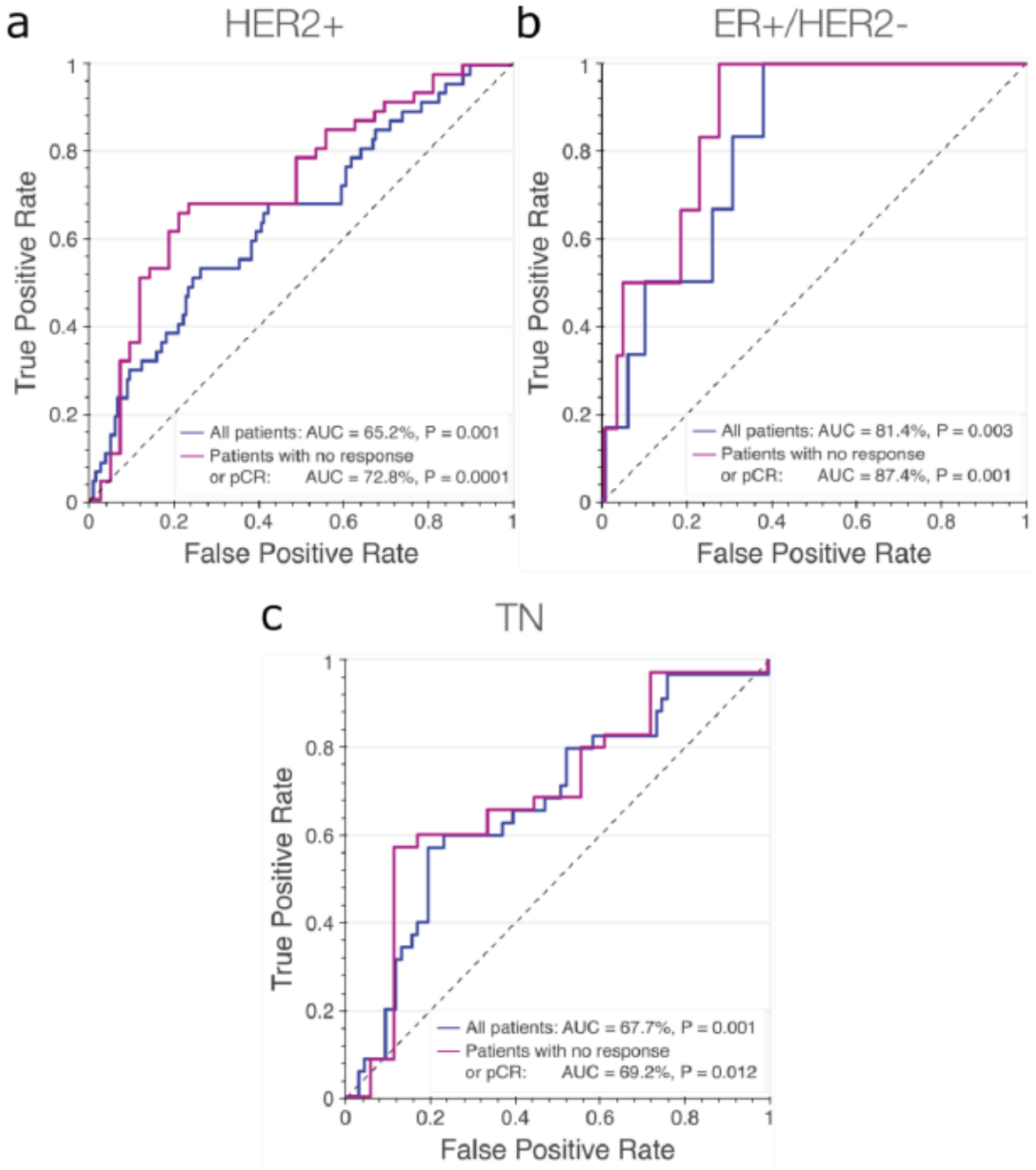


Figure 87. Evaluation of Chemo-prAIdict Breast outcome prediction performance in all patients compared to those who did not respond or exhibited complete response to Neoadjuvant Chemotherapy (excluding patients with partial response defined in the refined labels). We show the results for patients with a) HER2+, b) ER+/HER2- and c) TN tumours in external validation. Legends in the figures provide details on the Area Under the Curve (AUC) and associated p-values (P).

This was further validated in **Figure 87**, where we evaluated model performance (AUC) using only the most biologically distinct classes: complete responders versus non-responders, excluding partial responders. In this stricter dichotomy, Chemo-prAIdict Breast demonstrated higher discriminative performance across all subtypes, supporting the model's capacity to identify underlying chemo-sensitivity more robustly when extremes of response are considered.

Interestingly, the score distributions between partial responders and non-responders were less separable, which may reflect either overlapping biological characteristics or the fact that the model was not specifically trained to discriminate within this intermediate class. Nonetheless, the ability of the model to distinguish complete responders from the other groups, particularly when prediction scores trend with increasing pathological regression, suggests that Chemo-prAIdict Breast captures clinically and biologically meaningful differences in treatment response potential.

These findings form the conceptual foundation for the next phase of our work. If the pCR score reflects intrinsic tumour chemo-sensitivity, then incorporating it into post-NAC survival prediction could enhance the identification of high-risk residual disease—not solely based on morphological features of the surgical specimen, but informed by the tumour's baseline biological behaviour. In the next section, we assess this hypothesis by integrating Chemo-prAIdict Breast scores into our OS and DFS prediction models based on surgical WSIs.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

In this section, we introduced Chemo-prAIdict Breast, a deep learning pipeline that predicts pathological complete response (pCR) from diagnostic biopsy slides in early breast cancer. The model consistently outperformed a clinical model across HER2+, ER+/HER2−, and TNBC subtypes, with particularly strong results in the challenging ER+/HER2− group.

Beyond binary pCR classification, our deep learning model captured a continuum of treatment response, correlating prediction scores with non-, partial, and complete responders.

This indicates the model's ability to detect subtle, biologically relevant features of chemo-sensitivity from histology alone.

Importantly, the model's score offers more than a proxy for pCR. It also provides a morphology-derived estimate of intrinsic treatment response. This insight lays the groundwork for the next section, where we integrate biopsy-derived chemo-sensitivity with post-NAC histology to refine survival prediction and better stratify residual disease.

5.3.3 Combining morpho-molecular correlates and direct OS and DFS prediction pipeline

In this section, we investigate whether integrating morpho-molecular features into our survival prediction pipeline could improve risk stratification in early breast cancer patients with residual disease after NAC. The underlying hypothesis is that inferred morpho-molecular correlates could provide complementary information and enhance the model's ability to predict patient outcomes.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset and study design

This experiment included all patients' post-NAC surgical specimen WSIs sourced from the PRIMUNEO and CGFL Neoadj cohorts. It also utilises the TCGA BRCA dataset to train the HRD and PAM50 predictors, as described in Chapter 4.3.2. A full description of dataset composition and preprocessing can be found in **Chapter 2 and chapter 4.2**.

Methodology

We trained two separate methods for predicting HRD and PAM50, as detailed in **Chapter 4.3.2**. For each method, we generated three models using cross-validation. These models were then applied to the surgical specimen's whole slides images from the PRIMUNEO and CGFL datasets, and the resulting HRD and PAM50 prediction scores were saved. For the survival prediction tasks, we employed the same methodology described in as defined in **Chapter 5.3.1.1**, with the modification of concatenating the predicted HRD and PAM50 scores as additional features to the reduced embeddings (the output of the MLP's N-1 layer). In the experimental section, we assess the optimal approach for concatenating these features. Notably, for HRD scores, we have three values (one from each cross-validation model), while for PAM50, we predict four classes, resulting in a total of 12 output values.

We average the WSI-level predictions of the 3 models trained during the cross-validation. Then, we evaluate with the mean AUC metric, C-index Harrell and the metric used in DeepHit, C-t index, and the weighted versions of the AUC and C-t-index metrics. For the OS

task we evaluate from years 3 to 5, and for the DFS task from years 2 to 5, and show the average of these results. We also compute the p-value of the AUC and weighted AUC at each year using a one-sided Mann-Whitney U test.

RESULTS

In the following set of experiments, we evaluate different options for aggregating the HRD and PAM50 previously computed outputs. In **Figures 88, 89, 92 and 93** we show the best method to this point, which uses the CoxPH as the survival method, using as input the mid-task reduced ViT-S16 embeddings, which come from image information only. We systematically compare with concatenating 1) all the HRD and PAM50 raw features, 2) a mean ensemble within the cross-validation models (1 value for HRD, and 4 for PAM50), and 3) adding the mean value for HRD and PAM50, but first running the PAM50 predictions through a softmax operation in the internal and external validation.

DFS

Figures 88 and 89 present our current optimal results, with a weighted AUC of 69.2% and a Harrell's C-Index of 65.8% for image information alone. These figures also compare these results with those obtained by concatenating various features derived from HRD and PAM50 predictions using a previous method. We observe that in most cases, incorporating these features enhances the initial performance. Specifically, either concatenating the raw values or using an ensemble of a single HRD value and a 4-class score for PAM50 improves performance by 0.3% to 0.4% in the internal validation and by 1.2% to 1.4% in the external validation. Figures 3 and 4 provide a detailed breakdown of the performance by subtype for the Mean Ensemble HRD and PAM50 concatenation method in both internal and external validation setups.

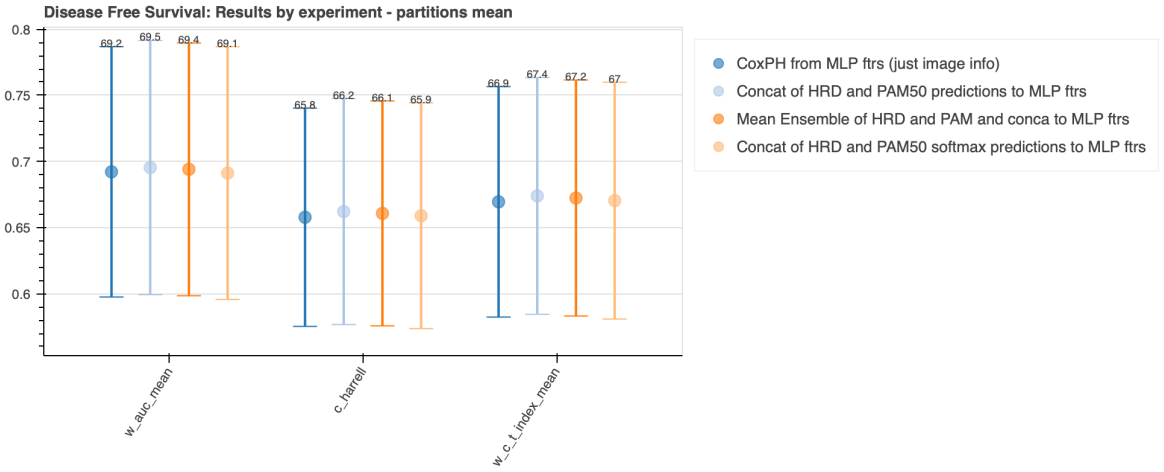


Figure 88. Disease-Free Survival (DFS) performance in the internal validation. Comparison of survival models using image-derived features alone versus models augmented with homologous recombination deficiency (HRD) and PAM50 subtype information. Different aggregation strategies for HRD and PAM50 outputs are assessed within the internal validation setup.

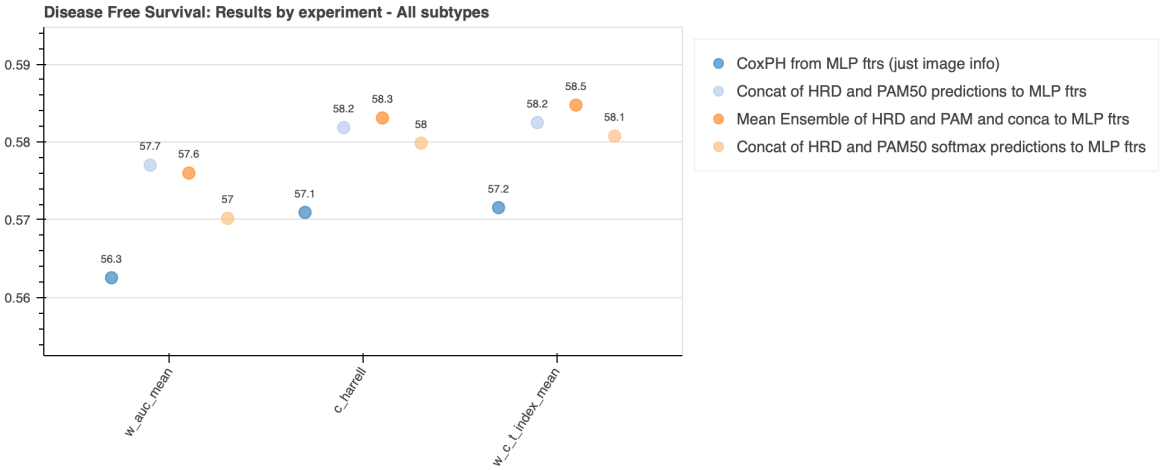


Figure 89. Disease-Free Survival (DFS) performance in the external validation. Evaluation of the same HRD and PAM50 aggregation strategies in the external validation cohort, comparing their generalisation performance against the image-only baseline.

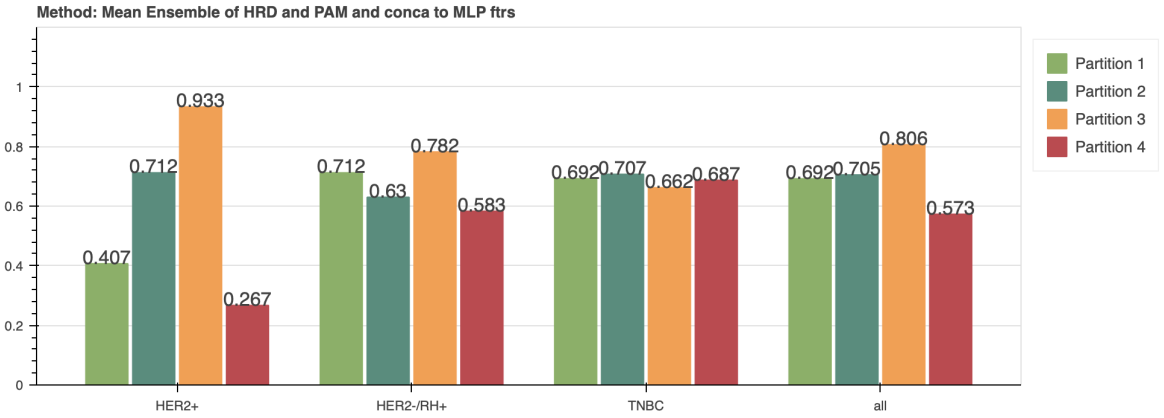


Figure 90. Subtype-level DFS performance of the Mean Ensemble HRD and PAM50 concatenation approach in the internal validation. Weighted AUC results are stratified by molecular subtype and partition, highlighting intra-cohort variability in model performance.

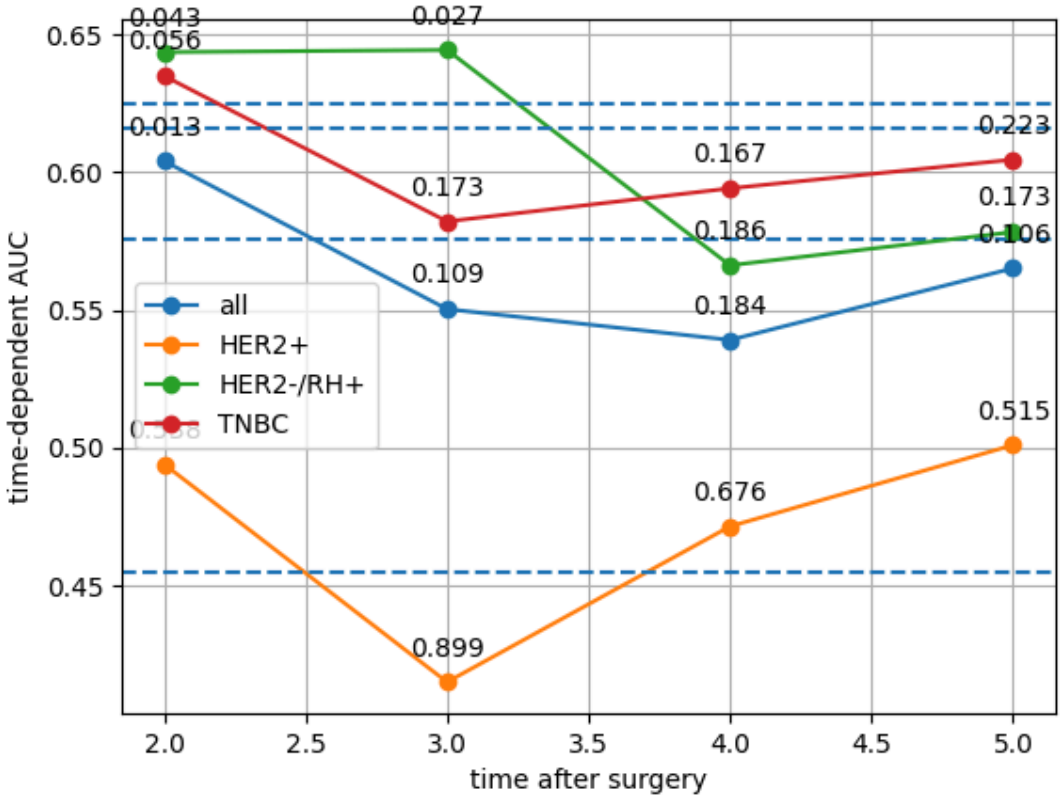


Figure 91. Subtype-level DFS performance of the Mean Ensemble HRD and PAM50 concatenation approach in the external validation. Time-dependent weighted AUC curves computed for each molecular subtype, illustrating the evolution of model discrimination capability across post-surgical years.

OS

As in the DFS subsection, **Figures 92 and 93** highlight our best current results, with the image-only model achieving a weighted AUC of 73.9% and a Harrell's C-Index of 71.7%. Incorporating the averaged HRD score and the 4-class PAM50 classification consistently enhances performance—by 0.5% to 0.8% in internal validation and by 1.5% to 2.9% in external validation. Figures 6 and 7 further break down performance by subtype using the Mean Ensemble method for HRD and PAM50 concatenation. Notably, the HER2-/HR+ subtype shows stable performance between 60% and 70% across both validation sets. In contrast, other subtypes that perform very well in internal validation show a drop of around 20% in AUC when evaluated externally, highlighting challenges in model generalisation across cohorts.

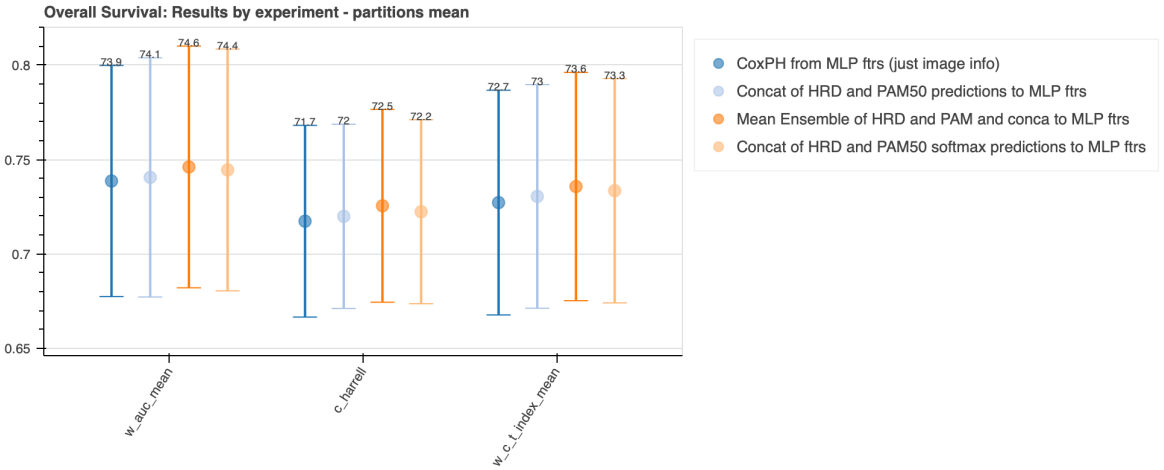


Figure 92. Overall Survival (OS) performance in the internal validation. Comparison between image-only models and those incorporating concatenated HRD and PAM50-derived features using multiple aggregation schemes in the internal validation setup.

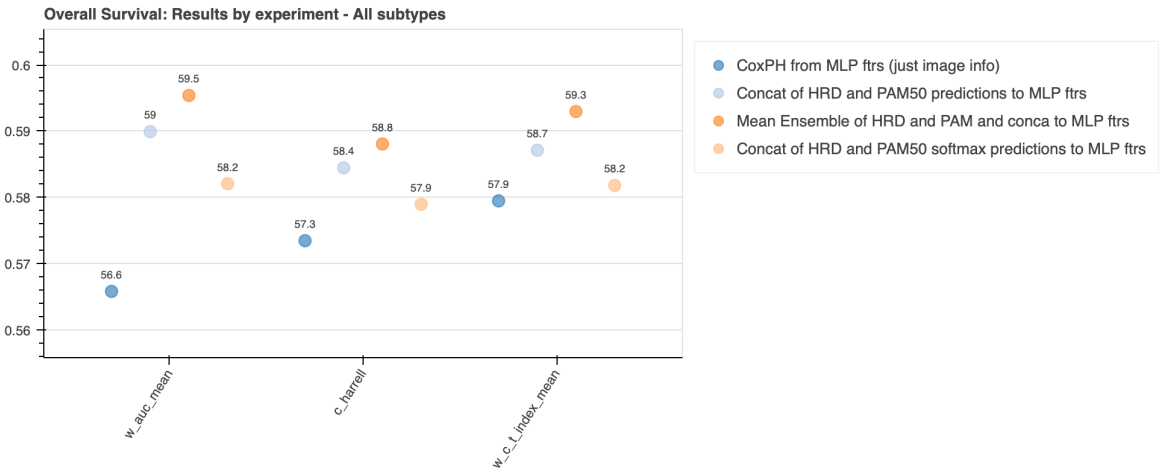


Figure 93. Overall Survival (OS) performance in the external validation. External validation of models integrating HRD and PAM50 features, compared to the image-only baseline across the main survival evaluation metrics.

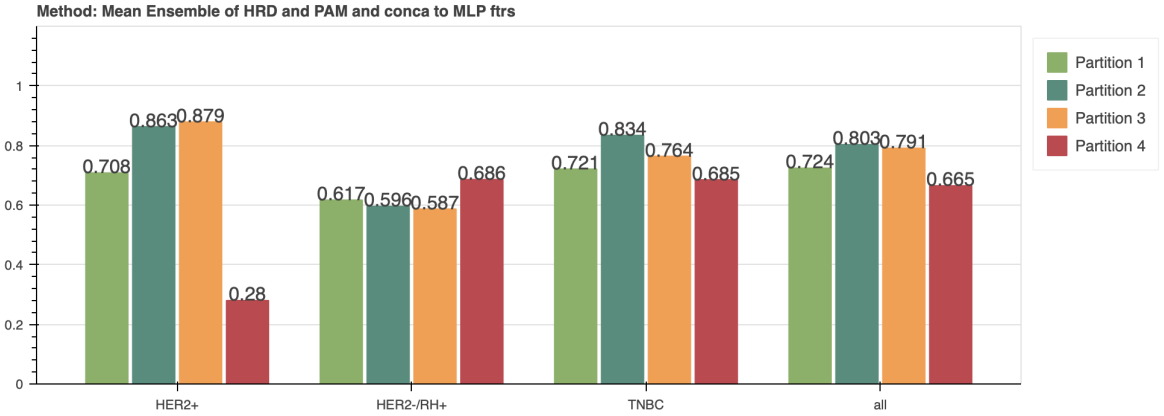


Figure 94. Subtype-level OS performance of the Mean Ensemble HRD and PAM50 concatenation approach in the internal validation. Weighted AUC scores are stratified by molecular subtype and internal partition, assessing intra-cohort heterogeneity.

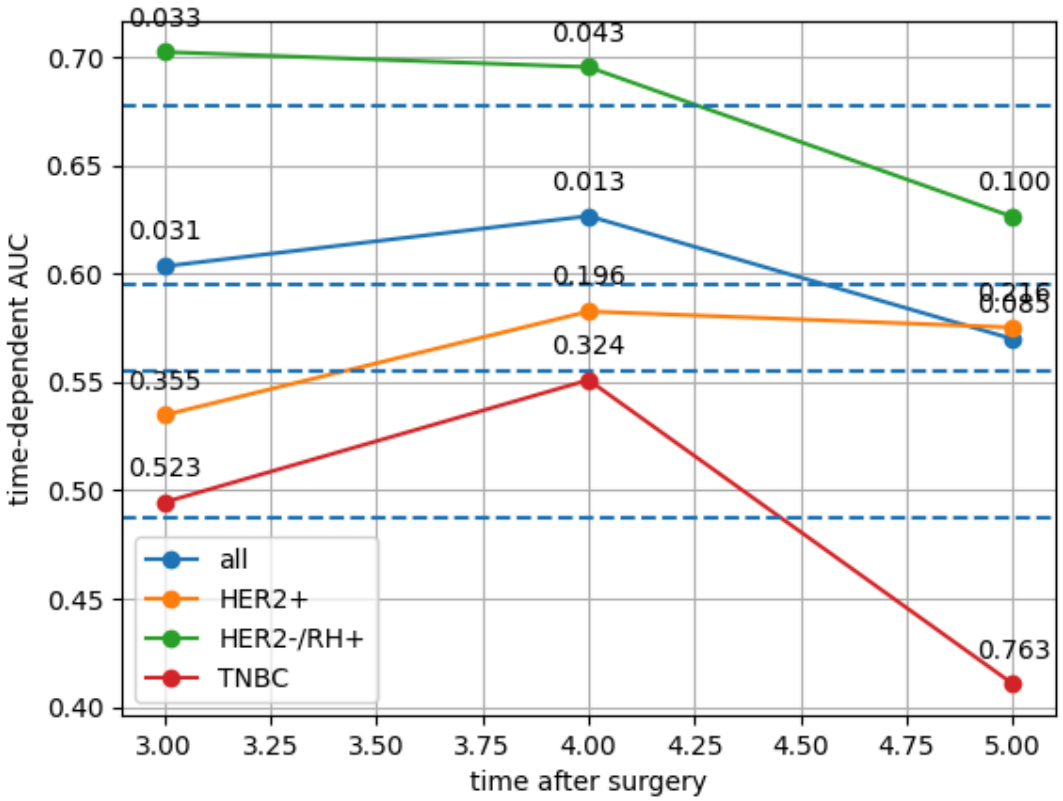


Figure 95. Subtype-level OS performance of the Mean Ensemble HRD and PAM50 concatenation approach in the external validation. Time-dependent weighted AUC results

by molecular subtype, illustrating model generalisation and survival discrimination over follow-up time.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

Integrating the aggregated predictions from the HRD and PAM50 models, which were trained on TCGA data and applied to the PRIMUNEO and CGFL patient cohorts, into the Whole Slide Image (WSI) embedding information resulted in a consistent performance enhancement for both Overall Survival (OS) and Disease-Free Survival (DFS) tasks in internal and external validations. While a modest improvement was observed, there remains significant room for further enhancement, particularly in external validation, where performance currently exceeds approximately 60%.

5.3.4 Combining chemosensitivity score and post-NAC OS/DFS prediction pipeline

In this experiment, we evaluated whether incorporating the chemo-sensitivity score predicted by Chemo-prAIdict Breast could improve the prediction of overall survival (OS) and disease-free survival (DFS) from post-NAC surgical specimens. The rationale stems from previous findings showing that the pCR prediction score captures intrinsic tumour response potential, even in cases where residual disease remains (**Chapter 5.3.2**). We hypothesised that integrating this biologically informative score into the post-surgical survival pipeline could improve prognostic accuracy, in a manner comparable to integrating inferred molecular signatures such as PAM50 and HRD (**Chapter 5.3.3)**.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset and study design

This experiment included all patients for whom both diagnostic biopsy and post-NAC surgical specimen WSIs were available, sourced from the PRIMUNEO and CGFL Neoadj cohorts. A full description of dataset composition and preprocessing can be found in **Chapter 2**.

Methodology

As described in **Chapter 5.3.2**, we generated pCR prediction scores for each patient using the Chemo-prAIdict Breast model trained on biopsy WSIs, following the external validation setup described previously. For inference, we used the complete ensemble of 18 model weights: 9 based on EfficientNetB7 and 9 based on ViT-S/16, corresponding to three independent 3-fold cross-validation runs per architecture. The average output across runs produced two pCR-related features per patient (one per architecture).

These pCR scores, derived solely from biopsy histology, were incorporated into the surgical WSI-based survival pipeline by appending them to the intermediate 64-dimensional feature vector generated by the MLP (as defined in **Chapter 5.3.1.1**). All survival models were

trained using the same hyperparameter settings and evaluation metrics as in previous sections (as defined in **Chapter 5.3.1.1**).

RESULTS

Internal Validation

Figures 96 and 97 show that adding the pCR score to the WSIM + HRD + PAM50 baseline model led to consistent performance gains in internal validation for both DFS and OS. This improvement was most pronounced in the HER2+ subtype (**Figures 98 and 99**), where pCR is strongly associated with long-term outcomes and is a known surrogate for survival.

For HER2−/HR+ and TNBC subtypes, the added value of the pCR score was less consistent. In certain partitions, particularly where class imbalance was more pronounced, performance showed minor reductions compared to the WSIM + HRD + PAM50 model alone. These findings suggest that while the pCR prediction captures relevant biological signals in HER2+ disease, it may be less discriminative in other subtypes, possibly due to greater intratumoural heterogeneity or the lower predictive value of pCR itself in luminal tumours.

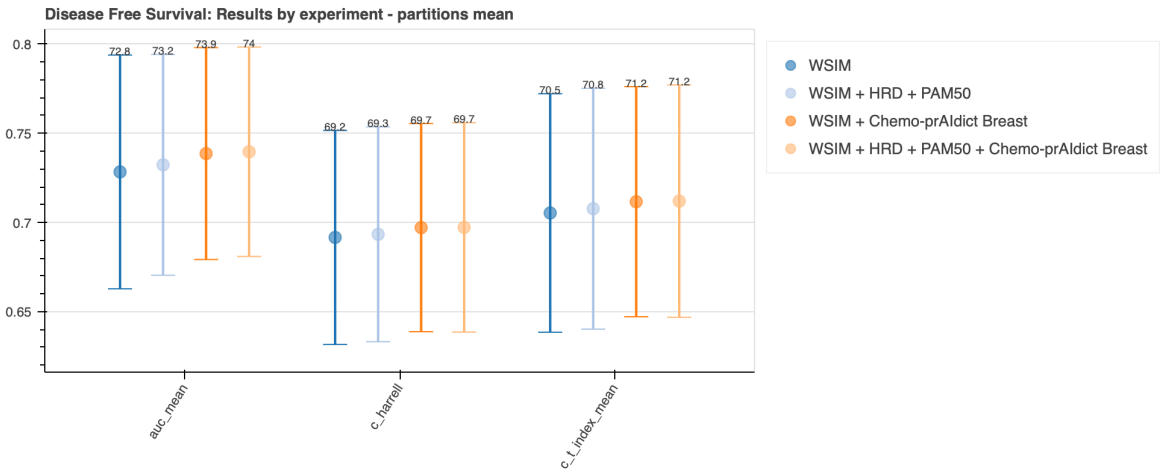


Figure 96. Comparison in the DFS performance in AUC, C-index Harrell and C-t-index in the internal validation.

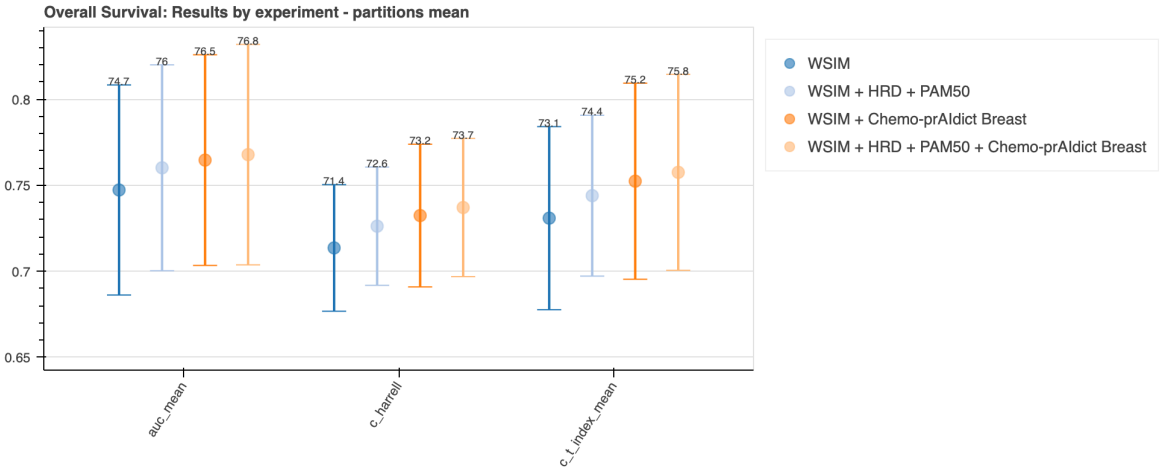


Figure 97. Comparison in the OS performance in AUC, C-index Harrell and C-t-index in the internal validation.

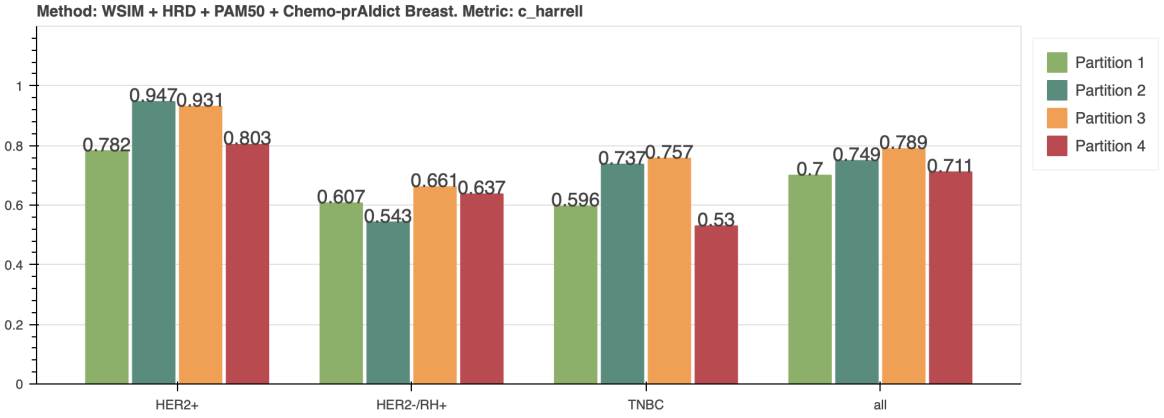


Figure 98. WSIM + HRD + PAM50 + Chemo-prAIdict Breast detailed performance in each molecular subtype and partition of the internal validation measured with the C-index Harrell in the DFS task.

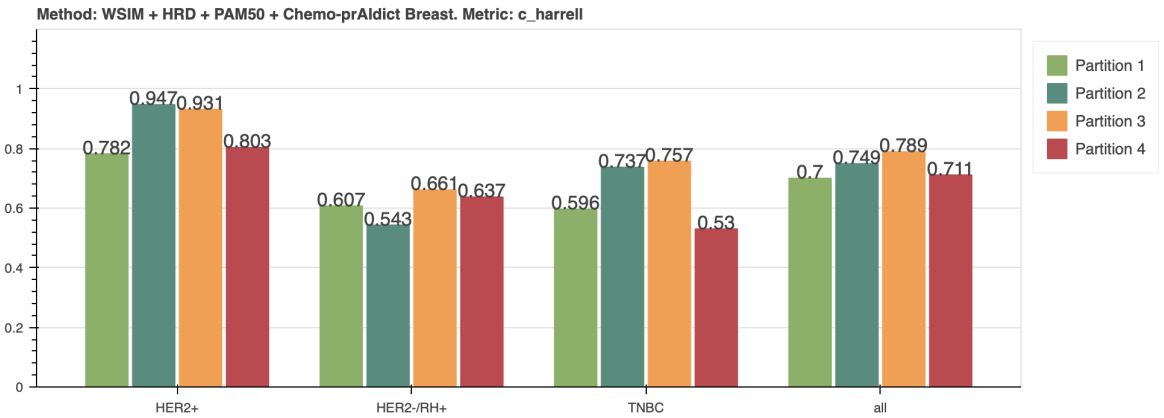


Figure 99. WSIM + HRD + PAM50 + Chemo-prAIdict Breast detailed performance in each molecular subtype and partition of the internal validation measured with the C-index Harrell in the OS task.

External Validation

In contrast to internal results, the external validation (**Figures 100 and 101**) revealed a marked drop in performance across all models that included the pCR score. While the pCR-augmented model modestly outperformed WSIM alone for DFS in a few partitions (**Figure 100**), it consistently underperformed in OS prediction (**Figure 101**). Subtype-specific evaluations (**Figures 102–105**) further showed that this performance decline was evident across most molecular subtypes and timepoints.

This suggests that the predictive value of the pCR score did not generalise across cohorts. One possible explanation lies in the potential domain shift between biopsy processing and clinical protocols across centres, as well as the model's reliance on histological features that may not transfer seamlessly outside the PRIMUNEO training domain.

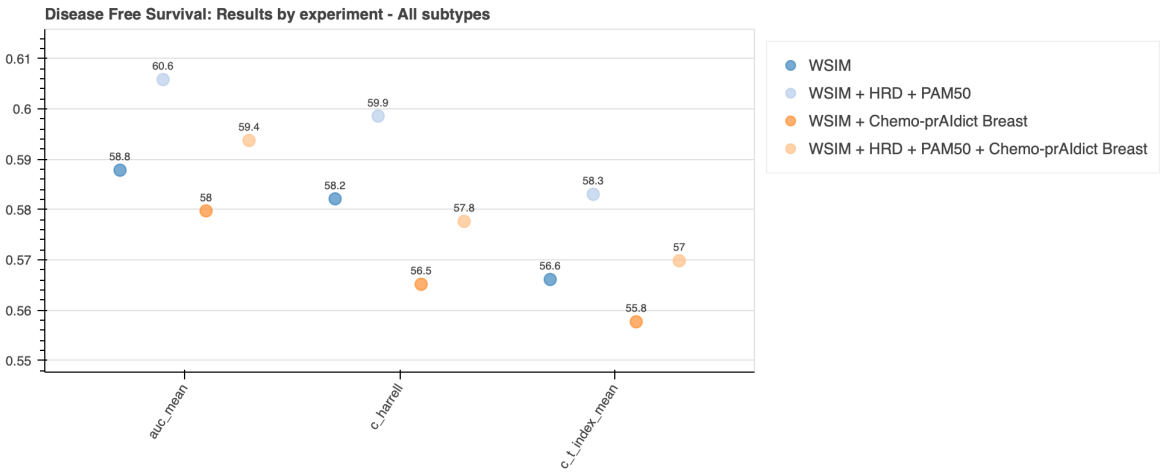


Figure 100. Comparison in the DFS performance in AUC, C-index Harrell and C-t-index in the external validation.

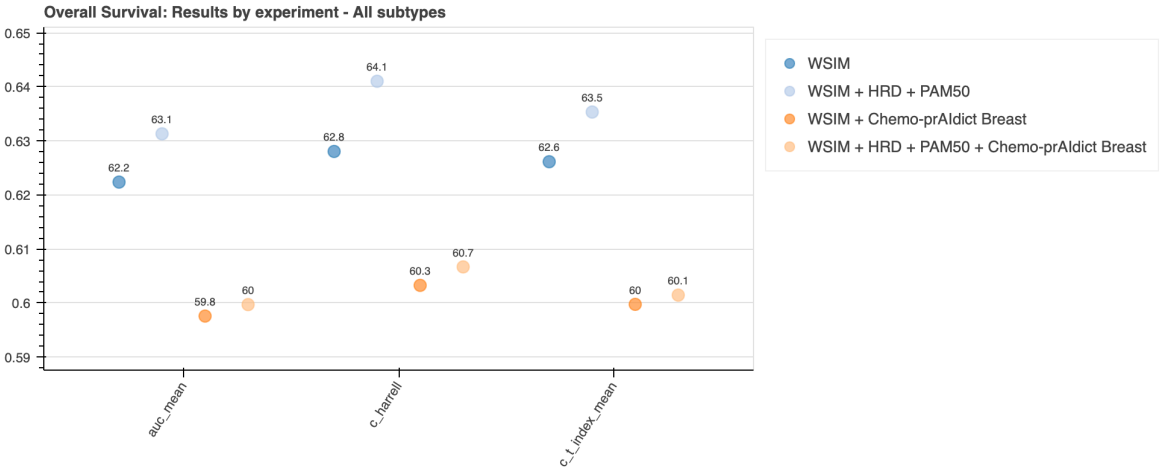


Figure 101. Comparison in the OS performance in AUC, C-index Harrell and C-t-index in the external validation.

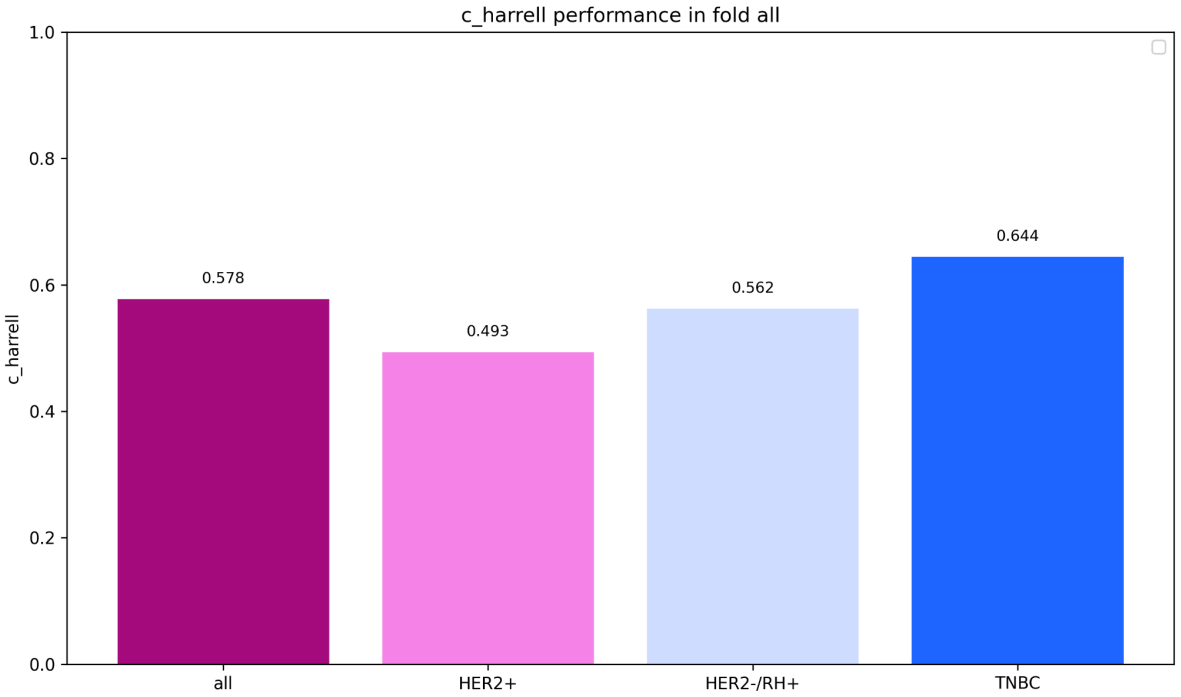


Figure 102. WSIM + HRD + PAM50 + Chemo-prAIdict Breast detailed performance in the external validation for each molecular subtype measured with the C-index Harrell in the DFS task.

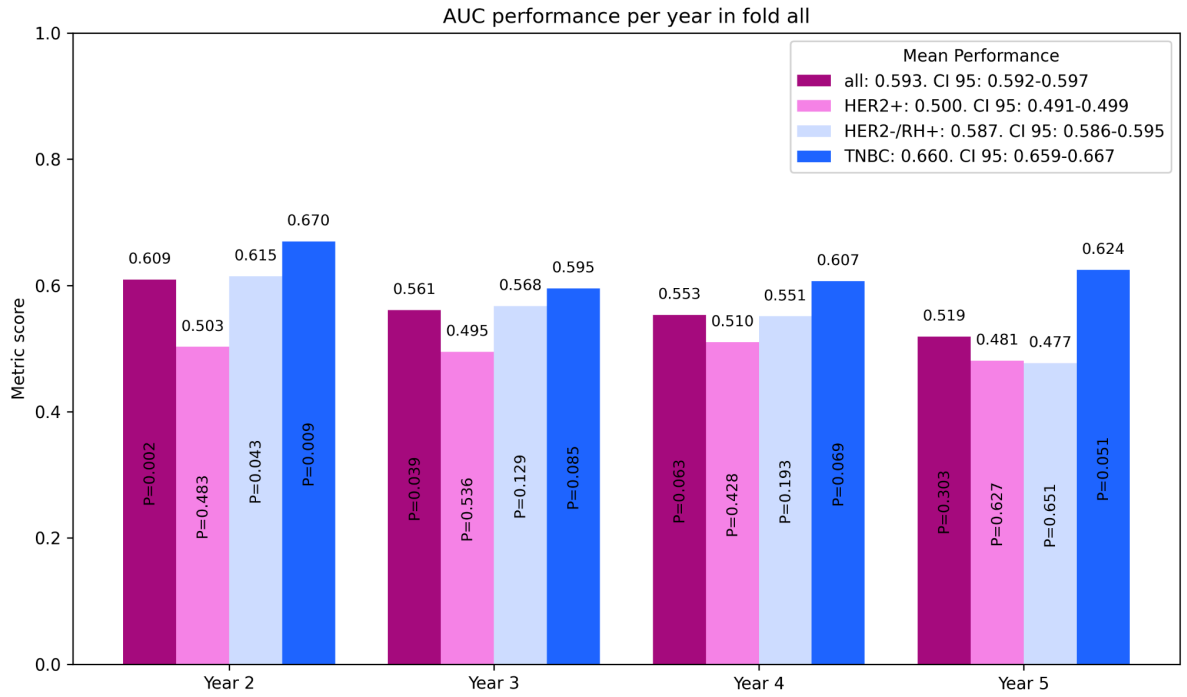


Figure 103. WSIM + HRD + PAM50 + Chemo-prAIdict Breast detailed performance in the external validation for each molecular subtype and year measured with the AUC.

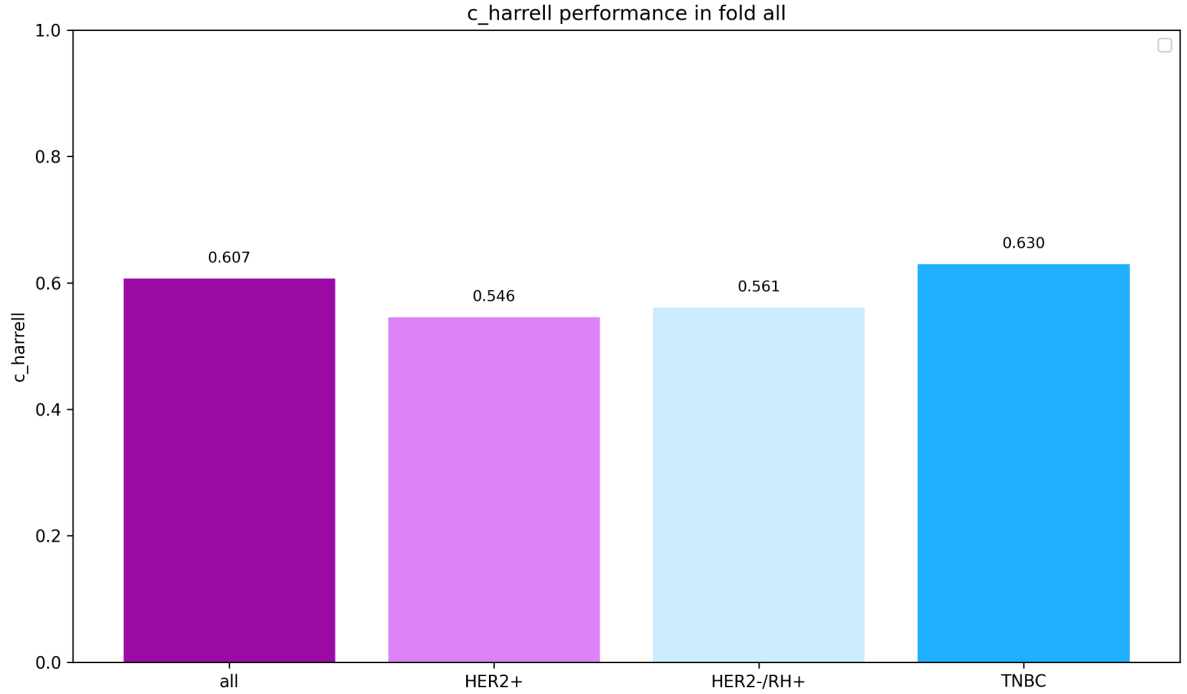


Figure 104. WSIM + HRD + PAM50 + Chemo-prAIdict Breast detailed performance in the external validation for each molecular subtype measured with the C-index Harrell in the OS task.

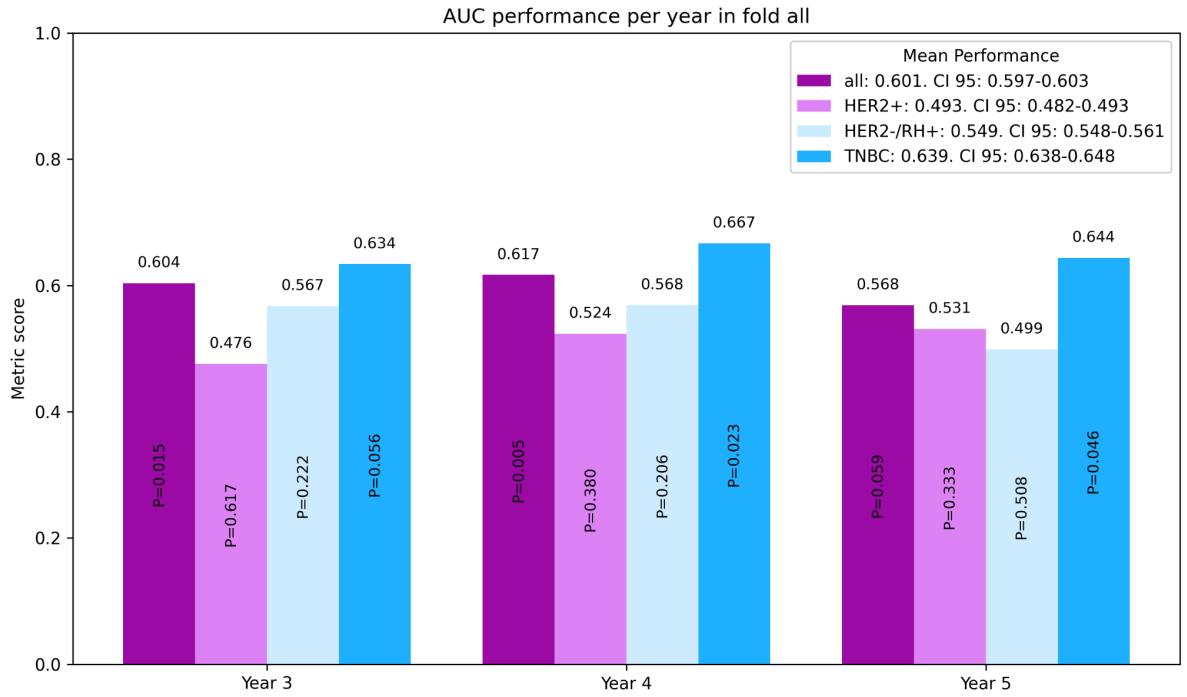


Figure 105. WSIM + HRD + PAM50 + Chemo-prAIdict Breast detailed performance in the external validation for each molecular subtype and year measured with the AUC.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

These experiments demonstrate the potential and limitations of incorporating biopsy-derived chemo-sensitivity scores into post-NAC survival prediction models. In internal validation, particularly for HER2+ patients, the inclusion of the pCR score improved both OS and DFS predictions, supporting its value as a proxy for intrinsic tumour response. This aligns with the clinical role of pCR as a surrogate endpoint in HER2+ disease and validates the biological relevance of the Chemo-prAIdict Breast score.

However, in external validation, the integration of pCR scores led to diminished performance, suggesting that the model may not yet generalise across real-world clinical cohorts or histology domains. This highlights the importance of rigorous domain adaptation, model calibration, and possibly the joint training of multi-modal models to harmonise feature representations between biopsy- and surgery-based predictors.

Future work should investigate how to optimise the integration of inferred features, such as pCR, HRD, and PAM50, within survival models. Joint modelling frameworks, domain adaptation techniques, or attention-based fusion strategies may be necessary to fully realise the benefit of combining multi-temporal and morpho-molecular information.

5.3.5 Predicting survival curves

In this section, we assess the ability of our models to stratify patients into high- and low-risk groups using histology and clinical data, focusing on three survival endpoints: overall survival (OS), disease-free survival (DFS), and, for the first time, invasive DFS (iDFS). iDFS, endorsed by STEEP and used in trials like monarchE, is a clinically relevant endpoint that excludes non-invasive recurrences[164,165].

We compare the performance of image-based models (e.g., WSIM), clinical models (CM), and their combinations across these outcomes. As model outputs are continuous risk scores, we define thresholds to enable binary stratification for clinical interpretation via Kaplan–Meier curves and hazard ratios.

Choosing a threshold is a key challenge: it must be interpretable, statistically sound, and generalisable across molecular subtypes (HER2+, HER2−/HR+, TNBC) and datasets. This section thus evaluates not only predictive performance but also the robustness and clinical relevance of deep learning–based risk stratification.

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

Dataset and study design

This experiment included all patients' post-NAC surgical specimen WSIs sourced from the PRIMUNEO and CGFL Neoadj cohorts. It also utilises the TCGA BRCA dataset to train the HRD and PAM50 predictors, as described in Chapter 4.3.2. A full description of dataset composition and preprocessing can be found in **Chapter 2 and chapter 4.2**.

Chapter 2 shows the patient distribution in the PRIMUNEO and CGFL Neoadj datasets for the OS and DFS task. **Figures 106 and 107** show the quantity of patients that present the event, do not present it or are censored over time in months and years for both datasets in the invasive Disease Free Survival task.

iDFS

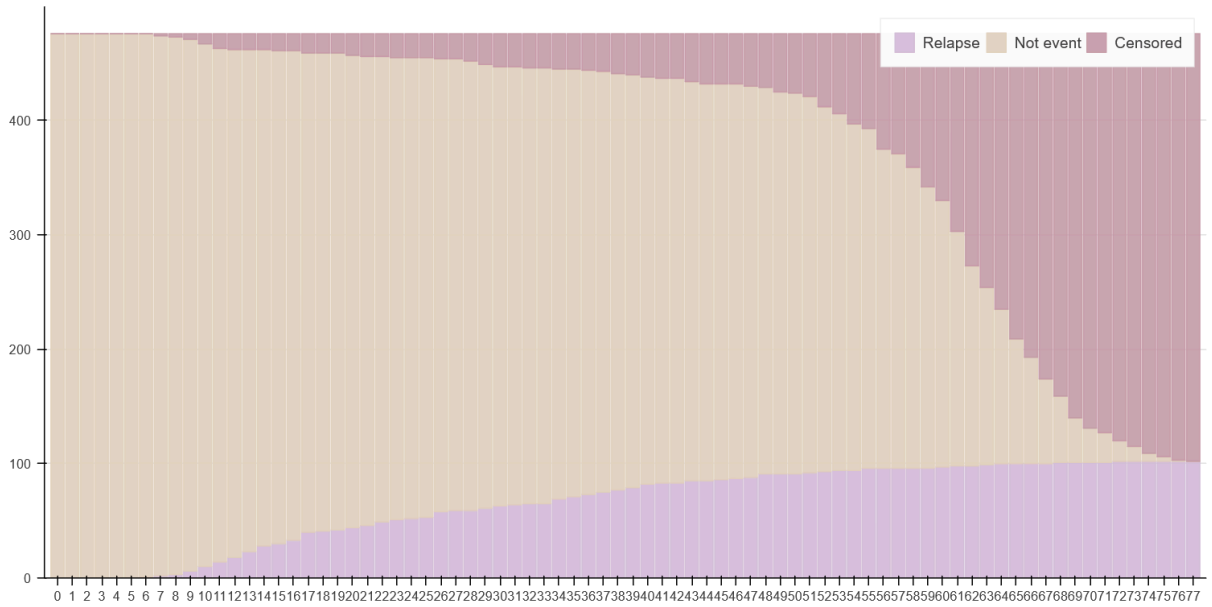


Figure 106. Number of patients that present or not the event (relapse, i.e. metastasis), or are censored each month in the PRIMUNEO dataset.

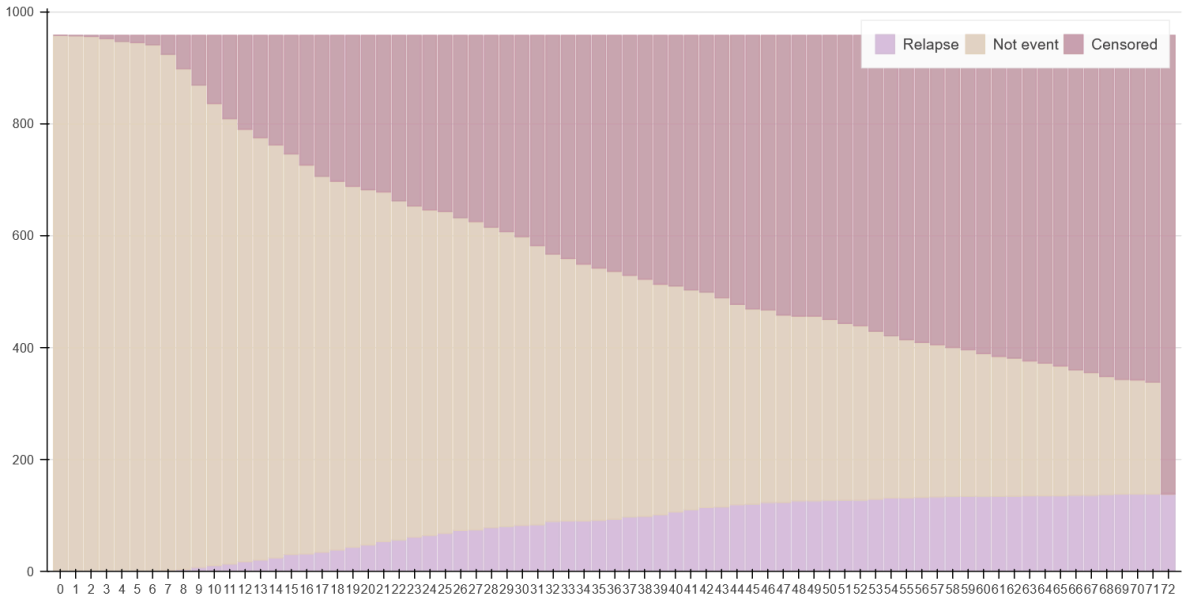


Figure 107. Number of patients that present or not the event (relapse, i.e. metastasis), or are censored each month in the CGFL Neoadj dataset.

Additionally, we show the censored patients and the ones that present the event for iDFS in PRIMUNEO (**Table 26**) and CGFL Neoadj (**Table 27**). We also show the data distribution in each of the partitions stratified by the molecular subtype in the internal validation (**Figure 108**) for the iDFS task. We kept the same partitions as in previous chapters, i.e. the same laboratory centers for training and testing.

Table 26. Patients that present the event or are censored in the iDFS task stratified by molecular subtype in the PRIMUNEO dataset.

Cancer Subtype	iDFS (censored)	iDFS (relapse)	Total Patients
HER2+	94	22	116
HER2-/RH+	153	38	191
TNBC	91	40	131

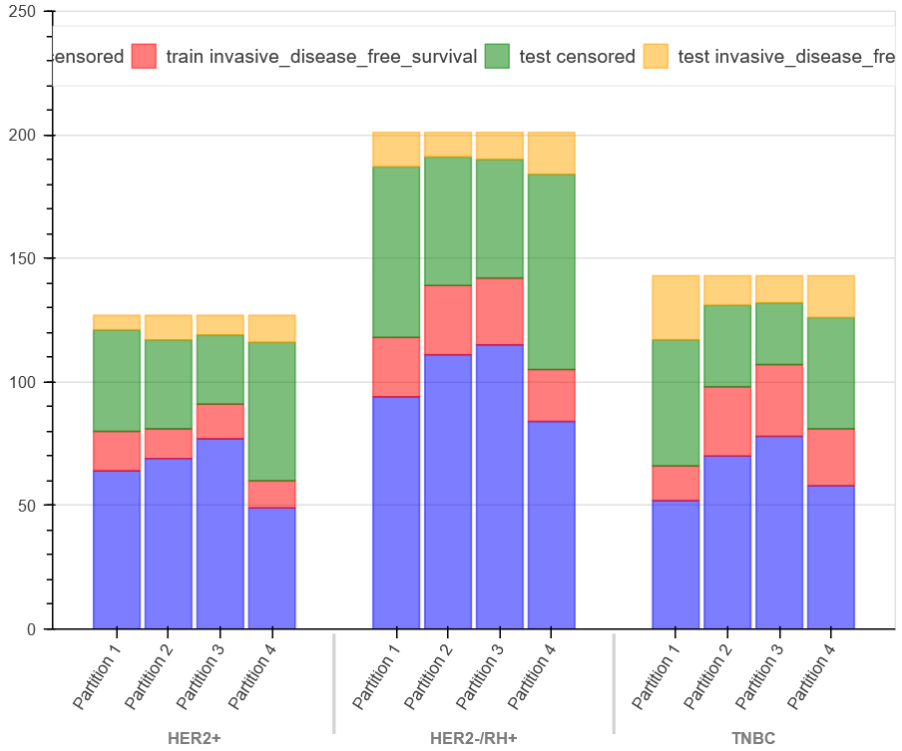


Figure 108. Distribution of invasive relapse events across partitions in the internal validation study. Invasive Disease-Free Survival (iDFS) data distributions are shown for each molecular subtype (HER2+, ER+/HER2–, and TNBC) across the four training and testing partitions used for cross-validation. Each stacked bar represents the number of patients contributing to the survival analysis, distinguishing between censored cases (blue and orange) and patients with observed events (red and green). The consistent height of bars across partitions demonstrates balanced sampling and equitable representation of censored and event cases between training and testing subsets, ensuring methodological robustness for downstream survival modelling.

Table 27. Patients that present the event or are censored in the iDFS task stratified by molecular subtype in the CGFL Neoadj dataset.

Cancer Subtype	iDFS (censored)	iDFS (relapse)	Total Patients
HER2+	172	48	220
HER2-/ RH+	113	43	156
TNBC	74	40	114

Methodology

The objective of this experiment is to identify a robust and interpretable method for determining a threshold based on risk scores from the internal validation set that can be used to stratify patients in the external validation cohort into two risk groups for survival curve comparison. Additionally, we want to compare the survival curves along the image-based models and the clinical-based models in each molecular subtype.

We use the slide-level risk scores predicted by WSIM, CM, WSIM combined with HRD/PAM50, and WSIM combined with pCR across all partitions of the internal validation set to compute thresholds for each method and molecular subtype using various statistical approaches. These thresholds are then applied to the predicted risk scores in the CGFL external validation dataset to stratify patients into two groups: high-risk and low-risk. We subsequently generate Kaplan-Meier (KM) survival curves for these groups and perform the log-rank test to assess whether the survival distributions differ significantly (**Figure 109**), aiming to demonstrate that the low-risk group consistently exhibits better survival outcomes. Additionally, we report other comparative metrics such as the hazard ratio and the probability of experiencing an event within five years.

Let $h_i(t)$ be the hazard ratio of group i at time t, then:

$$H_0 : h_1(t) = h_2(t)$$
$$H_A : h_1(t) = ch_2(t), \quad c \neq 1$$

Figure 109. Log-rank test null and alternative hypothesis.

Evaluation

Evaluation is performed separately for each molecular subtype (HER2+, HER2−/HR+, TNBC), and across all three survival endpoints (OS, DFS, iDFS). For each model and subtype, we present the results of the mean in the internal validation as the threshold for each subtype.

These results are summarised using a consistent layout across figures: three columns representing molecular subtypes and a row for each method:

- Whole Slide Image model (WSIM) survival curves on external validation
- WSIM + HRD/PAM50 survival curves on external validation
- WSIM + pCR survival curves on external validation
- Clinical Model (CM) survival curves on external validation
- WSIM + CM survival curves on external validation
- WSIM + PAM50/HRD + PCR + CM survival curves on external validation

The clinical model is described in Chapter 3.3.4.

RESULTS

Using the mean threshold for patient stratification (**Figures 110-112)**, we observe varying levels of success across survival endpoints and molecular subtypes. For overall survival (OS), the WSIM model achieves significant stratification in both HER2+ and TNBC subtypes, while WSIM + PAM50/HRD shows broader robustness, with significant results across all three subtypes (HER2+, HER2−/HR+, and TNBC). Neither the clinical model (CM) nor WSIM + CM shows significance for HER2+ or luminal tumours, though WSIM + CM retains discriminative power in TNBC.

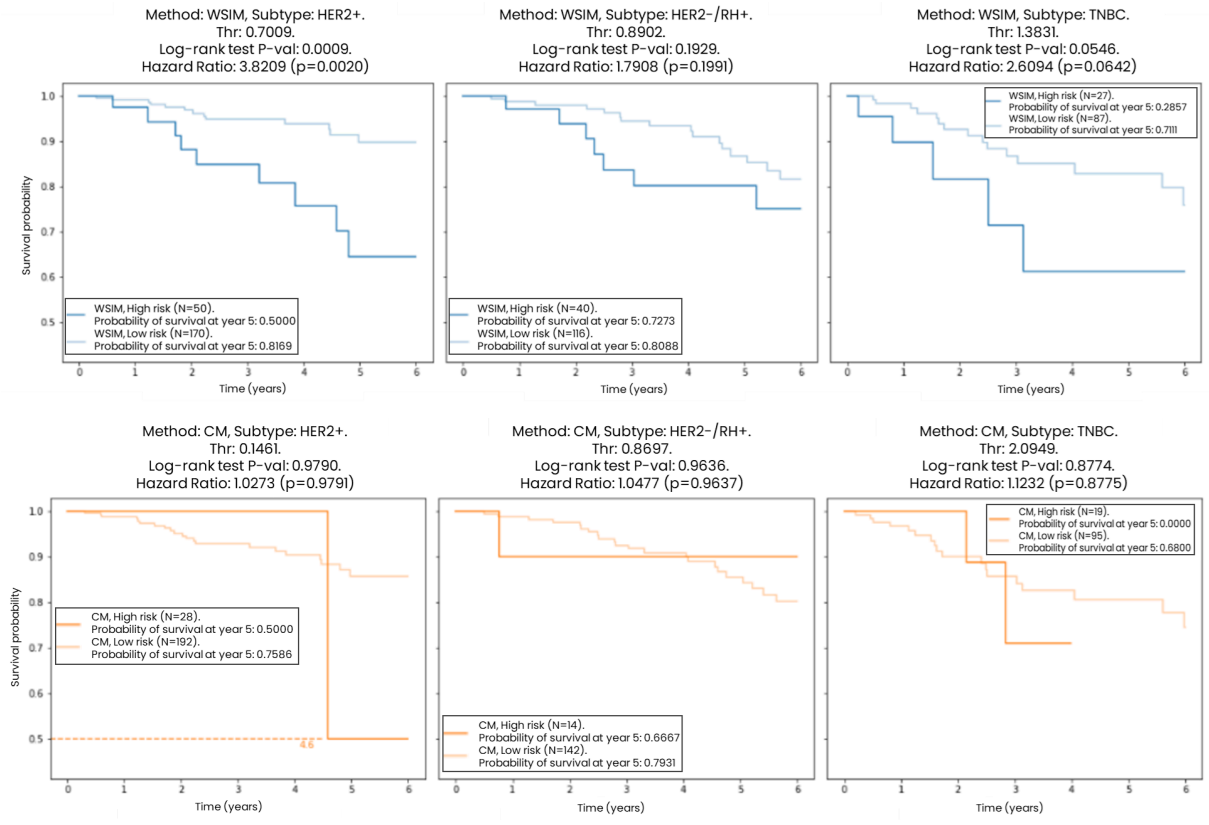
In the disease-free survival (DFS) setting, only the WSIM + PAM50/HRD model demonstrates significant stratification in TNBC, while no other model achieves significance in any subtype. This highlights the challenge of predicting DFS and suggests that morphological and clinical signals may be less discriminative for this endpoint using a mean-based threshold.

For invasive DFS (iDFS), stratification performance improves slightly. WSIM + PAM50/HRD again performs best, achieving significance in both HER2−/HR+ and TNBC subtypes. The WSIM + pCR and WSIM + CM models also show significant stratification for

TNBC, indicating that this subtype benefits most consistently from multimodal integration. However, HER2+ tumours remain difficult to stratify across all models and endpoints.

In summary, the WSIM + PAM50/HRD model achieves very promising results, especially for TNBC where it consistently emerges as a significant predictor of prognosis for OS, DFS and iDFS. It also works well for luminal tumours in the OS and DFS tasks. HER2+ subtypes, however, require further model refinement or alternative biomarker integration to achieve reliable clinical separation using the mean threshold strategy.

OS



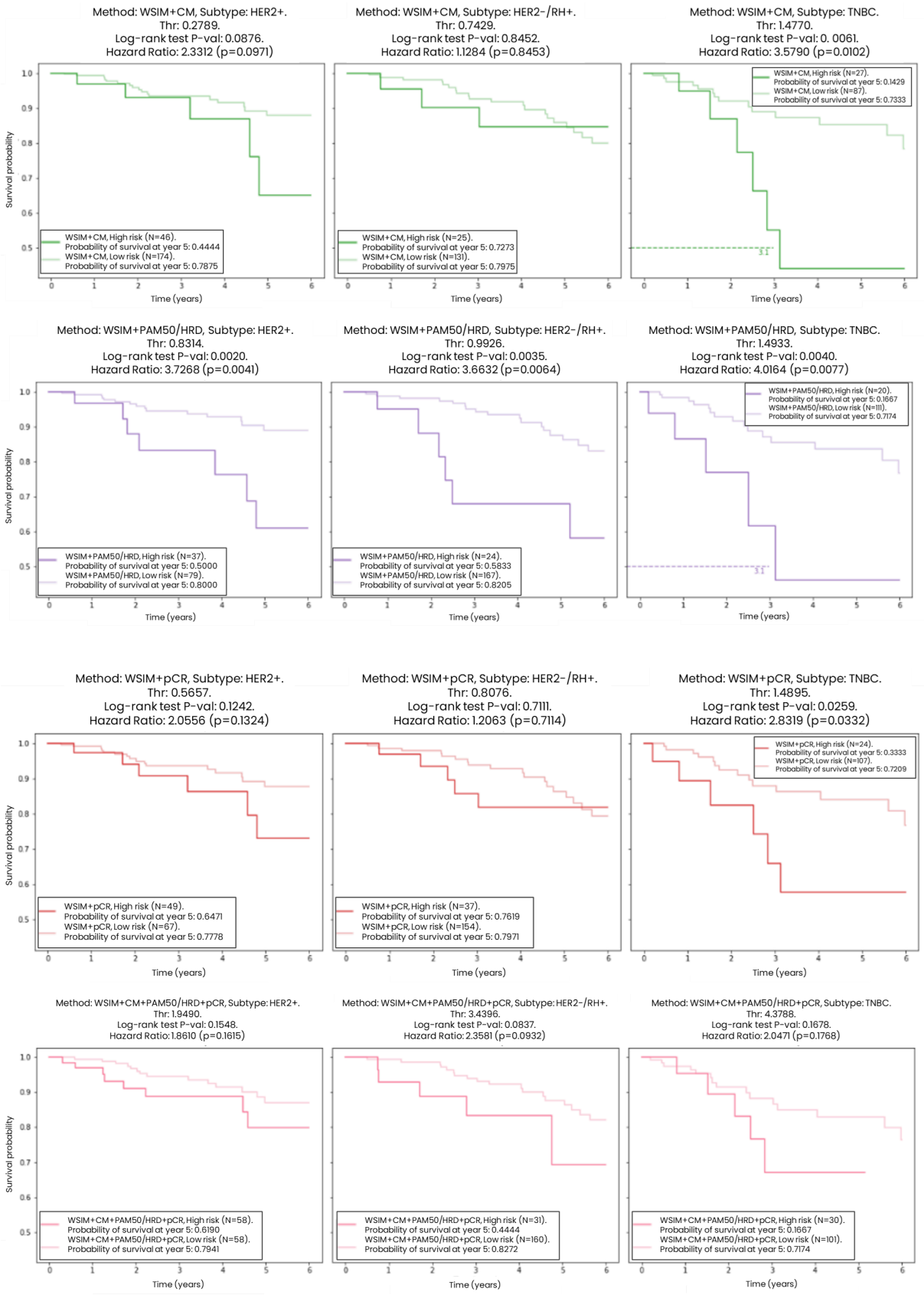
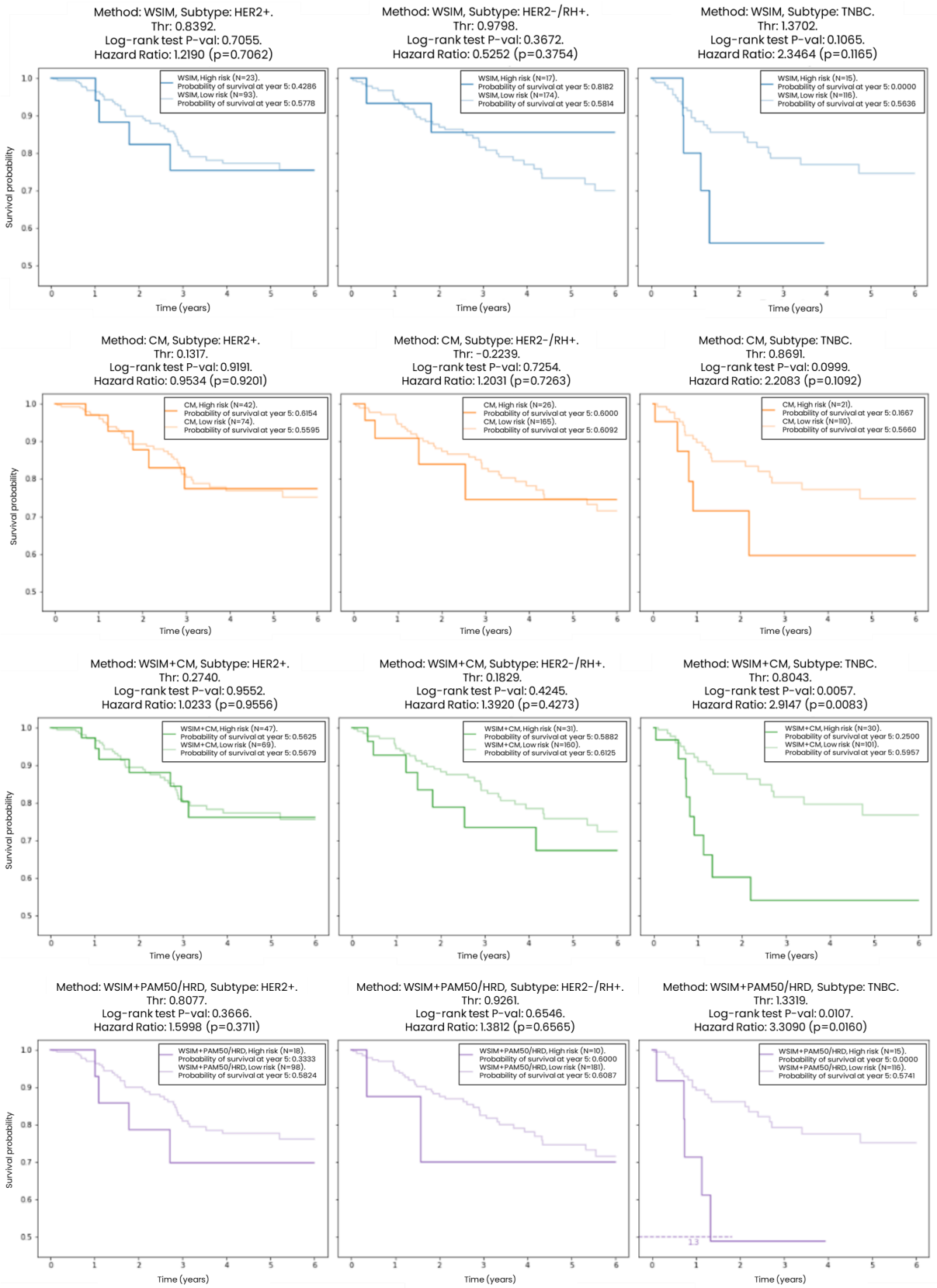


Figure 110. OS survival curves for WSIM, CM, WSIM+CM, WSIM + HRD/PAM50, WSIM + PCR score, WSIM + CM + HRD/PAM50 + PCR score.

DFS


Method: WSIM, Subtype: HER2+.
Thr: 0.8392.
Log-rank test P-val: 0.7055.
Hazard Ratio: 1.2190 (p=0.7062)
WSIM, High risk (N=23).
Probability of survival at year 5: 0.4286
WSIM, Low risk (N=93).
Probability of survival at year 5: 0.5778
Method: WSIM, Subtype: HER2-/RH+.
Thr: 0.9798.
Log-rank test P-val: 0.3672.
Hazard Ratio: 0.5252 (p=0.3754)
WSIM, High risk (N=17).
Probability of survival at year 5: 0.8182
WSIM, Low risk (N=174).
Probability of survival at year 5: 0.5814
Method: WSIM, Subtype: TNBC.
Thr: 1.3702.
Log-rank test P-val: 0.1065.
Hazard Ratio: 2.3464 (p=0.1165)
WSIM, High risk (N=15).
Probability of survival at year 5: 0.0000
WSIM, Low risk (N=116).
Probability of survival at year 5: 0.5636
Method: CM, Subtype: HER2+.
Thr: 0.1317.
Log-rank test P-val: 0.9191.
Hazard Ratio: 0.9534 (p=0.9201)
CM, High risk (N=42).
Probability of survival at year 5: 0.6154
CM, Low risk (N=74).
Probability of survival at year 5: 0.5595
Method: CM, Subtype: HER2-/RH+.
Thr: -0.2239.
Log-rank test P-val: 0.7254.
Hazard Ratio: 1.2031 (p=0.7263)
CM, High risk (N=26).
Probability of survival at year 5: 0.6000
CM, Low risk (N=165).
Probability of survival at year 5: 0.6092
Method: CM, Subtype: TNBC.
Thr: 0.8691.
Log-rank test P-val: 0.0999.
Hazard Ratio: 2.2083 (p=0.1092)
CM, High risk (N=21).
Probability of survival at year 5: 0.1667
CM, Low risk (N=110).
Probability of survival at year 5: 0.5660
Method: WSIM+CM, Subtype: HER2+.
Thr: 0.2740.
Log-rank test P-val: 0.9552.
Hazard Ratio: 1.0233 (p=0.9556)
WSIM+CM, High risk (N=47).
Probability of survival at year 5: 0.5625
WSIM+CM, Low risk (N=69).
Probability of survival at year 5: 0.5679
Method: WSIM+CM, Subtype: HER2-/RH+.
Thr: 0.1829.
Log-rank test P-val: 0.4245.
Hazard Ratio: 1.3920 (p=0.4273)
WSIM+CM, High risk (N=31).
Probability of survival at year 5: 0.5882
WSIM+CM, Low risk (N=160).
Probability of survival at year 5: 0.6125
Method: WSIM+CM, Subtype: TNBC.
Thr: 0.8043.
Log-rank test P-val: 0.0057.
Hazard Ratio: 2.9147 (p=0.0083)
WSIM+CM, High risk (N=30).
Probability of survival at year 5: 0.2500
WSIM+CM, Low risk (N=101).
Probability of survival at year 5: 0.5957
Method: WSIM+PAM50/HRD, Subtype: HER2+.
Thr: 0.8077.
Log-rank test P-val: 0.3666.
Hazard Ratio: 1.5998 (p=0.3711)
WSIM+PAM50/HRD, High risk (N=18).
Probability of survival at year 5: 0.3333
WSIM+PAM50/HRD, Low risk (N=98).
Probability of survival at year 5: 0.5824
Method: WSIM+PAM50/HRD, Subtype: HER2-/RH+.
Thr: 0.9261.
Log-rank test P-val: 0.6546.
Hazard Ratio: 1.3812 (p=0.6565)
WSIM+PAM50/HRD, High risk (N=10).
Probability of survival at year 5: 0.6000
WSIM+PAM50/HRD, Low risk (N=181).
Probability of survival at year 5: 0.6087
Method: WSIM+PAM50/HRD, Subtype: TNBC.
Thr: 1.3319.
Log-rank test P-val: 0.0107.
Hazard Ratio: 3.3090 (p=0.0160)
WSIM+PAM50/HRD, High risk (N=15).
Probability of survival at year 5: 0.0000
WSIM+PAM50/HRD, Low risk (N=116).
Probability of survival at year 5: 0.5741
Survival probability
Time (years)

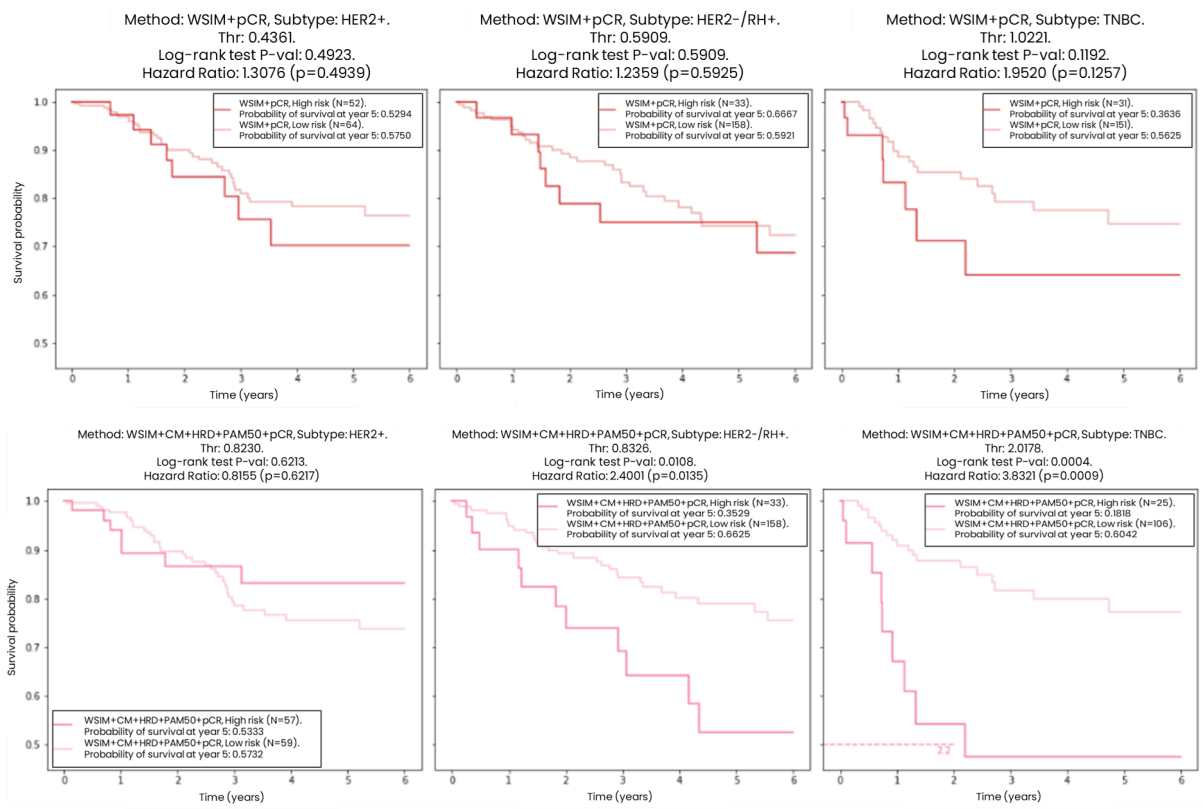
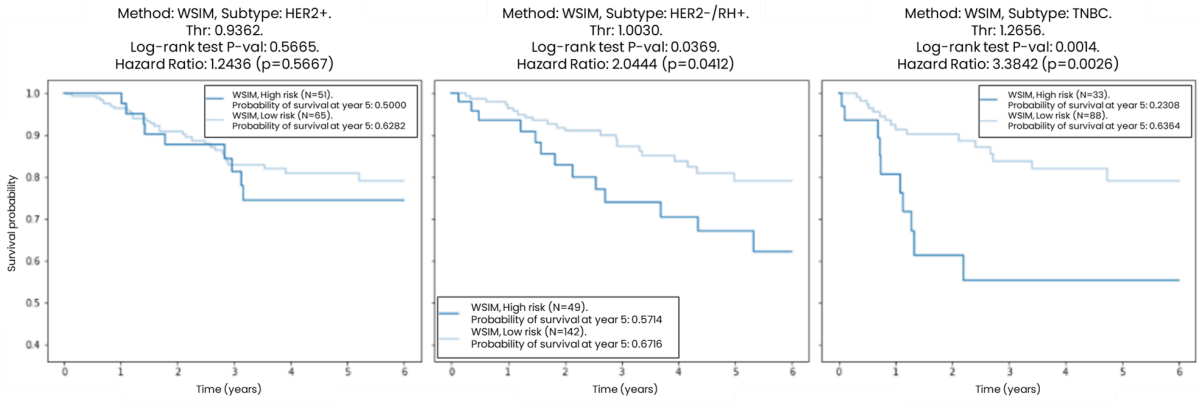
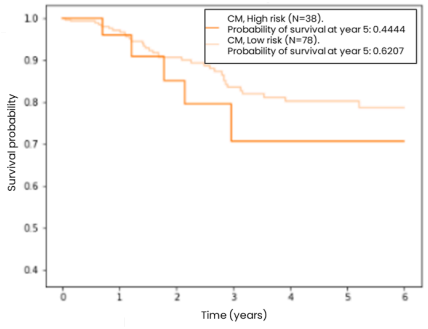


Figure 111. OS survival curves for WSIM, CM, WSIM+CM, WSIM + HRD/PAM50, WSIM + PCR score, WSIM + CM + HRD/PAM50 + PCR score.

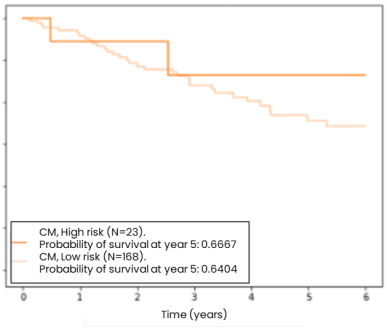
iDFS



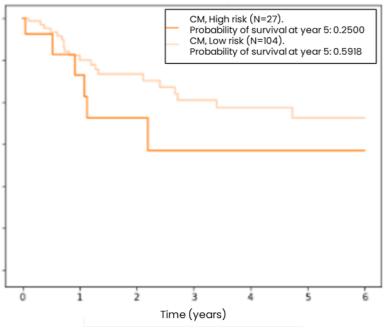
Method: CM, Subtype: HER2+.
Thr: -0.2093.
Log-rank test P-val: 0.3953.
Hazard Ratio: 1.4993 (p=0.3981)



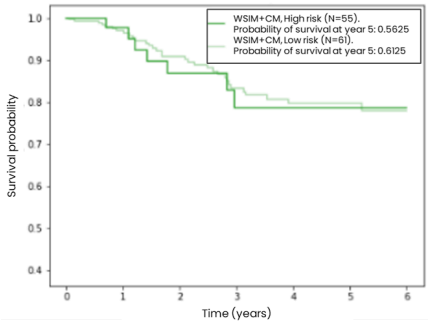
Method: CM, Subtype: HER2-/RH+.
Thr: -0.3566.
Log-rank test P-val: 0.5595.
Hazard Ratio: 0.6555 (p=0.5623)



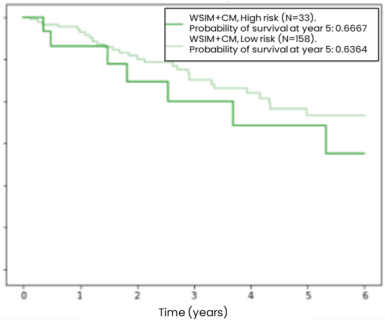
Method: CM, Subtype: TNBC
Thr: 0.6343.
Log-rank test P-val: 0.2017.
Hazard Ratio: 1.7999 (p=0.2086)



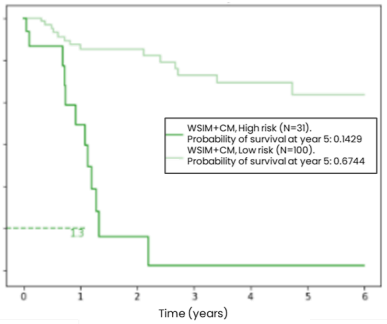
Method: WSIM+CM, Subtype: HER2+.
Thr: 0.2595.
Log-rank test P-val: 0.8672.
Hazard Ratio: 1.0730 (p=0.8660)



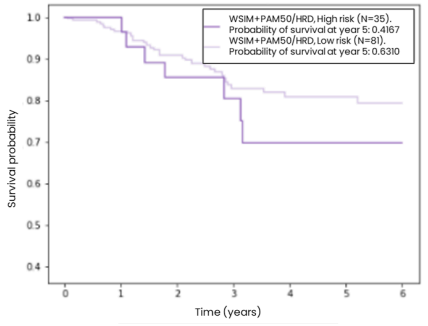
Method: WSIM+CM, Subtype: HER2-/RH+.
Thr: 0.2814.
Log-rank test P-val: 0.3910.
Hazard Ratio: 1.4342 (p=0.3940)



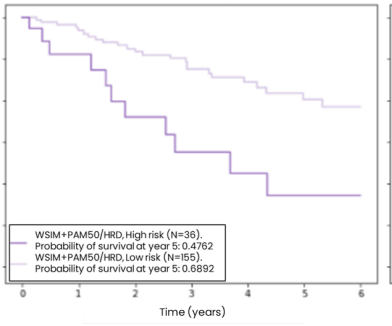
Method: WSIM+CM, Subtype: TNBC.
Thr: 0.8488.
Log-rank test P-val: 0.0000.
Hazard Ratio: 6.0050 (p=0.0000)



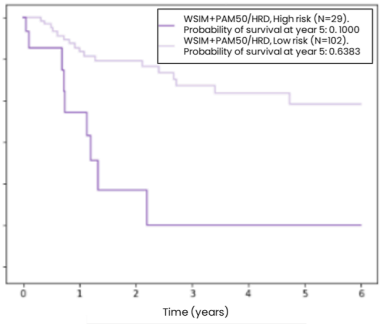
Method: WSIM+PAM50/HRD, Subtype: HER2+.
Thr: 0.9328.
Log-rank test P-val: 0.3161.
Hazard Ratio: 1.5160 (p=0.3191)



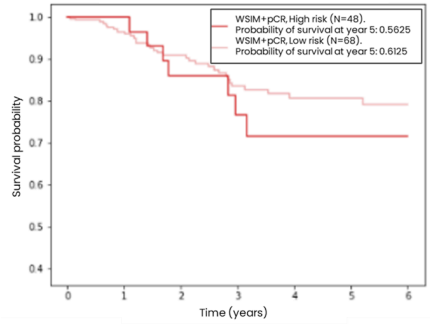
Method: WSIM+PAM50/HRD, Subtype: HER2-/RH+.
Thr: 0.9874.
Log-rank test P-val: 0.0062.
Hazard Ratio: 2.6124 (p=0.0084)



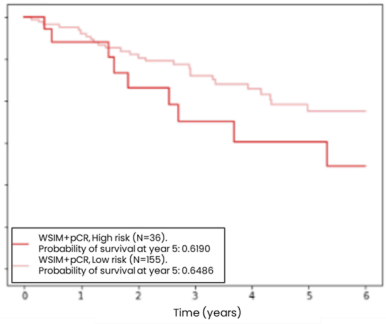
Method: WSIM+PAM50/HRD, Subtype: TNBC.
Thr: 1.2718.
Log-rank test P-val: 0.0003.
Hazard Ratio: 4.0093 (p=0.0009)



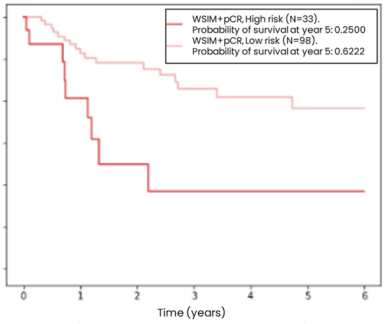
Method: WSIM+pCR, Subtype: HER2+.
Thr: 0.7481.
Log-rank test P-val: 0.4704.
Hazard Ratio: 1.3506 (p=0.4715)



Method: WSIM+pCR, Subtype: HER2-/RH+.
Thr: 0. 9280.
Log-rank test P-val: 0.1579.
Hazard Ratio: 1.7154 (p=0.1633)



Method: WSIM+pCR, Subtype: TNBC
Thr: 1.3330.
Log-rank test P-val: 0.0071.
Hazard Ratio: 2.9021 (p=0.0101)



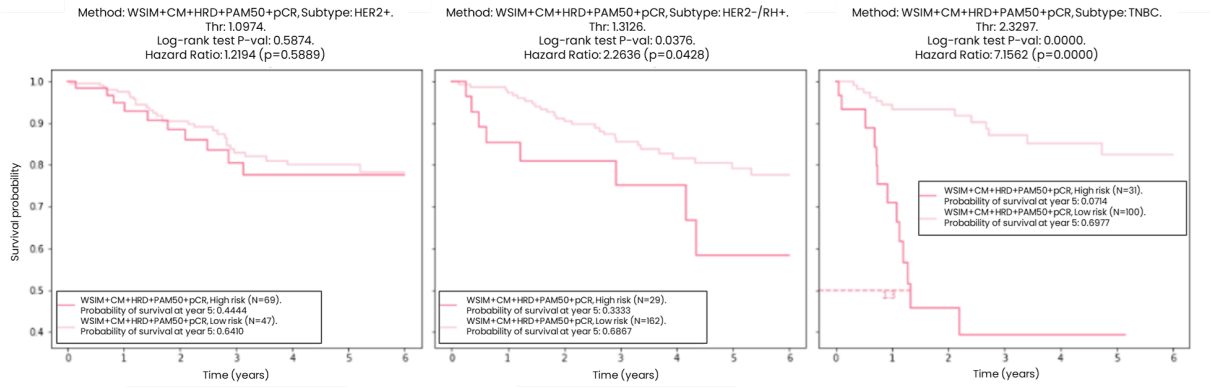


Figure 112. iDFS survival curves for WSIM, CM, WSIM+CM, WSIM + HRD/PAM50, WSIM + PCR score, WSIM + CM + HRD/PAM50 + PCR score.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

In this section, we assessed the ability of image-based and clinical models to stratify patients into high- and low-risk groups using predicted risk scores for three survival endpoints: overall survival (OS), disease-free survival (DFS), and invasive DFS (iDFS).

Using thresholds derived from internal validation and applied to the external CGFL cohort, we observed clear differences in stratification performance across subtypes and model configurations. The WSIM + PAM50/HRD model provided the most consistent and robust separation between risk groups, particularly in triple-negative breast cancer (TNBC), where statistically significant results were achieved across all three endpoints (OS, DFS, iDFS). This configuration also demonstrated meaningful stratification in luminal tumours, especially for iDFS.

In contrast, the clinical model alone showed limited discriminative power, achieving significance only for luminal DFS and iDFS under a single threshold. The WSIM + CM configuration improved performance in TNBC but did not significantly enhance prediction in other subtypes. The WSIM + pCR model, incorporating chemo-sensitivity inferred from the diagnostic biopsy, also achieved significant, albeit threshold-sensitive, results in TNBC,

suggesting that baseline chemosensitivity retains prognostic value in highly treatment-responsive disease.

Across all models, HER2-positive tumours proved the most challenging to stratify, regardless of the survival endpoint. This likely reflects both the biological and therapeutic homogeneity induced by HER2-targeted regimens, resulting in fewer morphologically distinguishable subgroups detectable through histology or clinical data alone. It further underscores the need for integrating HER2-specific molecular or immune-related biomarkers to enhance prognostic discrimination in this population.

Interestingly, although the C-index and time-to-event metrics of the survival models remained moderate, the Kaplan–Meier analyses revealed sharp and statistically robust separations between high- and low-risk groups. This apparent discrepancy highlights a common property of survival modelling: while censored and heterogeneous follow-up data can hinder temporal calibration, models may still accurately capture the relative ranking of patient risk. In other words, the networks effectively learned the order of prognostic severity rather than the exact timing of recurrence or death. Such behaviour explains why stratification analyses outperform direct time-to-event predictions and confirms that the learned risk scores encode biologically meaningful gradients of aggressiveness. Importantly, this also suggests that tackling the harder survival task, rather than a simplified binary endpoint such as 5-year recurrence, may produce more generalisable and interpretable prognostic representations.

5.3.6 patches analysis for identifying novel biomarkers linked to OS and DFS.

To determine whether the image regions driving the neural networks' predictions correspond to biologically interpretable morphologies, we performed a pathologist-guided analysis of the highest-scoring regions of interest identified by the model

METHODOLOGY (SPECIFIC MATERIAL & METHODS)

To explore the morphological basis of the predictions made by our neural networks, we conducted a pathologist-driven interpretability analysis on the regions most strongly contributing to each model's decision. Two separate networks were considered:

(1) Chemo-prAIdict Breast, trained on diagnostic biopsy slides to predict tumour chemosensitivity, and

(2) a post-NAC residual disease network, trained on surgical specimens to predict prognosis and disease-free survival (DFS).

Only slides containing residual invasive disease were included in this analysis. Slides corresponding to complete pathological response (pCR) were systematically excluded to ensure that the morphological review focused exclusively on residual tumour areas, which constitute the histological substrate relevant to recurrence and survival modelling.

For each network, the 20 highest-scoring and 20 lowest-scoring whole-slide images were selected based on their respective model prediction scores (i.e., those most confidently classified as chemosensitive vs chemoresistant for the biopsy model, and as low-risk vs high-risk residual disease for the post-NAC model). It is important to note that these scores reflect the model's own assessment and do not correspond to the pathological ground truth.

From each selected slide, we extracted the 20 top-ranked patches (regions with the highest model attention), yielding 400 high-confidence patches per predicted category for each framework. These regions thus represent the areas most influential to the network's internal decision-making, focusing the review on biologically relevant tumour zones rather than artefactual or stromal regions.

Three expert breast pathologists (CP, LA, HPM; 5, 40, and 35 years of experience) independently examined all patches, being aware of the slide-level prediction category (chemosensitive/chemoresistant or good/poor prognosis) but not the individual model scores or clinical outcomes. This setup enabled an exploratory, inductive morphological analysis free from predefined scoring grids, in which each expert independently identified recurring histological motifs overrepresented in each class. Individual observations were then discussed in a joint review session to reach qualitative consensus.

This design was chosen a priori to avoid trivial contrasts that could arise if comparing the lowest-scoring regions from chemoresistant slides, often lacking tumoural content, with tumour-rich high-scoring regions. By focusing instead on the most predictive, tumour-containing regions across both categories, this approach ensures that the interpretability analysis targets true morpho-biological correlates of model behaviour rather than differences in tissue composition or sampling artefacts.

RESULTS

While quantitative performance metrics confirmed that both neural networks could discriminate between highly chemosensitive and chemoresistant tumours (biopsy framework) and between low- and high-risk residual disease (post-NAC framework), these global scores provided limited insight into the underlying morphological cues. The pathologist-guided review therefore aimed to bridge the models' internal representations with human-recognisable histopathological features, assessing whether the networks relied on biologically meaningful structures or spurious correlations.

By comparing the most predictive tumoural regions across the highest and lowest predicted classes, the reviewers identified morphological patterns systematically associated with model-defined chemosensitivity or poor prognosis. The subsequent consensus analysis provided a qualitative validation layer, confirming that the features emphasised by the networks could be interpreted within an established biological and clinical framework. The analyses were conducted separately for the two models: first on diagnostic biopsies (chemosensitivity), and then on post-NAC surgical specimens (residual disease prognosis).

5.3.6.1 Initial tumour biopsy patch analysis

Across molecular subtypes, the reviewers consistently identified distinct morphologies between the extreme categories predicted by the neural network predicting chemo-sensitivity based on the initial biopsy's WSI (*chemo-prAIdict Breast*) (**Figure 113**).

In HER2-positive tumours (**Fig. 113a**), *chemosensitive* cases showed high tumour cellularity with cohesive sheets of large, nucleolated epithelial cells, brisk mitotic activity, and prominent lymphocytic infiltrates within a reactive stroma. *Resistant* counterparts were characterised by sparse cellularity, cord-like tumour strands embedded in dense fibrotic stroma, and a paucity of lymphocytes.

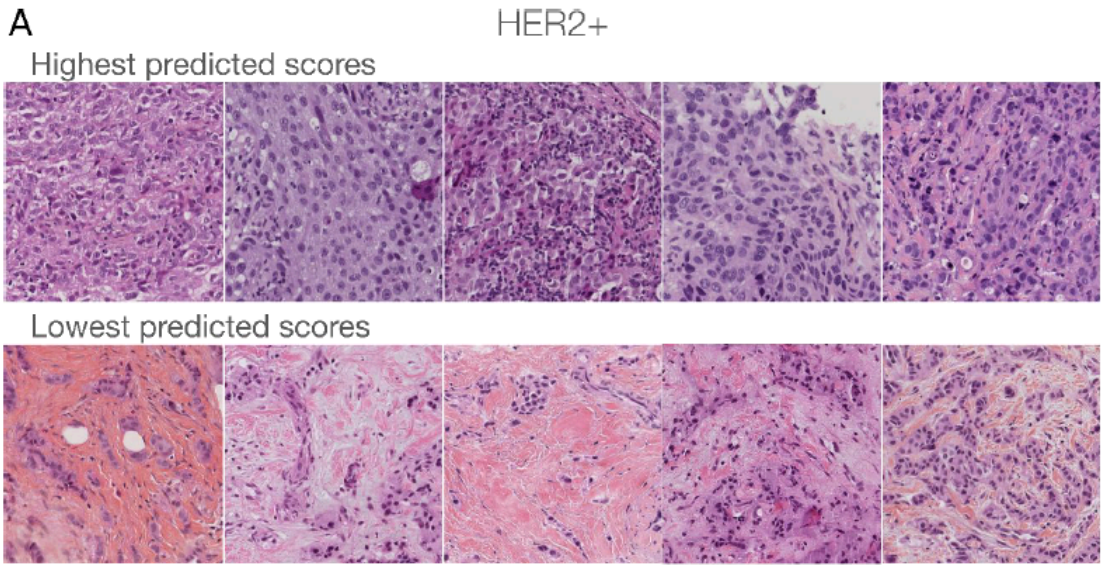


Figure 113.A. HER2-amplified tumours. Representative biopsy patches with the highest and lowest predicted chemosensitivity scores by the *Chemo-prAIdict Breast* model. High-scoring regions show densely cellular areas with active proliferation, whereas low-scoring regions display fibrotic and hypocellular patterns.

In luminal tumours (**Fig. 113b**), *chemosensitive* areas exhibited enlarged nuclei with conspicuous nucleoli, heterogeneous chromatin, and moderate TILs, whereas *resistant* areas displayed low tumour density, small regular nuclei, and collagen-rich interstitium. The presence of TILs in both groups suggests that inflammatory infiltrate alone is not a discriminant feature within this subtype.

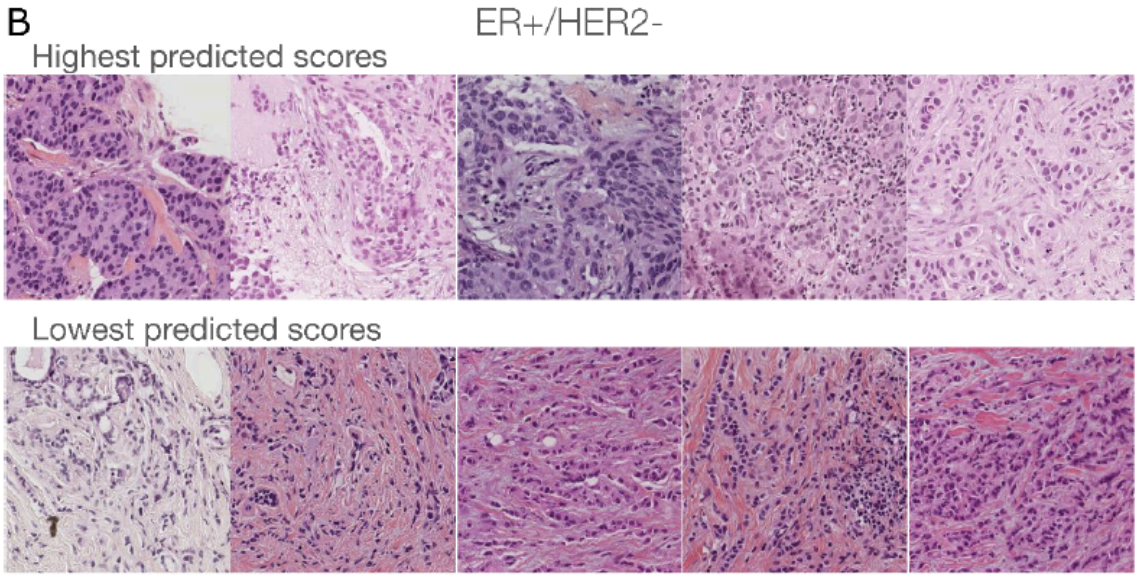


Figure 113.b. ER+/HER2- tumours. Representative biopsy patches from luminal cases showing contrasting morphologies between high and low predicted chemosensitivity by the model.

In triple-negative cancers (**Fig. 113c**), *chemosensitive* patches revealed marked pleomorphism, high nuclear-cytoplasmic ratio, and lymphocyte-rich stroma, while *chemoresistant* regions showed tumour islands with a cordonal or trabecular architecture surrounded by dense collagen and limited inflammation.

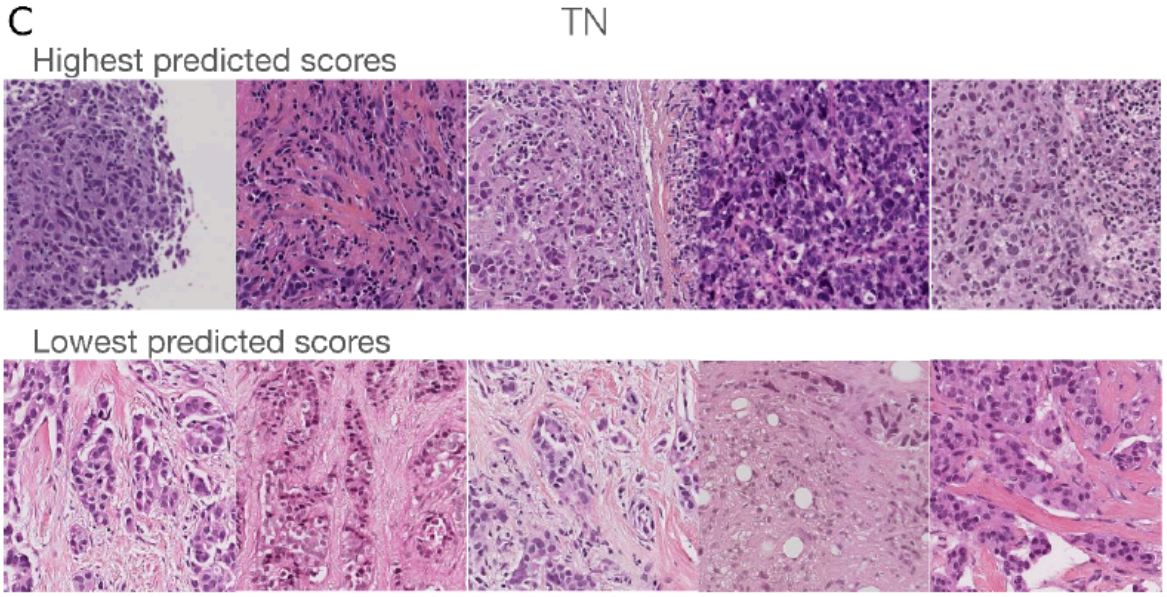


Figure 113.c. TNBC tumours. Representative biopsy patches from triple-negative breast cancers with the most extreme predicted chemosensitivity scores, illustrating morphological differences captured by the model.

Overall, the features highlighted by pathologists overlap closely with those implicitly captured by the network attention maps, suggesting that Chemo-prAIdict Breast learned morphologies concordant with recognised indicators of chemosensitivity, such as proliferative activity and immune response intensity.

3.6.2 Residual disease patch analysis

In the resection specimens following neoadjuvant chemotherapy, reviewers identified features correlating with high-risk residual disease and shorter disease-free survival (DFS). High-risk slides were characterised by a larger residual invasive component, frequent necrosis, high-grade cytology with pronounced nuclear atypia, and occasional apocrine differentiation. Conversely, low-risk residual disease frequently displayed dense fibrosis, limited invasive tumour, and frequent areas of in-situ carcinoma only.

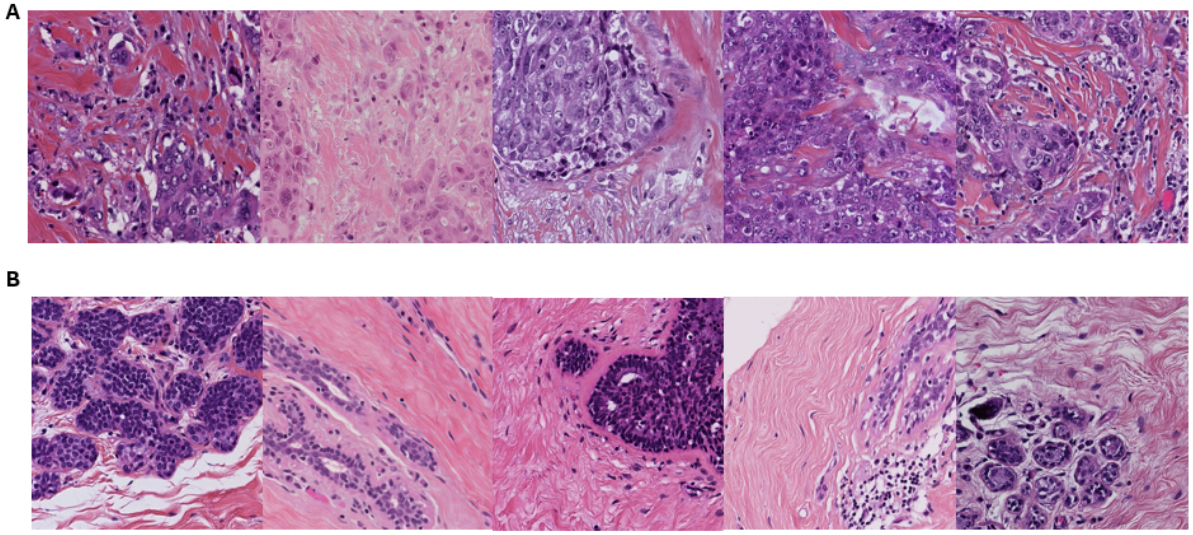


Figure 114.A. HER2-amplified tumours. Representative post-neoadjuvant surgical specimens classified as high-risk residual disease (A) and low-risk (B) according to the neural network.

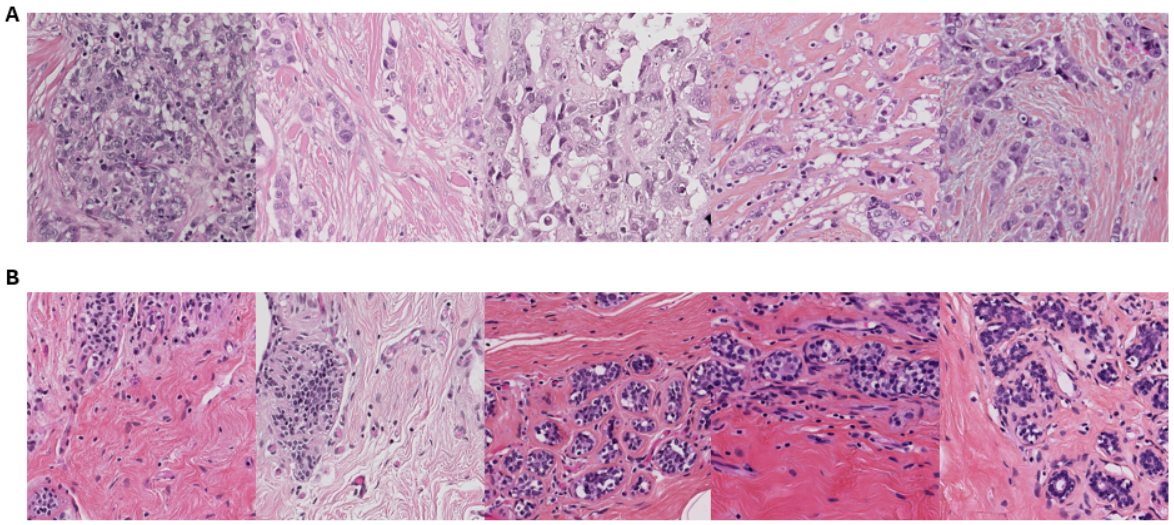


Figure 114.B. ER+/HER2- tumours. Representative post-neoadjuvant surgical specimens classified as high-risk residual disease (A) and low-risk (B) according to the neural network.

In luminal (ER+/HER2−) tumours, an additional observation emerged during expert review: several low-risk surgical slides predicted by the neural network showed a predominance of normal-appearing acini within dense, homogeneous stromal tissue. This finding suggests a potential overrepresentation of preserved or reparative glandular structures in post-NAC specimens classified as low-risk, likely reflecting extensive stromal regression and partial replacement of tumoural areas by benign epithelial components. Although a few scattered atypical epithelial cells were noted, their unequivocal malignant nature could not be confirmed without immunohistochemical validation. In contrast, high-risk luminal slides consistently displayed looser, oedematous stroma intermingled with clearly atypical epithelial clusters, consistent with residual invasive carcinoma. This morphological dichotomy could highlight both the impact of treatment-induced tissue remodelling and the interpretative limitations of purely H&E-based assessment in these contexts.

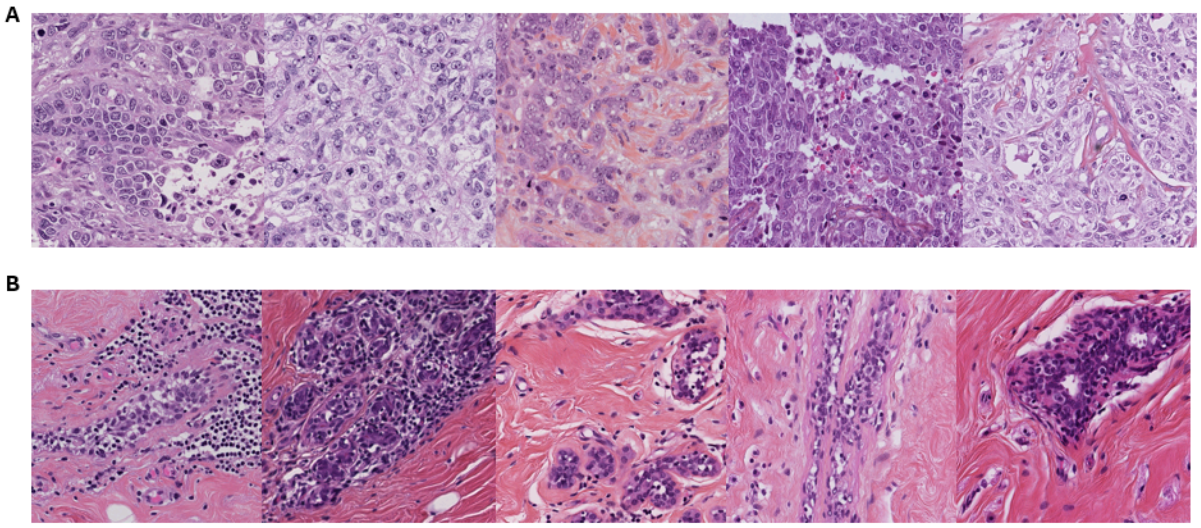


Figure 114.C. TNBC tumours. Representative post-neoadjuvant surgical specimens classified as high-risk residual disease (A) and low-risk (B) according to the neural network.

Pathologists emphasised that these findings are morphologically consistent with aggressive residual phenotypes described in the literature (high residual tumour cellularity and necrosis correlate with poor prognosis [166,167], while fibrotic regression patterns are typical of better responders. The observation of normal acinar structures within low-risk luminal slides further supports the notion that stromal architecture and epithelial regeneration patterns contribute significantly to the model's learned representations of favourable prognosis. The model's ability to capture these distinctions suggests potential value as a digital grading system for residual disease risk, complementing the limitations of current post-NAC staging frameworks.

CONCLUSION, FINDINGS AND FUTURE DIRECTIONS

Taken together, these pathologist-guided analyses provide an interpretable bridge between the predictions of our neural networks and established histopathological knowledge. Across both diagnostic biopsies and post-NAC resections, the model consistently emphasised features recognised as markers of tumour aggressiveness or regression suggesting that the network's decision patterns align with biologically meaningful morphology rather than abstract image statistics.

Interestingly, certain morphological patterns recurred across both pre- and post-treatment settings: fibrosis, cellularity, and nuclear atypia emerged as central discriminants of both

chemosensitivity and residual risk. This convergence supports the notion that neural networks may be recognising persistent micro-architectural phenotypes reflective of tumour biology rather than treatment-induced artefacts. However, the apparent overlap also raises conceptual questions:

- Are these stable “intrinsic” features that pre-exist and predict response, or are they dynamic markers reflecting tumour adaptation to therapy?
- Could fibrosis, observed as favourable post-therapy but unfavourable pre-therapy, represent different biological processes (reactive stroma vs treatment-induced scarring)?

In luminal tumours, the model’s low-risk predictions were frequently associated with slides containing predominantly normal-appearing acini within dense fibrotic stroma. This unexpected finding raises the possibility that the network, in part, identifies residual benign or reparative epithelial components as indicators of favourable outcome. Although this interpretation aligns with the concept of stromal regression and tissue normalisation following effective therapy, the exact nature of these acinar structures remains uncertain. Future validation with immunohistochemical markers (such as cytokeratin panels or Ki-67) will be necessary to confirm whether these regions correspond to truly residual tumour cells, quiescent epithelial remnants, or benign ductal regeneration. Such multimodal correlation could help refine the interpretability of deep learning models in treated breast cancer specimens and delineate genuine biological signals from histology.

While qualitative agreement between expert interpretation and network saliency reinforces trust in the model, caution remains warranted. Deep networks may amplify correlated but non-causal cues (e.g. tissue preparation artefacts, biopsy depth, or tumour grade). Future work combining histological, molecular, and spatial-omics annotations will be necessary to disentangle predictive morphology from confounding context and to confirm whether the features identified truly mediate treatment response or simply mirror it.

Chapter 6 - General Discussion

The fundamental clinical challenge addressed by this thesis is the stratification of the residual disease after neoadjuvant chemotherapy (NAC) in early breast cancer. Patients with a pathological complete response (pCR) have excellent outcomes, whereas those with residual invasive disease remain at substantially higher risk of recurrence and death. Contemporary frameworks (e.g., RCB, cellularity, nodal status) capture only part of this heterogeneity and rely on manual assessment. At the same time, comprehensive molecular assays that could refine risk prediction are not universally accessible, may be delayed, and do not always map cleanly to histological phenotype. This work therefore asks a focused question: to what extent can routine H&E morphology, read by deep learning under biologically informed supervision, support pre- and post-treatment risk stratification and guide therapeutic decisions ?

6.1 What the results collectively show

Across chapters, three conclusions emerge that are directly relevant to post-NAC risk:

- **Biologically structured supervision**

When labels reflect functional biology (variants rather than genes; PAM50/HRD rather than undifferentiated endpoints) morphological learning sharpens. This is why MultiVarNet's variant-aware design yields measurable, though moderate, gains, and why PAM50/HRD features improve survival stratification from surgical slides. The granularity and biological validity of the target define the ceiling of morphology-based prediction.

- **Morpho-molecular signals are context-bound.**

Cross-cancer transfer and domain-adversarial learning show that the same mutation does not induce a uniform morphological signature across tissues; exceptions (e.g., IDH1 across LGG/GBM) are lineage-consistent rather than pan-cancer. For post-NAC stratification, this means models perform best when trained and validated within the precise clinicopathologic context of use (subtype, scanner/stain pipeline, treatment era).

- **Risk ordering is robust even when time calibration is modest.**

Survival models showed only moderate C-index/time-to-event accuracy, yet their Kaplan–Meier separations were strong, indicating that the networks captured relative risk ranking more reliably than exact event timing in censored, real-world data. Practically, this supports the use of model-derived risk scores as stratification tools (who is high vs low risk), which is exactly what post-NAC decisions require, while acknowledging limits for precise temporal prediction.

Implications for residual disease stratification

Taken together, these findings support a two-step, morphology-driven pathway that aligns with clinical workflow:

- Before NAC (biopsy): a baseline, histology-only estimate of intrinsic chemosensitivity (Chemo-prAIdict Breast) can suggest de-escalation (when high pCR likelihood to standard backbones) or escalation/trial referral (when low), particularly in TNBC where morphology and treatment response are tightly coupled.
- After NAC (surgery): a histology-based residual-risk score complements RCB by capturing diffuse features (necrosis, pleomorphism, stromal patterns) associated with poor or favourable prognosis. Here, the consistent KM separation, even under censoring, argues for practical utility in who to escalate (e.g., capecitabine, olaparib for BRCA1/2, T-DM1 in HER2+) and who may be spared.

Pathologist-guided review shows that the networks focus on credible histological correlates (cellularity, nuclear atypia, necrosis, fibrosis) rather than artefacts. Notably, some features display directional reversals across the treatment timeline: dense fibrosis is adverse at baseline (barrier to drug penetration) but favourable post-therapy (marker of regression). Morphology is not merely static structure but can look like a trajectory of tissue response, and the learned representations appear sensitive to that trajectory. Future work should test whether these high-attention regions map to immune activation, stromal reprogramming, or DNA-damage response using spatial-omics or multiplex proteomics.

Chapter III's benchmarking delivers a cautionary message for the field: on controlled benchmarks, highly parameterised, end-to-end models can look excellent yet fail on external data, whereas foundation-model features + Cox generalise more reliably and are easier to

audit. In a setting like post-NAC stratification where decisions carry direct therapeutic consequences, parsimony, transparency, and calibration matter more than incremental AUC gains on familiar datasets. This is the rationale for treating these systems as analytical tools under fixed splits, nested CV, and external validation, on a path toward regulatory evaluation.

The thesis therefore positions histology-based deep learning not as a replacement for molecular assays, but as a triage and decision-support layer that is:

(i) available at scale,

(ii) biologically oriented via variant/signature supervision, and

(iii) clinically aligned with pre- and post-NAC decision points.

Its utility is likely greatest where outcome heterogeneity is high and therapy is not already homogenising biology (e.g., TNBC vs HER2+), and where time-to-insight matters (baseline planning, adjuvant escalation).

While these contributions clarify how and when morphology can inform therapy, they also reveal important constraints (biological, technical, and regulatory) that shape real-world performance and translation. We therefore next detail the threats to validity and the steps we took to mitigate them.

6.2 Limitations

Performance and external validity

Although both neural-network frameworks achieved encouraging discrimination, especially for ER+/HER2- chemoresistance prediction and prognosis (OS) stratification of the residual disease using PAM50/HRD virtual molecular signatures, absolute performance remains moderate and below what would be required for clinical deployment.

This reflects both the biological complexity of treatment response and the technical constraints of our datasets. The external validation cohort was monocentric and digitised using the same scanner models as the training set, limiting our ability to estimate robustness across staining, scanning, and institutional domains. None of the systems used were CE-marked or subject to medical-device conformity assessment, which underscores that our results represent analytical research findings rather than a clinically validated product.

Future evaluations should therefore include multi-scanner, multi-centre studies with independent site calibrations and prospective external testing to meet regulatory reproducibility standards.

Regimen heterogeneity and temporal drift

All patients in our training and validation cohorts were treated before 2022, i.e. before the widespread adoption in France of KEYNOTE-522-type regimens combining anthracycline–taxane chemotherapy with carboplatin and pembrolizumab in triple-negative breast cancer (TNBC), and before routine dual HER2 blockade (trastuzumab + pertuzumab) in HER2-positive disease. No patient received neoadjuvant carboplatin, pembrolizumab, or pertuzumab. Consequently, our models estimate intrinsic chemosensitivity to the anthracycline–taxane backbone, the pharmacological foundation that persists within all modern regimens, but they cannot directly predict the effect of immune or targeted agents currently gold standard treatment protocols for early breast cancer patients.

Given that current pCR rates in TNBC now reach ~65 % with carboplatin + pembrolizumab and that dual HER2 blockade markedly improves response, our results should be interpreted as baseline predictors that might inform escalation or de-escalation around the chemotherapy core rather than as models of total regimen response. Prospective retraining on modern cohorts will be necessary to ensure clinical relevance.

Molecular-context dependence of morpho-molecular signatures

While the *MultiVarNet* experiments demonstrate that morpho-molecular signatures appear unique to each gene–variant–tissue context, these signatures are not isolated phenomena. They exist within broader *molecular ecosystems* defined by co-mutations and co-expression patterns that influence phenotype. For example, *XIRP2* is frequently co-mutated with *KRAS*, modifying downstream cytoskeletal dynamics and possibly altering the resulting morphology, but this interaction was not captured by our single-variant models.

Furthermore, the limited transferability of morpho-molecular signatures across tumour types likely reflects not only tissue-specific programmes and microenvironmental influences, but also mutation clonality and subclonality. Highly clonal driver mutations may produce more robust and detectable histopathological correlates, whereas subclonal variants, common in many cancers, can generate heterogeneous morphological signals that are more difficult for

deep learning models to consistently identify. This added layer of complexity further supports our shift from the pursuit of universal models toward well-defined, context-specific companion diagnostics.

Hence, variant-specific morphology should be viewed as one projection of a multidimensional molecular fingerprint, rather than a stand-alone determinant. Capturing these interactions will require a *MultiVarNet 2.0* that models co-mutational context, transcriptomic modules, and protein expression, using multimodal learning or graph-based architectures integrating variant co-occurrence priors.

Biological and clinical label limitations

The histological signals captured by the networks are constrained by the precision of the labels that supervised them. For mutation prediction, the binary annotation “mutated / wild-type” aggregates diverse functional effects and inevitably includes both driver and passenger mutations. For treatment response, the endpoint pCR versus RD is regimen-specific and time-dependent; it conflates distinct biological mechanisms such as immune activation, apoptosis, and fibrosis.

Importantly, our models do not directly predict chemosensitivity. Chemo-prAIdict Breast and the residual-disease network predict prognosis (DFS and OS) or the probability of pathological complete response under a specific regimen, rather than the true differential treatment effect. A genuine prediction of chemosensitivity would require randomised data with a control arm (e.g., placebo or standard-of-care versus experimental treatment) and the modelling of an interaction term (treatment arm × predicted score) to estimate the individual benefit of therapy. In the absence of such counterfactual data, our predictions remain correlative surrogates of treatment response. Future studies aiming at true personalised prediction of therapeutic benefit will need access to randomised controlled trial cohorts with appropriate control arms.

Our labels therefore act as noisy surrogates of the biological reality. Future studies should refine endpoints toward mechanistic correlates, such as immune-related pathologic response (irPR) scores or residual-cancer burden (RCB) combined with molecular response signatures, and pair mutation prediction with functional readouts (protein loss, RNA expression) to mitigate label noise.

Model generalisation and overfitting

The methodological benchmarking performed in Chapter III revealed a recurring pattern: in controlled datasets such as TCGA, sophisticated deep architectures (multi-task survival networks, end-to-end WSIs) can reach very high performance, but when applied to unseen, real-world data, these same models overfit severely. In contrast, simpler pipelines, notably foundation-model feature extraction followed by Cox proportional hazards regression, generalised more reliably. This finding tempers the optimism of much of the recent methodological literature: apparent improvements on public benchmarks often reflect data familiarity rather than generalisable learning.

It highlights that model simplicity, transparency, and clinical plausibility may outweigh incremental AUC gains from highly parameterised architectures. For the field, this advocates a shift from benchmark-driven performance inflation to prospective, cross-site validation as the gold standard.

Interpretability limits

Although the pathologist-guided review confirmed that the networks rely on recognisable histological motifs, interpretability remains qualitative. No causal inference can be drawn regarding whether these features mediate or merely correlate with treatment response.

The next step will be to combine digital-morphology predictions with spatial-transcriptomic, multiplex-immunofluorescence, or proteomic imaging data to test whether the high-attention regions correspond to active biological processes (immune activation, stromal reprogramming, DNA-damage response). This integration would enable causal triangulation between morphology, molecular pathways, and therapeutic outcome.

In summary, despite the encouraging results obtained for variant prediction, chemosensitivity, and residual-disease stratification, the present work remains an analytical proof of concept. Its models perform well within their technical confines but require broader, contemporary, and multi-institutional validation before clinical translation. The findings delineate the true boundaries of morphology-based AI: promising as a triage and discovery tool, but limited by regimen evolution, contextual molecular complexity, and real-world heterogeneity. Future development should therefore prioritise modern treatment cohorts, multimodal integration,

explicit uncertainty quantification, and biological causal validation to transform these analytical signals into reliable clinical instruments.

6.3 Final perspective: from universal models to companion diagnostics

A central insight emerging from this thesis is that the promise of histology-based deep learning is inherently limited by strong context dependence. Across all experiments, from variant prediction to survival modelling, we consistently observed that morpho-molecular signatures and prognostic patterns are highly tissue-, subtype-, and even treatment-specific. These findings do not support broad pan-cancer generalisability. On the contrary, they strongly advocate for a context-specific, tissue-specific companion diagnostic approach, in which models are developed, validated, and deployed for well-defined clinical indications, scanners, and patient populations.

This shift is not merely technical but fundamental. Even with domain-adversarial training, cross-cancer transfer largely failed, except in biologically coherent cases such as IDH1 mutations in gliomas. The limited transferability likely reflects not only tissue context and microenvironment, but also mutation clonality/subclonality and treatment-induced remodelling. Consequently, the field should move away from the pursuit of universal models toward carefully scoped, indication-specific tools analogous to companion diagnostics in molecular pathology. This perspective aligns with current regulatory expectations and offers a more realistic and safer path toward clinical translation[159].

This evolution has deep implications. First, it limits interpretability: if a model's weights are cohort-dependent, then its highlighted features cannot be assumed to represent general morphological principles. Second, it complicates clinical translation: deploying AI will likely require CE-marked or FDA-cleared models tied to specific scanner–stain–cohort combinations, as already seen in pathology products cleared by the FDA since 2021. Third, it challenges the economic and organisational model of digital pathology. Rather than open, general-purpose algorithms, we may see a proliferation of vendor-specific, proprietary companion tools co-developed with pharmaceutical companies, integrated into closed ecosystems that replicate the economic model of targeted therapies. This raises questions about interoperability, reproducibility, data ownership, and long-term sustainability—issues

now explicitly acknowledged by regulatory agencies (EMA 2023 AI Roadmap; UK MHRA 2023 AIaMD Framework).

In that sense, the next frontier of computational pathology will not only be scientific, but strategic: balancing innovation with reproducibility, and open science with regulatory compliance. Progress will depend on the creation of transparent data standards, federated validation networks, and economically sustainable models for updating and auditing algorithms across centres. Without this, histology-based AI risks fragmenting into a patchwork of narrow, non-interoperable systems, scientifically impressive, yet clinically brittle.

This shift, however, highlights a profound asymmetry between molecular and AI-based diagnostics. In molecular biology, one can design a hypothesis-driven assay within a controlled laboratory setting (for instance, identifying a prognostic gene signature) and then validate it prospectively in a phase III clinical trial or through non-inferiority studies (as in Oncotype DX or MammaPrint). The development and validation are sequential, each step refining a fixed, interpretable test. By contrast, deep-learning models depend on large, well-annotated datasets for training, which themselves must possess the same level of standardisation and clinical precision as a phase III trial. This means that, in practice, the data required to train the algorithm already presuppose the existence of a controlled, trial-like infrastructure. True clinical validation would then require a second, equally rigorous prospective trial, effectively doubling the cost and complexity compared to traditional companion diagnostics. Such a model is economically and logistically unsustainable for most academic or independent AI developers, and will likely restrict regulatory-grade model development to industrial or pharmaceutical partnerships.

In summary, this thesis defines both the scientific potential and the structural limits of deep learning in pathology.

Morphology encodes genuine biological information, but the translation of that information into clinical practice will require accepting that deep-learning models are not universal interpreters of disease, they are contextual instruments, whose validity must be demonstrated within the precise boundaries of their intended use.

Recognising this shift toward a companion-diagnostic era is essential to guide future research, regulation, and business models for digital pathology that are scientifically credible, clinically safe, and economically viable.

References

1. Balss, J. et al. Analysis of the IDH1 codon 132 mutation in brain tumors. Acta Neuropathol. (Berl.) 116, 597–602 (2008).
2. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. Nat. Med. 19, 1423–1437 (2013).
3. Pusztai, L. et al. Event-free survival by residual cancer burden with pembrolizumab in early-stage TNBC: exploratory analysis from KEYNOTE-522. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. 35, 429–436 (2024).
4. Geyer, C. E. et al. Overall survival in the OlympiA phase III trial of adjuvant olaparib in patients with germline pathogenic variants in BRCA1/2 and high-risk, early breast cancer. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. 33, 1250–1268 (2022).
5. Masuda, N. et al. Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. N. Engl. J. Med. 376, 2147–2159 (2017).
6. Survival with Trastuzumab Emtansine in Residual HER2-Positive Breast Cancer | New England Journal of Medicine. https://www.nejm.org/doi/full/10.1056/NEJMoa2406070.
7. Viale, G. & Fusco, N. Pathology after neoadjuvant treatment – How to assess residual disease. Breast Off. J. Eur. Soc. Mastology 62, S25–S28 (2021).
8. Cardoso, F. et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. 30, 1194–1220 (2019).
9. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. 71, 209–249 (2021).
10. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. Lancet Oncol. 13, 1141–1151 (2012).
11. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. The Lancet 360, 187–195 (2002).
12. The Endogenous Hormones and Breast Cancer Collaborative Group. Endogenous Sex Hormones and Breast Cancer in Postmenopausal Women: Reanalysis of Nine Prospective Studies. JNCI J. Natl. Cancer Inst. 94, 606–616 (2002).

13. Duffy, S. W. et al. Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial. Lancet Oncol. 21, 1165–1172 (2020).
14. Duffy, S. et al. Annual mammographic screening to reduce breast cancer mortality in women from age 40 years: long-term follow-up of the UK Age RCT. Health Technol. Assess. 24, 1–24 (2020).
15. Duffy, S. W. et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer 126, 2971–2979 (2020).
16. Rosai and Ackerman's Surgical Pathology. (Elsevier, Philadelphia, PA, 2018).
17. Street, W. Breast Cancer Facts & Figures 2019-2020.
18. Giuliano, A. E., Edge, S. B. & Hortobagyi, G. N. Eighth Edition of the AJCC Cancer Staging Manual: Breast Cancer. Ann. Surg. Oncol. 25, 1783–1785 (2018).
19. Amat, S. et al. Scarff-Bloom-Richardson (SBR) grading: a pleiotropic marker of chemosensitivity in invasive ductal breast carcinomas treated by neoadjuvant chemotherapy. Int. J. Oncol. https://doi.org/10.3892/ijo.20.4.791 (2002) doi:10.3892/ijo.20.4.791.
20. Allison, K. H. et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. J. Clin. Oncol. 38, 1346–1366 (2020).
21. Wolff, A. C. et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. Arch. Pathol. Lab. Med. 142, 1364–1382 (2018).
22. Swain Sandra M. et al. Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer. N. Engl. J. Med. 372, 724–734 (2015).
23. Denkert, C. et al. Clinical and molecular characteristics of HER2-low-positive breast cancer: pooled analysis of individual patient data from four prospective, neoadjuvant clinical trials. Lancet Oncol. 22, 1151–1161 (2021).
24. Mark, C., Lee, J. S., Cui, X. & Yuan, Y. Antibody–Drug Conjugates in Breast Cancer: Current Status and Future Directions. Int. J. Mol. Sci. 24, 13726 (2023).
25. Oxford Textbook of Oncology. (Oxford university press, Oxford, 2016).
26. Cortazar, P. et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. The Lancet 384, 164–172 (2014).
27. Spring, L. M. et al. Pathologic Complete Response after Neoadjuvant Chemotherapy and Impact on Breast Cancer Recurrence and Survival: A Comprehensive Meta-analysis. Clin. Cancer Res. 26, 2838–2848 (2020).

28. Lehmann, B. D. et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. 121, 2750–2767 (2011).
29. Minckwitz, G. von et al. Neoadjuvant carboplatin in patients with triple-negative and HER2-positive early breast cancer (GeparSixto; GBG 66): a randomised phase 2 trial. Lancet Oncol. 15, 747–756 (2014).
30. Schmid, P. et al. Event-free Survival with Pembrolizumab in Early Triple-Negative Breast Cancer. N. Engl. J. Med. 386, 556–567 (2022).
31. Asselain, B. et al. Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. Lancet Oncol. 19, 27–39 (2018).
32. Agostinetto, E. et al. Post-Neoadjuvant Treatment Strategies for Patients with Early Breast Cancer. Cancers 14, 5467 (2022).
33. Von Minckwitz, G. et al. Trastuzumab Emtansine for Residual Invasive HER2-Positive Breast Cancer. N. Engl. J. Med. 380, 617–628 (2019).
34. Ma, J., Chan, J. J., Toh, C. H. & Yap, Y.-S. Emerging systemic therapy options beyond CDK4/6 inhibitors for hormone receptor-positive HER2-negative advanced breast cancer. Npj Breast Cancer 9, 74 (2023).
35. Wang, X. et al. Recent progress of CDK4/6 inhibitors' current practice in breast cancer. Cancer Gene Ther. 1–9 (2024) doi:10.1038/s41417-024-00747-x.
36. Caparica, R. et al. Post-neoadjuvant treatment and the management of residual disease in breast cancer: state of the art and perspectives. Ther. Adv. Med. Oncol. 11, 1758835919827714 (2019).
37. Dieci, M. V., Arnedos, M., Delaloge, S. & Andre, F. Quantification of residual risk of relapse in breast cancer patients optimally treated. The Breast 22, S92–S95 (2013).
38. Dogan, S. et al. Landscape and evolution of therapeutic research for breast cancer patients. Breast Cancer Res. Treat. 138, 319–324 (2013).
39. Yau, C. et al. Residual cancer burden after neoadjuvant chemotherapy and long-term survival outcomes in breast cancer: a multicentre pooled analysis of 5161 patients. Lancet Oncol. 23, 149–160 (2022).
40. Mittendorf, E. A. et al. The Neo-Bioscore Update for Staging Breast Cancer Treated With Neoadjuvant Chemotherapy: Incorporation of Prognostic Biologic Factors Into Staging After Treatment. JAMA Oncol. 2, 929 (2016).

41. Nielsen, T. O. et al. Assessment of Ki67 in Breast Cancer: Updated Recommendations From the International Ki67 in Breast Cancer Working Group. JNCI J. Natl. Cancer Inst. 113, 808–819 (2021).
42. Denkert, C. et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol. 19, 40–50 (2018).
43. Pease, A. M., Riba, L. A., Gruner, R. A., Tung, N. M. & James, T. A. Oncotype DX® Recurrence Score as a Predictor of Response to Neoadjuvant Chemotherapy. Ann. Surg. Oncol. 26, 366–371 (2019).
44. Cescon, D. W. et al. Therapeutic Targeting of Minimal Residual Disease to Prevent Late Recurrence in Hormone-Receptor Positive Breast Cancer: Challenges and New Approaches. Front. Oncol. 11, 667397 (2022).
45. Hamy, A.-S. et al. Prognostic value of the Residual Cancer Burden index according to breast cancer subtype: Validation on a cohort of BC patients treated by neoadjuvant chemotherapy. PLoS ONE 15, e0234191 (2020).
46. Yoshioka, T. et al. Prognostic significance of pathologic complete response and Ki67 expression after neoadjuvant chemotherapy in breast cancer. Breast Cancer 22, 185–191 (2015).
47. Mittendorf, E. A. et al. Validation of a Novel Staging System for Disease-Specific Survival in Patients With Breast Cancer Treated With Neoadjuvant Chemotherapy. J. Clin. Oncol. 29, 1956–1962 (2011).
48. Licata, L. et al. Oncotype DX results increase concordance in adjuvant chemotherapy recommendations for early-stage breast cancer. Npj Breast Cancer 9, 1–7 (2023).
49. Dubsky, P. et al. The EndoPredict score provides prognostic information on late distant metastases in ER+/HER2− breast cancer patients. Br. J. Cancer 109, 2959–2964 (2013).
50. Soliman, H. et al. MammaPrint guides treatment decisions in breast Cancer: results of the IMPACt trial. BMC Cancer 20, 81 (2020).
51. Wallden, B. et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med. Genomics 8, 54 (2015).
52. Loibl, S. et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann. Oncol. 35, 159–182 (2024).

53. Ignatiadis, M., Lee, M. & Jeffrey, S. S. Circulating Tumor Cells and Circulating Tumor DNA: Challenges and Opportunities on the Path to Clinical Utility. Clin. Cancer Res. 21, 4786–4800 (2015).
54. Stecklein, S. R. et al. ctDNA and residual cancer burden are prognostic in triple-negative breast cancer patients with residual disease. Npj Breast Cancer 9, 1–8 (2023).
55. Nader-Marta, G. et al. Circulating tumor DNA for predicting recurrence in patients with operable breast cancer: a systematic review and meta-analysis. ESMO Open 9, 102390 (2024).
56. Munoz-Arcos, L. S. et al. Latest advances in clinical studies of circulating tumor cells in early and metastatic breast cancer. in International Review of Cell and Molecular Biology vol. 381 1–21 (Elsevier, 2023).
57. Trapp, E. et al. Presence of Circulating Tumor Cells in High-Risk Early Breast Cancer During Follow-Up and Prognosis. JNCI J. Natl. Cancer Inst. 111, 380–387 (2019).
58. Mateo, J. et al. Delivering precision oncology to patients with cancer. Nat. Med. 28, 658–665 (2022).
59. Trosman, J. R., Weldon, C. B., Kelley, R. K. & Phillips, K. A. Challenges of Coverage Policy Development for Next-Generation Tumor Sequencing Panels: Experts and Payers Weigh In. J. Natl. Compr. Canc. Netw. 13, 311–318 (2015).
60. Gondos, A. et al. Genomic testing among patients (pts) with newly diagnosed advanced non-small cell lung cancer (aNSCLC) in the United States: A contemporary clinical practice patterns study. J. Clin. Oncol. 38, 9592–9592 (2020).
61. Aziz, Z. et al. Cost-Effectiveness of Liquid Biopsy for Colorectal Cancer Screening in Patients Who Are Unscreened. JAMA Netw. Open 6, e2343392 (2023).
62. Cruz-Roa, A. et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Sci. Rep. 7, 46450 (2017).
63. Falk, T. et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods 16, 67–70 (2019).
64. Shahin, A. I., Guo, Y., Amin, K. M. & Sharawi, A. A. White blood cells identification system based on convolutional deep neural learning networks. Comput. Methods Programs Biomed. 168, 69–80 (2019).
65. Schmauch, B. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat. Commun. 11, 3877 (2020).

66. Morel, L.-O., Derangère, V., Arnould, L., Ladoire, S. & Vinçon, N. Preliminary evaluation of deep learning for first-line diagnostic prediction of tumor mutational status. Sci. Rep. 13, 6927 (2023).

67. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in Advances in Neural Information Processing Systems vol. 25 (Curran Associates, Inc., 2012).

68. Reichling, C. et al. Artificial intelligence-guided tissue analysis combined with immune infiltrate assessment predicts stage III colon cancer outcomes in PETACC08 study. Gut 69, 681–690 (2020).

69. Crowell, E. F. et al. CytoProcessorTM: A New Cervical Cancer Screening System for Remote Diagnosis. Acta Cytol. 63, 215–223 (2019).

70. Cireşan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. Med. Image Comput. Comput.-Assist. Interv. MICCAI Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 16, 411–418 (2013).

71. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. ArXiv150302531 Cs Stat http://arxiv.org/abs/1503.02531 (2015).

72. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv190511946 Cs Stat http://arxiv.org/abs/1905.11946 (2020).

73. Courtiol, P., Tramel, E. W., Sanselme, M. & Wainrib, G. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. ArXiv180202212 Cs Stat http://arxiv.org/abs/1802.02212 (2020).

74. Amores, J. Multiple instance classification: Review, taxonomy and comparative study. Artif. Intell. 201, 81–105 (2013).

75. Deep learning of feature representation with multiple instance learning for medical image analysis | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/6853873.

76. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25, 1301–1309 (2019).

77. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based Deep Multiple Instance Learning. ArXiv180204712 Cs Stat http://arxiv.org/abs/1802.04712 (2018).

78. Ehteshami Bejnordi, B. et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318, 2199–2210 (2017).
79. Jarkman, S. et al. Generalization of Deep Learning in Digital Pathology: Experience in Breast Cancer Metastasis Detection. Cancers 14, 5424 (2022).
80. Lu, M. Y., Chen, R. J., Wang, J., Dillon, D. & Mahmood, F. Semi-Supervised Histology Classification using Deep Multiple Instance Learning and Contrastive Predictive Coding. Preprint at https://doi.org/10.48550/arXiv.1910.10825 (2019).
81. Ghaffari Laleh, N. et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med. Image Anal. 79, 102474 (2022).
82. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A. & Courtiol, P. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. ArXiv201203583 Cs Eess http://arxiv.org/abs/2012.03583 (2020).
83. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. ArXiv191105722 Cs http://arxiv.org/abs/1911.05722 (2020).
84. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. Preprint at http://arxiv.org/abs/2002.05709 (2020).
85. Caron, M. et al. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. in Advances in Neural Information Processing Systems vol. 33 9912–9924 (Curran Associates, Inc., 2020).
86. Alfasly, S. et al. Foundation Models for Histopathology—Fanfare or Flair. Mayo Clin. Proc. Digit. Health 2, 165–174 (2024).
87. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. Nat. Med. 30, 850–862 (2024).
88. Lu, M. Y. et al. A visual-language foundation model for computational pathology. Nat. Med. 30, 863–874 (2024).
89. Therneau, T. M. & Grambsch, P. M. Modeling Survival Data: Extending the Cox Model. (Springer, New York, NY, 2000). doi:10.1007/978-1-4757-3294-8.
90. Wulczyn, E. et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLOS ONE 15, e0233678 (2020).
91. Sullivan, L. Survival Analysis. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival_print.html

92. Cox, D. R. Regression Models and Life-Tables. J. R. Stat. Soc. Ser. B Methodol. 34, 187–220 (1972).

93. Zhu, X., Yao, J. & Huang, J. Deep convolutional neural network for survival analysis with pathological images. in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 544–547 (2016). doi:10.1109/BIBM.2016.7822579.

94. Katzman, J. L. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol. 18, 24 (2018).

95. Zhu, X., Yao, J., Zhu, F. & Huang, J. WSISA: Making Survival Prediction From Whole Slide Histopathological Images. in 7234–7242 (2017).

96. Meier, A. et al. Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. J. Pathol. Clin. Res. 6, 273–282 (2020).

97. Lee, C., Zame, W., Yoon, J. & Van Der Schaar, M. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proc. AAAI Conf. Artif. Intell. 32, (2018).

98. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat. Genet. 45, 1113–1120 (2013).

99. Waks, A. G. & Winer, E. P. Breast Cancer Treatment. JAMA 321, 316 (2019).

100. Basch, E. et al. Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. JAMA 318, 197–198 (2017).

101. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U. S. A. 98, 10869–10874 (2001).

102. Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19, 403–410 (1991).

103. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. U. S. A. 115, E2970–E2979 (2018).

104. Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018).

105. Spring, L. M. et al. Cyclin-dependent kinase 4 and 6 inhibitors for hormone receptor-positive breast cancer: past, present, and future. The Lancet 395, 817–827 (2020).

106. Loibl, S. et al. Palbociclib for Residual High-Risk Invasive HR-Positive and HER2-Negative Early Breast Cancer—The Penelope-B Trial. J. Clin. Oncol. 39, 1518–1530 (2021).
107. Prat, A. et al. Ribociclib plus letrozole versus chemotherapy for postmenopausal women with hormone receptor-positive, HER2-negative, luminal B breast cancer (CORALLEEN): an open-label, multicentre, randomised, phase 2 trial. Lancet Oncol. 21, 33–43 (2020).
108. Delaloge, S. et al. Survival outcomes after neoadjuvant letrozole and palbociclib versus third generation chemotherapy for patients with high-risk oestrogen receptor-positive HER2-negative breast cancer. Eur. J. Cancer 166, 300–308 (2022).
109. De Caluwé, A. et al. Neo-CheckRay: radiation therapy and adenosine pathway blockade to increase benefit of immuno-chemotherapy in early stage luminal B breast cancer, a randomized phase II trial. BMC Cancer 21, 899 (2021).
110. Geyer, C. E. et al. Long-term efficacy and safety of addition of carboplatin with or without veliparib to standard neoadjuvant chemotherapy in triple-negative breast cancer: 4-year follow-up data from BrighTNess, a randomized phase III trial. Ann. Oncol. 33, 384–394 (2022).
111. von Minckwitz, G. et al. Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. N. Engl. J. Med. 377, 122–131 (2017).
112. Matikas, A. et al. Survival Outcomes, Digital TILs, and On-treatment PET/CT During Neoadjuvant Therapy for HER2-positive Breast Cancer: Results from the Randomized PREDIX HER2 Trial. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 29, 532–540 (2023).
113. Ramshorst, M. S. van et al. Neoadjuvant chemotherapy with or without anthracyclines in the presence of dual HER2 blockade for HER2-positive breast cancer (TRAIN-2): a multicentre, open-label, randomised, phase 3 trial. Lancet Oncol. 19, 1630–1640 (2018).
114. Schneeweiss, A. et al. 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). Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. 24, 2278–2284 (2013).
115. Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat. Commun. 12, 4423 (2021).
116. Chen, X., Xie, S. & He, K. An Empirical Study of Training Self-Supervised Vision Transformers. Preprint at https://doi.org/10.48550/arXiv.2104.02057 (2021).

117. Chen, X. & He, K. Exploring Simple Siamese Representation Learning. Preprint at https://doi.org/10.48550/arXiv.2011.10566 (2020).
118. Caron, M. et al. Emerging Properties in Self-Supervised Vision Transformers. ArXiv210414294 Cs http://arxiv.org/abs/2104.14294 (2021).
119. Punn, N. S. & Agarwal, S. BT-Unet: A self-supervised learning framework for biomedical image segmentation using Barlow Twins with U-Net models. Preprint at https://doi.org/10.48550/arXiv.2112.03916 (2022).
120. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. in Proceedings of the 38th International Conference on Machine Learning 12310–12320 (PMLR, 2021).
121. Grill, J.-B. et al. Bootstrap your own latent: A new approach to self-supervised Learning. Preprint at https://doi.org/10.48550/arXiv.2006.07733 (2020).
122. Bardes, A., Ponce, J. & LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. Preprint at https://doi.org/10.48550/arXiv.2105.04906 (2022).
123. Yang, Z. et al. Deep learning-based overall survival prediction in patients with glioblastoma: An automatic end-to-end workflow using pre-resection basic structural multiparametric MRIs. Comput. Biol. Med. 185, 109436 (2025).
124. He, T., Huang, L., Li, J., Wang, P. & Zhang, Z. Potential Prognostic Immune Biomarkers of Overall Survival in Ovarian Cancer Through Comprehensive Bioinformatics Analysis: A Novel Artificial Intelligence Survival Prediction System. Front. Med. 8, 587496 (2021).
125. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
126. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550 (2014).
127. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat. Biotechnol. 38, 737–746 (2020).
128. Nichols, R. V. et al. High-throughput robust single-cell DNA methylation profiling with sciMETv2. Nat. Commun. 13, 7627 (2022).
129. Flaherty, K. T. et al. Molecular Landscape and Actionable Alterations in a Genomically Guided Cancer Clinical Trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH). J. Clin. Oncol. 38, 3883–3894 (2020).

130. Jager, V. D. de et al. Developments in predictive biomarker testing and targeted therapy in advanced stage non-small cell lung cancer and their application across European countries. Lancet Reg. Health – Eur. 38, (2024).
131. Kerr, K. M. et al. The evolving landscape of biomarker testing for non-small cell lung cancer in Europe. Lung Cancer 154, 161–175 (2021).
132. Laleh, N. G. et al. Deep Learning for interpretable end-to-end survival (E-ESurv) prediction in gastrointestinal cancer histopathology. in Proceedings of the MICCAI Workshop on Computational Pathology 81–93 (PMLR, 2021).
133. Lee, S. H. & Jang, H.-J. Deep learning-based prediction of molecular cancer biomarkers from tissue slides: A new tool for precision oncology. Clin. Mol. Hepatol. 28, 754–772 (2022).
134. Murchan, P. et al. Deep Learning of Histopathological Features for the Prediction of Tumour Molecular Genetics. Diagnostics 11, 1406 (2021).
135. Skoulidis, F. et al. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. N. Engl. J. Med. 384, 2371–2381 (2021).
136. Jänne, P. A. et al. Adagrasib in Non–Small-Cell Lung Cancer Harboring a KRASG12C Mutation. N. Engl. J. Med. 387, 120–131 (2022).
137. Liu, B., Zhou, H., Tan, L., Siu, K. T. H. & Guan, X.-Y. Exploring treatment options in cancer: tumor treatment strategies. Signal Transduct. Target. Ther. 9, 175 (2024).
138. Glodzik, D. et al. Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. Nat. Commun. 11, 3747 (2020).
139. Dumas, N. et al. Inter-Semantic Domain Adversarial in Histopathological Images. Preprint at https://doi.org/10.48550/arXiv.2201.09041 (2022).
140. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes | Journal of Clinical Oncology. https://ascopubs.org/doi/10.1200/JCO.2008.18.1370.
141. Ray-Coquard, I. et al. Olaparib plus Bevacizumab as First-Line Maintenance in Ovarian Cancer. N. Engl. J. Med. 381, 2416–2428 (2019).
142. Yann, C. Actualisation 2023 : utilité clinique des signatures génomiques dans le cancer du sein RH+/HER2- de stade précoce. (2023).
143. Tutt, A. N. J. et al. Adjuvant Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer. N. Engl. J. Med. 384, 2394–2405 (2021).
144. Robson, M. et al. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. N. Engl. J. Med. 377, 523–533 (2017).

145. Litton, J. K. et al. Talazoparib in Patients with Advanced Breast Cancer and a Germline BRCA Mutation. N. Engl. J. Med. 379, 753–763 (2018).
146. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).
147. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012).
148. A, J., R, Z., H, G., M, F. & A, M. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. JCO Clin. Cancer Inform. 3, (2019).
149. Shen, Y., Chen, J.-Q. & Li, X.-P. Differences between lung adenocarcinoma and lung squamous cell carcinoma: Driver genes, therapeutic targets, and clinical efficacy. Genes Dis. 12, 101374 (2025).
150. Huang, T. et al. Distinguishing Lung Adenocarcinoma from Lung Squamous Cell Carcinoma by Two Hypomethylated and Three Hypermethylated Genes: A Meta-Analysis. PLoS ONE 11, e0149088 (2016).
151. Wang, X., Zheng, K. & Hao, Z. In-depth analysis of immune cell landscapes reveals differences between lung adenocarcinoma and lung squamous cell carcinoma. Front. Oncol. 14, (2024).
152. Dolezal, J. M. et al. Deep learning generates synthetic cancer histology for explainability and education. Npj Precis. Oncol. 7, 49 (2023).
153. Song, S. et al. Identification of immune-related gene signature for predicting prognosis in uterine corpus endometrial carcinoma. Sci. Rep. 13, 9255 (2023).
154. Nojima, S. et al. Deep Learning-Based Differential Diagnosis of Follicular Thyroid Tumors Using Histopathological Images. Mod. Pathol. 36, (2023).
155. Coudray, N. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. 24, 13 (2018).
156. Choudhary, A., Tong, L., Zhu, Y. & Wang, M. D. Advancing Medical Imaging Informatics by Deep Learning-Based Domain Adaptation. Yearb. Med. Inform. 29, 129–138 (2020).
157. Ganin, Y. et al. Domain-Adversarial Training of Neural Networks. in Domain Adaptation in Computer Vision Applications (ed. Csurka, G.) 189–209 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-58347-1_10.
158. Dumas, N. et al. Inter-Semantic Domain Adversarial in Histopathological Images. Preprint at https://doi.org/10.48550/arXiv.2201.09041 (2022).

159. Echle, A. et al. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br. J. Cancer 124, 686–696 (2021).
160. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer 1, 789–799 (2020).
161. Adam, G. et al. Machine learning approaches to drug response prediction: challenges and recent progress. Npj Precis. Oncol. 4, 19 (2020).
162. ai, kaiko et al. Towards Large-Scale Training of Pathology Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2404.15217 (2024).
163. Filiot, A., Jacob, P., Kain, A. M. & Saillard, C. Phikon-v2, A large and public feature extractor for biomarker prediction. Preprint at https://doi.org/10.48550/arXiv.2409.09173 (2024).
164. Johnston, S. R. D. et al. Abemaciclib Combined With Endocrine Therapy for the Adjuvant Treatment of HR+, HER2-, Node-Positive, High-Risk, Early Breast Cancer (monarchE). J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 38, 3987–3998 (2020).
165. Hudis, C. A. et al. Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 25, 2127–2132 (2007).
166. Symmans, W. F. et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 25, 4414–4422 (2007).
167. Sledge, G. W. et al. MONARCH 2: Abemaciclib in Combination With Fulvestrant in Women With HR+/HER2- Advanced Breast Cancer Who Had Progressed While Receiving Endocrine Therapy. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 35, 2875–2884 (2017).
168. Ju Y, A large-scale snapshot of intratumor heterogeneity in human cancer. Cancer Cell, 39, 463-465.
169. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017 Feb 9;168(4):613-628. doi: 10.1016/j.cell.2017.01.018. PMID: 28187284.
170. Morel et al. Chemo-prAIdict Breast: A deep learning solution for predicting residual disease on biopsies of breast cancer patients treated with neoadjuvant chemotherapy, European Journal of Cancer, Volume 234, 2026,116222, ISSN 0959-8049, https://doi.org/10.1016/j.ejca.2026.116222.

.

THE END